Epigenetic alterations associated with EMT within ER positive breast cancers

Jørgen Ankill

Master's thesis

Molecular Biology

60 Study points

Department of Biosciences

Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

December 2018

Epigenetic alterations associated with EMT within ER positive breast cancers

Jørgen Ankill

60 Study points

Department of Biosciences
Faculty of Mathematics and Natural Sciences
University of Oslo

Department of Cancer Genetics
Institute for Cancer Research
The Norwegian Radium hospital

December 2018

UiO : University of Oslo

Oslo University Hospital

Epigenetic alterations associated with EMT within ER positive breast cancers
Master's thesis, Winter 2018

Department of Cancer Genetics
The Norwegian Radium Hospital
Main supervisor: Thomas Fleischer

Department of Biosciences
The Faculty of Mathematics and Natural Sciences
University of Oslo
Co-supervisor: Ragnhild Eskeland

Abstract

Breast cancer is the most prevalent type of cancer affecting females in world today and poses a serious public health problem worldwide. Even though the overall survival has increased significantly the last decades, the high incidence of breast cancer signifies the importance of improvements in diagnostics and treatment. Like all types of cancers, breast cancer is a result of accumulation of genetic and epigenetic alterations that leads to repression of tumor suppressor genes and activation of oncogenes.

Over the past decades, several studies have highlighted alterations in DNA methylation patterns as hallmark events in many cancer types including breast cancer (1). Among the major types of breast cancers, the estrogen receptor positive tumors which accounts for 70 % of all breast cancers tend to display more pronounced changes in their DNA methylation landscape than the ER negative tumors when compared to normal adjacent tissue (2-4). It is well known that alterations in DNA methylation may affect the expression of genes, depending on where the changes occur. For instance, changes in DNA methylation at CpGs in *cis*-regulatory regions such as promotors tend to repress the expression of its associated gene. Furthermore, CpGs as far as 100 kb away from the transcription start site have been demonstrated to be associated with gene expression (5). Therefore, CpGs in intergenic and enhancer regions may play key roles in breast cancer pathogenesis through the regulation of expression of their associated genes. Enhancer methylation has been shown experimentally to be associated with gene expression, and these genomic regions are considered to be the most differentially methylated genomic regions during carcinogenesis and cancer progression (5, 6). Enhancers are known to bind cell type specific proteins called transcription factors (TFs) which are proteins involved in the regulation of gene transcription. However, the role of DNA methylation at enhancer regions regarding TF binding and breast cancer pathogenesis is still not fully understood.

Genome-wide expression-methylation quantitative trait loci (emQTL) analysis have previously been shown to identify significant correlations between the level of DNA methylation at CpG sites and gene expression due to intertumoral heterogeneity within ER positive and ER negative tumors (7). It has also been shown to be a valuable tool in the identification of key gene regulatory networks involved in breast cancer pathogenesis (7). To

take this further, the same approach was applied to the ER positive breast tumors only, to investigate whether any differences within the ER positive tumors in respect to DNA methylation and gene expression could be observed.

The study resulted in the identification of CpG-gene pairs in which the level of DNA methylation was significantly correlated with gene expression. Hierarchical clustering of the significant associations led to the discovery of a previously undiscovered cluster of CpG-gene associations. Gene set enrichment analysis indicated an enrichment of the genes in EMT-related processes, while the CpGs were highly enriched in enhancer regions. The CpGs in this EMT-cluster was divided into CpG-cluster A and CpG-cluster B based on whether their mean methylation value was more or less than 0.5 respectively. The CpGs in both clusters were shown to be differentially methylated among the ER positive tumors. Further characterization of the CpG-clusters by ChIP-seq peaks enrichment analysis revealed that CpG-cluster A CpGs were enriched within ChIP-seq peaks of TFs associated with EMT such as TEAD1, FOSL1, TWIST1, SIX2, YAP1 and PPARG. To investigate whether the difference in DNA methylation was associated with any phenotypic feature associated with EMT in the tumors, an EMT score was utilized and correlated with the mean methylation of CpG-cluster A CpGs. The mean methylation was negatively correlated with the EMT score, meaning that lower methylation was associated with more mesenchymal-like characteristic of the tumor samples. These findings suggest that EMT-related CpG-gene pairs discovered in this study are associated with gene regulatory networks wired by EMT related TFs through a relationship between DNA methylation at their target DNA binding regions in enhancers, and gene expression of their target genes. This study highlights the CpGs identified as potential contributors to EMT-related cancer pathogenesis in the ER positive breast tumors and constitute interesting regions for further investigations. However, these *in silico* findings still require better validation *in vitro*. The identification of cancer-causing epigenetic changes may open up possibilities of targeted treatment by utilization of technologies such as CRISPR to edit epigenetic cancer-causing mutations to inhibit tumor growth in the future.

Acknowledgements

The work presented in this master thesis was carried out at the Department of Cancer Genetics at The Norwegian Radium Hospital, during the period of January 2017 to December 2018 under the supervision of Thomas Fleischer.

After 5 years at university there are so many people who deserves my greatest thanks. First of all, I would like to express my gratitude to my supervisor Thomas Fleischer for giving me the opportunity to work with this project. Thanks for all the support, encouragement and guidance that I got during my master's thesis and for pushing me out of my comfort zone to become better. Much of my academic growth is thanks to you.

I would also like to thank professor Vessela Kristensen for including me in the research group. I have felt so welcome from the first day I started. I also would like to thank you and Thomas for including me in the group retreats and team building activities.

I would also like to give thanks to Grethe Irene Grenaker Alnæs and Marie Fongaard for supervising me in the lab. I have learned so much from working with you, and I am greatly appreciating it. In addition, I would like to thanks Daniel Nebdal for your technical support. And thanks to the rest of the members of the group for being so friendly, helpful and willing to discuss.

Finally, I would like to express my gratitude to my family and friends for the love and support over all these years. And a special thanks to my mother for encouraging me to reach my academic goals. I wish you could still be here and celebrate with us.

Oslo, December 2018

Jørgen Ankill

List of Abbreviations

ASCAT	Allele-Specific Copy Number Analysis of Tumors
CGI	CpG island
ChIA-PET	Chromatin Interaction Analysis with Paired-End-Taq
ChIP-seq	Chromatin Immunoprecipitation Sequencing
CIBERSORT	Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts
CRISPR	Clustered regulatory interspaced short palindromic repeats
DCIS	Ductal carcinoma *in situ*
DNMT	DNA methyltransferase
emQTL	Expression-methylation quantitative trait loci
EMT	Epithelial-mesenchymal transition
ER	Estrogen receptor
FDR	False discovery rate
GSEA	Gene set enrichment analysis
GO	Gene ontology
HER2	Human epidermal growth factor receptor 2
LCIS	Lobular carcinoma *in situ*
MET	Mesenchymal-epithelial transition
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MSigDB	Molecular Signatures Database
NK	Natural killer cells
OSL2	Oslo2
PAM50	Prediction Analysis Microarray 50
PCR	Polymerase chain reaction
PR	Progesterone receptor
ROR	Risk-of-recurrence
TCGA	The Cancer Genome Atlas
TET	Ten-eleven Translocation
TF	Transcription factor

Table of contents

1 Introduction

1.1 Cancer

Cancer is a general term used to describe a condition in which the cells in a body acquire the ability to grow and divide in an uncontrollable way. Eventually, as a tumor progress the cells may start to destroy nearby healthy tissues and spread to distal parts of the body and establish secondary tumors. If left untreated, the disease will eventually become a major burden for the body and will cause death. Cancer may affect anyone independent of age, lifestyle and familial history. In Norway, 36 % of men and 30 % of females are expected to be affected by cancer prior to the age of 75 (8).

Tumors are highly heterogenic and consists of a diverse bulk of cells with distinct molecular signatures and responses to cancer treatment. The unique biology of each tumor reflects the importance of research on the developmental process of the disease, risk factors associated with increased predisposition, classification of patients, treatments and the side effects associated with the treatments.

1.1.1 The hallmarks of cancer

For a cell to become cancerous, it must acquire some traits in a multistep process that allows it to gain a selective advantage compared to normal cells, as described by Douglas Hanahan and Robert A. Weinberg in year 2000 (9). Six essential alterations in the cell physiology must occur so it can develop into a malignant neoplasm. First, the cell must become **self-sufficient in growth signals**. Production and release of growth promoting signals are tightly regulated in normal cells to ensure homeostasis of cell number and tissue structure, and they are essential for the normal cells to shift from a quiescent to an active proliferative state. Molecular events to achieve this involves alteration of extracellular signals, transcellular transducers or intracellular circuits. In addition to growth factors, normal tissue cells express multiple antiproliferative signals that keep normal cells in a quiescent state to maintain tissue homeostasis. Cancer cells must be able to **block antigrowth signals** to divide. Antigrowth signals can block proliferation by forcing the cell out of the active proliferative cycle and into a quiescent state or by inducing the cell to enter a postmitotic state associated with gain of specific differentiation-associated traits. Achievement of this involves the alterations of

factors negatively regulating cell proliferation, such as tumor suppressor genes. Cancer cells must also be able to **avoid programmed cell death** which is triggered in response to various physiological stresses. Normal cells can only divide a limited number of times before it enters cell senescence or apoptosis, a mechanism associated with telomere shortening. To overcome this, the majority of human cancers have induced the expression of telomerase enzymes that lengthens the telomeres at the chromosome DNA ends thereby providing the **ability of cancer cells to replicate limitlessly.** For an emerging tumor to thrive, the ability to **induce angiogenesis** is essential to provide vital nutrients to ensure cancer cell survival in a densely packed tumor. Sooner or later in the tumor development, cancer cells acquire the ability to **metastasize** to other parts of the body and **invade distinct tissues to establish secondary tumors**.

In addition to these six hallmarks, increased understanding of the disease and cancer research progression have led to the addition of four additional hallmark capabilities reviewed in the updated paper of hallmarks of cancer by Hanahan and Weinberg that was published in 2013.

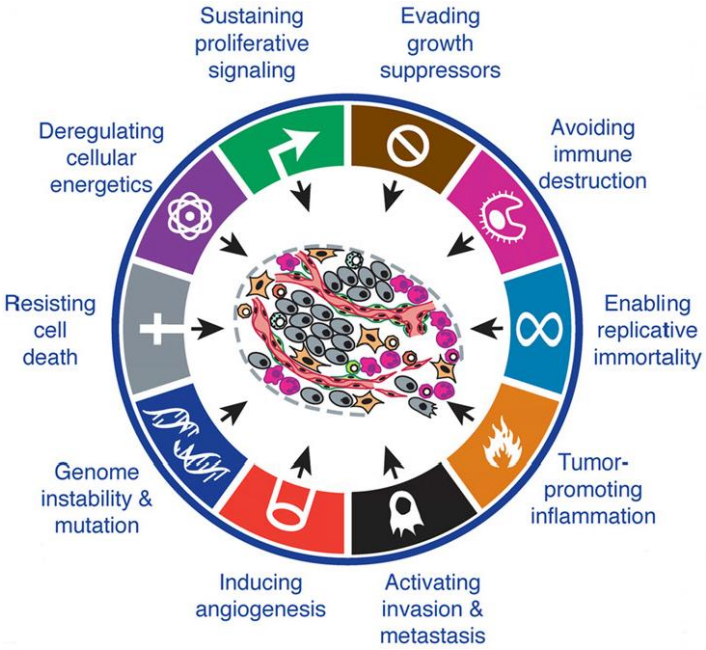


Figure 1. The hallmarks of cancer. The ten hallmarks of cancer common to most malignancies according to the updated paper of Hannahan and Weinberg published in 2013. These ten hallmarks include: the ability to sustain proliferative signaling, evade growth suppressors, avoiding immune destruction, enable replicative immortality, induce tumor promoting inflammation, deregulation of cellular energetics, resist cell death, genome instability and mutations, induce angiogenesis and activate invasion and metastasis. The order in which these traits are acquired differs from cancer to cancer. Reprinted from Hallmarks of Cancer (10).

Genomic instability and increased mutation rate are observed in cancer and is caused by increased sensitivity of cancer cells to mutagenic agents and/or the alteration of cell machinery components responsible for genomic maintenance. Moreover, cancer cells tend to display a **deregulation of the cellular energetics**. In many cases, cancer cells limit their energy metabolism largely to glycolysis and increases the glucose import into the cytoplasm, as first described by Otto Warburg (11-13).

These two last emerging hallmarks involves **tumor promoting inflammation**, and the **ability to evade immune destruction**. The immune surveillance theory propose that the immune system is continuously monitoring cells and tissues in the body and may eliminate highly immunogenic incipient cancer cells. This selection process of cells known as immunoediting leaves behind weakly immunogenic cancer cells that may avoid immune destruction and advance to become a solid tumor. Tumor tissues are known to contain immune cells in various densities with similar characteristics of non-neoplastic inflammation (14). Immune cell infiltration has been known for antitumoral response, but paradoxically they may also enhance tumorigenesis. Immune cells may supply the tumor microenvironment with biological molecules such as survival factors, growth factors, proangiogenic factors and extracellular matrix modifying enzymes. These factors may attract inflammatory cells, such as neutrophils, that can release highly mutagenic chemicals such as reactive oxygen species to further enhance carcinogenesis (15). In addition, such factors may facilitate angiogenesis and epithelial-mesenchymal transition (EMT) leading to invasion, metastasis and increased resistance to apoptosis (15-20). The ten hallmarks of cancer described by Hanahan and Weinberg is summarized in Figure 1.

1.2 Breast cancer

1.2.1 Incidence and survival

Breast cancer (Lat: *Cancer mammae*) is the most frequently occurring cancer type in women worldwide today with an estimated 1.67 million new cancer cases diagnosed in 2012 (25% of all cancers). The same year, breast cancer claimed 522,000 lives and was thereby ranked as the fifth cause of cancer death in the world (21). The global incidence of breast cancer increased from 641,000 in 1980 to 1,643,000 cases in 2010, which is a 3.1 % increase in annual rate (22). In Norway, 32,827 people were diagnosed with cancer in 2016, in which 3,402 of these incidences represented breast cancer. Females represents the great majority of these cases as only 31 of the affected were males (23).

The breast cancer incidence has increased significantly the last decades (Figure 2), but also the five-year overall survival (24). This is probably partly due to early disease detection and improvement in diagnostics and treatments. At the same time as the incidence have increased, the mortality of the disease has decreased from about 30 deaths to less than 25 deaths per 100,000 people in 2015. The observable decrease in mortality after 1996 may be partly due to the mammography-screening program that started in 1996 in Norway.

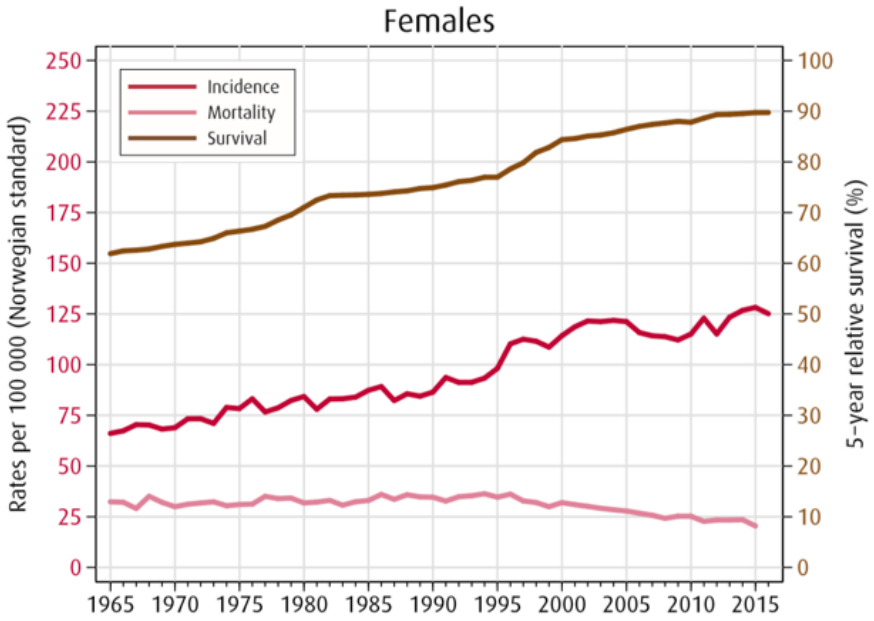


Figure 2. Trends in incidence, mortality rates and 5-year relative survival for Norwegian females from 1965 to 2015. Reprinted from Cancer in Norway 2016, The Cancer Registry of Norway (23).

1.2.2 Risk factors contributing to breast cancer development

It is usually not possible to pinpoint the exact cause of why someone develop cancer. But intensive research the last decades have elucidated several factors that may predispose a person to cancer development which includes genetic (inherent) factors and extrinsic factors.

It is well known that familial history is an important genetic factor of breast cancer predisposition, even though most women developing breast cancer have no familial history of the disease. Only about 13 % of women diagnosed with breast cancer had a first-degree female relative with breast cancer (25). The most commonly mutated genes responsible for the familial cases of breast cancer includes the tumor suppressor genes *BRCA1* and *BRCA2*, in which their gene products are involved in repair of damaged DNA (26). Females with somatic mutations in *BRCA1* are more likely to obtain additional genetic alterations that may lead to cancer. The life time risk of females with *BRCA1* or *BRCA2* mutations is estimated to be approximately 75 % (27). In sporadic breast cancers however, somatic mutations in these genes are rarely detected (28). In addition to familial *BRCA1* and *BRCA2* mutations, several rare hereditary syndromes are associated with breast cancer susceptibility such as Cowden syndrome caused by germline *PTEN* mutations (29), Li-Fraumeni syndrome with germline mutations of the tumor suppressor gene *TP53* (30), Peutz-Jeghers syndrome with mutations in the germline of the genes *STK11/LKB1* (31), Ataxia-telangiectasia caused by germline mutations in the *ATM* gene (32).

In addition to genetic predisposition, researchers have revealed that immigrants from low breast cancer risk countries have an increased risk of getting breast cancer towards that of the destination countries (33). This strongly emphasize the impact of environmental factors and lifestyle on breast cancer predisposition. Life style factors such as overweight, diet, alcohol consumption and smoking can affect the risk of developing breast cancer. Alcohol consumption is showed to be associated with the risk of getting breast cancer. The exact mechanism is still a bit unclear, but the breast seems to be more susceptible to the carcinogenic effects of alcohol for patients with moderate to high intake (34). Studies have shown conflicting results between dietary fat and breast cancer risk (35-38). However, there is substantial evidence for the link between overweight and breast cancer risk. Overweight and obese postmenopausal women have 1.5 times and 2 times, respectively, larger risk compared to normal weight women to develop breast cancer (39). This link between overweight and

breast cancer risk is likely to be caused by estrogen hormones. The main source of estrogen in postmenopausal women is the adipose tissue (40).

1.2.3 Breast anatomy and breast cancer progression

The breasts are paired structures found on the anterior thoracic wall and extend from the second rib superiorly to the sixth rib inferiorly (41). Most of the female breasts are made up of adipose tissue extending from the collarbone, down to the underarm and across the middle of the ribcage. The adipose tissue is surrounded by a network of ligaments, fibrous connective tissue, lymph vessels, lymph nodes, blood vessels and nerves (41, 42). The female breasts are exocrine glands consisting of milk producing cavities called alveoli that clusters into groups called lobules (41). Each breast contains about 20 milk secreting lobules (43). The lobules are connected to the nipples by lactiferous ducts that allows passage of the milk to the nipples. The majority (~75 %) of breast cancers forms at this site (44). Ducts and lobules are surrounded by suspensory ligaments that functions as structural support for the breasts. The major characteristics of the anatomy of the adult female breast is illustrated in Figure 3.

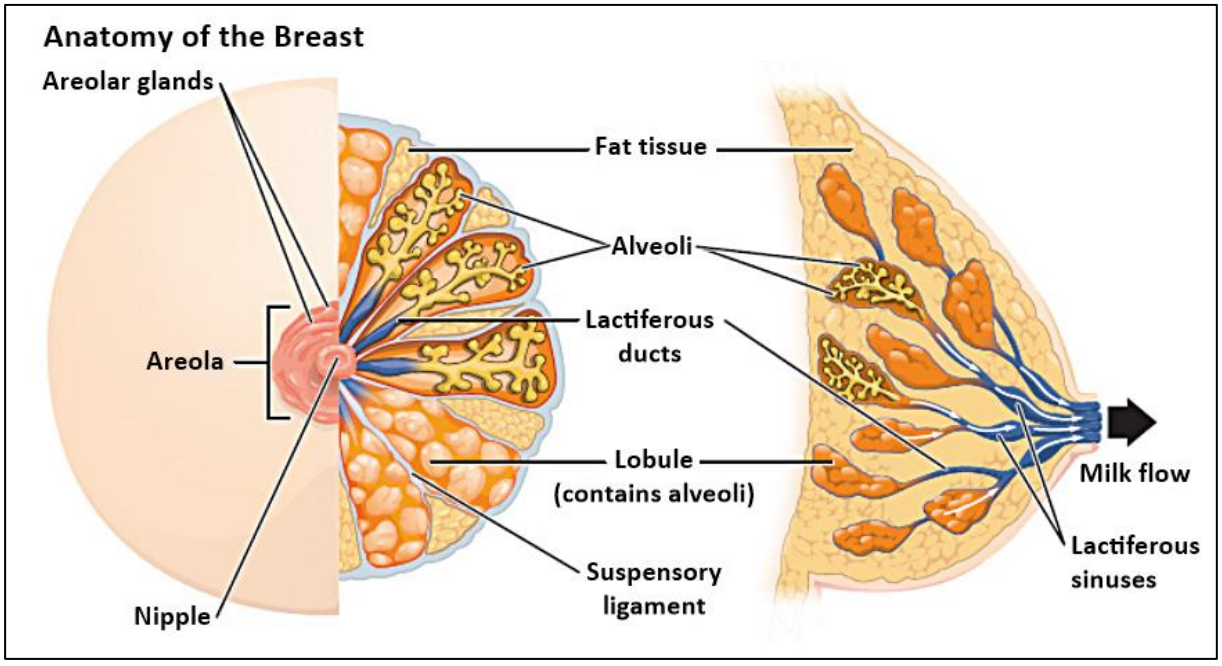


Figure 3. Anatomy of the adult female breast. The lobules produce milk that is transported to the nipple. Adipose tissue surrounds the functional tissue of the breasts. Adapted from (45).

Breast cancer development is a multistep process altered by genetic- and epigenetic changes as well as a changing tumor microenvironment. Such changes may alter the growth of cells surrounding the ducts or lobules inside the breasts. The stepwise conversion from normal epithelial tissue to an invasive carcinoma is described in Figure 4. Under normal conditions, the ducts and lobules in the breasts are surrounded by two layers cells, the outer luminal epithelial cells and the inner myoepithelial cells lining the basement membrane. More than two layers of cells in the ducts or lobules describes a condition called lobular hyperplasia or ductal hyperplasia. The hyperplasia may eventually display histological abnormal characteristics including the accumulation of normal looking luminal epithelial cells within the lobules or ducts, a condition known as atypical hyperplasia. The atypical hyperplasia lesion is not yet defined as cancerous, but the condition is associated with 4-5 times increase in breast cancer risk (46). The cells may further progress by increased proliferation leading to a condition called carcinoma *in situ*, more specific ductal carcinoma *in situ* (DCIS) or lobular carcinoma *in situ* (LCIS). Another form of breast cancer is breast sarcomas. Breast sarcomas accounts for less than 1% of all breast cancers (47).

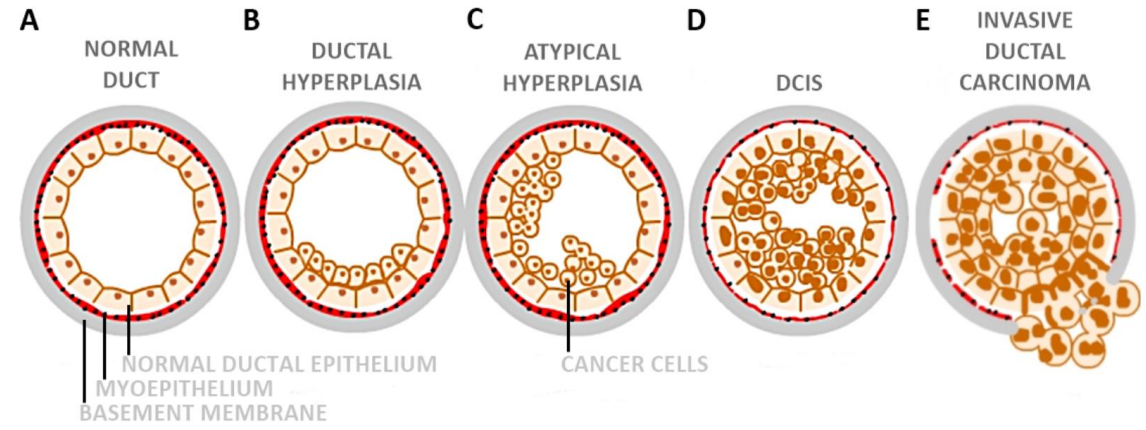


Figure 4. Breast cancer progression. (**A**) represents a normal duct in the breast of a healthy individual. During the early steps of breast cancer abnormal tissue patters occur characterized by histological abnormalities (**B-C**). Increased cell proliferation may eventually lead to DCIS (**D**). If left untreated, the DCIS may become invasive (**E**). Modified from (48).

A DCIS or LCIS is defined as a non-invasive neoplasm of epithelial cells separated from the breast stroma by the basement membrane. If left untreated a significantly proportion of these neoplasms will undergo a biological process termed the invasion-metastasis cascade. Early stages of this cascade involve the gain of mesenchymal cell characteristics of epithelial-like

tumor cells by EMT. Over time the mesenchymal-like tumor cells will be able to break through the basement membrane and disseminate from the primary tumor and migrate into the surrounding tissues (49). Eventually the cancer cells will intravasate into the lumen of blood vessels and lymphatic vessels. At this stage the area of abnormal cell growth is defined as an invasive ductal- or invasive lobular carcinoma. Breast cancer tumor cells tend to intravasate into the lymphatic vessels due to their high permeability (50). Tumor cell selection determines which cell that survives the transport through the vasculature. The tumor cell may extravasate into the parenchyma of distant organs and initially survive in the foreign microenvironment by initiating their proliferative characteristics and develop into a secondary tumor (51).

1.2.4 Breast cancer classification

Breast cancer is a heterogenous disease with diverse morphological features, variable clinical outcome and variable response to therapeutic treatment. Classification of breast cancer is essential to understand the underlying biological mechanisms driving the disease. Breast cancer classification is widely used in the clinic to determine the optimal treatment for each patient, and to provide information about treatment response as well as the expected prognosis. By identifying patients with a worse overall prognosis, one may be able to early consider the need for a more aggressive treatment.

TNM-classification

TNM classification is used to determine the cancer stage by combining information about primary tumor size (T), regional lymph node involvement (N) and distant metastases (M) as shown in Table 1. The guidelines for TNM staging is published in the 7th edition of the AJCC Cancer Staging Manual (52). Tumors are categorized as either non-detectable (T0), carcinoma *in situ* (Tis), smaller than 20 mm in greatest dimension (T1), larger than 20 mm but smaller than 50 mm (T2) or larger than 50 mm (T3). T4 is characterized as a tumor of any size with extension to the chest wall and/or the skin. N0 represents no detection of tumor cells within the lymph nodes, while N1, N2 and N3 display increased number of regional lymph nodes with increasing distance from the primary tumor. No distant metastases are represented as M0, while M1 represents metastasis detection.

The TNM value combinations are combined to give an overall stage from 0 to IV, in which stage IV is the highest stage with a more advanced cancer than for lower stages. Table 1 summarizes the characteristics of the different stages of breast cancer. Stage 0 is characterized by the presence of *carcinoma in situ*. Stage I disease is a T1 tumor with N0 or detection of micrometastases in the lymph nodes. Stage II disease can be T0-T3 tumor with no or little spread to the lymph nodes. Stage III includes T0-T4 tumors with more pronounced spread to lymph nodes. Stage IV disease represents any T or N and the detection of distant metastases. The prognosis is usually not so good for higher stages (53). Sometimes the stages are subdivided into A and B (53) as annotated in Table 1.

Table 1. Summary of TNM classification.

Stage	T	N	M
Stage 0	Tis	N0	M0
Stage IA	T1*	N0	M0
Stage IB	T0	N1mi	M0
	T1*	N1mi	M0
Stage IIA	T0	N1**	M0
	T1*	N1**	M0
	T2	N0	M0
Stage IIB	T2	N1	M0
	T3	N0	M0
Stage IIA	T0	N2	M0
	T1*	N2	M0
	T2	N2	M0
	T3	N1	M0
	T3	N2	M0
Stage IIB	T4	N0	M0
	T4	N1	M0
	T4	N2	M0
Stage IIIC	Any T	N3	M0
Stage IV	Any T	Any N	M1

*T1 includes T1mi

**T0, T1 nodal micrometastases are only excluded from stage IIA and are classified as Stage IB.

Histological grade

Tumor grading is used as an indicator of how likely a tumor is to grow and spread. Tumor grade is based on the tumor tissue morphology compared to normal tissue under a microscope. The grade may help to predict disease prognosis and treatments that may be beneficial for the patient. A scoring system is used to set tumor grade based on three features determined by tumor examination including the appearance of the glandular and tubular formation, nuclear pleomorphism and mitotic count (54). Each category is given a score between 1 and 3. A total score of 3-5 is considered to be grade 1, a score of 6-7 is grade 2 and a score of 8-9 is a grade 3 (55). Grade 1 tumors are slow growing and normal breast cell looking, while grade 3 tumors tend to grow faster, be poorly differentiated and look different from normal breast tissue. Grade 2 tumors are usually moderately differentiated with an intermediate appearance of grade 1 and grade 3 tumors.

Hormone receptor status

There are two main types of hormone receptors are typically involved in breast cancer which includes the estrogen receptor (ER) and progesterone receptor (PR) (56). Both receptors are steroid hormone receptors mainly found in cytosol. The steroid hormone estradiol can diffuse through the cell membrane and bind to ER, which leads to receptor migration into the nucleus where it dimerizes and binds to DNA. DNA binding causes activation of estrogen-responsive genes (57). There are two different forms of the estrogen receptor, α and β, which are encoded by two different genes called *ESR1* and *ESR2*, respectively (58). The clinical relevance of the ER β-form is still unknown (59), so the abbreviation ER will later be referred to as the estrogen receptor α-isoform. The PR receptor also exists in two isoforms; PRA and PRB. PRA is a truncated version of PRB but they share similar functions (60). Breast cancers can be classified based on hormone receptor status of cells evaluated from biopsy or surgery. A tumor is classified as estrogen receptor positive (ER+) if the estrogen receptor is upregulated, and ER negative (ER-) if the receptor is downregulated in the tumor cells. Around 70 % of all breast tumors are ER positive tumors (4). The same principle is true for the progesterone receptor.

HER2

Human epidermal growth factor receptor 2 (HER2) is a transmembrane protein receptor with tyrosine kinase activity encoded by the *ERBB2* gene. HER2 overexpression is observed in 20-30 % of all breast cancers and is often caused by *ERBB2* amplification (61). Growth factor binding to the HER2 receptor leads to dimerization and autophosphorylation in which the tyrosine residues on the cytoplasmic domains becomes phosphorylated. This modification leads to activation of various signaling pathways such as the mitogen-activated protein kinase (MAPK) pathway involved in cell proliferation and the phosphoinositide 3-kinase pathway involved in cell survival (62). HER2 amplification is associated with metastasis and reduced survival (63).

Classification by gene expression

Development of cDNA microarray technology made it possible to classify breast cancer patients based on gene expression pattern. Perou et al. (64) described one such approach and made to classify 65 breast cancer specimens into intrinsic subtypes associated with ER status and other ER related genes. Parker et al (65) formalized gene expression subtyping to include

five subtypes derived from the expression of 50 genes: luminal A, luminal B, basal-like, HER2-enriched and normal-like (64, 66).

The luminal A subtype is the most commonly occurring breast cancer subtype and accounts for 50 % of all invasive breast cancers (67, 68). Luminal A tumors tend to be ER positive and/or PR positive, HER2 negative and to have a low Ki67 index (proliferating cell nuclear antigen necessary for cell proliferation) (67). They tend to display low histological grade, low degree of nuclear pleomorphism and are associated with a good prognosis (69, 70). Luminal A tumors tend to express high levels of ER (*ESR1*), luminal epithelial markers such as cytokeratin 8 (*KRT8*) and cytokeratin 18 (*KRT18*) and other genes associated with ER function such as forkhead box protein A1 (*FOXA1*), GATA-binding protein 3 (*GATA3*) and zinc transporter ZIP6 (*SLC39A6*) (71).

Luminal B tumors in contrast tend to display a more aggressive phenotype with characteristics including higher histological grade, variable expression of HER2 (positive or negative), higher proliferative Ki67 index compared to luminal A and a worse prognosis (68, 72). The main difference between the luminal subtypes is the increased expression of proliferation-related genes such as lysosome-associated transmembrane protein 4-beta (*LAPTM4B*), avian myeloblastosis viral oncogene homolog (*MYB*), nuclease sensitive element binding protein 1 (*YBX1*) and cyclin E1 (*CCNE1*). Tumors of the luminal subtypes are considered to be among the most differentiated tumor subtypes (73, 74).

HER2-enriched breast cancer tumors are commonly ER-/PR- and HER2 positive. Morphologically, HER2-enriched breast cancer tumors tend to display high proliferation and high histological- and nuclear grade. In addition to high expression of HER2, genes associated with the HER2 pathway tend to be upregulated as well (67). The overall survival associated with the HER2-enriched subtype is similar to that of luminal B (75).

Basal-like tumors usually express low levels of ER, PR and HER2, a term referred to as triple-negative tumors in the pathology. Basal-like tumors are associated with high histological grade, high nuclear grade, high Ki67 index and worse survival than the luminal B subtype (69, 76, 77). The tumors tend to have a high expression of cytokeratin 5 (*KRT5*), cytokeratin 14 (*KRT14*) and cytokeratin 17 (*KRT17*) (78).

Normal-like tumors tend to be hormone receptor positive (ER+ and/or PR+), HER2 negative and have expression profiles similar to normal breast tissue (79). Normal-like breast cancers have an intermediate prognosis (69).

These five subtypes may be predicted by the differential expression of 50 genes by prediction Analysis Microarray (PAM50) (65). An emerging PAM50-based subtype classifier and risk model now included in the international clinical practice guidelines is the Prosigna® PAM50 assay made by NanoString. Prosigna® has shown promising results in classifying breast cancer patients into prognostic groups. The Prosigna® assay is a genomic test that analyzes the activity of 50 PAM50 genes. Based on the activity of these genes a risk of recurrence (ROR) score is estimated, allowing the categorization of patients into a low, intermediate and high-risk groups (80). A study published in 2017 by Ohnstad et al. (81) showed that PAM50 intrinsic subtype- and ROR score classification improves classification of breast cancer subtypes into prognostic groups. It provides a more precise indication of future recurrence risk and may improve the basis for adjuvant treatment decisions (81).

TP53

TP53 is a tumor suppressor gene encoding tumor protein p53, which is involved in response to cell stress by inducing pathways leading to cell cycle arrest, DNA repair and apoptosis. About 30 % of all breast carcinomas contains *TP53* mutations, and more than 75 % of the mutations is caused by missense mutations (82, 83). *TP53* status have been shown to have a high prognostic value in which mutations in *TP53* is associated with worse disease-free and overall survival (84).

Other molecular classifications

DNA copy number alterations (CNAs) occurs when larger portions of the genome are duplicated or deleted. This may alter the expression of genes as the DNA fragment to be duplicated or deleted may contain genes. Duplications and deletions may also disrupt proximal or distal regulatory regions which may alter the properties of the cell. Copy number alterations have previously been shown to be associated with cancer progression (85). Breast cancer classification based on DNA copy number alterations have previously been performed. The classification was based on complex rearrangements (complex arm aberration index; CAAI) and whole-arm gains and losses (whole-arm aberration index; WAAI). By applying

CAAI and WAAI, eight subgroups were identified with distinct prognosis (86). Another approach classified breast cancers based on the integration of DNA copy number and gene expression data of loci where gene expression was affected by DNA copy number alterations. This approach revealed ten integrative clusters associated with different prognosis. One subgroup mainly consisting of luminal A and luminal B tumors was identified with poor prognosis and was associated with *cis*-acting CNAs at 11q13/14. Another subgroup devoid of CNAs was associated with good prognosis (87).

1.2.5 Breast cancer treatment

The main goal of breast cancer treatment is to completely cure the disease, and if this is unachievable, to provide a prolonged life for the breast cancer patient and at the same time maintaining a good quality of life. The primary treatment for most patients with breast cancer is surgery. There are two main types of surgery; breast-conserving surgery (lumpectomy) in which only part of the breast is removed including the cancer and surrounding normal tissue, and mastectomy in which the entire breast is removed including all breast tissue (88). Chemotherapy, radiation therapy, hormone therapy or targeted therapy may be included in addition, both before surgery (neoadjuvant treatment) and after surgery (adjuvant treatment).

Chemotherapy targets cells growing and dividing rapidly such as cancer cells but also affect other rapidly dividing normal cells which may contribute to its side effects. Chemotherapy is usually preferred at the early stage of invasive breast cancer and advanced-stage breast cancer (89). Radiation therapy use high-energy radiation to damage cancer cells as well as the nearby normal cells (90). Since cancer cells are less organized than healthy cells, DNA damage caused by the radiation therapy is harder for the cancer cell to repair than for normal healthy cells. The aim of hormone treatment is to either lower the estrogen level in the body or to block the action of estrogen on breast cancer cells in hormone-receptor-positive breast cancers (91). Estrogen is a major contributor to the growth of hormone-receptor-positive breast cancers (92). Receptor status determination is important when deciding treatment options for the patient, as hormone therapy will not affect hormone receptor negative tumors. For these patients, chemotherapy is often suggested (93). Hormone receptors such as ER and PR binds estrogen and progesterone respectively and promotes tumor growth (60, 94). By knowing the receptor status of a breast cancer one can use drugs specifically targeting these receptors either by lowering of hormone level or to block the hormone from binding to these receptors

in hormone positive tumors. The FDA approved humanized monoclonal antibody drug trastuzumab (Herceptin) is an example of targeted cancer therapy which targets HER2 positive breast cancers by blocking growth signals (95). Breast cancer classification is essential to map potential targets of drug treatments that will benefit the patient the most.

1.3 Epigenetics

Cell types within multicellular organisms may accomplish highly distinct functions and display very different morphology, but still they share the same genome. Cell types within and organism are distinct from one another because they synthesize and accumulate different sets of RNA and protein molecules. This phenomenon can be explained by epigenetics. The term epigenetics can be defined as regulatory mechanisms that influence gene expression without altering the DNA sequence (97). Epigenetic modifications are essential for normal development and tissue specific gene expression in mammals (98). Changes in epigenetics have the potential to alter gene expression by several mechanisms including histone modification, nucleosome positioning and DNA methylation (Figure 5). Improper modifications may have adverse health effects and lead to diseases such as cancer (99). Some of these alterations can be transferred between generations (100).

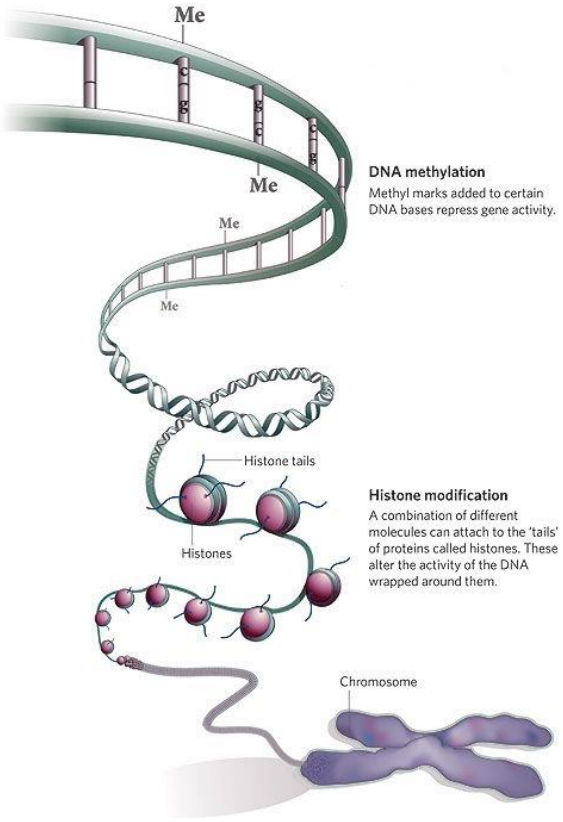


Figure 5. Epigenetic modifications of the DNA. The cytosine bases of DNA can either be methylated or unmethylated while the histone tails can be modified by chemical groups. The chromatin can be densely- or loosely packaged. Reprinted from (96).

1.3.1 Chromatin

Chromatin is the compact organization of DNA within the cell nucleus and consisting of DNA and proteins. The main role of chromatin is to effectively package DNA in the nucleus, reinforce DNA during cell cycle, prevent DNA damage and control DNA replication and

transcription. There are two states of chromatin; euchromatin which is a lightly packed form of chromatin associated with transcriptional activity, and tightly packed heterochromatin associated with gene silencing and protection of chromosome integrity (101). The fundamental unit of chromatin is the nucleosome that consists of a core particle and an internucleosomal region linking the core particles together. A nucleosome consists of DNA wrapped around a histone octamer composed of a H3/H4 tetramer flanked by two H2A-H2B dimers (102). Each histone has a protruding N-terminal tail (H2A and H2B also have a C-terminal tail). The nucleosome is the first level of organization of chromatin and look like "beads on a string" forming fibers of approximately 10 nm with a DNA packing ratio (length of DNA/length of unit) around 6. The 10-nm fibers are condensed into 30-nm fibers in a solenoid like structure stabilized by the linker histone H1 that binds to DNA entry and exit sites of the nucleosome with a DNA packing ratio of 35-40 (102, 103).

Histones tails are subjected to various types of post-translational modifications including methylation, acetylation, phosphorylation, ubiquitylation, sumoylation, ADP ribosylation, deamination and proline isomerization (104). Acetylation at amino acid position 4, 9, 14 at H3 lysines and at position 5, 16 at H4 lysines are common targets of acetylation associated with transcriptional activation (105, 106). Histone acetylation is catalyzed by histone acetylase enzymes and reversed by histone deacetylases. Phosphatases are involved in histone tail phosphorylation a modification associated with regulation of transcription and DNA damage response. E.g. Phosphorylation of the H2A(X) histone at serine at position 139 triggers DNA-damage response pathways eventually leading to non-homologous end joining, homologous recombination or replication-coupled DNA repair (107). Serine phosphorylation on H3 at amino acid position 10 and 32 have been associated with transcription of the proto-oncogenes *c-fos*, *c-jun* and *c-myc* (108-110). The modification of histone tails by methylation mainly occurs at lysines or arginines and is linked to both transcriptional activation and inactivation. Lysines may be mono-, di- or tri-methylated while arginine may be mono-, symmetrically- or asymmetrically demethylated in a process catalyzed by histone methyltransferases (111). H3K9me3, H3K27me3 and H3K79me3 are associated with transcriptional repression while H3K4me3, H3K36me3 and H4K20me3 are associated with transcriptional activation (112, 113).

Histone modifications may exert their effects either by directly influencing the local structure of chromatin or by interacting with effector molecules. Acetylation and phosphorylation of

histones disrupts the electrostatic interactions between the positively charged histone and the negatively charged DNA. This causes a less compact chromatin structure which may allow the transcriptional machinery to access DNA and initiate transcription (111). In addition, many chromatin-associated factors have domains that recognize specific histone tail modifications. For instance, the tandem chromodomains of chromodomain-helicase-DNA-binding protein 1 can bind trimethylated H3K4. The protein is an ATP-dependent remodeling enzyme involved in nucleosome repositioning (114). Acetylated histones may be recognized by the bromodomain of chromatin remodeling complexes such as Swi2/Snf2 complex to loosen the chromatin structure, thereby allowing transcription to occur (115). Some remodeling complexes are also associated with gene repression by changing the local chromatin structure, blocking the transcription factor machinery from binding (116). One such remodeling complex is the mammalian ISWI chromatin remodeling ATPase SNF2H (117).

1.3.2 DNA methylation

DNA methylation is the process of adding a methyl group ($-CH_3$) to the C-5 position of the pyrimidine ring of cytosine in DNA as indicated in Figure 6. The process is facilitated by a DNA methyltransferase enzyme (DNMT) (118). There are two different types of DNMTs involved in DNA methylation in mammals. DNMT1 is involved in the maintenance of the methylation pattern of DNA, mainly by methylating the unmethylated DNA strand of hemimethylated double stranded DNA after DNA replication. In contrast to this, DNMT3 prefer *de novo* methylation of DNA (119, 120). However, mounting evidences indicates that DNMT3 may also play a role in maintenance methylation during replication (120).

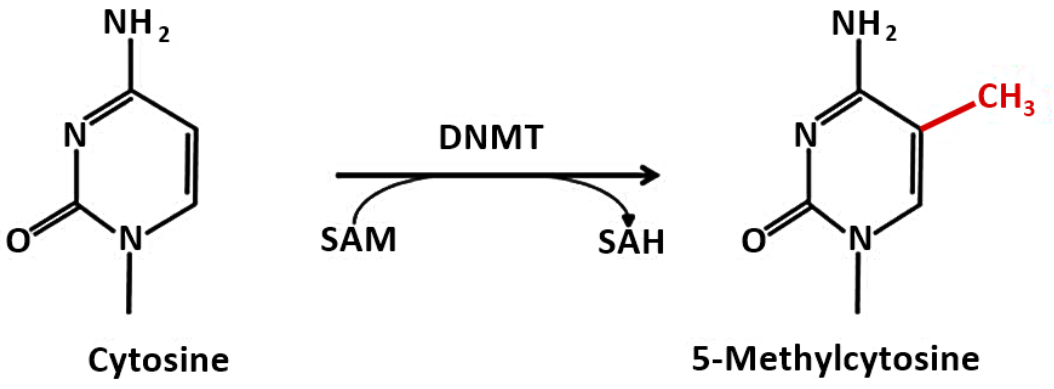


Figure 6. Cytosine methylation. DNMT catalyzes the reaction in which a methyl group is donated from S-adenosyl-L-methionine to the 5 position of the cytosine pyrimidine ring.

DNA methylation occurs primarily at CpG sites in DNA. CpG sites are parts of DNA where the cytosine (C) nucleotide is followed by a guanine (G) nucleotide. The p in CpG stands for the phosphate bond connecting these two nucleotides together. In the human genome, many CpGs are found enriched in certain regions called CpG islands (CGI). More than half of the human genes contain CGIs, while the rest of the genome generally is depleted of CpGs (121). These CGIs are enriched in regulatory regions such as promotors (122). Promotor methylation is associated with the repression of gene expression. For gene transcription to occur, the promoter region needs to be easily accessible for transcription factors and other components of the transcription machinery. Transcription factors will generally not bind to methylated promoters unless they have a methyl-CpG binding domain (123). In addition, methylated CpGs may attract proteins that can bind to the methyl groups of the promotor to recruit repressing remodeling complexes (119, 124). A consequence of this alteration is chromatin compaction which is associated with gene silencing (125, 126).

DNA methylation of cytosines in DNA is a reversible process. Demethylation involves enzymes such as ten-eleven translocation enzymes (TET). TET enzymes are hydroxylases that can oxidize 5-methylcytosine (5mCs) to 5-hydroxymethylcytosine (5hmC) or further oxidize it to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). Alternatively, 5mC could be deaminated by AID/APOBEC family member enzymes to create a thymidine base, or on 5hmC to produce 5-hydroxymethyluracil (5hmU) or 5-hydroxymethyluracil (5hmU), all which may be recognized as base mismatches (106, 127). Such mismatches may be replaced by unmethylated cytosines by glycosylase enzymes in a process called base excision repair (127). The exact mechanisms behind demethylation of 5-methylcytosines to cytosine is not fully understood.

1.3.3 DNA methylation in cancer

Epigenetic features such as DNA methylation and chromatin states are often found to be altered in cancer cells (128). Appropriate DNA methylation is essential for development and appropriate cell function. Abnormalities in the DNA methylation pattern may lead to various diseases, such as cancer (129). Both global hypomethylation and DNA hypermethylation is observed in cancer. Global hypomethylation contributes to genomic instability. For instance, normal cells are highly methylated at satellite sequences and repetitive DNA sequences (E.g. LINE, SINE and Alu elements) within the genome, which maintains genomic stability and

integrity (130, 131). In many tumors, the loss of DNA methylation has activated transposable elements (132). Transposable elements may integrate at random sites within the genome which may cause mutations and genomic instability. In addition, loss of DNA methylation at promoters may activate oncogenes. Contradictory to global hypomethylation, certain genes are inactivated by hypermethylation of CGIs in their regulatory regions (133).

1.3.4 DNA methylation in breast cancer

Several studies have reported that DNA methylation may be an early event in breast carcinogenesis, which may lead to oncogene activation and silencing of tumor suppressor genes (134-137). More than 100 genes have been reported to be aberrantly methylated in breast carcinomas and many of them have been associated with gene repression. *BRCA1* have been found to be frequently inactivated in sporadic breast cancers due to promoter hypermethylation. Another gene, the *CDKN2A* tumor suppressor gene tend to be more frequently hypermethylated among the ER negative breast cancers than the ER positives (138). *CDKN2A* is a member of a family of proteins that binds to and block the cyclin D/cdk4 complex activity and induces G1 cell cycle arrest in cells with a functional retinoblastoma protein (139). In addition to hypermethylation, global hypomethylation have been associated with tumor stage, size and grade (140). Although hypomethylation of *cis*-regulatory regions in cancer is much less frequent than hypermethylation of CGIs overlapping promoters, a few genes have been reported to be frequently hypomethylated in breast cancers (140, 141). Some of them includes *IL-10* (142), *MDR1* (135) and *CDH3* (143), in which hypomethylation in their promoter region have been associated with increased gene expression. In addition to promoters, several published papers the past years have elucidated the link between aberrant DNA methylation at enhancers to tumor progression and plasticity (6, 144, 145). Enhancers are *cis*-regulatory regions known to regulate gene expression through the binding of cell-type specific TFs that can recognize specific sequences of DNA (146, 147). TF binding to enhancers is also known to be influenced by DNA methylation at the TF binding sites (148).

In the last decade, technological advances in high-density microarray technology and high-throughput DNA sequencing has made extensive genome-wide analyses achievable. This enables more comprehensive studies of the role of DNA methylation in breast cancer. Numerous papers have been published investigating global DNA methylation differences in relation to breast cancer subtypes and clinical features (3, 149, 150). One study identified a

significant association between the methylation of *RECK*, *SFRP2*, *UAP1L1*, *ACADL*, *ITR* and *UGT3A1* and relapse-free survival (151). Another made a model based on five biomarkers that could distinguish HER2 overexpressing subtypes (Luminal B and ERBB2 positive) and basal-like tumors. In HER2 overexpressing tumors *NPY*, *HS3ST2*, *RASSF1*, *FGF2* and Let-7a were hypermethylated, while basal-like tumors displayed lack of methylation (2). In addition, large scale methylation analyses have shown that breast cancers can be classified into three clusters. Each cluster were associated with different ER status, *TP53* mutation status, molecular subtype and overall survival (2, 3, 136, 152, 153). A DNA methylation signature (SAM40) has also been developed that segregated luminal A patients based on their prognosis, thereby identifying one subgroup that may benefit from more aggressive treatment and another one that would benefit from less (154). DNA methylation is a robust biomarker, more stable than RNA and proteins and is therefore an appealing target for development of new approaches for diagnosis and prognosis of breast cancer. Since DNA methylation is critical for gene expression, DNA methylation may provide an additional layer of information that may provide better breast cancer classification and clinical information in the future.

1.4 Genome-wide expression-methylation quantitative trait loci analysis

Genome-wide expression methylation quantitative trait loci (emQTL) analysis is a bioinformatical approach used to identify and characterize significant correlations between the level of DNA methylation at CpG sites and gene expression (emQTLs). Fleischer, Tekpli et al. (7) demonstrated this method for the first time in 2017 and revealed a hitherto unknown connection between the epigenome, transcription factor activity and gene expression in breast cancer. DNA methylation at enhancers at ERα, FOXA1 and GATA3 binding regions was found to be a breast cancer subtype specific feature (7).

Significant correlations between the level of DNA methylation and gene expression were identified using Pearson correlation and Bonferroni correction. The Bonferroni corrected p-values were clustered by unsupervised clustering to identify their biological relevance. From this they discovered two distinct bi-clusters of CpG-gene associations with different biological characteristics; Cluster 1 genes were found to be enriched in processes related to immune response and Cluster 2 genes were enriched in processes associated with estrogen response. In addition, ChromHMM segmentation data from the MCF7 cell line revealed that the CpGs from Cluster 1 and Cluster 2 were enriched in enhancer regulatory regions within

the genome. Since enhancers are known carry DNA sequences (motifs) recognized by cell type specific transcription factors, they sought for ChIP-seq peaks enriched within a close proximity to Cluster 1 and Cluster 2 CpGs. ChIP-seq data from the MCF7 cell line revealed that Cluster 1 CpGs were significantly enriched in TF binding regions of TFs involved in immune cell homeostasis such as RUNX1 (155), FLI1 (156) and ERG (157, 158). Cluster 2 CpGs were found to be enriched in TF binding regions of TFs associated with estrogen signaling such as FOXA1 (159), GATA3 (160) and ERα (161). They are TFs well known to play key roles in breast cancer pathogenesis (160, 162).

Further investigation of the level of DNA methylation for the CpGs in Cluster 2 regarding histopathological features and molecular classification was performed by unsupervised clustering of Cluster 2 CpGs. The level of methylation clearly distinguished the ER positive tumors from the ER negative tumors. Two CpG sub-clusters appeared as well, CpG-Cluster 2A and CpG-Cluster 2B. CpG-Cluster 2A CpGs were enriched in the binding regions of ERα, FOXA1 and GATA3 and showed lower methylation in the ER positive tumors compared to ER negative tumors. The methylation pattern in CpG-Cluster 2B CpGs showed inverse methylation pattern. In addition, the methylation pattern was compared to the normal tissue CpGs in CpG-Cluster 2A and CpG-Cluster 2B and showed that CpG-Cluster 2A CpGs are hypomethylated in ER positive tumors while CpG-Cluster 2B CpGs are hypermethylated in ER positive tumors. Overall, this suggested that the methylation patterns of CpGs in Cluster 2 are features acquired during carcinogenesis.

Unsupervised clustering of the expression level of the genes in Cluster 2 was also performed, almost perfectly separating ER positive tumors from ER negative tumors. Two gene sub-clusters were identified with differential expression according to ER status. Further investigation of the Cluster 2 genes revealed that 32% of the genes in Cluster 2 were paired with a minimum of one CpG locally (within ±10 kb window). This suggested that the genes of Cluster 2 are locally regulated through DNA methylation of enhancers that carries transcription factor binding regions for TFs such as FOXA1, GATA3 and ERα. CpG-Cluster 2A CpGs with a low methylation level in ER positive tumors were locally paired with genes with high expression in ER positive patients, and the CpGs with a low methylation in ER negative tumors were locally paired with genes with high expression in ER negative patients. Gene knockdown experiments of FOXA1 and GATA3 in addition to Global Run On sequencing data revealed that 67% of genes in Cluster 2 were targets of ERα, FOXA1 and

GATA3. These target genes were significantly higher expressed in ER positive tumors, and the gene expression was higher in the Luminal A and Luminal B subtypes versus Normal-like and Basal-like breast cancer subtypes, thereby highlighting the link between Cluster 2 and estrogen signaling. Further assessment of the link between DNA methylation at enhancers and the expression of target genes was performed using ChIA-PET Pol2 data sets. ChIA-PET Pol2 data sets contains experimentally defined data containing information about long-range chromatin interaction genome wide. Cluster 2A-CpGs in emQTL with Cluster 2A genes were enriched in ChIA-PET Pol2 loops and provided additional evidence for the regulation of the expression of target genes through DNA enhancer methylation containing transcription factor binding regions.

Finally, unsupervised clustering of the genes in Cluster 1 associated with immune processes was performed. The level of lymphocyte infiltration was assessed using the *in silico* nanodissection algorithm (http://nano.princeton.edu/) to quantify the level of lymphocyte infiltration based on gene expression data. The unsupervised clustering of Cluster 1 genes did not segregate the breast cancer patients based on the PAM50 subtype or ER status, but it did segregate the patients based on the level of lymphocyte infiltration in the tumor sample.

1.5 EMT in breast cancer

Multiple lines of evidence the past decades have suggested that epithelial cancers can transform into a more mesenchymal-like phenotype in a process called epithelial-to-mesenchymal transition (EMT). Several studies have highlighted EMT as an important contributor to cancer progression, metastasis and drug resistance (163-168). Tools to study EMT in cancer may therefore provide insight into the development of cancer during tumorigenesis as well as the molecular mechanisms behind EMT. Such knowledge may contribute to improved treatments of mesenchymal cancers.

EMT is a biological process in which an epithelial cell undergoes multiple biochemical changes that enables it to lose apical-basal cell polarity and cell-cell adhesion to assume a mesenchymal-like migration prone phenotype. EMT eventually leads to tumor cell dissemination from the primary site, allowing invasion of malignant breast cancer cells into secondary sites in a reversible process called mesenchymal-epithelial transition (MET)(49). There are three main types of EMT programs; type 1 is involved in embryogenesis,

gastrulation and neural crest formation. Type 2 is associated with wound healing and tissue regeneration, and type 3 is related to cancer malignancy, invasion and metastasis (169). In breast cancer, EMT has been found associated with cancer stem cell properties including abilities of self-renewal, resistance to chemotherapy and expression of stem cell associated $CD44^{+}/CD24^{-}$ antigenic profile, thereby contributing to a more aggressive breast cancer phenotype (170-173). Today, stemness and EMT are considered to be functionally interconnected via gene expression (174).

EMT is orchestrated by a set of pleiotropically acting TFs such as Twist, Snail, Slug and Zeb1/2 (9). Other EMT-related TFs includes TEAD1 (175), FOSl1 (175), SIX2 (176), YAP1 (175) and PPARG (177). The EMT-associated TFs are expressed in various combination in malignant tumors and they have been shown experimentally to be important during tumor cell invasion. Ectopically overexpression of several of these transcription factors has been observed to elicit metastasis (9, 178-180). For instance, Snail overexpression and binding to the E-cadherin promotor have been shown to strongly suppress the expression of E-cadherin (181). Loss of E-cadherin expression is a major event during EMT (182). Twist1 overexpression has been shown to induce EMT and to reduce tumor cell proliferation (183, 184). Interestingly, accumulating evidences suggests that EMT attenuates proliferation (185-189). Other characteristics of EMT includes downregulation of cytokeratin, ZO-1 and upregulation of mesenchymal markers such as N-cadherin, fibronectin, vimentin and FOXC2 (190, 191).

2 Aims

An essential part of cancer research today is to identify and understand the molecular mechanisms driving the tumorigenesis towards malignancy. Identification of such causal pathways may enable the identification of new potential therapeutic targets that may inhibit cancer progression or hopefully cure the disease once and for all. Besides the major goal of all cancer research, that is to completely cure the disease with minimal side effects for the patient, the more specific goal of this project has been to:

- Identify differences within ER positive breast cancers in respect to DNA methylation levels of CpGs and gene expression by genome-wide expression-methylation quantitative trait loci (emQTL) analysis.
- Understand how DNA methylation and transcription factors contributes to carcinogenesis in luminal breast cancers.
- Identify CpG-gene pairs that represents candidates as cancer promoting alterations.
- Validate the *in silico* findings in a cell line model system

3 Materials and methods

3.1 Patient materials

All molecular profiles utilized in this study were from ER positive primary breast tumors of luminal A- or luminal B subtype.

3.1.1 Oslo2

The Oslo2 (OSL2) breast cancer cohort is a consecutive study aiming to collect material from breast cancer patients with primary operable disease (T1-T2) in several south-eastern Norwegian hospitals. Patient inclusion to the study started in 2006 and is still ongoing. The cohort consists of gene expression, DNA methylation and clinical data from more than 300 sporadic breast tumors (7, 192). All patients have provided written consent for use of the material for research purposes (193). Gene expression data and clinical data can be obtained from GEO with access number GSE58215. DNA methylation profiles can be found at GEO with access number GSE84207.

3.1.2 TCGA

The Cancer Genome Atlas (TCGA) is a publicly funded project that aims to discover and catalogue cancer-causing genomic alterations in order to create a comprehensive genomic "atlas" of cancer profiles (194). The cohort consisting of data from more than 500 patients with both sporadic and familial breast cancer disease with gene expression (RNA-seq) and DNA methylation data profiles. The DNA methylation data profiles used in this study was generated using Illumina HumanMethylation450K. The molecular data for TCGA is publicly available and can be obtained from the TCGA data portal. Level 3 DNA methylation and RNA-seq data were utilized in this study.

3.1.3 METABRIC

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort is a collaboration project between Canada and the United Kingdom that aims to classify breast tumors into additional subcategories based on molecular signatures that may help to

determine optimal course of treatment (87, 195). The cohort consists of more than 1900 fresh frozen breast cancer samples with different grade, stage and molecular markers. Gene expression data from METABRIC can be obtained from EGAD with access number EGAD00010000210.

3.2 Statistical computing and bioinformatical analyses using R

All the analyses were conducted using the R software (R version 3.4.0) (196). R is a free software environment used for statistical computation and graphics. R provides collection of statistical and graphical techniques including classical statistical testing, linear and non-linear modeling, classification and clustering. R contains a core set of packages that is included in the installation of R, but more than 15,000 user-created packages are available and can be downloaded and applied in the program.

3.3 Statistical tests and principles

3.3.1 Correlation analysis and linear regression

Correlation analysis is a statistical method used to investigate whether there is a possible linear association between two continuous variables. A correlation coefficient represents the extent of the association between the variables. There are two main types of correlation coefficients. Spearman's rank correlation coefficient is used when both variables are skewed, or ordinal and extreme values are present. Spearman's correlation is used to determine the direction and strength of a monotonic relationship between two variables. A Spearman correlation rho value of +1 or -1 indicates that the variable is a perfect monotone function of the other variable. Pearson's product-moment coefficient is used when both variables being studied are normally distributed. In contrast to Spearman correlation that considers a monotone relationship, Pearson correlation determines the direction and strength of a linear relationship between two variables. A correlation coefficient of +1 or -1 suggests a perfect linear relationship, while a correlation coefficient of zero indicates the absence of a linear relationship between two continuous variables. For Spearman- and Pearson correlation; a positive correlation is associated with a positive correlation coefficient (i.e. if the value of a variable goes up, the value of the other variable also tends to). On the other hand, a negative

correlation is characterized by a negative correlation coefficient (i.e. if the value of a variable goes up, the value of the other variable tends to go down). Correlation analyses must be performed by care as a statistical relationship between two variables does not necessarily imply a causal relationship between them.

3.3.2 Hierarchical clustering

Hierarchical clustering analysis is a method used to cluster groups of features with similar characteristics into clusters. There are two types of hierarchical clustering strategies that exists; divisive clustering and agglomerative clustering (197). By divisive clustering or top down clustering, each observation starts in their own cluster and splits are performed repetitively as one moves down the hierarchy. Agglomerative or bottom up clustering is the most commonly used approach in which all observations are assigned to their own cluster and each cluster pairs are merged as one moves up the hierarchy.

To determine which clusters that should be combined (agglomerative) or where the clusters should be split (divisive), a measure of dissimilarity of the observations is required (198). To determine this an appropriate distance metric and linkage criterion is applied. Distance metrices are different functions that defines the distance between data points and the choice influences the shape of the clusters. Some commonly used distance metrices includes Euclidean distance, Manhattan distance, Binary and Maximum distance. After this the linkage criterion determines the distance between each cluster. Average linkage is a possible choice as an agglomerative linkage criterion in which the distance between two clusters is the average distance between each datapoint in one cluster to every datapoint in the other cluster. The output of hierarchical clustering is usually presented in a dendrogram. The length of the branches usually represents the similarity between the samples. Dendrograms are often combined with heatmaps. Heatmaps are visual representations of data, in which each spot in the heatmap represents the value of the measured variable for each sample.

3.3.3 Kruskal-Wallis test

Kruskal-Wallis test is a rank-based non-parametric statistical test determining if there is a statistically significant difference between a categorical independent variable with two or more groups and a continuous variable. Non-parametric means that the test does not assume

normal distribution. A significant Kruskal-Wallis test indicates that there is a statistical difference between at least two of the groups, but it cannot imply which ones. Differences between two groups can be tested by a Mann-Whitney test.

3.3.4 Boxplots

Boxplots are non-parametric visual representations depicting groups of numerical data by their quartiles. In addition, boxplots may contain extended lines vertically from the boxes that indicates variability exterior the lower and upper quartiles. The line crossing the box is the median. Any outliers are commonly displayed as individual points in the boxplot. The interquartile range defined by the space between the first and third quartile in the boxplot indicates the degree of data skewness and dispersion.

3.3.5 Scatterplots

Scatterplots are mathematical diagrams used to display the relationship between two different quantitative variables. The relationship between each variable is displayed as points, in which the horizontal axis position is determined by one variable and the vertical axis position is determined by the other variable. Scatterplots can visually display various kinds of correlations between variables with a specified confidence interval.

3.4 Molecular subclassification of tumors into PAM50 subtypes

PAM50 molecular subclassification of tumor samples was performed by utilization of the R package *genefu* (function *molecular.subtyping*). The *genefu* R package uses gene expression data from the PAM50 gene set to identify the PAM50 subtype of the tumors(199).

3.5 Genome-wide correlation analysis

Genome-wide correlation analysis was performed using Pearson correlation to identify significant CpG-gene associations between the level of DNA methylation at CpGs and gene expression for the ER positive tumors with a luminal A and luminal B breast cancer disease from the OSL2 breast cancer cohort (n=177). Correlation coefficients were estimated using the R function *cor*. The significance of the correlation coefficients was calculated

independently and was based on t-distribution of the test statistic: $t = r\sqrt{n-2}/\sqrt{1-r^2}$ in which n is the number of samples and r is the correlation coefficient. The two-sided p-value was then calculated with the R function *pt*.

The genome-wide correlation analysis was not limited by any distance parameters, which means that CpGs could be associated with genes expressed *in cis* (same chromosome) or *in trans* (different chromosome). 173,654 CpGs with an interquartile range of more than 0.1 and all genes (18,551) were included in the analysis. An association was considered to be significant if the Bonferroni corrected p-value was less than 0.05 (nominal p-value < 1.55e-11). Validation of significant associations in OSL2 was performed by reanalyzing the associations in the ER positive tumor samples from the TCGA breast cancer cohort with luminal A and luminal B disease (n=304). Associations with a Bonferroni corrected p-value less than 0.05 (nominal p-value < 6.41e-08) were considered to be significant. Only associations confirmed in both datasets were included in the further analysis.

3.6 Hierarchical clustering analysis of emQTLs

Hierarchical clustering of the validated associations from the genome-wide correlation analysis was performed. Only CpGs and genes with one or more significant association were included in the analysis. The data matrix of p-values was converted to a binary form and clustered using binary as distance metric (function *designdist*, R package *vegan*) and average linkage as linkage criterion (function *hclust*).

3.7 Bi-cluster identification

DBSCAN is density-based clustering method that can be used to identify bi-clusters within a dataset. A cluster is a dense region of points surrounded by a low-density region of points. Low-density regions are required to separate clusters. Points in these sparse areas are usually considered to be noise or border points. The DBSCAN algorithm is based on connecting points within a distance threshold (epsilon) and will only connect points that satisfies a density criterion (minpts). A cluster is then defined as all density-connected points including points within the distance threshold. Bi-cluster identification was performed using the R package *DBSCAN*, setting the distance threshold to 15 and density criterion to 1.

3.8 Gene set enrichment analysis

Gene set enrichment analysis (GSEA) is a computational method used to identify classes of genes that are over-represented in a large set of genes, and that may be associated with a particular biological phenotype. Gene set enrichment analysis was performed using the molecular Signatures Database v6.0 (MSigDB; http://software.broadinstitute.org/gsea/index.jsp). Gene overlaps were computed against the hallmark gene set collection (H) and gene ontology (GO) gene set collection (C5). The C5 gene set collection is derived from gene ontology (GO) annotations based on GO terms and their association to human genes. GO annotations are statements that describes the function of specific genes, using concepts of Gene Ontology. The GO terms belong to one of three GO ontologies; biological process, molecular function and cellular component, and the collection is grouped into smaller GO collections accordingly. The hallmark gene set collection consists of a collection of 50 refined gene sets derived from many founder gene sets, each representing a specific biological state or process. Founder gene sets refers to the original overlapping gene sets in which the hallmark gene sets were derived from (C1-C6 collection) (200). MSigDB uses the hypergeometric distribution to calculate the probability of over-representation which is presented as p-values. In addition, false discovery rate (FDR) is estimated by multiple testing correction using the Benjamini and Hochberg method (201).

3.9 ChromHMM segmentation

ChromHMM is a multivariate hidden Markov model-based machine learning software that can be used to characterize chromatin states within the genomes of one or more cell types. Multiple chromatin datasets (e.g. ChIP-seq data) of different histone modification combinations associated with different chromatin states can be used by the software to learn to discover *de novo* re-occurring patterns in a cell type of interest. ChromHMM segmentation data for the MCF7 cell line was obtained from the work of Taberlay et al. (202) They collected the data by using ChIP-seq to generate signatures of key histone modifications (H3K4me1, H3K4me3, H3K27me3, H3K27ac) and regulatory factors (CTCF, RNA Pol II). They annotated the MCF7 genome into nine distinct chromatin states (heterochromatin, enhancer, enhancer + CTCF, promoter, promoter + CTCF, promoter_poised, repressed, transcribed and CTCF) based on the multivariate hidden Markov model (202).

In this study, the ChromHMM segmentation data was used to investigate whether the emQTL-CpGs were enriched within any particular regulatory region within the genome. This would suggest whether the emQTL-CpGs may have regulatory functions in the cell. Fold enrichment and statistical significance of CpGs in different genomic locations were calculated as the ratio between the frequency of emQTL-CpGs or CpGs in a cluster, located within a particular genomic segment type over the expected frequency for the same of CpGs from the Illumina HumanMethylation450k array or all hg19-CpGs. The significantce of the enrichment was estimated by hypergeometric test (R function *phyper*).

3.10 Heatmap generation of CpG-methylation profiles

Heatmap generation was accomplished using the R package *pheatmap* (203) to visualize the level of DNA methylation of the selected CpGs for the tumor samples. The CpGs were divided into two clusters; CpG-cluster A (mean methylation value>0.5; n=770) and CpG-cluster B (mean methylation value≤0.5; n=427). The CpGs in the rows of the heatmap were ordered such that the mean methylation value of CpG-cluster A CpGs decreased down the y-axis of the heatmap. A similar approach was applied to the columns of the heatmap in which the tumor samples were ordered by their mean methylation values of CpG-cluster A CpGs such that the mean methylation value decreased from the left of the heatmap towards the right side. The patients were then divided into three equally large patient groups: patient group A (n=59), patient group B (n=59) and patient group C (n=59).

3.11 ChIP-seq peaks enrichment analysis

ChIP-seq experiments was not performed in this thesis but was instead obtained from publicly available sources based on ChIP-seq experiments performed by other scientists. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a method that can be used for genome-wide mapping of DNA-protein interactions within any sequenced genome. This technique usually involves formaldehyde treatment of a desired cell type *in vivo* which leads to the reversible crosslinking of DNA-binding proteins to DNA. The cells are then lysed, and the chromatin is isolated and sheared into pieces of around 200-1000 bp in size (204). An antibody specific for the DNA-binding protein of interest can then be used to immunoprecipitated the DNA-protein complex. The crosslinks between DNA and protein is

reversed and the DNA is purified. The resulting DNA fragments can then be sequenced and aligned to a reference genome. A peak-calling algorithm can be used to identify areas within the genome that are enriched with aligned reads to determine the specific DNA-binding proteins binding loci (205). Mapping of DNA-protein interactions and epigenetic marks within the genome is essential for the understanding of transcriptional regulation. Precise maps for binding sites of DNA binding proteins such as TFs is important to elucidate the gene regulatory networks that underlies various biological processes (206).

ChIP-seq peak regions for the human reference genome hg19 is publicly available and was downloaded in narrowPeak format from the ReMap 2018 catalog (http://tagc.univmrs.fr/remap/). Data with merged peaks from 346 cell lines were utilized in this study. ChIP-seq peak enrichment was determined using the hypergeometric test (R function *phyper*) with Illumina Infinium HumanMethylation450 Bead Chip CpGs as background. False discovery rate (FDR) was estimated by the Benjamini-Hochberg method using the R function *p.adjust*.

3.12 Characterization of tumor samples using gene signatures

Development of gene expression signatures revealing the characteristics of biological samples is an area of active research today. In recent years, numerous gene signatures associated with different biological phenotypes and biological processes have been developed to differentiate tumors based on their gene expression.

3.12.1 EMT score

An EMT score was calculated using a pan-cancer EMT signature (207) to determine whether the tumors displayed epithelial-like or mesenchymal-like characteristics. The EMT signature consists of 77 genes (Appendix A, Table 6) of which 25 are associated with an epithelial-like phenotype and the remaining 52 genes are associated with a mesenchymal-like phenotype.

The EMT score was calculated by taking the mean expression of mesenchymal marker genes (M) subtracted by the mean expression of epithelial marker genes (E). A positive EMT score will therefore be associated with a mesenchymal phenotype, and a negative EMT score will be associated with an epithelial phenotype.

$$S_i = \frac{1}{|G_M|}\sum_{g \in G_M} e_{gi} - \frac{1}{|G_E|}\sum_{g \in G_E} e_{gi}$$

Formula 1. The EMT score (S_i) equation. Annotating the standardized expression for gene g in sample i to be e_{gi}, the set of M markers in the signature is G_M (Total of $|G_M|$ genes) and the set of E markers is G_E (total of $|G_E|$ genes).

3.12.2 Stemness score

A stemness score was calculated for each tumor sample by taking the mean expression of 11 genes known from the literature to be associated with a stem cell-like phenotype. This includes *NANOG* (208), *SOX2* (209), *POU5F1* (210), *BMI1* (211), *CD44* (212), *HOXB4* (213), *KIT* (214), *HOXA9* (215) , *HOXA10* (215), *MEIS1* (216) and *TIE2* (217). The mean expression values of the stemness genes for each tumor sample was then centered.

3.12.3 Proliferation score

A proliferation score was generated using the 11-gene proliferation score contained within the PAM50 assay by taking the mean expression for each tumor sample (218). These genes includes *BIRC5*, *CCNB1*, *CDC20*, *CDCA1*, *CEP55*, *KNTC2*, *MKI67*, *PTTG1*, *RRM2*, *TYMS* and *UBE2C* (219).

3.13 ChIA-PET data

Chromatin Interaction Analysis with Paired-End-Taq (ChIA-PET) sequencing is an analysis method that can be used for genome-wide mapping of long-rage chromatin interactions bound by protein factors (220). Mapping of such long-range chromatin loops may provide insight into transcriptional regulation of specific genes linked to human diseases. In this method, the DNA binding proteins are cross-linked to DNA by formaldehyde treatment prior to chromatin isolation. DNA fragments tethered to each chromatin complex are then connected with DNA linkers by proximity ligation before the Paired-End Tags are extracted and sequenced. The resulting ChIA-PET sequences are then mapped to a reference genome to identify remote chromosomal regions brought together by protein factors (220, 221).

ChIA-PET data for long range chromatin interactions in the MCF7 cell line was obtained from ENCODE (220). An emQTL was considered to be in a ChIA-PET loop if the genomic

distance between the CpG and the transcription start site of the gene was in the same genomic interval as for the corresponding Pol2 loops.

3.14 D492 and D492M cell lines

The human breast epithelial compartment comprises two distinct linages; the myoepithelial linage and the luminal epithelial linage. A putative precursor of the luminal epithelial compartment has been identified and isolated by Thorarinn Gudjonsson et al. (222). The D492 cell line is an immortalized human breast epithelial derived cell line with stem cell properties that can differentiate into luminal cells and myoepithelial cells in culture. Co-culture of D492 with breast endothelial cells have been demonstrated to generate spindle-like colonies of D492 cells with EMT-like characteristics such as enhanced ability to migrate (190, 223). Such spindle-like cells derived from D492 have been isolated by Sigurdsson et al. (190) and is referred to as the D492M cell line. D492M characteristics includes the lack of epithelial markers such as E-cadherin, keratins and the presence of mesenchymal markers such as N-cadherin, fibronectin, vimentin and FOXC2 (190). In addition, D492M display stem cell associated characteristics such as increased $CD44^{high}/CD24^{low}$ ratio, anchorage independent growth and increased resistance to apoptosis (190). The D492 and D492M cell lines were obtained from the Department of Tumor Biology at the Norwegian Radium Hospital.

3.14.1 Identification of candidate CpGs for pyrosequencing assays

The selection of candidate CpG targets for DNA methylation status analysis by pyrosequencing was based on the several criteria. (**1**) The CpG must be found in CpG-cluster A (**2**) and have a mean methylation value that is lower in patient group C compared to patient group A. (**3**) The CpGs must be located within enhancer regions defined by the ChromHMM data from the MCF7 breast cancer cell line (**4**) and be found within ChIP-seq peaks of TFs associated with EMT. In addition, (5) the CpG must be in long-range chromatin interaction loop (ChIA-PET Pol2 loops) with a gene associated with EMT. Genes were considered to be involved in EMT if they were present in the HALLMARK_EPITHELIAL_MESENCHYM-AL_TRANSITION gene set from the MSigDB v6.0 (http://software.broadinstitute.org /gsea/index.jsp) or from the publicly available Epithelial-Mesenchymal Transition Gene Database (dbEMT; http://dbemt .bioinfo-minzhao.org/index.html).

DNA sequences upstream and downstream from the candidate CpGs (±100 bp) in the human genome (GRCh37/ hg19) was obtained from the University of California Santa Cruz genome browser at http://hgdownload.soe.ucsc.edu/golden Path/hg19/chromosomes/ (Downloaded: 13. August 2018).

3.14.2 DNA isolation from the D492 and D492M cell lines

AllPrep DNA/RNA/Protein Mini Kit from QIAGEN was used to isolate genomic DNA from the human D492 and D492M cell lines. In total, 350 µl RLT buffer was added to the cell pellets for resuspension. The DNA isolation was performed according to the manufacturers protocol (December 2014). DNA yield was measured by NanoDrop One spectrophotometer (Nanodrop One, Thermo Fisher scientific).

3.14.3 Bisulfite conversion

Bisulfite sequencing is a method involving bisulfite conversion of DNA to determine the exact position of 5-methylcytosines. Bisulfite conversion is the process of converting cytosine to uracil in single stranded DNA treated with sodium bisulfite, in a process in which 5-methylcytosine residues remains unaffected. Subsequently, the DNA sequence of interest is amplified by PCR with primers specific to the bisulfite converted target sequence. QIAGEN EpiTect Fast Bisulfite Conversion Kit was used for complete bisulfite conversion and cleanup of 150 ng DNA. Following the manufacturers protocol (May 2012 edition) 40 µl DNA solution was added to the reaction mixture containing 85 µl bisulfite solution and 15 µl DNA protection buffer. Bisulfite conversion was performed using a thermal cycler (LifeECO Thermal Cycler, BIOER v2.01) according to the manufacturers protocol followed by DNA cleanup of the bisulfite converted DNA.

3.14.4 PCR

Polymerase chain reaction (PCR) is a method used to exponentially amplify specific segments of DNA. PCR is a highly sensitive method in which only trace amounts of DNA are needed to generate enough copies necessary to be analyzed in the lab. Each PCR assay requires template DNA, nucleotides (adenine, thymine, cytosine and guanine), primers and DNA polymerase enzymes. The components are mixed in PCR tubes and placed in a PCR machine (thermal

cycler) that amplifies DNA in three main steps. In the first step the reaction mixture is usually heated above the melting point of the complementary DNA strands of the target DNA, allowing double stranded DNA to denature. Then the temperature is lowered which allows the primers to anneal to the target complementary DNA segments. In the last step the temperature is raised to a temperature in which the DNA polymerase optimally can extend the primers by adding nucleotides to the new emerging DNA strand that are complementary to the DNA template strand. The number of copied DNA molecules rises exponentially for each repeated cycle (224).

PCR of bisulfite converted DNA was performed using the PyroMark PCR Kit from Qiagen. Following the manufacturers protocol (May 2009), $MgCl_2$ was added to the reaction mixture to a final concentration of 5.0 mM for each reaction. Primers for PCR were designed in PyroMark Assay Design software version 2.0 (Qiagen) using DNA sequences of ±100bp from the target CpGs in hg19 reference genome. The designed primer sequences among with the primer-specific temperature utilized during PCR can be found in Appendix B, Table 7. The rest of the program for the thermal cycler was set according to the manufacturers protocol.

3.14.5 Pyrosequencing

Pyrosequencing is a quantitative sequence-by-synthesis system commonly used for analyzing methylation status of CpG sites within an amplicon in real time. The method is based on the sequential addition of nucleotides to template DNA in a specific order. An apyrase enzyme is responsible for the continuous degradation of unincorporated nucleotides between the additions of a new nucleotides. A DNA polymerase catalyzes the addition of the nucleotides to template DNA. When correct nucleotide is added to the DNA template, a pyrophosphate is released and converted to ATP by an ATP sulfurylase enzyme. The luciferase enzyme converts ATP and luciferin to oxyluciferin in the presence of O_2 in a process that generates a light signal. The light intensity is proportional to the number of incorporated nucleotides. The methylated cytosine bases will correspond to the signal from C, and the extra thymidine bases will correspond to the unmethylated cytosines.

Primers for methylation analysis were designed in PyroMark Assay Design software version 2.0 (Qiagen) using DNA sequences of ±100bp from the target CpGs in hg19 human reference genome. The primer sequences used for pyrosequencing can be found in Appendix C, Table

7. Pyrosequencing was carried out in the PyroMark Q96 ID system using the PyroMark Q96 software version 2.5.8. 25µl amplified PCR product was added to each PCR-plate well containing a mixture of 2 µl Streptavidin Sepharose HP beads, 40 µl 1x Binding buffer and 13 µl Milli-Q H_2O. A master mix for all CpG assays was then prepared by adding 11.2 µl 1x Annealing buffer to 0.8 µl primer (10µM) and added to the pyrosequencing plate. The rest of the experiment was performed according to the PyroMark® Q96 ID User Manual 2016.

3.15 Tumor purity estimation by ASCAT

A tumor biopsy extracts tumor cells from the tumor itself, but also non-cancerous cells surrounding and infiltrating the tumor. Such non-cancerous components of the tumor may influence the data obtained from the tumor samples and can alter the biological interpretation of the results if not taken into consideration. In the recent years, several bioinformatical approaches have been developed to assess the purity of tumor samples, including Allele-Specific Copy Number Analysis of Tumors (ASCAT) (225-227).

ASCAT is a bioinformatical approach using single nucleotide polymorphism array data to dissect the allele-specific copy number of solid tumors, and at the same time estimating and adjusting for both non-aberrant cell admixture and tumor ploidy (226). This method allows the calculation of genome-wide allele specific copy-number profiles that reveals differences in aberrant tumor cell fraction, ploidy, losses, gains, copy number-neutral events and loss of heterozygosity. The ASCAT data used in this study to estimate tumor purity in the OSL2 breast cancer cohort was obtained from the work of Ragle Aure et al (228).

3.16 *In silico* nanodissection

The tumor microenvironment surrounding the tumor tissue consists of a heterogenous population of cell types infiltrating the tumor. Cell heterogeneity within a tissue may have a considerable effect on their gene expression profiles. Each cell type contains a wide diversity of cell types, each expressing a distinct repertoire of genes thereby making them different from one another. By identifying gene expression patterns unique to such infiltrating cells, one may be able to estimate the relative proportion of these cells in tumor tissue based on gene expression data from the bulk tumor.

In silico nanodissection algorithm v1.0 (http://nano.princeton.edu/) was used to predict lymphocyte infiltration. It is a genome-scale iterative machine learning approach used to predict human genes with cell-linage specific expression. The breast collection data (May 2013) containing 17,940 genes measured on 622 arrays was estimated for overlap with genes specifically expressed in lymphocytes (n=476), and not expressed in mammary epithelium (n=79) and mammary gland (n=777). Only genes with a probability of more than 65 % to be positive lymphocyte-specific standard genes as opposed to mammary epithelium and mammary gland were included for further analysis. Each OSL2 sample was scored for lymphocyte infiltration by the mean expression of the lymphocyte-specific genes.

3.17 CIBERSORT

In recent years several new improved deconvolution tools able to identify cell types within complex tissues have emerged with higher accuracy of prediction than before (229-232). One such deconvolution tool is Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts (CIBERSORT) which is a computational deconvolution tool used to estimate the relative fraction of diverse cell subsets. CIBERSORT uses gene signatures from the cell type of interest in addition to supervised learning frameworks through linear support vector regression to estimate the relative proportions of the cell type in the tumor tissue. Negative support vector regression coefficients are set to 0 while the remaining regression coefficients are normalized to sum to 1. The CIBERSORT software comes with one leukocyte gene signature matrix (LM22) on its own containing 547 genes that can distinguish 22 human hematopoietic cell phenotypes including natural killer (NK) cells, plasma cells, memory B cells, several types of T cells and myeloid subsets (233). The tool is publicly available and can be accessed from https://cibersort.stanford.edu/. CIBERSORT have been proven to outperform many other methods in respect to noise, closely related cell types, and unknown mixture content (233, 234).

CIBERSORT was performed to estimate the relative proportions of the 22 human hematopoietic cell types from the LM22 leukocyte gene signature matrix. The value of the relative cell type fraction was then correlated by Pearson's correlation with the mean methylation of the CpG-cluster A CpGs. False discovery rate was adjusted by Benjamini-Hochberg procedure.

3.18 Survival analysis in METABRIC

Survival analysis was performed by using survival data and gene expression data from patients from the METABRIC cohort. The overall survival was considered in regard to EMT score and mean expression of EMT-cluster genes. The luminal A and luminal B tumors were independently separated into two equally large groups based on the median of the variables.

4 Results

4.1 Identification and validation of significant CpG-gene associations

Genome-wide correlation analysis between the level of DNA methylation of CpG sites and gene expression in the OSL2 discovery cohort (n=177) lead to the discovery of 778,976 significant CpG-gene associations (emQTLs), in which 497,445 associations were validated in the independent TCGA breast cancer cohort (n=304). Due to missing CpG methylation values or gene expression values, 1,491 of the non-validated associations could not be tested. The validated associations included significant associations between the gene expression level of 2,991 genes and the DNA methylation level of 15,029 CpGs. The biological relevance of the emQTLs were investigated by hierarchical clustering. The clustering led to the discovery of two major clusters of significant CpG-gene associations (Figure 7). Cluster 1 consisted of 412 genes and 4,477 CpGs while Cluster 2 consisted of 453 genes and 1,197 CpGs.

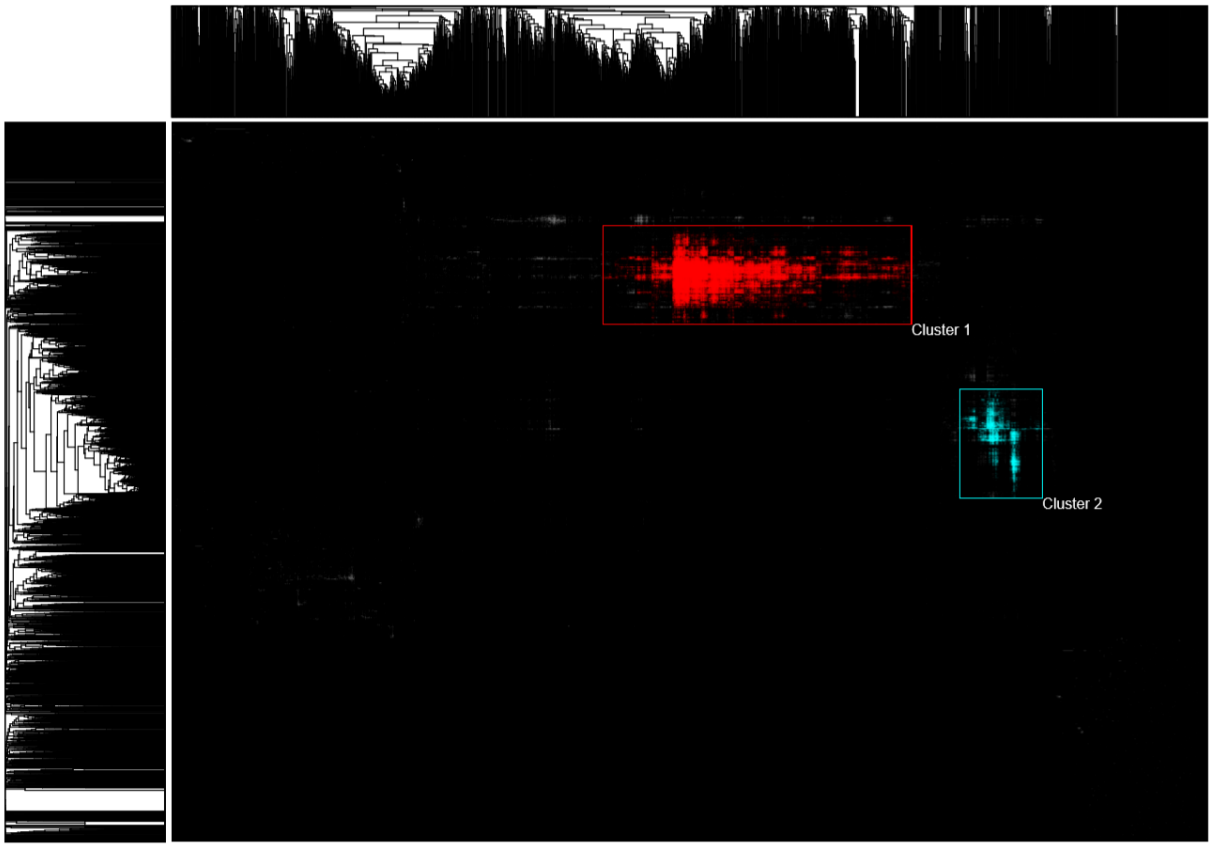


Figure 7. Unsupervised clustering of the Bonferroni corrected p-values from the genome-wide correlation analysis revealed two major clusters of CpG-gene associations. Rows represent genes (n=2,991) and columns represent CpGs (n=15,029). Colored and grey spots represent significant CpG-gene associations.

4.2 Biological characterization of the emQTL clusters

Gene set enrichment analysis showed high overlap between Cluster 1 genes and GO gene sets such as GO_IMMUNE_SYSTEM_PROCESS, GO_IMMUNE_RESPONSE, GO_REGULATION_ OF_IMMUNE_SYSTEM_PROCESS and GO_POSITIVE_REGULATION_OF_IMMUNE_ SYSTEM_PROCESS (Figure 8A). Cluster 2 genes were enriched in gene sets involved in EMT and cell-cell adhesion, such as GO_EXTRACELLULAR_MATRIX, HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION, GO_PROTEINACEOUS_EXTRACELLULAR_MATRIX and GO_EXTRACELLULAR_STRUCTURE_ORGANIZATION (Figure 8B). The immune cluster was first described by Fleischer, Tekpli et al. 2017 (7), and 83.2% of their immune cluster genes were found to overlap with the immune cluster discovered in this study. Therefore, the further project mainly focuses on the EMT-cluster.

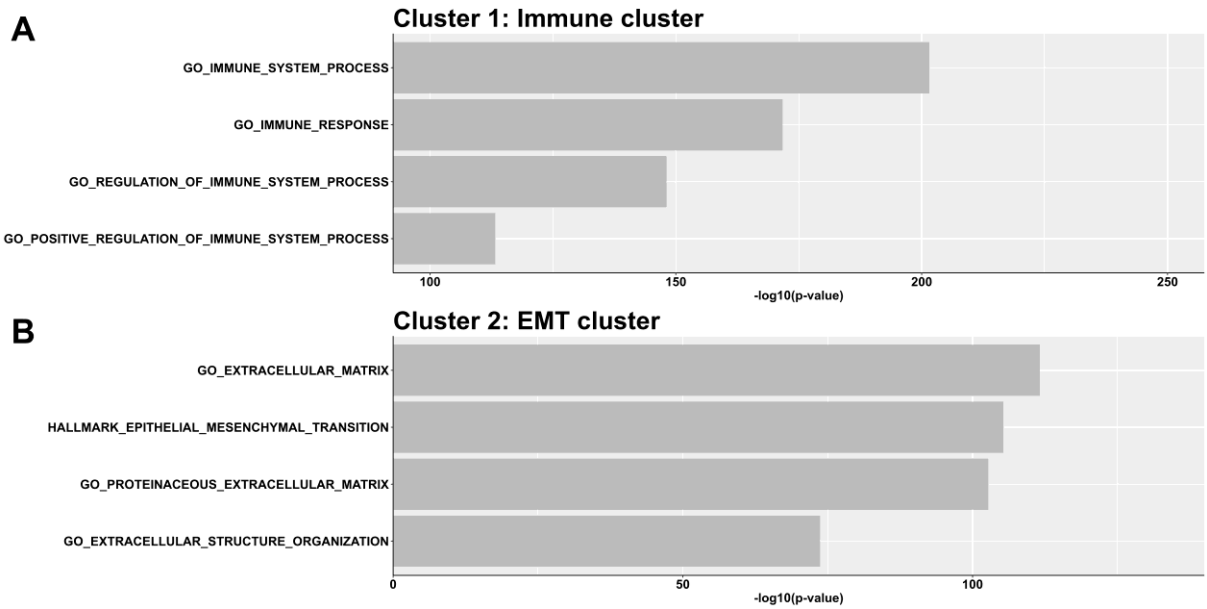


Figure 8. Characterization of the genes in the emQTL clusters. A Gene set enrichment analysis of the Cluster 1 genes (n=412) and the **B** Cluster 2 genes (n=453) using the Molecular Signatures Database (H and C5 gene set collections). The length of the bars represents the log-transformed Benjamini-Hochberg corrected p-values obtained by hypergeometric distribution.

4.3 Enrichment of emQTL-CpGs within ChromHMM-MCF7 regulatory regions

Further characterization of the EMT-cluster involved the utilization of ChromHMM segmentation data to investigate whether the emQTL- and EMT-cluster CpGs were enriched within any functional genomic region of the genome (Figure 9). The emQTL-CpGs were found to be enriched in CTCF+Enhancer, heterochromatin, transcribed and enhancer regions(p-values=1.05e-06, 2.46e-23, 5.76e-29 and <1.0e-30 respectively), while the EMT-cluster CpGs were significantly enriched in heterochromatin, transcribed and enhancer regions (p-values=7.28e-14, 6.33e-6 and 3.41e-22 respectively).

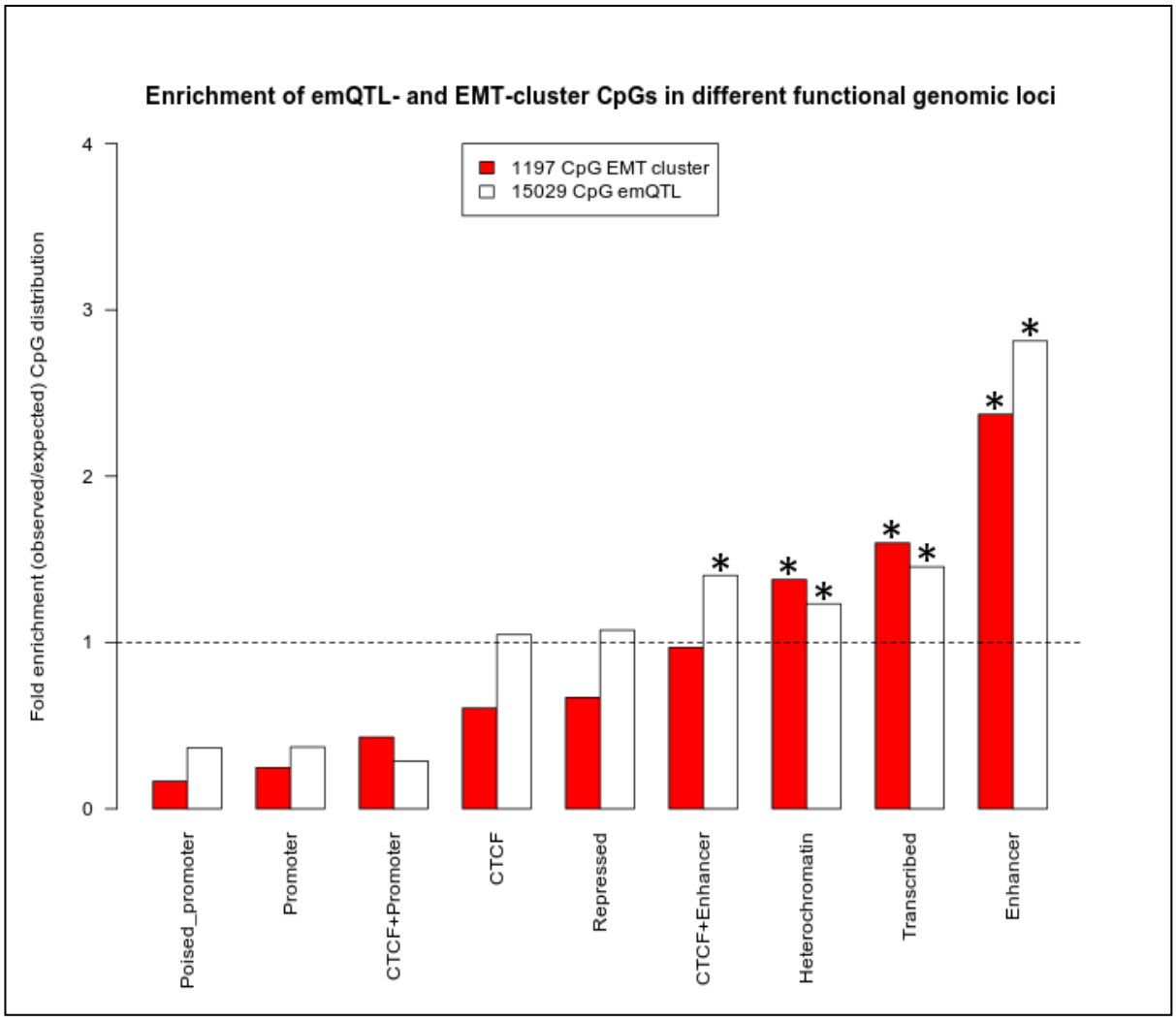


Figure 9. Enrichment of emQTL-CpGs within ChromHMM regulatory regions. Barplot showing the fold enrichment of the emQTL-CpGs and EMT-cluster CpGs in ChromHMM annotated genomic regions. Fold enrichment was calculated as the ratio between the frequency of emQTL-CpGs or EMT-cluster CpGs, located within a particular segment type over the expected frequency for the same of CpGs from the Illumina HumanMethylation450 array or all hg19-CpGs. Statistically significant enrichment ($p<0.05$) was determined by hypergeometric test and is marked with an asterisk.

4.4 Methylation profiles of the EMT-cluster CpGs in OSL2

A heatmap was generated to visually assess the DNA methylation level of the EMT-cluster CpGs for the 177 OSL2 ER positive luminal tumors (Figure 10A). A significant difference in the mean methylation of the CpG-cluster A CpGs and CpG-cluster B CpGs were observed between patient group A, B and C (Figure 10B-C). The same methylation pattern was observed for the same CpGs in the ER positive breast tumors from the TCGA breast cancer cohort (Appendix C, Figure 17A) and the difference in mean methylation of CpG-cluster A CpGs and CpG-cluster B CpGs between patient groups were significant (Appendix C, Figure 17B-C).

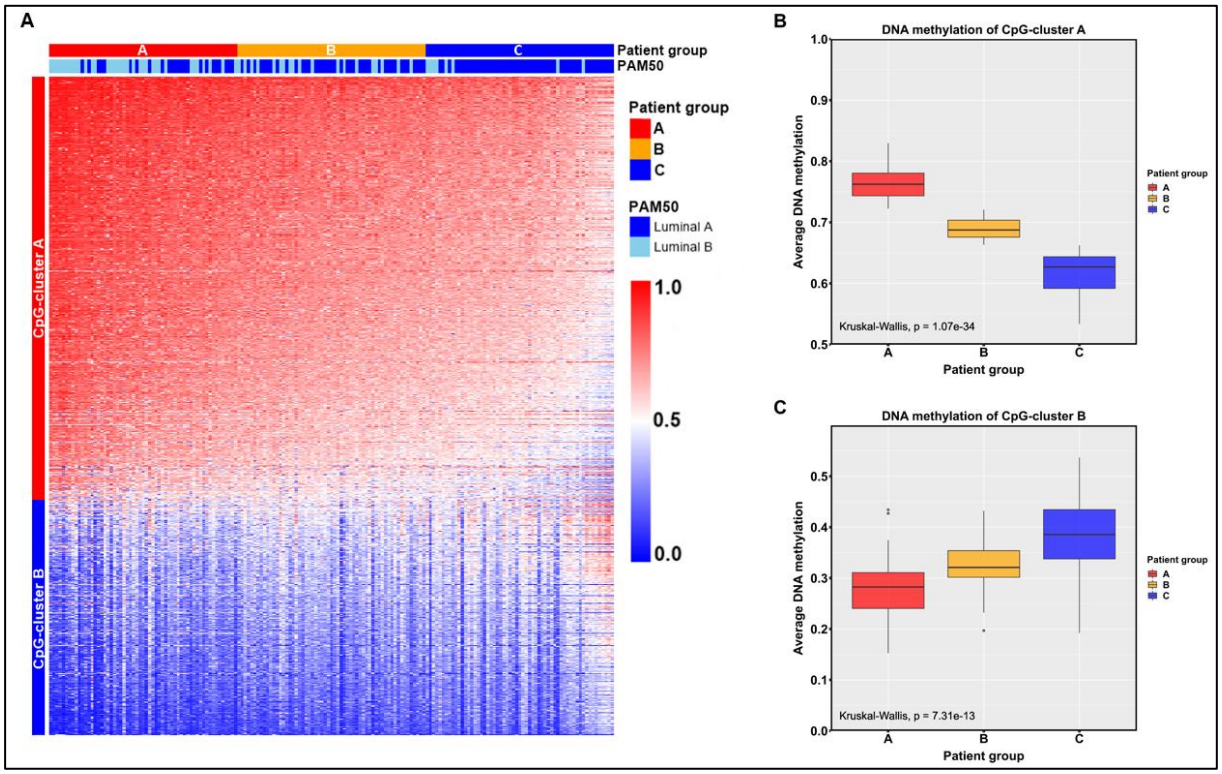


Figure 10. A DNA methylation level of the 1,197 EMT-cluster CpGs for the 177 ER positive breast tumor samples from OSL2. The columns show the tumor samples annotated with PAM50 subtype and patient group. The rows represent the EMT-cluster CpGs which are annotated as either CpG-cluster A CpGs (mean methylation value>0.5; n=770) or CpG-cluster B CpGs (mean methylation value<0.5; n=427). The CpGs in the rows of the heatmap were ordered such that the mean methylation value of CpG-cluster A CpGs decreased down the y-axis of the heatmap. The columns were ordered by the mean methylation of CpG-cluster A CpGs such that the value decreased from the left side of the heatmap towards the right side of the heatmap. Red dots represent methylation values close to 1 while blue spots represent methylation values close to 0. White spots represent an intermediate value close to 0.5. Figure **B** and **C** shows boxplots of the mean DNA methylation of CpG-cluster A CpGs and CpG-cluster B CpGs respectively, in patient group A, B and C.

4.5 ChIP-seq peaks enrichment analysis of CpG-cluster A and CpG-cluster B CpGs

ChIP-seq peaks enrichment analysis using ChIP-seq data revealed that CpG-cluster A CpGs are enriched within ChIP-seq peaks of TFs associated with EMT, such as TEAD1, FOSL1, TWIST1, SIX2, YAP1 and PPARG (Table 2). The most significantly enriched ChIP-seq peaks around the CpG-cluster B CpGs were associated with TFs such as FOXA1, GATA3, TLE3 and among other TFs such as ESR1 and PGR (Table 3).

Table 2. TFs with binding regions enriched at CpGs in CpG-cluster A. q represents the number of TF binding regions that overlaps with CpGs in CpG-cluster A and m represents the number of TF binding regions that overlap with all CpGs in the Illumina Human methylation 450K array. The rows are ordered by p-value and TFs with a fold enrichment less than 1.5 are not included.

TF	q	m	Fold enrichment	P-value	FDR corrected p-value
TEAD1	92	17472	3.32	7.78e-24	3.27e-21
NFIC	143	41072	2.20	2.49e-19	5.22e-17
FOXO1	137	46135	1.87	4.05e-13	4.25e-11
TP73	30	4264	4.44	6.17e-12	5.19e-10
FOSL1	88	26009	2.13	1.55e-11	1.04e-09
MYOD1	96	29569	2.05	1.74e-11	1.04e-09
CEBPB	212	88548	1.51	9.00e-11	4.73e-09
TWIST1	118	43548	1.71	4.89e-09	2.28e-07
OTX2	11	1033	6.71	1.55e-07	6.51e-06
SIX2	37	10848	2.15	7.55e-06	2.64e-04
LHX2	12	1893	4.00	1.49e-05	4.83e-04
YAP1	39	12357	1.99	2.58e-05	7.75e-04
PPARG	70	27058	1.63	3.03e-05	8.49e-04

Table 3. TFs with binding regions enriched at CpGs in CpG-cluster B. q represents the number of TF binding regions that overlaps with CpGs in CpG-cluster B and m represents the number of TF binding regions that overlap with all CpGs in the Illumina Human methylation 450K array. The rows are ordered by p-value and TFs with a fold enrichment less than 1.5 are not included.

TF	q	m	Fold enrichment	P-value	FDR corrected p-value
FOXA1	298	147683	2.29	1.56e-63	6.62e-61
GATA3	212	83908	2.87	1.39e-53	2.94e-51
TLE3	41	1554	30.00	7.83e-48	1.11e-45
FOXA2	180	79747	2.57	8.06e-37	8.55e-35
GATA2	201	98429	2.32	6.66e-36	5.65e-34
NRIP1	105	30339	3.94	9.56e-35	6.75e-33
PGR	179	83218	2.45	7.96e-34	4.82e-32
AR	284	187912	1.72	1.02e-31	5.43e-30
ESR1	222	127133	1.99	2.58e-30	1.21e-28
AHR	63	12175	5.88	4.46e-30	1.89e-28
ZNF217	26	1483	19.93	1.28e-26	4.94e-25
DAXX	69	18219	4.31	6.04e-25	2.14e-23
PIAS1	127	56576	2.55	2.60e-24	8.47e-23

4.6 EMT- and stemness score associated with CpG-cluster A methylation

The link between DNA methylation of CpG-cluster A and tumor phenotype in relation to EMT was investigated by utilizing an EMT score and a stemness score and correlating them with the mean methylation value of CpG-cluster A CpGs. The EMT score was found to be correlated with the mean methylation of CpG-cluster A CpGs (Figure 11A, p-value = 8.9e-12), the same was true for the stemness score (Figure 11B, p-value = 2.5e-07). In addition, the EMT score was significantly higher in the luminal A tumors compared to luminal B tumors (Figure 11C, p-value=0.00099).

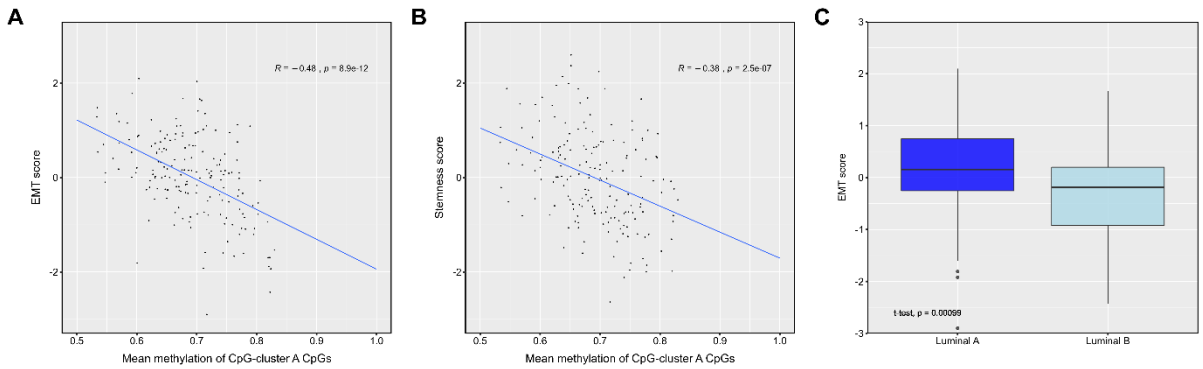


Figure 11. Scatterplots showing the Pearson's correlation between the mean methylation of CpG-cluster A CpGs versus EMT score **A** and stemness score **B**. **C** shows a boxplot of the overall difference in EMT score between the luminal A (n=115) and luminal B subtype (n=62).

4.7 Proliferation score and *ESR1* expression associated with CpG-cluster A methylation

From current literature one knows that the Luminal B subtype is mainly distinct from the luminal A subtype when it comes to proliferation. The heatmap showed that there was an uneven distribution of the tumor subtypes across the x-axis (Figure 10A). In addition, the luminal A tumors were shown to have a significantly higher EMT score than luminal B tumors, and many studies have suggested that EMT attenuates proliferation. Therefore, it was reasonable to investigate the correlation in mean methylation of CpG-cluster A with the proliferative phenotype of the tumor samples as well. The proliferation score was correlated with the mean methylation of CpG-cluster A CpGs. A significant positive correlation was observed (Figure 12A, p-value = 2.5e-06). The expression of *ESR1* was also correlated with the mean methylation of CpG-cluster A CpGs as it is known to be a contributor to tumor cell

proliferation in ER positive breast cancers. The mean methylation of CpG-cluster A CpGs was positively correlated with the expression of *ESR1* (Figure 12B, p-value=2.8e-05).

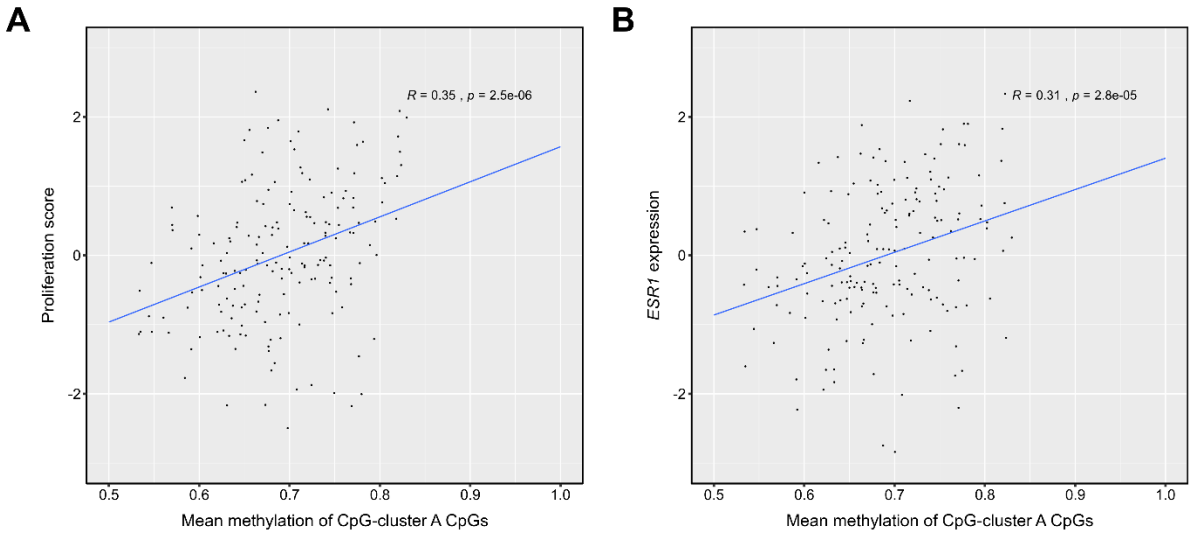


Figure 12. Scatterplots showing the Pearson correlation between the mean methylation of CpG-cluster A CpGs versus proliferation score **A** and *ESR1* expression **B**.

4.8 ASCAT and *in silico* nanodissection associated with CpG-cluster A methylation

Considering the DNA methylation data, several cell types within the tumor sample may contribute to the resulting methylation value. The infiltration of non-tumor cells such as lymphocytes may differ significantly from one tumor to another, and the quantity of these non-tumor cells may lead to an altered methylation value of CpGs that does not represent the signal from the tumor itself. In order to take this into account in the analysis, tumor purity from ASCAT was correlated with the mean methylation of CpG-cluster A CpGs (Figure 13A). A significant but low correlation between the mean methylation of CpG-cluster A CpGs and tumor purity was observed (p-value=0.00063). The estimated lymphocyte infiltration in the tumor samples based on the expression of lymphocyte specific genes predicted by *in silico* nanodissection showed no significant correlation between mean methylation of CpG-cluster A CpGs (Figure 13B, p-value=0.36).

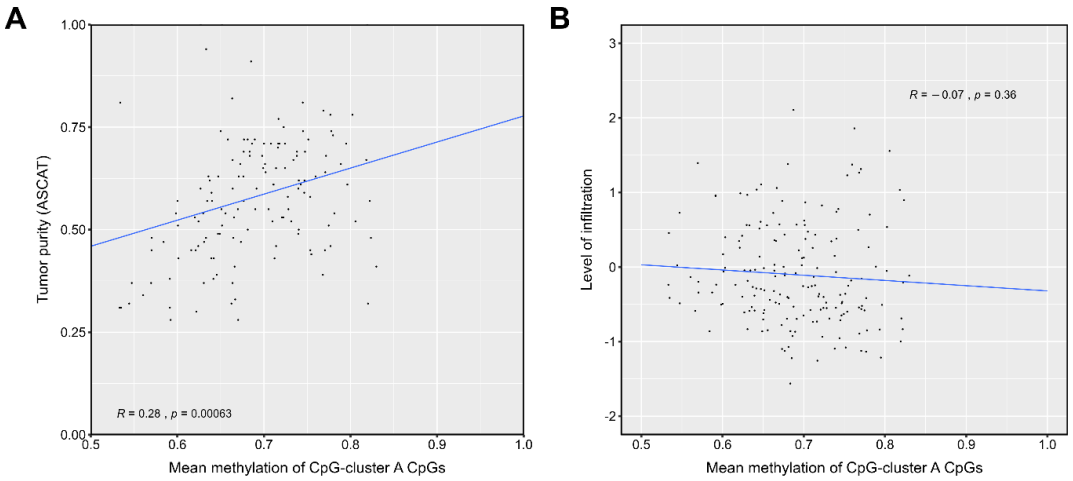


Figure 13. Scatterplots showing the correlation between tumor purity **A** and the level of lymphocyte infiltration **B** with respect to mean methylation of CpG-cluster A CpGs.

4.9 Hematopoietic cell type composition associated with CpG-cluster A methylation

CIBERSORT was performed to estimate the relative proportion of 22 human hematopoietic cell types from the LM22 leukocyte gene signature matrix and correlated using Pearson correlation with the mean methylation of CpG-cluster A CpGs (Table 4). This could indicate whether leucocyte infiltration potentially could be a factor contributing to the reduced methylation of CpG-cluster A CpGs and the EMT-phenotype. The results showed a significant negative correlation between the relative proportion of resting CD4+ memory T cells, resting mast cells and monocytes with the mean methylation of CpG-cluster A CpGs (Table 4, p-value=7.31e-08, 9.03e-04 and 3.08e-03 respectively). In addition, a significant positive correlation between these variables were discovered for activated mast cells and resting NK cells and neutrophils. The CIBERSORT estimate of relative proportion of the 22 human hematopoietic cell types in the OSL2 tumors can be found in Appendix D, Table 8.

Table 4. Overview of the significant Pearson correlations between the relative proportions of the human hematopoietic cell types correlated with mean methylation of CpG-cluster A. The correlation coefficient and p-values are annotated in addition to the FDR corrected p-values obtained by Benjamini-Hochberg procedure.

Cell type	Correlation coefficient	P-value	FDR corrected p-value
T cells CD4+ memory resting	-0.39118	7.31e-08	1.61e-06
Mast cells resting	-0.24735	9.03e-04	9.93e-03
Mast cells activated	0.22124	3.08e-03	2.26e-02
NK cells resting	0.20811	5.44e-03	2.83e-02
Monocytes	-0.19964	7.72e-03	2.83e-02
Neutrophils	0.20010	7.58e-03	2.83e-02

4.10 Survival analysis

Univariate survival analysis was performed in METABRIC by using Kaplan-Meier modeling to investigate the link between the EMT score and mean expression of the EMT-cluster genes, with overall survival. In the first analysis overall survival was considered in respect to the EMT score of the patient tumors. No statistical significant correlation within the luminal A subtype was found (Figure 14A, p-value=0.86), but a significant difference in overall survival was observed in the luminal B (Figure 14B, p-value=0.034). In the second analysis, overall survival was investigated regarding mean expression of the EMT-cluster genes. No significant difference in overall survival was observed within the luminal A- or luminal B subtypes (Figure 14C-D, p-value=0.75 and 0.091 respectively), but the survival trend was similar to that observed for the EMT score. The data utilized for the survival analysis in METABRIC can be found in Appendix E, Table 9.

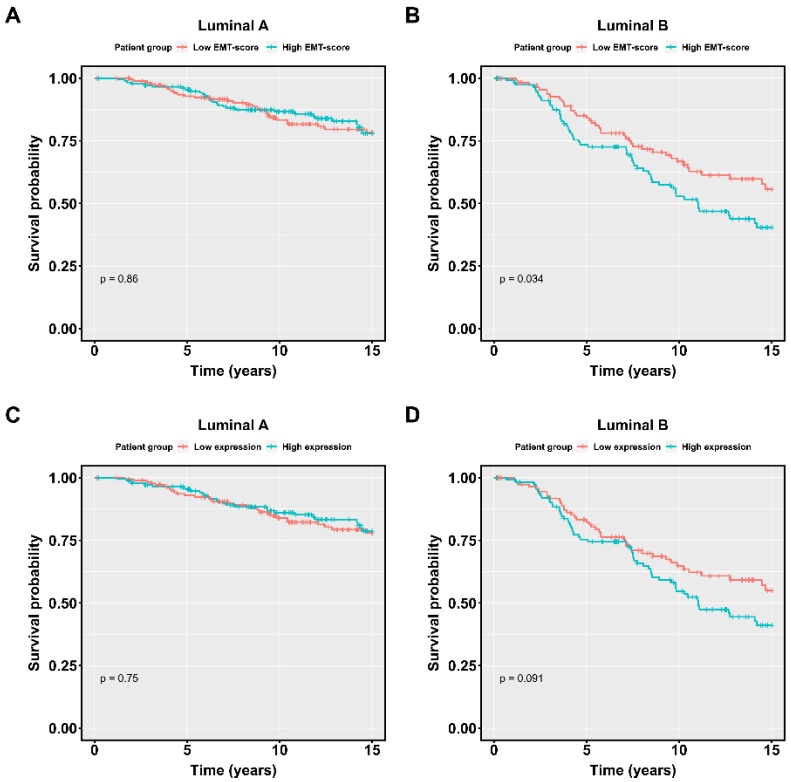


Figure 14. Survival analysis in METABRIC. A and **B** shows the difference in overall survival of luminal A (**A,** n=354) and luminal B (**B**, n=230) patients in respect to the EMT score. The two patient groups reflect whether the EMT score for the tumor sample was above or beneath the EMT score median. **C**, **D** displays the difference in overall survival in luminal A (**C**, n=354) and luminal B (**D**, n=230) patients based on their mean expression of the EMT cluster genes. The patients were divided into two groups based on whether their mean expression of EMT cluster genes was lower or higher than the median.

4.11 Generation of a correlation matrix

To summarize the *in silico* findings, a correlation matrix displaying the correlations between the main variables considered in this thesis was constructed (Figure 15). The plot presents the correlation between the variables: mean methylation of CpG-cluster A CpGs, proliferation score, *ESR1* expression, tumor purity, stemness score, EMT score and the level of lymphocyte infiltration. Some of the correlations with the highest correlation coefficient include the negative correlation between mean methylation of CpG-cluster A CpGs with EMT score, and stemness score, the negative correlation between EMT score and proliferation score, and the negative correlation between tumor purity and lymphocyte infiltration. A positive correlation was observed between the EMT score and the stemness score. An overview of data used to generate the correlation matrix can be found in Appendix F, Table 10.

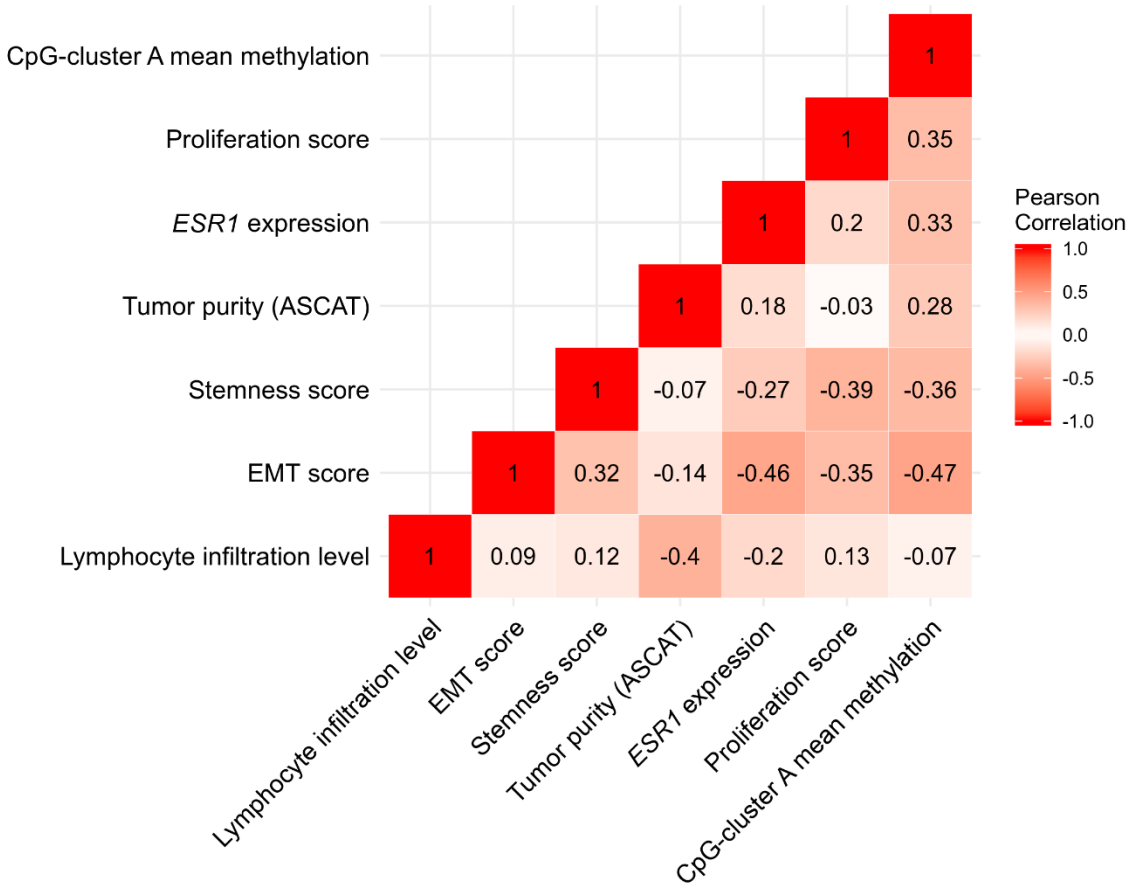


Figure 15. Correlation matrix showing the correlation between the main variables considered in this thesis. Rows and columns indicate variable type. Each number in the boxes display the correlation coefficient for each pair of variables independent of whether the correlation is positive or negative. A red color indicates high correlation, while a white color indicates no or low correlation.

4.12 Identification of differentially methylated CpGs in ChIA-PET Pol2 loops

As observed from the previous results, the CpGs in CpG-cluster A were enriched within enhancers and ChIP-seq peaks of TFs associated with EMT. Moreover, the methylation of these CpGs were lower in patient group C compared to patient group A. The EMT score was also correlated with mean methylation of CpG-cluster A CpGs. It was therefore of interest to identify potential CpGs in which the level of methylation was associated with EMT.

Several of the CpGs in the EMT cluster were found to be located near the binding sites of EMT-related TFs and to be in ChIA-PET Pol2 loops with genes known to promote EMT in cancer cells. In addition, all CpGs identified were located within ChromHMM-MCF7 enhancers. Six of these CpGs were selected from CpG-cluster A CpGs and are illustrated in Table 5. The methylation values for each CpG was correlated with the expression of the looped gene.

Table 5. The table shows the CpGs selected from CpG-cluster A that had a reduced methylation in patient group C compared to patient group A. TFs with DNA binding sites found to overlap the specific CpG determined from ChIP-seq data is also annotated in addition to which gene it is looped to. The p-values included are obtained by correlating the methylation level of the CpGs with the expression level of the genes they are in ChIA-PET loops with. FDR corrected p-values were estimated using Benjamini-Hochberg correction.

Transcription factor	Probe	Gene in ChIA-PET loop	Mean methylation difference	p-value	FDR corrected p-value	Correlation coefficient
CEBPB	cg06947286	*PDLIM4*	0.216	2.65e-08	1.50e-07	-0.4031
FOSL1	cg06947286	*PDLIM4*	0.216	2.65e-08	1.50e-07	-0.4031
PPARG	cg06947286	*PDLIM4*	0.216	2.65e-08	1.50e-07	-0.4031
TWIST1	cg10233454	*LRP1*	0.177	3.00e-07	1.28e-06	-0.3737
TEAD1	cg20909017	*ITGA5*	0.138	7.44e-04	2.11e-03	-0.2512
CEBPB	cg16888565	*TPM1*	0.205	1.00e-03	2.13e-03	-0.2452
TWIST1	cg16888565	*TPM1*	0.205	1.00e-03	2.13e-03	-0.2452
CEBPB	cg12232146	*PHLDA1*	0.136	1.05e-01	1.28e-01	-0.1222
FOSL1	cg12232146	*PHLDA1*	0.136	1.05e-01	1.28e-01	-0.1222
PPARG	cg12232146	*PHLDA1*	0.136	1.05e-01	1.28e-01	-0.1222
TEAD1	cg12232146	*PHLDA1*	0.136	1.05e-01	1.28e-01	-0.1222
YAP1	cg12232146	*PHLDA1*	0.136	1.05e-01	1.28e-01	-0.1222
CEBPB	cg05223441	*VEGFA*	0.161	2.54e-01	2.54e-01	0.0862
TEAD1	cg05223441	*VEGFA*	0.161	2.54e-01	2.54e-01	0.0862
YAP1	cg05223441	*VEGFA*	0.161	2.54e-01	2.54e-01	0.0862

4.12.1 Identification of differential methylation in the D492 and D492M cell line by pyrosequencing.

The methylation status of the differentially methylated CpGs in Table 5 were investigated in two different cell lines; the D429 with epithelial characteristics and D492M with mesenchymal characteristics. Differential DNA methylation of the target CpGs was investigated by pyrosequencing. The results are summarized in Figure 16. No prominent difference in DNA methylation of target CpGs were observed between the D492 and D492M cell line. Cg06947286, cg05223441, cg12232146, cg16888565 and cg20909017 seems to be unmethylated in the majority of DNA in the samples for both cell lines. The most pronounces difference in methylation of the target CpGs in the cell lines is observed at cg10233454 and cg20909017. Cg10233454 was the only methylated CpG, in both cell lines. Appendix G, Table 11 contains the pyrograms for each target CpG.

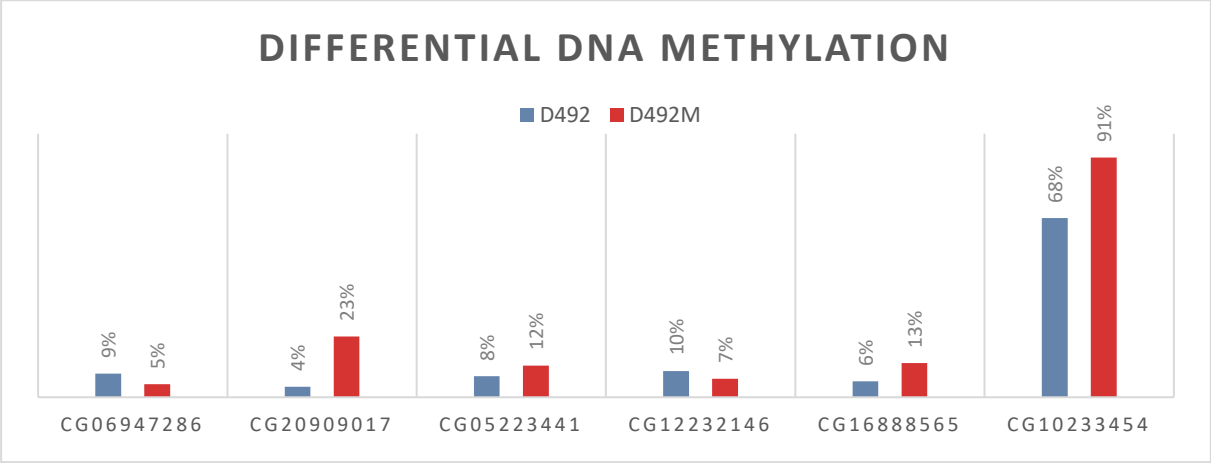


Figure 16. DNA methylation at the target CpGs. The figure shows the level of methylation for each target CpG selected in the D492 (blue bars) and the D492M (red bars).

5 Discussion

EMT is believed to play a key role in cancer progression in both pre-invasive and invasive state. However, much of the work on studying EMT in breast cancers up to date have mainly focused on the non-luminal subtypes and metastatic cancers such as the basal-like breast cancers. These are considered to be among the most aggressive and deadly breast cancers and tend to display mesenchymal-like characteristics, higher chemotherapy resistance and higher abilities to metastasize compared to the other breast cancer subtypes (77, 169). The reversibility and plasticity of EMT and MET suggest that epigenetic changes may play a pivotal role in this process, but only a limited number of studies have interpreted the dynamics of EMT-related epigenetic alterations. In this study we were able to identify alterations in DNA methylation associated with EMT in ER positive breast cancers *in silico*. These findings are important in order to understand the biological process underlying EMT and provide insight into the role of EMT in breast cancer pathogenesis. The identification of these EMT-related epigenetic alterations may open for future possibilities of epigenetic therapy for mesenchymal-like tumors.

5.1 Biological considerations

Genome-wide emQTL analysis have previously been shown to identify significant correlations between the level of DNA methylation at CpG sites and gene expression due to intertumoral heterogeneity within ER positive and ER negative breast tumors and to be a valuable tool in the identification of key gene regulatory networks involved in breast cancer pathogenesis (7). To take this further, the same approach was applied to the ER positive breast tumors only, to investigate whether any differences within the ER positive tumors in respect to DNA methylation and gene expression could be observed. By testing all possible Pearson's correlations between DNA methylation level of CpGs and gene expression prior to hierarchical clustering of the associations, we were able to discover two major clusters associated with two very distinct biological processes. Both processes are well known to be involved in breast cancer pathogenesis. One of the clusters were enriched for genes associated with immune response, while the other was associated with genes involved in EMT. The major cluster of interest for this study was the EMT-cluster as the immune cluster had already been described by Fleischer, Tekpli et al (7).

Several papers published in recent years have elucidated the link between aberrant DNA methylation at enhancers to tumor progression and plasticity (6, 144, 145). Interestingly, the EMT-cluster CpGs were found to be significantly enriched in distal regulatory regions in the genome overlapping with ChromHMM-MCF7 enhancers. A further subdivision of the EMT-cluster CpGs into two clusters in respect to DNA methylation also showed significant differences in average DNA methylation between the patient groups, thereby highlighting these CpGs as putative drivers of carcinogenesis within a subgroup of ER positive breast cancers.

The differential methylation of the CpG-cluster A and CpG-cluster B CpGs observed between the patient groups raised two major hypotheses. The first hypothesis was that the difference in DNA methylation was a result of tumor infiltration by non-tumor cells such as immune cells. When considering DNA methylation and gene expression data it is important to keep in mind that tumors are surrounded by a highly heterogenous populations of non-tumor cells such as immune cells. The degree of infiltration by non-tumor cells such as immune cells are known to vary from tumor to tumor and is also known to be subtype specific (235). Cell types infiltrating the tumor may affect the quantitative data obtained from tumor biopsies such as DNA methylation and gene expression data. For instance, if some cell-types infiltrating the tumor were unmethylated at CpG-cluster A CpGs or methylated at CpG-cluster B CpGs, this may affect the overall DNA methylation level of these CpGs depending on the extend of the infiltration. The ASCAT data showed a low correlation between tumor purity and the mean methylation levels of the CpG-cluster A CpGs, and *in silico* nanodissection showed no correlation at all. However, *in silico* nanodissection can only estimate lymphocyte infiltration, but it cannot imply which cell type. This again, may affect the output. To study even more cell types *in silico*, CIBERSORT was performed. This deconvolution tool has been shown to outperform many other methods in respect to unknown mixture content, noise and closely related cell types. Interestingly, the relative proportion of resting memory T cells, resting mast cells and monocytes were correlated with the mean methylation of CpG-cluster A CpGs, thereby indicating that these cell types may potentially affect the methylation level of the CpG-cluster A CpGs.

The second hypothesis is that the difference in DNA methylation was a result of tumor cell heterogeneity, in which there exists different populations of tumor cells within the tumor that are differentially methylated at the EMT-cluster CpG sites. This theory implies that the extend

of the reduction in DNA methylation observed between the patient groups in the CpG-cluster A CpGs is relative to the proportion of the tumor that is demethylated at these CpG sites.

The previous results showed an enrichment of the EMT-cluster CpGs in ChromHMM-MCF7 enhancers. One way the EMT-cluster CpGs may contribute to tumor progression and plasticity is through the regulation of TF binding to their associated enhancers, which again may regulate the expression of genes associated with EMT. Enhancers are known to regulate gene expression through the binding of cell type specific TFs that can recognize specific DNA sequences within the enhancers (146). TF binding to enhancers is also known to be influenced by DNA methylation at TF binding sites (148). However, little is known about the causality between DNA methylation and TF binding, and the answer seems to be more complex than previously anticipated by scientists. According to the traditional view, TFs tend to bind non-methylated DNA motifs in open chromatin regions. *In vivo* experiments investigating the association between hypermethylation and TF binding has shown that this tend to be the case for most TFs (148, 236, 237). Methylation of DNA itself may function as a physical barrier of TF binding, and even more important, affect chromatin organization through interaction with other factors associated with histone modifications, polycomb complexes, nucleosome positioning and chromatin remodeling proteins (119, 124). DNA methylation have previously been reported to induce compactization and increased rigidity of DNA which may suppress nucleosome structure dynamics and consequently lead to reduced TF accessibility and gene silencing (113, 125, 126). Alternatively, specific proteins called methyl-CpG binding domain proteins can bind methylated DNA motifs either in a sequence dependent or sequence independent fashion, thereby competing off TFs by their higher affinity to methylated CpGs (124, 148). Newly emerging scenarios challenges this view by suggests that some TFs lacking methyl-CpG binding domains are able to interact with methylated DNA (238-241). However, very little is still known about TFs with affinity for methylated DNA. Only a few TFs have been published to be implicated in this scenario so far (148).

In the subsequent analysis, enrichment of ChIP-seq peaks within the EMT-cluster CpGs was assessed. Such an analysis could indicate whether the EMT-cluster CpGs were targets of TFs associated with specific cellular states or processes. Interestingly, the ChIP-seq peaks enrichment analysis revealed that CpG-cluster A CpGs were enriched within binding sites of TFs associated with EMT such as TEAD1, FOSL1, TWIST1, SIX2, YAP1 and PPARG while CpG-cluster B CpGs were enriched within the binding sites of FOXA1, GATA3 among other

TFs such as ESR1. ESR1, GATA3 and FOXA1 are TFs already known to play key roles in breast cancer pathogenesis (160, 162). They are considered to be within the same interacting pathway and are major drivers of growth in ER positive breast cancers (160).

To elucidate the role of DNA methylation and TF activity in context of tumor phenotype, several scoring systems were applied based on gene signatures associated with specific cellular states or processes. Many of the TFs associated with CpG-cluster A binding were related to EMT while several of the most enriched TFs associated with CpG-cluster B binding were involved in cell proliferation. The EMT score and proliferation score were used to investigate these two features further. Intriguingly, the EMT score was positively correlated with the mean methylation of CpG-cluster A CpGs. This may suggest that reduced methylation at these sites allows access of EMT-related TFs to bind to the enhancers leading to the induction of transcription of EMT-related genes. Moreover, this EMT-feature was significantly more prominent in the luminal A subtype compared to the luminal B. EMT is a process well known to be linked with a stemness phenotype in human tumors, but still, little is known about the mechanisms wiring these two mechanisms together (172, 173). A stemness score was made as a second layer of proof for the EMT-phenotype concept and it could also support the second hypothesis; that the change in mean methylation of CpG-cluster A CpGs and CpG-cluster B CpGs is due to tumor heterogeneity. Consistent with this hypothesis and current literature, the EMT score and stemness score were positively correlated. The stemness score was also correlated with the mean methylation of CpG-cluster A CpGs.

What is interesting about these result is that luminal breast tumors in general are considered to be among the most differentiated tumors as they commonly are derived from more committed progenitor cells compared to the more mesenchymal-like and aggressive basal-like tumors (73, 74). However, several studies have reported evidences of dedifferentiation to occur in luminal tumors during breast cancer progression. One study highlighted EMT as a possible mechanism behind the dedifferentiation observed in several breast cancer tumors (242). In yet another study a dedifferentiation-like process was observed in which a part of a previously luminal tumor transformed into a basal-like carcinoma (ER-/PR-/HER2-) with myoepithelial characteristics (243). Since breast cancers are driven mainly by aberrant hormone-dependent pathways, some studies have the recent years investigated whether the loss of estrogen receptor expression may result in dedifferentiation from an epithelial to a mesenchymal phenotype in ER positive breast cancers (244, 245). Interestingly, siRNA-mediated silencing

of ER in the MCF7 cell line have been observed to promote morphological changes, increased motility and increased expression of vimentin (244). This is highly consistent with our results, as the expression of *ESR1* was found anticorrelated with the EMT- and stemness score.

Accumulating evidence suggests that proliferation and EMT are antagonistic features. EMT have been shown to attenuate proliferation in many, but not all systems (185-189). This appears in a recent published paper demonstrating that breast cancer stem cells located at the invasive part of the tumor primarily are quiescent, while the more central regions of the tumors are proliferative and retain the ability to transit between these two states (246). Moreover, an antiproliferative drug called cisplatin has been demonstrated to induce EMT (247). All this together is in agreement with the results from this study, as the proliferation score was negatively correlated with the EMT score.

As observed, the *in silico* findings showed that mean methylation of CpG-cluster A CpGs was negatively correlated with the EMT score of the ER positive breast tumors. In other words, higher EMT score is associated with lower mean methylation of CpG-cluster A CpGs. In an attempt to validate these *in silico* findings in a biological model system, the D492 and D492M cell lines were utilized as they are shown to display an epithelial-like and mesenchymal-like phenotype, respectively. The idea behind this was to investigate if the CpGs of the EMT-cluster was differentially methylated in the two cell lines. This could have elucidated the link between the DNA methylation of these CpGs and the EMT phenotype in the cell lines. Unfortunately, no pronounced difference in DNA methylation of the target CpGs were detected and surprisingly, seven of the eight CpGs were unmethylated at these CpGs. This was quite unexpected, as the reduced methylation of these CpGs were associated with tumors with a more mesenchymal-like phenotype. There are several possible explanations for this. First of all, the D492 and D492M cell lines do not directly represent ER positive breast cancer. Therefore, it is possible that the EMT signaling is regulated by different mechanisms in this cell type than ER positive tumors, and the CpGs identified here are not important in this specific cell line. Indeed, contrary to most of the ER positive tumors, the D492 cells were unmethylated in all but one of the measured CpGs. In addition, all the CpGs selected as targets for pyrosequencing were found within enhancers, but enhancer methylation will minimally affect expression of EMT-related genes if the promotor is highly methylated. Moreover, it may be that DNA methylation of these CpGs are only partly explaining the expression of these EMT-related genes, and that other factors such as the level of expression

of the TFs may be more important factors contributing to the expression of those genes. Therefore, the methylation status of the promotors and enhancers upstream in this regulatory pathway regulating the genes encoding these TFs may be more direct contributors to EMT. From here, the rest of the discussion will discuss the methods used in this master thesis, step by step. In the end, a conclusion is presented including thoughts and future perspectives.

5.2 Methodological considerations

5.2.1 Patient material

Gene expression and DNA methylation data from ER positive sporadic breast tumors utilized in the emQTL analysis were obtained from the OSL2 discovery cohort. All emQTLs discovered were reanalyzed in ER positive tumors from the independent TCGA breast cancer cohort. It is important to add that in contrast to OSL2, the determination of whether the breast tumor was of sporadic or familial disease has not been considered in TCGA. In other words, the TCGA breast cancer cohort is likely to consist of breast tumors of both sporadic and familial disease. About 10 % of all breast cancers are caused by inherited genetic factors (26). However, this should not affect the results from the emQTL validation in TCGA, as only already existing emQTLs from sporadic ER positive breast cancers will be confirmed. Any familial breast cancer specific emQTLs will therefore not be present.

5.2.2 emQTL analysis, hierarchical clustering and cluster characterization

The emQTL analysis requires several considerations to be made during the workflow that may affect the quality and precision of the result. Early in the workflow the emQTLs were clustered by hierarchical clustering, an approach requiring the choice of distance metric and linkage criterion. Binary is a distance metric commonly used in contexts requiring a decision to be made, that is in this case whether a CpG-gene association is significant or not. Significant associations in the correlation matrix of the p-values were set to 1, while non-significant associations were set to 0. However, such a conversion will cause information loss, as the binary classification does not provide information about the strength of the associations. Hierarchical clustering performed for such large datasets utilized in this study requires extensive computational power which may be a limitation for such large analyses. Therefore, the conversion to binary numbers was necessary. The selected linkage criteria for

the hierarchical clustering was average linkage as it provided the most dense and isolated clusters, thereby including minimal CpGs and genes lacking emQTLs within the defined clusters.

Subsequently in the emQTL analysis, DBSCAN was used to identify emQTL bi-clusters. The benefits of using this algorithm is that DBSCAN can discover any number of clusters with varying shape, size and shape but also detect and ignore outliers in the data. This algorithm requires two parameters to be set: a distance threshold (epsilon) and a density criterion (minpts). DBSCAN is very sensitive to the choice of epsilon. If epsilon is too small, a sparse cluster could be labeled as noise, and if epsilon is to large, then dense clusters could be merged together. However, since the clusters identified by DBSCAN were very isolated and dense, the choice of epsilon was not very sensitive and provided similar output.

One disadvantage is that most of these considerations are subjective and may provide slightly different results depending on the viewer. However, since the emQTL analysis is such a statistical powerful method including a large number of breast cancer tumors with extensive molecular profiles and uses strict p-value thresholds determined with Bonferroni correction, the major biological findings will be conserved.

5.2.3 ChromHMM segmentation

ChromHMM segmentation data provided information of the genomic locations of different chromatin states occurring throughout the MCF7 genome. ChromHMM is one of many other methods available that can be used to identify co-occurrence of chromatin marks. Some other methods are ChromaSig (248) and Segway (249). ChromaSig in contrast to Segway and ChromHMM does not provide genome-wide segmentation, and therefore such data would be inadequate to use since the emQTL analysis is genome-wide. Segway in contrast to ChromHMM provides a finer genome segmentation and handles missing data better, but the disadvantage of this method is that it requires more chromatin marks than ChromHMM to perform. ChromHMM segmentation data was preferred since the data was already generated for an appropriate ER positive breast cancer cell line (MCF7) by Taberlay et al (202), and ChromHMM provides satisfying resolution that covers the approximate size of nucleosomes (~200 bp).

ChromHMM segmentation data was utilized to determine if the emQTL-cluster CpGs or EMT-cluster CpGs were enriched within any particular functional genomic region. However, even though the ChromHMM-MCF7 enhancers should accommodate ER positive breast cancers, it is important to keep in mind that tumors are highly dynamic and may harbor very different biology. These data are indicatory and are the closest estimates we have for the location of regulatory regions in ER positive breast cancers.

5.2.4 Gene set enrichment analysis

The Molecular Signatures Database is one of the most comprehensive and widely used database for gene set enrichment analysis. Performing gene set enrichment analysis requires the consideration of which gene sets to include in the analysis. For this study, gene overlap was computed against the hallmark gene set collection (H) and gene ontology gene set collection (C5). The GO gene set collection (C5) was used to investigate the detailed characteristics of the genes such as molecular function, the cellular component where they exert their functions and the biological process. When performing gene set enrichment analysis, one should limit the inclusion of gene sets. Redundancy is a common issue and occurs when gene sets share a large portion of their genes. This may lead to another type of redundancy occurring when gene sets partly overlapping in which their annotations refer to the same or similar biological process. A consequence of this redundancy is that GSEA generates a long list of statistically significant results with many occurrences of the almost same biological process. Such gene sets may dominate on the top of the result and hide away other relevant findings further down on the list. Overrepresentation of top gene sets reflecting the same biological process may skew the tail of the observed distribution of enrichment scores, leading to an increased significance of scoring of the gene set on the top that represents the same signal (200). The hallmark gene sets describe well defined biological processes and prevents challenges such as redundancy by emphasizing genes displaying coordinate expression from prior knowledge.

5.2.5 ChIP-seq peaks enrichment analysis

Due to the limited mapping of TF specific binding sites within cell lines closely resembling ER positive breast cancers such as the MCF7, merged ChIP-seq peaks from 346 different cell lines were used. The advantage of using such data is that it provides information about

potential TF binding sites across several human cell lines derived from different parts of the body. Tumors are highly heterogenous, and since they are so different, merged peaks from several human cell lines may cover a broader landscape of protein-DNA interactions that may occur within the genome of a cancer cell.

5.2.6 Characterization of tumor samples using gene signatures

Tumor characterization can be performed by assessing the expression of gene signatures uniquely associated with a specific altered or unaltered biological process. In this study, a pan-cancer EMT gene signature was used to investigate the phenotype of the tumor samples. The gene signature was derived from 11 types of cancers including breast cancer. Instead of being just a pure quantitatively measurement of one given phenotype the scoring takes into consideration the concerted expression of epithelial and mesenchymal signature genes. The tumor samples were scored by using 11-genes from the literature known to be associated with a stem cell-like phenotype. Several stem cell gene signatures exist, but several were found to contain genes associated with stem cell differentiation and proliferation and will therefore be of limited value for the purpose of this thesis as current literature indicates that EMT may lead to a dedifferentiation like process to occur in luminal tumors. However, the genes used to score the tumors includes strong candidate genes supported by many studies to be associated with stem cell-like characteristics. The proliferation score used in this thesis was based on the 11-gene proliferation score contained within the PAM50 assay and have been widely used as a measure of proliferation based on gene expression data from breast cancer tumors.

5.2.7 ChIA-PET data

Mapping of genomic interactions occurring between regulatory regions and coding genes within the genome may provide insight into the mechanisms governing the expression of genes associated with disease. One of the most basic unsupervised methods used to predict promoter-enhancer interactions have been performed by simply selecting the enhancer closest to the promoter. However, only around 40 % of enhancers interacts with the nearest promotor (250, 251). ChIA-PET is an unbiased method used to map precise protein-mediated DNA-DNA interactions occurring within the genome. One major disadvantage with ChIA-PET is that it can only be used to identify global interactions mediated by one selected protein for each experiment. This makes the mapping of the interactome costly and time consuming.

However, extensive mapping of genomic interactions has already been applied to the MCF7 breast cancer cell line, and the utilization of these data was therefore preferred.

5.2.8 Validation of differential methylation in D492 and D492M

To validate the *in silico* findings in a biological model system two proper cell lines were require; one cell line with epithelial-like characteristics and another with mesenchymal-like characteristics. For the purpose of this study, D492 and D492M were selected. One major disadvantage using these cell lines is that they are derived from normal cells and will therefore not reflect the extensive epigenetic and genetic alterations occurring during carcinogenesis and cancer progression. A better cell line with epithelial characteristics could be for instance the MCF7 breast cancer cell line. And the more optimal model for ER positive breast cancer with mesenchymal characteristics could be a cell line derived from MCF7 with mesenchymal-like characteristics. Noteworthy, such a cell line already exists and would be of major interest for validation of the *in silico* findings from this study in the future (252). However, the experiment with the D492 and D492M was performed since the cell lines had many of the desired characteristics, they were easily accessible and the experiments was of low cost.

5.2.9 ASCAT, *in silico* nanodissection and CIBERSORT

Existing methods for tumor purity estimation exists which are based on gene expression, DNA methylation and copy number data from either high-throughput DNA sequencing or SNP arrays (227). For this analysis, ASCAT data was used as an estimate of tumor purity for the OSL2 samples. ASCAT have been shown to yield accurate estimates of tumor purity, and have even been shown in some cases to better estimate tumor purity than by pathological examination and other techniques (253). Since tumor purity was correlated with mean methylation of CpG-cluster A, it was of interest to identify potential cell types that could be responsible for this.

There are many *in silico* methods that can be used to estimate tumor infiltration. Two of these are *in silico* nanodissection and CIBERSORT. *In silico* nanodissection utilizes a small set of gene markers of varying quality and a compendia of gene expression data and uses support vector machines within an iterative framework to find cell type specific genes. The output

from *in silico* nanodissection provides a list of genes predicted to be specific for lymphocytes and an estimated probability that those genes are lymphocyte specific. For this analysis, only genes with a probability of more than 65 % of being lymphocyte specific was included to estimate the level of lymphocyte infiltration. CIBERSORT is another method that can be used to estimate the relative proportions of cells, but this method uses a defined signature gene matrix of the immune cell types as input that is based on differential gene expression analysis. CIBERSORT is a robust deconvolution tool that have shown to outperform several other *in silico* deconvolution methods and should be a better measurement of tumor infiltration.

5.2.10 Survival analysis

Several studies have highlighted EMT as an important contributor to cancer progression, metastasis and drug resistance (163-165). Therefore, it was reasonable to investigate whether the EMT score was associated with overall patient survival. Overall survival regarding mean expression of the EMT-cluster genes was also included in the analysis.

Due to the size of the METABRIC cohort and the extensive follow-up of the patients, this cohort was the preferred to use for the survival analysis. No difference in overall survival was observed in the luminal A subtype regarding EMT score and mean expression of the EMT-cluster genes. Patients with luminal A breast cancer subtype have in general a very good prognosis, and differences within this subtype may be therefore hard to find (254). No significant difference in overall survival regarding mean expression of EMT-cluster genes were discovered in the luminal B tumors. However, the EMT score was found to be associated with survival in luminal B tumors. Noteworthy, the EMT score and mean expression of the EMT-cluster genes showed a similar pattern in survival. Another analysis that would be interesting to investigate in the future is the prognostic value of DNA methylation of the EMT-cluster CpGs.

6 Conclusions and future perspectives

Today it is well known that one major feature separating the ER positive breast tumors from one another is proliferation, and this feature tend to be more pronounced in the luminal B breast cancer subtype. However, in this master thesis it has been demonstrated that EMT may be another major feature separating the ER positive breast tumors. This feature was found to be more prominently displayed in the luminal A subtype, and to be an antagonistic feature of proliferation. In agreement with previously published studies, EMT seemed to be dependent on the expression level of *ESR1*.

Genome-wide emQTL analysis lead to the discovery of CpG-gene associations associated with EMT. As observed, there was a low but significant correlation between the level of DNA methylation of CpG-cluster A CpGs with the relative proportion of memory T-cells, mast cells and monocytes which may affect the methylation of these CpGs in some degree. But it should not be decisive as the EMT score and mean methylation value of CpG-cluster A CpGs were correlated with the stemness score that includes genes associated with a stem cell-like phenotype and are related to functions such as self-renewal and regulation of differentiation. These are features not to expect from immune cells. In addition, the CpGs in CpG-cluster A was enriched in ChIP-seq peaks of TFs associated with EMT. Further investigation of the impact of the cell types correlated with mean methylation of CpG-cluster A CpGs on DNA methylation level of CpG-cluster A CpGs measured from the tumor samples would be valuable to study in the future.

The CpGs of the EMT-cluster were shown to be enriched in enhancer regions. In addition, the level of DNA methylation seemed to be connected to the EMT-phenotype thereby proposing the existence of gene regulatory networks connecting these two factors together. Indeed, CpG-cluster A CpGs were found to be highly enriched within ChIP-seq peaks associated with EMT related TFs. A few of these CpGs were picked out for validation in D492 and D492M without any success, but there are many factors that could be optimized in a future study. Another cell line such as the breast cancer derived epithelial MCF7 cell line and a mesenchymal-like MCF7 derived cell line may be better choices to validate the link between DNA methylation of the CpG-cluster A CpGs and EMT-phenotype in the future. In addition, genome-wide DNA methylation- and gene expression profiling of the cell lines could be more informal to interpret the results next time.

The CpGs identified in the EMT-cluster may be major contributors to the EMT related breast cancer pathogenesis and constitute interesting regions for further investigations. The identification of cancer-causing epigenetic changes will be of major interest and could open up possibilities of targeted treatment by utilization of technologies such as CRISPR to edit epigenetic cancer-causing mutations to inhibit tumor growth.

References

1. Jovanovic J, Ronneberg JA, Tost J, Kristensen V. The epigenetics of breast cancer. Molecular oncology. 2010;4(3):242-54.
2. Bediaga NG, Acha-Sagredo A, Guerra I, Viguri A, Albaina C, Ruiz Diaz I, et al. DNA methylation epigenotypes in breast cancer molecular subtypes. Breast cancer research : BCR. 2010;12(5):R77-R.
3. Holm K, Hegardt C, Staaf J, Vallon-Christersson J, Jönsson G, Olsson H, et al. Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. Breast cancer research : BCR. 2010;12(3):R36-R.
4. Lumachi F, Brunello A, Maruzzo M, Basso U, Basso SM. Treatment of estrogen receptor-positive breast cancer. Current medicinal chemistry. 2013;20(5):596-604.
5. Fleischer T, Frigessi A, Johnson KC, Edvardsen H, Touleimat N, Klajic J, et al. Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. Genome biology. 2014;15(8):435.
6. Bell RE, Golan T, Sheinboim D, Malcov H, Amar D, Salamon A, et al. Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. Genome research. 2016;26(5):601-11.
7. Fleischer T, Tekpli X, Mathelier A, Wang S, Nebdal D, Dhakal HP, et al. DNA methylation at enhancers identifies distinct breast cancer lineages. Nature communications. 2017;8(1):1379.
8. The Cancer Registry of Norway. 32 827 NYE KREFTTILFELLER I 2016 2017 [Available from: https://www.kreftregisteret.no/Generelt/Publikasjoner/Cancer-in-Norway/cancer-in-norway-2016/.
9. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70.
10. Douglas Hannahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. 2011.
11. Warburg O. On the origin of cancer cells. Science (New York, NY). 1956;123(3191):309-14.
12. Warburg O. On respiratory impairment in cancer cells. Science (New York, NY). 1956;124(3215):269-70.
13. Warburg O, Wind F, Negelein E. THE METABOLISM OF TUMORS IN THE BODY. The Journal of General Physiology. 1927;8(6):519-30.
14. Dvorak HF. Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing. The New England journal of medicine. 1986;315(26):1650-9.
15. Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. Cell. 2010;140(6):883-99.
16. DeNardo DG, Andreu P, Coussens LM. Interactions between lymphocytes and myeloid cells regulate pro- versus anti-tumor immunity. Cancer metastasis reviews. 2010;29(2):309-16.
17. Karnoub AE, Weinberg RA. Chemokine networks and breast cancer metastasis. Breast disease. 2006;26:75-85.
18. Qian BZ, Pollard JW. Macrophage diversity enhances tumor progression and metastasis. Cell. 2010;141(1):39-51.
19. Valdes F, Alvarez AM, Locascio A, Vega S, Herrera B, Fernandez M, et al. The epithelial mesenchymal transition confers resistance to the apoptotic effects of

transforming growth factor Beta in fetal rat hepatocytes. Molecular cancer research : MCR. 2002;1(1):68-78.
20. Vega S, Morales AV, Ocana OH, Valdes F, Fabregat I, Nieto MA. Snail blocks the cell cycle and confers resistance to cell death. Genes & development. 2004;18(10):1131-43.
21. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer. 2015;136(5):E359-86.
22. Forouzanfar MH, Foreman KJ, Delossantos AM, Lozano R, Lopez AD, Murray CJL, et al. Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. The Lancet.378(9801):1461-84.
23. Cancer Registry of Norway. Cancer in Norway 2016 - Cancer incidence, mortality, survival and prevalence in Norway. Cancer Registry of Norway. 2017.
24. The Cancer Registry of Norway. Brystkreft 15.01.2018 [Available from: https://www.kreftregisteret.no/Generelt/Fakta-om-kreft/Brystkreft-Alt2/.
25. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet (London, England). 2001;358(9291):1389-99.
26. Apostolou P, Fostira F. Hereditary breast cancer: the era of new susceptibility genes. BioMed research international. 2013;2013:747318-.
27. Narod SA, Salmena L. BRCA1 and BRCA2 mutations and breast cancer. Discovery medicine. 2011;12(66):445-53.
28. Xu CF, Solomon E. Mutations of the BRCA1 gene in human cancer. Seminars in cancer biology. 1996;7(1):33-40.
29. Erickson J, Lyon DE. Breast cancer in Cowden syndrome: manifestation of a familial cancer syndrome. Clinical journal of oncology nursing. 2010;14(1):33-5.
30. Nandikolla AG, Venugopal S, Anampa J. Breast cancer in patients with Li–Fraumeni syndrome – a case-series study and review of literature. Breast Cancer : Targets and Therapy. 2017;9:207-15.
31. Hearle N, Schumacher V, Menko FH, Olschwang S, Boardman LA, Gille JJ, et al. Frequency and spectrum of cancers in the Peutz-Jeghers syndrome. Clinical cancer research : an official journal of the American Association for Cancer Research. 2006;12(10):3209-15.
32. Janin N, Andrieu N, Ossian K, Laugé A, Croquette MF, Griscelli C, et al. Breast cancer risk in ataxia telangiectasia (AT) heterozygotes: haplotype study in French AT families. British Journal of Cancer. 1999;80(7):1042-5.
33. Kliewer EV, Smith KR. Breast cancer mortality among immigrants in Australia and Canada. Journal of the National Cancer Institute. 1995;87(15):1154-61.
34. Bagnardi V, Rota M, Botteri E, Tramacere I, Islami F, Fedirko V, et al. Light alcohol drinking and cancer: a meta-analysis. Annals of oncology : official journal of the European Society for Medical Oncology. 2013;24(2):301-8.
35. Knekt P, Albanes D, Seppanen R, Aromaa A, Jarvinen R, Hyvonen L, et al. Dietary fat and risk of breast cancer. The American journal of clinical nutrition. 1990;52(5):903-8.
36. Boyd NF, Stone J, Vogt KN, Connelly BS, Martin LJ, Minkin S. Dietary fat and breast cancer risk revisited: a meta-analysis of the published literature. Br J Cancer. 2003;89(9):1672-85.
37. Velie E, Kulldorff M, Schairer C, Block G, Albanes D, Schatzkin A. Dietary fat, fat subtypes, and breast cancer in postmenopausal women: a prospective cohort study. Journal of the National Cancer Institute. 2000;92(10):833-9.

38. Byrne C, Rockett H, Holmes MD. Dietary fat, fat subtypes, and breast cancer risk: lack of an association among postmenopausal women with no history of benign breast disease. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2002;11(3):261-5.
39. La Vecchia C, Giordano SH, Hortobagyi GN, Chabner B. Overweight, Obesity, Diabetes, and Risk of Breast Cancer: Interlocking Pieces of the Puzzle. The Oncologist. 2011;16(6):726-9.
40. Lorincz AM, Sukumar S. Molecular links between obesity and breast cancer. Endocrine-related cancer. 2006;13(2):279-92.
41. Jesinger RA. Breast anatomy for the interventionalist. Techniques in vascular and interventional radiology. 2014;17(1):3-9.
42. O'Connell RL, Rusby JE. Anatomy relevant to conservative mastectomy. Gland Surgery. 2015;4(6):476-83.
43. Pandya S, Moore RG. Breast development and anatomy. Clinical obstetrics and gynecology. 2011;54(1):91-5.
44. Arps DP, Healy P, Zhao L, Kleer CG, Pang JC. Invasive ductal carcinoma with lobular features: a comparison study to invasive ductal and invasive lobular carcinomas of the breast. Breast cancer research and treatment. 2013;138(3):719-26.
45. https://courses.lumenlearning.com/suny-contemporaryhealthissues/chapter/breasts/. The Breasts. ER services.
46. Page DL, Dupont WD, Rogers LW, Rados MS. Atypical hyperplastic lesions of the female breast. A long-term follow-up study. Cancer. 1985;55(11):2698-708.
47. Yin M, Mackley HB, Drabick JJ, Harvey HA. Primary female breast sarcoma: clinicopathological features, treatment and prognosis. Scientific reports. 2016;6:31497-.
48. RnCeus. Histology of DCIS [Available from: http://www.rnceus.com/dcis/sub.html.
49. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. The Journal of Clinical Investigation. 2009;119(6):1420-8.
50. Rahman M, Mohammed S. Breast cancer metastasis and the lymphatic system. Oncology Letters. 2015;10(3):1233-9.
51. Valastyan S, Weinberg RA. Tumor metastasis: molecular insights and evolving paradigms. Cell. 2011;147(2):275-92.
52. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Annals of surgical oncology. 2010;17(6):1471-4.
53. American Cancer Society. Cancer staging 2015 [Available from: https://www.cancer.org/treatment/understanding-your-diagnosis/staging.html.
54. medicine JH. Breast cancer & breast pathology 2015 [Available from: http://pathology.jhu.edu/breast/grade.php.
55. National Cancer Institute. Tumor Grade: National Cancer Institute; 2013 [Available from: https://www.cancer.gov/about-cancer/diagnosis-staging/prognosis/tumor-grade-fact-sheet.
56. Lim E, Metzger-Filho O, Winer EP. The natural history of hormone receptor-positive breast cancer. Oncology (Williston Park, NY). 2012;26(8):688-94, 96.
57. Deroo BJ, Korach KS. Estrogen receptors and human disease. The Journal of Clinical Investigation. 2006;116(3):561-70.
58. Ghali RM, Al-Mutawa MA, Al-Ansari AK, Zaied S, Bhiri H, Mahjoub T, et al. Differential association of ESR1 and ESR2 gene variants with the risk of breast cancer and associated features: A case-control study. Gene. 2018;651:194-9.

59. Sommer S, Fuqua SA. Estrogen receptor and breast cancer. Seminars in cancer biology. 2001;11(5):339-52.
60. Daniel AR, Hagan CR, Lange CA. Progesterone receptor action: defining a role in breast cancer. Expert review of endocrinology & metabolism. 2011;6(3):359-69.
61. Slamon DJ, Godolphin W, Jones LA, Holt JA, Wong SG, Keith DE, et al. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. Science (New York, NY). 1989;244(4905):707-12.
62. Iqbal N, Iqbal N. Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications. Molecular Biology International. 2014;2014:852748.
63. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science (New York, NY). 1987;235(4785):177-82.
64. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.
65. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. Journal of Clinical Oncology. 2009;27(8):1160-7.
66. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2009;27(8):1160-7.
67. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. World Journal of Clinical Oncology. 2014;5(3):412-24.
68. Makki J. Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. Clinical Medicine Insights Pathology. 2015;8:23-31.
69. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, 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. 2001;98(19):10869-74.
70. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast cancer research : BCR. 2010;12(5):R68-R.
71. Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, et al. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. Jama. 2006;295(21):2492-502.
72. Creighton CJ. The molecular profile of luminal B breast cancer. Biologics : targets & therapy. 2012;6:289-97.
73. Calvo J, Sanchez-Cid L, Munoz M, Lozano JJ, Thomson TM, Fernandez PL. Infrequent loss of luminal differentiation in ductal breast cancer metastasis. PloS one. 2013;8(10):e78097.
74. Skibinski A, Kuperwasser C. The origin of breast tumor heterogeneity. Oncogene. 2015;34(42):5309-16.
75. Fallahpour S, Navaneelan T, De P, Borgo A. Breast cancer survival by molecular subtype: a population-based analysis of cancer registry data. CMAJ Open. 2017;5(3):E734-E9.
76. Heitz F, Harter P, Lueck HJ, Fissler-Eckhoff A, Lorenz-Salehi F, Scheil-Bertram S, et al. Triple-negative and HER2-overexpressing breast cancers exhibit an elevated risk and an earlier occurrence of cerebral metastases. European journal of cancer (Oxford, England : 1990). 2009;45(16):2792-8.

77. Bertucci F, Finetti P, Birnbaum D. Basal breast cancer: a complex and deadly molecular subtype. Current molecular medicine. 2012;12(1):96-110.
78. Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(14):8418-23.
79. Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. American Journal of Cancer Research. 2015;5(10):2929-43.
80. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Medical Genomics. 2015;8:54.
81. Ohnstad HO, Borgen E, Falk RS, Lien TG, Aaserud M, Sveli MAT, et al. Prognostic value of PAM50 and risk of recurrence score in patients with early-stage breast cancer with long-term follow-up. Breast cancer research : BCR. 2017;19:120.
82. Bertheau P, Lehmann-Che J, Varna M, Dumay A, Poirot B, Porcher R, et al. p53 in breast cancer subtypes and new insights into response to chemotherapy. Breast (Edinburgh, Scotland). 2013;22 Suppl 2:S27-9.
83. Walerych D, Napoli M, Collavin L, Del Sal G. The rebel angel: mutant p53 as the driving oncogene in breast cancer. Carcinogenesis. 2012;33(11):2007-17.
84. Langerod A, Zhao H, Borgan O, Nesland JM, Bukholm IR, Ikdahl T, et al. TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. Breast cancer research : BCR. 2007;9(3):R30.
85. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nature genetics. 2013;45(10):1134-40.
86. Russnes HG, Vollan HKM, Lingjærde OC, Krasnitz A, Lundin P, Naume B, et al. Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. Science translational medicine. 2010;2(38):38ra47-38ra47.
87. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346-52.
88. American Cancer Society medical and editorial content team. Surgery for Breast Cancer: American Cancer Society; 2016 [Available from: https://www.cancer.org/cancer/breast-cancer/treatment/surgery-for-breast-cancer.html#references.
89. M. Fleegler F, Griggs, Jennifer, Reiner, Blanche, Reville, Barbara, F. Schnall, Sandra, Weiss, Marisa, Weissmann, Lisa,. Chemotherapy: breastcancer.org; 2018 [Available from: www.breastcancer.org/treatment/chemotherapy.
90. American Cancer Society medical and editorial content team. Radiation for Breast Cancer: American Cancer Society; 2017 [Available from: https://www.cancer.org/cancer/breast-cancer/treatment/radiation-for-breast-cancer.html.
91. American Cancer Society medical and editorial content team. Hormone Therapy for Breast Cancer: American Cancer Society; 2017 [Available from: https://www.cancer.org/cancer/breast-cancer/treatment/hormone-therapy-for-breast-cancer.html.
92. Russo J, Russo IH. The role of estrogen in the initiation of breast cancer. The Journal of steroid biochemistry and molecular biology. 2006;102(1-5):89-96.

93. Isakoff SJ. Triple Negative Breast Cancer: Role of Specific Chemotherapy Agents. Cancer journal (Sudbury, Mass). 2010;16(1):53-61.
94. Gross JM, Yee D. How does the estrogen receptor work? Breast cancer research : BCR. 2002;4(2):62-4.
95. Maximiano S, Magalhaes P, Guerreiro MP, Morgado M. Trastuzumab in the Treatment of Breast Cancer. BioDrugs : clinical immunotherapeutics, biopharmaceuticals and gene therapy. 2016;30(2):75-86.
96. Qiu J. Epigenetics: unfinished symphony. Nature. 2006;441(7090):143-5.
97. Gaal Z, Olah E. [Epigenetic regulatory mechanisms and their disorders in leukemia]. Magyar onkologia. 2014;58(2):99-107.
98. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2010;31(1):27-36.
99. Gronbaek K, Hother C, Jones PA. Epigenetic changes in cancer. APMIS : acta pathologica, microbiologica, et immunologica Scandinavica. 2007;115(10):1039-59.
100. Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. Nature. 2004;429(6990):457-63.
101. Grewal SI, Jia S. Heterochromatin revisited. Nature reviews Genetics. 2007;8(1):35-46.
102. Santisteban MS. [Structure of chromatin. I: Levels of DNA organization in the nucleus; nucleosome and chromatin fibres]. Pathologie-biologie. 1994;42(9):868-83.
103. Hergeth SP, Schneider R. The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle. EMBO Reports. 2015;16(11):1439-53.
104. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128(4):693-705.
105. Sterner DE, Berger SL. Acetylation of Histones and Transcription-Related Factors. Microbiology and Molecular Biology Reviews. 2000;64(2):435-59.
106. Krebs JE, Goldstein ES, Kilpatrick ST. Lewin's genes XII2018.
107. Fernandez-Capetillo O, Lee A, Nussenzweig M, Nussenzweig A. H2AX: the histone guardian of the genome. DNA repair. 2004;3(8-9):959-67.
108. Lau AT, Lee SY, Xu YM, Zheng D, Cho YY, Zhu F, et al. Phosphorylation of histone H2B serine 32 is linked to cell transformation. The Journal of biological chemistry. 2011;286(30):26628-37.
109. Chadee DN, Hendzel MJ, Tylipski CP, Allis CD, Bazett-Jones DP, Wright JA, et al. Increased Ser-10 phosphorylation of histone H3 in mitogen-stimulated and oncogene-transformed mouse fibroblasts. The Journal of biological chemistry. 1999;274(35):24914-20.
110. Choi HS, Choi BY, Cho YY, Mizuno H, Kang BS, Bode AM, et al. Phosphorylation of histone H3 at serine 10 is indispensable for neoplastic cell transformation. Cancer research. 2005;65(13):5818-27.
111. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Research. 2011;21(3):381-95.
112. He S, Tong Q, Bishop DK, Zhang Y. Histone methyltransferase and histone methylation in inflammatory T-cell responses. Immunotherapy. 2013;5(9):10.2217/imt.13.101.
113. Ma Y, Jacobs SB, Jackson-Grusby L, Mastrangelo MA, Torres-Betancourt JA, Jaenisch R, et al. DNA CpG hypomethylation induces heterochromatin reorganization involving the histone variant macroH2A. Journal of cell science. 2005;118(Pt 8):1607-16.
114. Sims RJ, 3rd, Chen CF, Santos-Rosa H, Kouzarides T, Patel SS, Reinberg D. Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4

via its tandem chromodomains. The Journal of biological chemistry. 2005;280(51):41789-92.
115. Hassan AH, Prochasson P, Neely KE, Galasinski SC, Chandy M, Carrozza MJ, et al. Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. Cell. 2002;111(3):369-79.
116. Razin A, Kantor B. DNA methylation in epigenetic control of gene expression. Progress in molecular and subcellular biology. 2005;38:151-67.
117. Alenghat T, Yu J, Lazar MA. The N-CoR complex enables chromatin remodeler SNF2H to enhance repression by thyroid hormone receptor. The EMBO journal. 2006;25(17):3966-74.
118. Robertson KD. DNA methylation and human disease. Nature reviews Genetics. 2005;6(8):597-610.
119. Jin B, Li Y, Robertson KD. DNA methylation: superior or subordinate in the epigenetic hierarchy? Genes & cancer. 2011;2(6):607-17.
120. Riggs AD, Xiong Z. Methylation and epigenetic fidelity. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(1):4-5.
121. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews Genetics. 2012;13(7):484-92.
122. Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(6):3740-5.
123. Blattler A, Farnham PJ. Cross-talk between site-specific transcription factors and DNA methylation states. The Journal of biological chemistry. 2013;288(48):34287-94.
124. Wade PA. Methyl CpG-binding proteins and transcriptional repression. BioEssays : news and reviews in molecular, cellular and developmental biology. 2001;23(12):1131-7.
125. Choy JS, Wei S, Lee JY, Tan S, Chu S, Lee TH. DNA methylation increases nucleosome compaction and rigidity. Journal of the American Chemical Society. 2010;132(6):1782-3.
126. Nisha P, Plank JL, Csink AK. Analysis of chromatin structure of genes silenced by heterochromatin in trans. Genetics. 2008;179(1):359-73.
127. Bhutani N, Burns DM, Blau HM. DNA Demethylation Dynamics. Cell. 2011;146(6):866-72.
128. Feinberg AP, Tycko B. The history of cancer epigenetics. Nature reviews Cancer. 2004;4(2):143-53.
129. Wajed SA, Laird PW, DeMeester TR. DNA Methylation: An Alternative Pathway to Cancer. Annals of Surgery. 2001;234(1):10-20.
130. Weisenberger DJ, Campan M, Long TI, Kim M, Woods C, Fiala E, et al. Analysis of repetitive element DNA methylation by MethyLight. Nucleic Acids Research. 2005;33(21):6823-36.
131. Zheng Y, Joyce BT, Liu L, Zhang Z, Kibbe WA, Zhang W, et al. Prediction of genome-wide DNA methylation in repetitive elements. Nucleic Acids Research. 2017;45(15):8697-711.
132. Anwar SL, Wulaningsih W, Lehmann U. Transposable Elements in Human Cancer: Causes and Consequences of Deregulation. International Journal of Molecular Sciences. 2017;18(5):974.
133. Kulis M, Esteller M. DNA methylation and cancer. Advances in genetics. 2010;70:27-56.
134. Brooks J, Cairns P, Zeleniuch-Jacquotte A. Promoter Methylation and the Detection of Breast Cancer. Cancer causes & control : CCC. 2009;20(9):1539-50.

135. Sharma G, Mirza S, Parshad R, Srivastava A, Datta Gupta S, Pandya P, et al. CpG hypomethylation of MDR1 gene in tumor and serum of invasive ductal breast carcinoma patients. Clinical biochemistry. 2010;43(4-5):373-9.
136. Ronneberg JA, Fleischer T, Solvang HK, Nordgard SH, Edvardsen H, Potapenko I, et al. Methylation profiling with a panel of cancer related genes: association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer. Molecular oncology. 2011;5(1):61-76.
137. Jeronimo C, Monteiro P, Henrique R, Dinis-Ribeiro M, Costa I, Costa VL, et al. Quantitative hypermethylation of a small panel of genes augments the diagnostic accuracy in fine-needle aspirate washings of breast lesions. Breast Cancer Res Treat. 2008;109(1):27-34.
138. Tao MH, Shields PG, Nie J, Millen A, Ambrosone CB, Edge SB, et al. DNA hypermethylation and clinicopathological features in breast cancer: the Western New York Exposures and Breast Cancer (WEB) Study. Breast Cancer Res Treat. 2009;114(3):559-68.
139. Medema RH, Herrera RE, Lam F, Weinberg RA. Growth suppression by p16ink4 requires functional retinoblastoma protein. Proceedings of the National Academy of Sciences of the United States of America. 1995;92(14):6289-93.
140. Soares J, Pinto AE, Cunha CV, Andre S, Barao I, Sousa JM, et al. Global DNA hypomethylation in breast carcinoma: correlation with prognostic factors and tumor progression. Cancer. 1999;85(1):112-8.
141. Hoffmann MJ, Schulz WA. Causes and consequences of DNA hypomethylation in human cancer. Biochemistry and cell biology = Biochimie et biologie cellulaire. 2005;83(3):296-321.
142. Son KS, Kang HS, Kim SJ, Jung SY, Min SY, Lee SY, et al. Hypomethylation of the interleukin-10 gene in breast cancer tissues. Breast (Edinburgh, Scotland). 2010;19(6):484-8.
143. Ribeiro AS, Albergaria A, Sousa B, Correia AL, Bracke M, Seruca R, et al. Extracellular cleavage and shedding of P-cadherin: a mechanism underlying the invasive behaviour of breast cancer cells. Oncogene. 2010;29(3):392-402.
144. Qu Y, Siggens L, Cordeddu L, Gaidzik VI, Karlsson K, Bullinger L, et al. Cancer-specific changes in DNA methylation reveal aberrant silencing and activation of enhancers in leukemia. Blood. 2017;129(7):e13-e25.
145. Kordowski F, Kolarova J, Schafmayer C, Buch S, Goldmann T, Marwitz S, et al. Aberrant DNA methylation of ADAMTS16 in colorectal and other epithelial cancers. BMC cancer. 2018;18(1):796.
146. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nature reviews Genetics. 2014;15(4):272-86.
147. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nature reviews Genetics. 2013;14(4):288-95.
148. Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation. Nature reviews Genetics. 2016;17(9):551-65.
149. Rønneberg JA, Fleischer T, Solvang HK, Nordgard SH, Edvardsen H, Potapenko I, et al. Methylation profiling with a panel of cancer related genes: Association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer. Molecular oncology. 2011;5(1):61-76.
150. Bediaga NG, Acha-Sagredo A, Guerra I, Viguri A, Albaina C, Ruiz Diaz I, et al. DNA methylation epigenotypes in breast cancer molecular subtypes. Breast cancer research : BCR. 2010;12(5):R77.

151. Hill VK, Ricketts C, Bieche I, Vacher S, Gentle D, Lewis C, et al. Genome-wide DNA methylation profiling of CpG islands in breast cancer identifies novel genes associated with tumorigenicity. Cancer research. 2011;71(8):2988-99.
152. The Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumors. Nature. 2012;490(7418):61-70.
153. Kamalakaran S, Varadan V, Giercksky Russnes HE, Levy D, Kendall J, Janevski A, et al. DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables. Molecular oncology. 2011;5(1):77-92.
154. Fleischer T, Klajic J, Aure MR, Louhimo R, Pladsen AV, Ottestad L, et al. DNA methylation signature (SAM40) identifies subgroups of the Luminal A breast cancer samples with distinct survival. Oncotarget. 2017;8(1):1074-82.
155. Beck D, Thoms JA, Perera D, Schutte J, Unnikrishnan A, Knezevic K, et al. Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. Blood. 2013;122(14):e12-22.
156. Starck J, Weiss-Gayet M, Gonnet C, Guyot B, Vicat JM, Morle F. Inducible Fli-1 gene deletion in adult mice modifies several myeloid lineage commitment decisions and accelerates proliferation arrest and terminal erythrocytic differentiation. Blood. 2010;116(23):4795-805.
157. Kruse EA, Loughran SJ, Baldwin TM, Josefsson EC, Ellis S, Watson DK, et al. Dual requirement for the ETS transcription factors Fli-1 and Erg in hematopoietic stem cells and the megakaryocyte lineage. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(33):13814-9.
158. Loughran SJ, Kruse EA, Hacking DF, de Graaf CA, Hyland CD, Willson TA, et al. The transcription factor Erg is essential for definitive hematopoiesis and the function of adult hematopoietic stem cells. Nature immunology. 2008;9(7):810-9.
159. Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS. FOXA1 is a critical determinant of Estrogen Receptor function and endocrine response. Nature genetics. 2011;43(1):27-33.
160. Theodorou V, Stark R, Menon S, Carroll JS. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. Genome research. 2013;23(1):12-22.
161. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature. 2012;481(7381):389-93.
162. Hurtado A, Holmes KA, Ross-Innes CS, Schmidt D, Carroll JS. FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nat Genet. 2011;43(1):27-33.
163. Singh A, Settleman J. EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer. Oncogene. 2010;29(34):4741-51.
164. De Craene B, Berx G. Regulatory networks defining EMT during cancer initiation and progression. Nature reviews Cancer. 2013;13(2):97-110.
165. Byers LA, Diao L, Wang J, Saintigny P, Girard L, Peyton M, et al. An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. Clinical cancer research : an official journal of the American Association for Cancer Research. 2013;19(1):279-90.
166. Micalizzi DS, Ford HL. Epithelial-mesenchymal transition in development and cancer. Future oncology (London, England). 2009;5(8):1129-43.

167. Wang Y, Zhou BP. Epithelial-mesenchymal Transition---A Hallmark of Breast Cancer Metastasis. Cancer hallmarks. 2013;1(1):38-49.
168. Knudsen ES, Ertel A, Davicioni E, Kline J, Schwartz GF, Witkiewicz AK. Progression of ductal carcinoma in situ to invasive breast cancer is associated with gene expression programs of EMT and myoepithelia. Breast Cancer Res Treat. 2012;133(3):1009-24.
169. Felipe Lima J, Nofech-Mozes S, Bayani J, Bartlett JMS. EMT in Breast Carcinoma-A Review. Journal of clinical medicine. 2016;5(7):65.
170. May CD, Sphyris N, Evans KW, Werden SJ, Guo W, Mani SA. Epithelial-mesenchymal transition and cancer stem cells: a dangerously dynamic duo in breast cancer progression. Breast cancer research : BCR. 2011;13(1):202-.
171. Kotiyal S, Bhattacharya S. Breast cancer stem cells, EMT and therapeutic targets. Biochemical and biophysical research communications. 2014;453(1):112-6.
172. Wu KJ, Yang MH. Epithelial-mesenchymal transition and cancer stemness: the Twist1-Bmi1 connection. Bioscience reports. 2011;31(6):449-55.
173. Fabregat I, Malfettone A, Soukupova J. New Insights into the Crossroads between EMT and Stemness in the Context of Cancer. Journal of clinical medicine. 2016;5(3):37.
174. Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial-mesenchymal transition. Nature reviews Molecular cell biology. 2014;15(3):178-96.
175. Zhao B, Ye X, Yu J, Li L, Li W, Li S, et al. TEAD mediates YAP-dependent gene induction and growth control. Genes & development. 2008;22(14):1962-71.
176. Wang C-A, Drasin D, Pham C, Jedlicka P, Zaberezhnyy V, Guney M, et al. Homeoprotein Six2 promotes breast cancer metastasis via transcriptional and epigenetic control of E-cadherin expression. Cancer research. 2014;74(24):7357-70.
177. Reka AK, Kurapati H, Narala VR, Bommer G, Chen J, Standiford TJ, et al. Peroxisome proliferator-activated receptor-gamma activation inhibits tumor metastasis by antagonizing Smad3-mediated epithelial-mesenchymal transition. Molecular cancer therapeutics. 2010;9(12):3221-32.
178. Micalizzi DS, Farabaugh SM, Ford HL. Epithelial-mesenchymal transition in cancer: parallels between normal development and tumor progression. Journal of mammary gland biology and neoplasia. 2010;15(2):117-34.
179. Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J, et al. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(35):15449-54.
180. Sun D, Li X, He Y, Li W, Wang Y, Wang H, et al. YAP1 enhances cell proliferation, migration, and invasion of gastric cancer in vitro and in vivo. Oncotarget. 2016;7(49):81062-76.
181. Cano A, Perez-Moreno MA, Rodrigo I, Locascio A, Blanco MJ, del Barrio MG, et al. The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. Nature cell biology. 2000;2(2):76-83.
182. Wong SHM, Fang CM, Chuah LH, Leong CO, Ngai SC. E-cadherin: Its dysregulation in carcinogenesis and clinical implications. Critical reviews in oncology/hematology. 2018;121:11-22.
183. Zhu QQ, Ma C, Wang Q, Song Y, Lv T. The role of TWIST1 in epithelial-mesenchymal transition and cancers. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine. 2016;37(1):185-97.

184. Tsai JH, Donaher JL, Murphy DA, Chau S, Yang J. Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. Cancer cell. 2012;22(6):725-36.
185. Hugo HJ, Gunasinghe NPAD, Hollier BG, Tanaka T, Blick T, Toh A, et al. Epithelial requirement for in vitro proliferation and xenograft growth and metastasis of MDA-MB-468 human breast cancer cells: oncogenic rather than tumor-suppressive role of E-cadherin. Breast cancer research : BCR. 2017;19(1):86-.
186. Hugo HJ, Pereira L, Suryadinata R, Drabsch Y, Gonda TJ, Gunasinghe NP, et al. Direct repression of MYB by ZEB1 suppresses proliferation and epithelial gene expression during epithelial-to-mesenchymal transition of breast cancer cells. Breast cancer research : BCR. 2013;15(6):R113.
187. Rubio CA. Further studies on the arrest of cell proliferation in tumor cells at the invading front of colonic adenocarcinoma. Journal of gastroenterology and hepatology. 2007;22(11):1877-81.
188. Chen S, Chen JZ, Zhang JQ, Chen HX, Yan ML, Huang L, et al. Hypoxia induces TWIST-activated epithelial-mesenchymal transition and proliferation of pancreatic cancer cells in vitro and in nude mice. Cancer letters. 2016;383(1):73-84.
189. Liu LZ, He YZ, Dong PP, Ma LJ, Wang ZC, Liu XY, et al. Protein tyrosine phosphatase PTP4A1 promotes proliferation and epithelial-mesenchymal transition in intrahepatic cholangiocarcinoma via the PI3K/AKT pathway. Oncotarget. 2016;7(46):75210-20.
190. Sigurdsson V, Hilmarsdottir B, Sigmundsdottir H, Fridriksdottir AJ, Ringner M, Villadsen R, et al. Endothelial induced EMT in breast epithelial cells with stem cell properties. PloS one. 2011;6(9):e23833.
191. Zeisberg M, Neilson EG. Biomarkers for epithelial-mesenchymal transitions. J Clin Invest. 2009;119(6):1429-37.
192. Oslo University Hopsital. Overall aims [Available from: https://www.ous-research.no/home/kgjebsen/home/15313.
193. Aure MR, Jernström S, Krohn M, Vollan HKM, Due EU, Rødland E, et al. Integrated analysis reveals microRNA networks coordinately expressed with key proteins in breast cancer. Genome Medicine. 2015;7(1):21.
194. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology. 2015;19(1A):A68-A77.
195. BC Cancer Agency. Department of Molecular Oncology 2013 [Available from: http://molonc.bccrc.ca/aparicio-lab/research/metabric/.
196. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2008.
197. Zhang Z, Murtagh F, Van Poucke S, Lin S, Lan P. Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting its visualization with R. Annals of Translational Medicine. 2017;5(4):75.
198. Sebastiani P, Perls TT. Detection of Significant Groups in Hierarchical Clustering by Resampling. Frontiers in Genetics. 2016;7:144.
199. Gendoo DM, Ratanasirigulchai N, Schroder MS, Pare L, Parker JS, Prat A, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics (Oxford, England). 2016;32(7):1097-9.
200. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell systems. 2015;1(6):417-25.
201. Liberzon A. A description of the Molecular Signatures Database (MSigDB) Web site. Methods in molecular biology (Clifton, NJ). 2014;1150:153-60.

202. Taberlay PC, Statham AL, Kelly TK, Clark SJ, Jones PA. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome research. 2014;24(9):1421-32.
203. Kolde R. pheatmap: Pretty Heatmaps. (2015). R package version 1.0.8. https://CRAN.R-project.org/package=pheatmap.
204. Mundade R, Ozer HG, Wei H, Prabhu L, Lu T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle. 2014;13(18):2847-52.
205. Xu J, Zhang Y. A Generalized Linear Model for Peak Calling in ChIP-Seq Data. Journal of Computational Biology. 2012;19(6):826-38.
206. Farnham PJ. Insights from genomic profiling of transcription factors. Nature reviews Genetics. 2009;10(9):605-16.
207. Mak MP, Tong P, Diao L, Cardnell RJ, Gibbons DL, William WN, et al. A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. Clinical cancer research : an official journal of the American Association for Cancer Research. 2016;22(3):609-20.
208. Yin X, Li YW, Zhang BH, Ren ZG, Qiu SJ, Yi Y, et al. Coexpression of stemness factors Oct4 and Nanog predict liver resection. Annals of surgical oncology. 2012;19(9):2877-87.
209. Song WS, Yang YP, Huang CS, Lu KH, Liu WH, Wu WW, et al. Sox2, a stemness gene, regulates tumor-initiating and drug-resistant properties in CD133-positive glioblastoma stem cells. Journal of the Chinese Medical Association : JCMA. 2016;79(10):538-45.
210. Zeineddine D, Hammoud AA, Mortada M, Boeuf H. The Oct4 protein: more than a magic stemness marker. American Journal of Stem Cells. 2014;3(2):74-82.
211. Paranjape AN, Balaji SA, Mandal T, Krushik EV, Nagaraj P, Mukherjee G, et al. Bmi1 regulates self-renewal and epithelial to mesenchymal transition in breast cancer cells through Nanog. BMC cancer. 2014;14:785.
212. Zhao W, Li Y, Zhang X. Stemness-Related Markers in Cancer. Cancer translational medicine. 2017;3(3):87-95.
213. Hong SH, Yang SJ, Kim TM, Shim JS, Lee HS, Lee GY, et al. Molecular integration of HoxB4 and STAT3 for self-renewal of hematopoietic stem cells: a model of molecular convergence for stemness. Stem cells (Dayton, Ohio). 2014;32(5):1313-22.
214. Ip CKM, Li S-S, Tang MYH, Sy SKH, Ren Y, Shum HC, et al. Stemness and chemoresistance in epithelial ovarian carcinoma cells under shear stress. Scientific Reports. 2016;6:26788.
215. Ferrell CM, Dorsam ST, Ohta H, Humphries RK, Derynck MK, Haqq C, et al. Activation of stem-cell specific genes by HOXA9 and HOXA10 homeodomain proteins in CD34+ human cord blood cells. Stem cells (Dayton, Ohio). 2005;23(5):644-55.
216. Kumar AR, Sarver AL, Wu B, Kersey JH. Meis1 maintains stemness signature in MLL-AF9 leukemia. Blood. 2010;115(17):3642-3.
217. Tang K-D, Holzapfel BM, Liu J, Lee TK-W, Ma S, Jovanovic L, et al. Tie-2 regulates the stemness and metastatic properties of prostate cancer cells. Oncotarget. 2016;7(3):2572-84.
218. Martin M, Prat A, Rodriguez-Lescure A, Caballero R, Ebbert MT, Munarriz B, et al. PAM50 proliferation score as a predictor of weekly paclitaxel benefit in breast cancer. Breast Cancer Res Treat. 2013;138(2):457-66.

219. Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor positive breast cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2010;16(21):5222-32.
220. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive Promoter-centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. Cell. 2012;148(1-2):84-98.
221. Fullwood MJ, Liu MH, Pan YF, Liu J, Han X, Mohamed YB, et al. An Oestrogen Receptor α-bound Human Chromatin Interactome. Nature. 2009;462(7269):58-64.
222. Gudjonsson T, Villadsen R, Nielsen HL, Ronnov-Jessen L, Bissell MJ, Petersen OW. Isolation, immortalization, and characterization of a human breast epithelial cell line with stem cell properties. Genes & development. 2002;16(6):693-706.
223. Sigurdsson V, Fridriksdottir AJ, Kjartansson J, Jonasson JG, Steinarsdottir M, Petersen OW, et al. Human breast microvascular endothelial cells retain phenotypic traits in long-term finite life span culture. In vitro cellular & developmental biology Animal. 2006;42(10):332-40.
224. Garibyan L, Avashia N. Research Techniques Made Simple: Polymerase Chain Reaction (PCR). The Journal of investigative dermatology. 2013;133(3):e6-e.
225. Su X, Zhang L, Zhang J, Meric-Bernstam F, Weinstein JN. PurityEst: estimating purity of human tumor samples using next-generation sequencing data. Bioinformatics (Oxford, England). 2012;28(17):2265-6.
226. Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(39):16910-5.
227. Zheng X, Zhang N, Wu HJ, Wu H. Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. Genome biology. 2017;18(1):17.
228. Aure MR, Vitelli V, Jernstrom S, Kumar S, Krohn M, Due EU, et al. Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. Breast cancer research : BCR. 2017;19(1):44.
229. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods. 2015;12:453.
230. Qiao W, Quon G, Csaszar E, Yu M, Morris Q, Zandstra PW. PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. PLoS computational biology. 2012;8(12):e1002838.
231. Zhong Y, Wan Y-W, Pang K, Chow LML, Liu Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC bioinformatics. 2013;14:89-.
232. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, et al. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PloS one. 2011;6(11):e27156-e.
233. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature methods. 2015;12(5):453-7.
234. Dannenfelser R, Nome M, Tahiri A, Ursini-Siegel J, Vollan HKM, Haakensen VD, et al. Data-driven analysis of immune infiltrate in a large cohort of breast cancer and its association with disease progression, ER activity, and genomic complexity. Oncotarget. 2017;8(34):57121-33.

235. Dannenfelser R, Nome M, Tahiri A, Ursini-Siegel J, Vollan HKM, Haakensen VD, et al. Data-driven analysis of immune infiltrate in a large cohort of breast cancer and its association with disease progression, ER activity, and genomic complexity. Oncotarget. 2017;8(34):57121-33.
236. Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schubeler D. Competition between DNA methylation and transcription factors determines binding of NRF1. Nature. 2015;528(7583):575-9.
237. Maurano MT, Wang H, John S, Shafer A, Canfield T, Lee K, et al. Role of DNA Methylation in Modulating Transcription Factor Occupancy. Cell reports. 2015;12(7):1184-95.
238. Rishi V, Bhattacharya P, Chatterjee R, Rozenberg J, Zhao J, Glass K, et al. CpG methylation of half-CRE sequences creates C/EBPalpha binding sites that activate some tissue-specific genes. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(47):20311-6.
239. Prokhortchouk A, Hendrich B, Jorgensen H, Ruzov A, Wilm M, Georgiev G, et al. The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. Genes & development. 2001;15(13):1613-8.
240. Quenneville S, Verde G, Corsinotti A, Kapopoulou A, Jakobsson J, Offner S, et al. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. Molecular cell. 2011;44(3):361-72.
241. Liu Y, Toh H, Sasaki H, Zhang X, Cheng X. An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. Genes & development. 2012;26(21):2374-9.
242. Scimeca M, Antonacci C, Colombo D, Bonfiglio R, Buonomo OC, Bonanno E. Emerging prognostic markers related to mesenchymal characteristics of poorly differentiated breast cancers. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine. 2016;37(4):5427-35.
243. Tajima S, Koda K. Dedifferentiation-like progression of breast carcinoma: report of a case showing transition from luminal-type carcinoma to triple-negative carcinoma with myoepithelial features. International journal of clinical and experimental pathology. 2015;8(2):2117-22.
244. Al Saleh S, Al Mulla F, Luqmani YA. Estrogen receptor silencing induces epithelial to mesenchymal transition in human breast cancer cells. PloS one. 2011;6(6):e20610-e.
245. Hawkins RA, Roberts MM, Forrest AP. Oestrogen receptors and breast cancer: current status. The British journal of surgery. 1980;67(3):153-69.
246. Liu S, Cong Y, Wang D, Sun Y, Deng L, Liu Y, et al. Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts. Stem cell reports. 2014;2(1):78-91.
247. Fang S, Yu L, Mei H, Yang J, Gao T, Cheng A, et al. Cisplatin promotes mesenchymal-like characteristics in osteosarcoma through Snail. Oncol Lett. 2016;12(6):5007-14.
248. Hon G, Ren B, Wang W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS computational biology. 2008;4(10):e1000201.
249. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods. 2012;9:473.

250. He B, Chen C, Teng L, Tan K. Global view of enhancer-promoter interactome in human cells. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(21):E2191-E9.
251. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455-61.
252. Guttilla IK, Phoenix KN, Hong X, Tirnauer JS, Claffey KP, White BA. Prolonged mammosphere culture of MCF-7 cells induces an EMT and repression of the estrogen receptor by microRNAs. Breast Cancer Res Treat. 2012;132(1):75-85.
253. Oesper L, Mahmoody A, Raphael BJ. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome biology. 2013;14(7):R80-R.
254. Ohnstad HO, Borgen E, Falk RS, Lien TG, Aaserud M, Sveli MAT, et al. Prognostic value of PAM50 and risk of recurrence score in patients with early-stage breast cancer with long-term follow-up. Breast cancer research : BCR. 2017;19(1):120.

Appendix

Appendix A: The pan-cancer EMT signature

Appendix B: PCR and pyrosequencing primers

Appendix C: DNA methylation profiles of EMT-cluster CpGs in TCGA

Appendix D: Relative cell-type infiltration in tumors estimated by CIBERSORT

Appendix E: METABRIC data for survival analysis

Appendix F: Overview of the data generated for OSL2

Appendix G: Pyrograms of the local DNA sequence around the target CpGs for D492 and D492M

Appendix A: The pan-cancer EMT signature

Table 6. The pan-cancer EMT signature. Each gene is annotated with a certain characteristic describing whether the gene is associated to an epithelial phenotype (E) or a mesenchymal phenotype (M).

Gene	Characteristic	Gene	Characteristic
ADAM12	M	FSTL1	M
ADAMTS12	M	GALNT3	E
ADAMTS2	M	GPC6	M
AEBP1	M	GPR56	E
ANGPTL2	M	GRHL2	E
ANTXR1	M	GYPC	M
AP1G1	E	HOOK1	E
ATP8B1	E	HTRA1	M
AXL	M	INHBA	M
BNC2	M	IRF6	E
CALD1	M	ITGA11	M
CDH1	E	LOXL2	M
CDH2	M	LRRC15	M
CDS1	E	MAP7	E
CGN	E	MARVELD2	E
CLDN4	E	MARVELD3	E
CMTM3	M	MMP2	M
CNOT1	E	MSRB3	M
CNRIP1	M	MYO5B	E
COL10A1	M	NAP1L3	M
COL1A1	M	NID2	M
COL1A2	M	OCLN	E
COL3A1	M	OLFML2B	M
COL5A1	M	PCOLCE	M
COL5A2	M	PDGFRB	M
COL6A1	M	PMP22	M
COL6A2	M	POSTN	M
COL6A3	M	PRSS8	E
COL8A1	M	SPARC	M
CTNND1	E	SPINT1	E
DACT1	M	SPOCK1	M
DYNC1LI2	E	SULF1	M
EMP3	M	SYT11	M
ERBB3	E	THBS2	M
ESRP1	E	VCAN	M
ESRP2	E	VIM	M
F11R	E	ZEB2	M
FAP	M		
FBN1	M		
FN1	M		

Appendix B: PCR and pyrosequencing primers

Table 7. Overview of the primers designed by PyroMark Assay Design software version 2.0 (Qiagen) for PCR amplification and pyrosequencing reactions. The utilized annealing temperatures for the respective PCR amplifications are also shown.

CpG	Gene	PCR primer forward	PCR primer reverse	Temperature PCR (ºC)	Pyrosequencing primer
cg05223441	*VEGFA*	GTTGATTAGAATTTTTTGGATTTTGTGG	TACTCTCTTACTCATAACCCCCAAC	55	TGGATTTTGTGGGTG
cg06947286	*PDLIM4*	AGAGTTGGTAGTATTTTAGTTATTATTGT	ACTCAACCAACACAAAAAATACATT	55	GTTATTATTGTTTTTAAGAAATTTT
cg10233454	*LRP1*	TTGAAGGAAATTAAAGATAGGTTTTAGT	AAACCCCTATCCCACAACAA	55	AATATTAGTTTTGATAGGAAG
cg12232146	*PHLDA1*	ATTATAGGTTTATTAGTAAGGATAGAAATT	CCCTCCATAAACCATAACTATCTATA	55	CCAAAAATACTAAATAACCCTTC
cg16888565	*TPM1*	GTGTGTTATTAATGGTATTTGGTTTTGTAT	ACATCCAAAAAAAATATAACTCTTCACAA	50	ACATCCAAAAAAAATATAACTCTTCACAA
cg20909017	*ITGA5*	AGTTAAAGGAATTGAATAGTTTGTGTAATA	ATATAACTCTCATTTCCCCTAAATCAC	55	GTGTAATATTTTTTGTTTTTGTTTG

Appendix C: DNA methylation profiles of EMT-cluster CpGs in TCGA

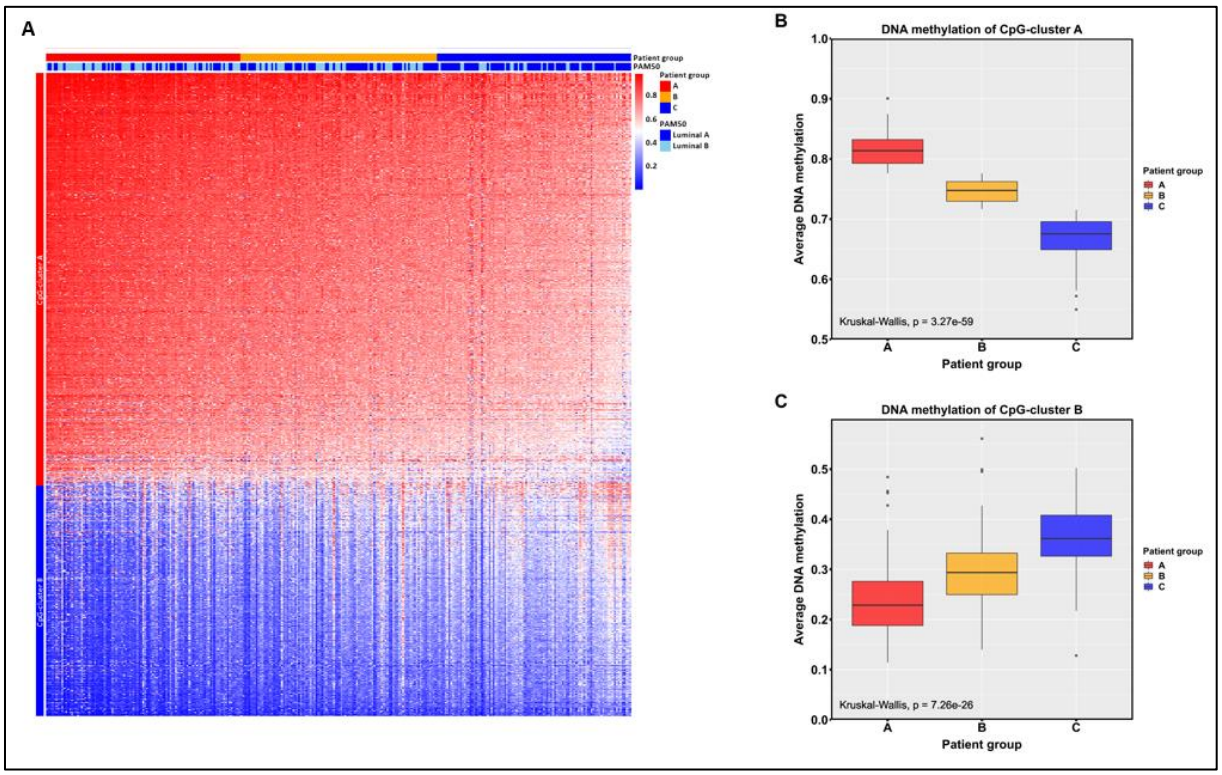


Figure 17. A DNA methylation level of the 1,197 EMT-cluster CpGs for the 304 ER positive breast tumor samples from the TCGA breast cancer cohort. The columns show the tumor samples from TCGA annotated with PAM50 subtype. The tumor samples were also divided into three groups; patient group A (n=101), patient group B (n=102) and patient group C (n=101) based on their mean methylation value of CpG-cluster A CpGs. The columns are ordered based on their mean methylation value of the CpG-cluster A CpGs such that this value decreases from the left of the heatmap towards the right. The rows of the EMT-cluster CpGs (n=1,197) are annotated as either CpG-cluster A CpGs (mean methylation value>0.5; n=770) or CpG-cluster B CpGs (mean methylation value<0.5; n=427). Red spots represent methylation values close to 1 while blue spots have a value close to 0. White spots have an intermediate value of these which is 0.5. Figure **B** and **C** shows the average DNA methylation of CpG-cluster A and CpG-cluster B CpGs respectively in-patient group A, B and C.

Appendix D: Relative cell-type infiltration in tumors estimated by CIBERSORT

Table 8. Overview of estimated relative proportion of the 22 human hematopoietic cell types in the ER positive OSL2 tumor samples (n=177).

Sample ID	B cells naive	B cells memory	Plasma cells	T cells CD8	T cells CD4 naive	T cells CD4 memory resting	T cells CD4 memory activated	T cells follicular helper	T cells regulatory Tregs	T cells gamma delta	NK cells resting	NK cells activated	Monocytes	Macrophages M0	Macrophages M1	Macrophages M2	Dendritic cells resting	Dendritic cells activated	Mast cells resting	Mast cells activated	Eosinophils	Neutrophils	P value	Pearson Correlation	RMSE
OSL2U-0442	0.0721	0.0000	0.0237	0.0000	0.2130	0.0000	0.0000	0.0000	0.0377	0.0000	0.0916	0.0000	0.0106	0.0000	0.0000	0.1370	0.0067	0.0494	0.0000	0.1426	0.0918	0.1238	0.67	0.0166	1.08
OSL2U-0090	0.0198	0.0200	0.0430	0.2023	0.1410	0.0000	0.2401	0.0000	0.0000	0.1148	0.0251	0.0000	0.0000	0.0007	0.0760	0.0173	0.0000	0.0483	0.0460	0.0000	0.0000	0.0055	0.00	0.3486	0.95
OSL2U-0358	0.0000	0.1321	0.0496	0.0000	0.1245	0.0000	0.0434	0.0000	0.0760	0.0952	0.0245	0.0000	0.0000	0.0365	0.0403	0.0000	0.0000	0.0000	0.3778	0.0000	0.0000	0.0000	0.01	0.2661	0.99
OSL2R-3002	0.0000	0.0000	0.0938	0.0000	0.3169	0.0000	0.0000	0.0000	0.0000	0.0000	0.1009	0.0000	0.0000	0.0486	0.0000	0.0000	0.0177	0.0000	0.4218	0.0000	0.0000	0.0002	0.25	0.0885	1.09
OSL2U-0407	0.0000	0.0000	0.1801	0.0000	0.0000	0.0071	0.0000	0.0000	0.0528	0.0000	0.1992	0.0000	0.0000	0.0000	0.0000	0.1189	0.0000	0.0000	0.0835	0.2033	0.0733	0.0819	0.20	0.0973	1.07
OSL2U-0059	0.0588	0.0000	0.0000	0.0000	0.1540	0.0000	0.0000	0.1575	0.0000	0.0000	0.2726	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0026	0.0000	0.3156	0.0390	0.0000	1.00	-0.1534	1.25
OSL2U-0534	0.0445	0.1124	0.0000	0.0840	0.0751	0.0000	0.1765	0.0479	0.0000	0.1112	0.0000	0.0338	0.0000	0.1106	0.1346	0.0400	0.0186	0.0000	0.0000	0.0000	0.0000	0.0107	0.00	0.4587	0.89
OSL2U-0334	0.0076	0.0092	0.0000	0.2511	0.0038	0.0000	0.2904	0.0000	0.0000	0.2436	0.0000	0.0003	0.0000	0.0494	0.0545	0.0319	0.0023	0.0019	0.0318	0.0000	0.0000	0.0222	0.00	0.3964	0.93
OSL2U-0262	0.0000	0.0790	0.0491	0.0000	0.0000	0.0858	0.0000	0.0000	0.0000	0.1617	0.0000	0.0029	0.0000	0.0000	0.0332	0.1218	0.0374	0.0128	0.3491	0.0113	0.0000	0.0559	0.01	0.3002	0.97
OSL2U-0289	0.0000	0.1329	0.1155	0.0000	0.0885	0.0000	0.0000	0.0000	0.1239	0.2865	0.0000	0.0000	0.0000	0.0000	0.0580	0.0000	0.0000	0.0000	0.0000	0.1302	0.0096	0.0550	0.98	-0.0385	1.12
OSL2U-0472	0.0000	0.0898	0.0000	0.0000	0.0629	0.0000	0.0000	0.0619	0.1295	0.0000	0.1103	0.0174	0.0000	0.0000	0.0000	0.0000	0.1377	0.0221	0.0000	0.2482	0.0452	0.0751	0.71	0.0122	1.10
OSL2U-0486	0.0000	0.0000	0.1921	0.0000	0.4598	0.0000	0.0000	0.0000	0.0000	0.0000	0.1377	0.0000	0.0000	0.0000	0.0000	0.0726	0.0000	0.0000	0.0000	0.0509	0.0870	0.0000	1.00	-0.1319	1.16
OSL2R-3030	0.1870	0.0000	0.0517	0.0116	0.0000	0.0000	0.0000	0.0461	0.0617	0.0000	0.0000	0.0785	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5255	0.0000	0.0050	0.0329	0.06	0.1611	1.08
OSL2U-0269	0.0000	0.0699	0.0000	0.0000	0.2268	0.0000	0.0000	0.0134	0.0753	0.1025	0.0551	0.0000	0.0000	0.2067	0.0000	0.0000	0.0476	0.0000	0.0000	0.2028	0.0000	0.0000	0.01	0.2985	0.97
OSL2U-0020	0.0232	0.0352	0.0000	0.0614	0.0260	0.0000	0.4156	0.0000	0.0000	0.1590	0.0000	0.0000	0.0256	0.0571	0.0364	0.0788	0.0000	0.0000	0.0758	0.0000	0.0000	0.0059	0.00	0.3532	0.94
OSL2U-0392	0.0000	0.0190	0.0432	0.2619	0.0000	0.0390	0.0834	0.1803	0.0000	0.3083	0.0000	0.0000	0.0000	0.0000	0.0368	0.0000	0.0000	0.0109	0.0000	0.0000	0.0172	0.0000	0.01	0.2828	1.02
OSL2U-0370	0.0000	0.1497	0.0000	0.0000	0.0919	0.0000	0.0000	0.1279	0.0000	0.0000	0.2104	0.0000	0.0693	0.0000	0.0000	0.0000	0.0223	0.0000	0.0000	0.2671	0.0615	0.0000	0.87	-0.0063	1.13
OSL2U-0383	0.0000	0.0037	0.0667	0.0000	0.2888	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0563	0.0221	0.0000	0.0000	0.4264	0.0195	0.0000	0.0000	0.0000	0.0000	0.1166	0.99	-0.0409	1.12
OSL2U-0101	0.0000	0.0273	0.0000	0.0000	0.2928	0.0000	0.0000	0.0000	0.1274	0.0000	0.1968	0.0000	0.0133	0.0000	0.0000	0.0755	0.0000	0.0000	0.0000	0.1767	0.0794	0.0109	1.00	-0.1562	1.19
OSL2U-0064	0.0000	0.0000	0.0082	0.0000	0.0956	0.1875	0.0000	0.0000	0.1306	0.0000	0.1501	0.0000	0.0390	0.0000	0.0000	0.0000	0.0000	0.0838	0.0000	0.1373	0.1680	0.0000	1.00	-0.1997	1.19
OSL2U-0364	0.1553	0.0000	0.0000	0.0000	0.2974	0.0000	0.0000	0.0000	0.0000	0.0000	0.1974	0.0000	0.0000	0.0000	0.0000	0.1452	0.0000	0.0000	0.0347	0.0000	0.0563	0.1136	1.00	-0.0745	1.13
OSL2A-6181	0.1214	0.0000	0.0873	0.1574	0.0365	0.0000	0.0909	0.1254	0.0000	0.0882	0.0000	0.0000	0.0000	0.0993	0.0182	0.0000	0.0473	0.0217	0.0000	0.1029	0.0000	0.0037	0.95	-0.0289	1.10
OSL2U-0488	0.0951	0.0000	0.0499	0.0000	0.1783	0.1592	0.0000	0.0000	0.0000	0.0000	0.2788	0.0000	0.0000	0.0000	0.0000	0.1232	0.0225	0.0000	0.0000	0.0930	0.0000	0.0000	1.00	-0.2422	1.23
OSL2U-0229	0.0521	0.0000	0.1866	0.0000	0.2499	0.0000	0.0000	0.0000	0.0800	0.0000	0.1090	0.0000	0.0000	0.3216	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007	0.0000	0.0000	0.94	-0.0207	1.11
OSL2U-0422	0.1004	0.0263	0.0370	0.0469	0.0000	0.0378	0.2216	0.0000	0.0059	0.1383	0.0000	0.0410	0.0000	0.0235	0.0839	0.0633	0.0000	0.0850	0.0703	0.0000	0.0000	0.0190	0.00	0.4955	0.87
OSL2U-0489	0.0937	0.0000	0.0000	0.0000	0.0099	0.0000	0.0000	0.1642	0.0000	0.0000	0.2378	0.0000	0.1095	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2332	0.0576	0.0941	1.00	-0.1165	1.19
OSL2U-0079	0.0000	0.1313	0.0000	0.0659	0.1412	0.0000	0.1228	0.0000	0.0793	0.1800	0.0000	0.0676	0.0000	0.0781	0.0403	0.0000	0.0000	0.0113	0.0640	0.0000	0.0054	0.0128	0.00	0.5564	0.83
OSL2U-0179	0.1153	0.0000	0.0000	0.0000	0.0000	0.2506	0.0000	0.0000	0.1073	0.0000	0.0000	0.0000	0.1400	0.0000	0.0000	0.0068	0.0000	0.0000	0.0000	0.2324	0.0737	0.0740	0.88	-0.0093	1.11
OSL2R-3052	0.0254	0.0000	0.0000	0.1753	0.0091	0.0000	0.0531	0.0652	0.0000	0.1712	0.0000	0.0438	0.0000	0.1138	0.0556	0.1163	0.0000	0.0446	0.0000	0.0481	0.0000	0.0785	0.01	0.3035	0.97
OSL2A-6133	0.0000	0.0560	0.0442	0.3605	0.0000	0.0000	0.0872	0.0978	0.0000	0.1694	0.0000	0.0278	0.0000	0.0542	0.0572	0.0000	0.0000	0.0006	0.0452	0.0000	0.0000	0.0000	0.01	0.2439	1.02
OSL2U-0342	0.0343	0.0030	0.0421	0.1541	0.1649	0.0360	0.1760	0.0000	0.0000	0.2589	0.0000	0.0000	0.0000	0.0285	0.0999	0.0000	0.0023	0.0000	0.0000	0.0000	0.0000	0.0000	0.00	0.5355	0.84
OSL2U-0027	0.0000	0.2112	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3041	0.0000	0.0105	0.0025	0.0000	0.2861	0.0941	0.0000	0.0019	0.0000	0.0895	0.0000	0.0000	0.0000	0.95	-0.0265	1.11
OSL2U-0325	0.0000	0.0000	0.0000	0.0000	0.0883	0.0649	0.0000	0.0000	0.0000	0.0000	0.1177	0.0000	0.0180	0.0754	0.0000	0.1093	0.0000	0.0199	0.0000	0.1737	0.0824	0.2504	0.06	0.1708	1.05
OSL2U-0379	0.0000	0.0577	0.1261	0.1405	0.0000	0.0000	0.1311	0.0176	0.0000	0.1264	0.0000	0.0148	0.0439	0.2013	0.0410	0.0123	0.0049	0.0482	0.0339	0.0000	0.0000	0.0002	0.06	0.1647	1.02
OSL2U-0544	0.0000	0.0000	0.0375	0.0000	0.2146	0.1479	0.0000	0.0000	0.2237	0.0103	0.1589	0.0000	0.0000	0.0365	0.0000	0.0000	0.0000	0.0000	0.0000	0.0961	0.0746	0.0000	1.00	-0.1670	1.20
OSL2U-0242	0.0675	0.0000	0.0000	0.0000	0.3338	0.0000	0.0000	0.0319	0.0000	0.0000	0.3730	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1520	0.0416	0.0000	1.00	-0.0991	1.23
OSL2U-0209	0.0000	0.2837	0.0000	0.1826	0.0244	0.0000	0.0000	0.0000	0.1552	0.2419	0.0000	0.0014	0.0000	0.0000	0.0000	0.0000	0.0198	0.0000	0.0760	0.0000	0.0149	0.0000	0.43	0.0423	1.11
OSL2U-0120	0.0145	0.0000	0.0000	0.0000	0.0859	0.1421	0.0000	0.0000	0.0000	0.0000	0.1197	0.0386	0.0000	0.0000	0.0000	0.4122	0.0209	0.0000	0.0000	0.0101	0.0670	0.0891	1.00	-0.0491	1.13
OSL2U-0519	0.0000	0.2729	0.0000	0.0000	0.1911	0.0000	0.0000	0.0000	0.0000	0.2174	0.0000	0.0000	0.0000	0.0000	0.0000	0.0749	0.0618	0.0000	0.0000	0.1348	0.0342	0.0129	1.00	-0.0906	1.14
OSL2U-0086	0.0000	0.0342	0.0000	0.1797	0.0000	0.0000	0.0000	0.1527	0.0448	0.0000	0.0000	0.2729	0.0000	0.0000	0.0000	0.0000	0.0881	0.0000	0.0000	0.1373	0.0561	0.0343	0.34	0.0673	1.15
OSL2U-0296	0.0000	0.1535	0.0000	0.0000	0.0774	0.0000	0.0000	0.0781	0.0000	0.0000	0.2320	0.0000	0.0000	0.0702	0.0000	0.1883	0.0169	0.0000	0.0000	0.1422	0.0000	0.0413	0.98	-0.0393	1.12
OSL2U-0171	0.0000	0.0000	0.0000	0.0000	0.0552	0.0000	0.0000	0.0000	0.5798	0.0000	0.0000	0.1261	0.0000	0.2388	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.00	-0.1146	1.22
OSL2R-3053	0.0000	0.0000	0.0535	0.0000	0.1659	0.0000	0.0000	0.0534	0.0000	0.0000	0.0000	0.2187	0.1077	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2986	0.0697	0.0324	0.95	-0.0225	1.19
OSL2A-6166	0.0000	0.0000	0.0000	0.0000	0.0137	0.0000	0.0000	0.0000	0.1795	0.0000	0.2572	0.0000	0.0060	0.0265	0.0000	0.0000	0.0153	0.0000	0.0000	0.2329	0.1742	0.0948	0.65	0.0185	1.14
OSL2U-0280	0.0000	0.1940	0.0000	0.0202	0.0344	0.0820	0.2016	0.0000	0.0000	0.1879	0.0000	0.0019	0.0000	0.0154	0.1497	0.0553	0.0000	0.0575	0.0000	0.0000	0.0000	0.0000	0.00	0.3330	0.96
OSL2U-0250	0.0000	0.0712	0.0000	0.2800	0.0000	0.0000	0.0368	0.0912	0.0796	0.0000	0.0000	0.0583	0.0000	0.0000	0.1368	0.1857	0.0000	0.0000	0.0328	0.0000	0.0000	0.0276	0.01	0.2548	0.99
OSL2U-0285	0.0000	0.0972	0.1026	0.1304	0.0000	0.0000	0.0032	0.0433	0.0024	0.2400	0.0000	0.0595	0.0000	0.1731	0.0377	0.0379	0.0000	0.0000	0.0728	0.0000	0.0000	0.0000	0.00	0.3332	0.95
OSL2U-0318	0.0000	0.0500	0.0000	0.0000	0.0000	0.0229	0.0000	0.2016	0.0000	0.0000	0.1574	0.0000	0.0282	0.0000	0.0000	0.0688	0.0989	0.0035	0.0000	0.1946	0.0918	0.0824	0.34	0.0651	1.07
OSL2U-0484	0.1456	0.0000	0.0000	0.0000	0.2139	0.0000	0.0000	0.0000	0.1103	0.0000	0.1538	0.0000	0.0112	0.0000	0.0000	0.0000	0.0172	0.0388	0.0000	0.1677	0.1415	0.0000	1.00	-0.0811	1.14
OSL2U-0232	0.0000	0.0000	0.1453	0.0000	0.2321	0.1338	0.0000	0.0000	0.0840	0.0582	0.1575	0.0000	0.0000	0.0000	0.0000	0.0000	0.0188	0.0000	0.0000	0.0868	0.0836	0.0000	1.00	-0.1555	1.18
OSL2U-0069	0.0000	0.0972	0.0000	0.0514	0.0000	0.1471	0.0000	0.0000	0.1449	0.1137	0.0000	0.0000	0.0000	0.1143	0.0805	0.0262	0.0905	0.0000	0.0859	0.0000	0.0000	0.0483	0.01	0.2277	0.99
OSL2U-0328	0.0000	0.0000	0.0407	0.1886	0.0000	0.0000	0.1202	0.0059	0.0000	0.2414	0.0000	0.0936	0.0000	0.0000	0.0641	0.0845	0.0291	0.0000	0.1306	0.0000	0.0000	0.0012	0.00	0.3298	0.97
OSL2U-0223	0.0000	0.1079	0.0000	0.0000	0.1928	0.0627	0.0000	0.0000	0.0000	0.0000	0.2491	0.0000	0.0000	0.0000	0.0000	0.0087	0.0000	0.0000	0.0000	0.0738	0.0000	0.3050	1.00	-0.0703	1.19
OSL2U-0083	0.0000	0.0745	0.1754	0.1623	0.1295	0.0000	0.0000	0.0809	0.0224	0.0000	0.1262	0.0000	0.0000	0.0216	0.0000	0.0549	0.0000	0.0000	0.0000	0.1522	0.0000	0.0000	0.67	0.0151	1.10
OSL2U-0291	0.0000	0.0000	0.0000	0.0000	0.0313	0.0281	0.0000	0.0878	0.0000	0.1575	0.0002	0.0000	0.0242	0.0631	0.0000	0.0000	0.0229	0.0000	0.4405	0.1285	0.0158	0.0000	0.01	0.2553	1.04
OSL2U-0260	0.0000	0.0000	0.0000	0.1040	0.0405	0.0000	0.0000	0.0000	0.3028	0.0000	0.0046	0.0000	0.0000	0.0000	0.0000	0.0156	0.0000	0.0000	0.0000	0.2815	0.1393	0.1116	1.00	-0.0887	1.18
OSL2U-0299	0.0000	0.1181	0.0000	0.0000	0.1661	0.0000	0.0000	0.0000	0.0000	0.0428	0.0000	0.0997	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4692	0.0831	0.0210	0.0000	0.08	0.1564	1.10
OSL2U-0174	0.1794	0.0000	0.0000	0.0000	0.0000	0.1809	0.0000	0.0000	0.1223	0.0000	0.1878	0.0000	0.0145	0.0000	0.0000	0.0000	0.0493	0.0000	0.0000	0.1506	0.0474	0.0679	1.00	-0.1145	1.16
OSL2U-0477	0.0000	0.2732	0.0000	0.0648	0.0119	0.0148	0.0904	0.0290	0.1079	0.2206	0.0000	0.0000	0.0000	0.1028	0.0783	0.0000	0.0000	0.0064	0.0000	0.0000	0.0000	0.0000	0.01	0.2450	1.00
OSL2U-0367	0.0000	0.2357	0.0000	0.0000	0.1669	0.0000	0.0000	0.0500	0.0000	0.0000	0.0000	0.0963	0.0000	0.0000	0.0000	0.0000	0.0608	0.0000	0.1097	0.1896	0.0909	0.0000	1.00	-0.0621	1.15
OSL2U-0243	0.1963	0.0067	0.0000	0.0000	0.3706	0.0000	0.0000	0.0000	0.0000	0.0455	0.0000	0.0662	0.0445	0.0000	0.0000	0.0000	0.0517	0.0000	0.1742	0.0000	0.0444	0.0000	0.76	0.0038	1.10
OSL2U-0548	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0519	0.1533	0.0000	0.1673	0.0000	0.0000	0.3795	0.0000	0.0000	0.0000	0.0000	0.0000	0.1844	0.0000	0.0636	0.04	0.1998	1.05

Sample ID	B cells naive	B cells memory	Plasma cells	T cells CD8	T cells CD4 naive	T cells CD4 memory resting	T cells CD4 memory activated	T cells follicular helper	T cells regulatory Tregs	T cells gamma delta	NK cells resting	NK cells activated	Monocytes	Macrophages M0	Macrophages M1	Macrophages M2	Dendritic cells resting	Dendritic cells activated	Mast cells resting	Mast cells activated	Eosinophils	Neutrophils	P-value	Pearson Correlation	RMSE
OSL2U-0553	0.0000	0.0000	0.0591	0.0000	0.2625	0.0000	0.0000	0.0000	0.0940	0.0000	0.2126	0.0000	0.0478	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1649	0.1592	0.0000	1.00	-0.1996	1.20
OSL2U-0491	0.0000	0.0000	0.0000	0.0576	0.0000	0.1862	0.0000	0.1262	0.0000	0.0000	0.0000	0.0804	0.0000	0.0000	0.0615	0.1911	0.0000	0.0133	0.0490	0.0000	0.0000	0.2348	0.73	0.0070	1.11
OSL2U-0197	0.0958	0.0000	0.0000	0.0000	0.1199	0.0000	0.0000	0.0000	0.1621	0.0000	0.0570	0.0000	0.0000	0.2508	0.0000	0.0000	0.0000	0.0044	0.1348	0.0317	0.0000	0.1435	0.76	0.0033	1.10
OSL2U-0400	0.0204	0.1272	0.0000	0.0000	0.2299	0.0000	0.1957	0.0000	0.0000	0.0973	0.0453	0.0000	0.0043	0.0364	0.1197	0.0601	0.0000	0.0599	0.0000	0.0000	0.0000	0.0037	0.04	0.2018	1.01
OSL2U-0053	0.0000	0.0741	0.0987	0.0000	0.0127	0.0000	0.0000	0.0647	0.0388	0.0000	0.0267	0.0000	0.0000	0.3561	0.0179	0.0325	0.0000	0.0000	0.0000	0.1376	0.0253	0.1150	0.06	0.1609	1.05
OSL2U-0096	0.0000	0.0000	0.0431	0.0000	0.0151	0.0000	0.0000	0.0890	0.0642	0.0000	0.0307	0.0000	0.0000	0.0242	0.0000	0.0795	0.1316	0.0000	0.4930	0.0166	0.0000	0.0131	0.00	0.3326	0.97
OSL2U-0393	0.0000	0.0442	0.0000	0.0000	0.2780	0.0000	0.0000	0.0000	0.0000	0.0000	0.1558	0.0000	0.0000	0.3203	0.0000	0.0615	0.0000	0.0000	0.0000	0.1402	0.0000	0.0000	0.88	-0.0121	1.13
OSL2U-0037	0.0000	0.1854	0.0211	0.0000	0.0224	0.4019	0.0000	0.0000	0.0000	0.0040	0.0172	0.0000	0.0087	0.0000	0.0000	0.0099	0.1591	0.0000	0.0539	0.1165	0.0000	0.0000	0.06	0.1719	1.02
OSL2U-0267	0.0000	0.0000	0.0000	0.0000	0.0687	0.0000	0.0000	0.0000	0.2339	0.0000	0.0000	0.0834	0.0135	0.0936	0.1256	0.0000	0.0728	0.0000	0.0000	0.2083	0.1002	0.0000	0.43	0.0436	1.11
OSL2U-0126	0.0000	0.1524	0.0000	0.2790	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0641	0.0000	0.0000	0.0000	0.0000	0.0302	0.0000	0.4324	0.0000	0.0000	0.0420	0.77	0.0028	1.14
OSL2U-0324	0.0000	0.0121	0.0000	0.0000	0.0963	0.0000	0.0000	0.0000	0.1895	0.0000	0.0819	0.0000	0.0000	0.3245	0.1615	0.0021	0.0252	0.0000	0.0000	0.1070	0.0000	0.0000	0.18	0.1092	1.08
OSL2U-0225	0.0708	0.0000	0.0000	0.0000	0.0550	0.0000	0.0800	0.0000	0.0000	0.0642	0.0400	0.0000	0.0462	0.3914	0.0167	0.1556	0.0000	0.0415	0.0370	0.0000	0.0000	0.0015	0.00	0.7149	0.72
OSL2U-0056	0.0000	0.0957	0.0000	0.0000	0.1647	0.0173	0.1505	0.0000	0.0600	0.0595	0.0018	0.0125	0.0000	0.0792	0.1087	0.0000	0.0000	0.0000	0.2502	0.0000	0.0000	0.0000	0.00	0.4723	0.88
OSL2U-0275	0.0000	0.0578	0.0000	0.1070	0.0000	0.0000	0.0000	0.0000	0.0614	0.1435	0.0000	0.1295	0.1180	0.1290	0.1041	0.0997	0.0000	0.0000	0.0258	0.0000	0.0000	0.0242	0.06	0.1616	1.05
OSL2U-0130	0.0170	0.0791	0.0078	0.0000	0.0000	0.0000	0.1061	0.2106	0.0358	0.0690	0.0000	0.0907	0.0000	0.1083	0.2368	0.0000	0.0242	0.0015	0.0130	0.0000	0.0000	0.0000	0.00	0.3250	0.97
OSL2A-6134	0.0282	0.0599	0.0000	0.0000	0.1459	0.0624	0.2988	0.0000	0.0000	0.0000	0.0446	0.0009	0.0388	0.0000	0.0000	0.0939	0.0034	0.0182	0.2032	0.0000	0.0000	0.0018	0.37	0.0506	1.06
OSL2U-0109	0.0000	0.0204	0.0003	0.0000	0.0000	0.0301	0.0000	0.0000	0.0330	0.1146	0.0305	0.0000	0.0000	0.1796	0.0000	0.2729	0.0000	0.0000	0.0000	0.0881	0.0072	0.2233	0.10	0.1452	1.07
OSL2U-0112	0.0000	0.1059	0.0000	0.0000	0.0000	0.1111	0.0000	0.0000	0.0628	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4385	0.1655	0.0000	0.1161	0.25	0.0874	1.15
OSL2U-0554	0.0000	0.0624	0.0787	0.0000	0.0000	0.2544	0.0000	0.0343	0.0407	0.0000	0.1177	0.0000	0.0613	0.0000	0.1007	0.0000	0.0000	0.0000	0.0992	0.0769	0.0738	0.0000	0.01	0.2125	1.00
OSL2U-0438	0.0000	0.0054	0.0395	0.2923	0.0000	0.0054	0.0000	0.0120	0.0000	0.1049	0.0000	0.0546	0.0048	0.0000	0.0936	0.0853	0.0335	0.0000	0.2231	0.0000	0.0000	0.0455	0.06	0.1668	1.04
OSL2U-0128	0.0000	0.1220	0.0000	0.0808	0.0000	0.0417	0.0000	0.0000	0.0459	0.0000	0.0000	0.0604	0.0282	0.0000	0.0133	0.1691	0.1383	0.0000	0.0000	0.1013	0.0399	0.1591	0.06	0.1672	1.02
OSL2U-0085	0.0000	0.0615	0.1546	0.0795	0.0000	0.0003	0.0200	0.0000	0.0347	0.1954	0.0000	0.0042	0.0170	0.1461	0.0712	0.0000	0.0020	0.0094	0.1546	0.0474	0.0019	0.0000	0.04	0.1978	1.01
OSL2U-0249	0.0000	0.0134	0.0000	0.0000	0.1689	0.0000	0.0000	0.0000	0.0442	0.0000	0.1160	0.1095	0.0568	0.0000	0.0000	0.0000	0.0000	0.0000	0.4912	0.0000	0.0000	0.0000	0.88	-0.0089	1.18
OSL2U-0248	0.0000	0.0946	0.1797	0.0115	0.0000	0.0212	0.1185	0.0134	0.0689	0.0000	0.0000	0.1160	0.0624	0.0000	0.1758	0.0000	0.0164	0.0263	0.0234	0.0719	0.0000	0.0000	0.00	0.4806	0.88
OSL2U-0268	0.0000	0.1288	0.0000	0.1008	0.0652	0.0000	0.0000	0.0000	0.0618	0.2571	0.0000	0.0194	0.0000	0.0000	0.0000	0.0000	0.2172	0.0000	0.0000	0.0859	0.0639	0.0000	0.88	-0.0106	1.11
OSL2U-0463	0.0251	0.0000	0.0000	0.0000	0.0509	0.1595	0.0000	0.0000	0.0597	0.0000	0.2376	0.0000	0.0160	0.0000	0.0000	0.0000	0.0218	0.0000	0.0000	0.2523	0.0130	0.1641	1.00	-0.1300	1.21
OSL2U-0186	0.0000	0.0000	0.0758	0.0000	0.4726	0.0000	0.0000	0.0198	0.0000	0.0000	0.1189	0.0000	0.0674	0.0000	0.0000	0.0000	0.0000	0.0290	0.0000	0.1624	0.0539	0.0000	1.00	-0.1529	1.18
OSL2A-6132	0.0000	0.0477	0.0646	0.2103	0.0000	0.0000	0.2011	0.0393	0.0114	0.0359	0.0000	0.0726	0.0031	0.0436	0.1526	0.0440	0.0367	0.0000	0.0287	0.0000	0.0000	0.0084	0.00	0.5215	0.85
OSL2U-0430	0.1857	0.0000	0.0000	0.0000	0.2616	0.0000	0.0000	0.0000	0.0000	0.0545	0.1780	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2031	0.0677	0.0494	1.00	-0.0937	1.17
OSL2U-0537	0.0000	0.0000	0.1205	0.0000	0.2810	0.0000	0.0000	0.0000	0.0081	0.0000	0.2089	0.0000	0.0000	0.0000	0.0000	0.0000	0.0311	0.0000	0.2060	0.1248	0.0196	0.0000	0.37	0.0525	1.10
OSL2U-0290	0.3112	0.0000	0.0000	0.0000	0.2154	0.0000	0.0000	0.0000	0.0442	0.0000	0.0000	0.0946	0.0000	0.0032	0.0000	0.0000	0.0000	0.0000	0.0000	0.0799	0.1293	0.1224	1.00	-0.1387	1.18
OSL2U-0227	0.0000	0.1667	0.0000	0.0000	0.2620	0.0000	0.0000	0.0000	0.0000	0.0000	0.3457	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2256	0.0000	0.0000	1.00	-0.1758	1.26
OSL2U-0182	0.0051	0.0000	0.0000	0.2600	0.0055	0.0000	0.0000	0.0000	0.3094	0.0662	0.0000	0.0000	0.0000	0.2680	0.0000	0.0000	0.0000	0.0000	0.0178	0.0000	0.0680	0.0000	0.95	-0.0261	1.14
OSL2U-0531	0.0101	0.0000	0.0000	0.0000	0.0000	0.2232	0.0000	0.0000	0.0190	0.0000	0.1128	0.0000	0.0000	0.3563	0.0000	0.0000	0.0000	0.0000	0.0000	0.1747	0.0000	0.1040	1.00	-0.1080	1.20
OSL2U-0159	0.0047	0.0872	0.0000	0.2903	0.0115	0.0000	0.2094	0.0000	0.0000	0.1131	0.0000	0.0148	0.0000	0.0518	0.0789	0.0387	0.0000	0.0358	0.0502	0.0000	0.0000	0.0135	0.00	0.5693	0.83
OSL2U-0045	0.0223	0.0000	0.0829	0.0000	0.0000	0.0089	0.0000	0.2300	0.0000	0.0000	0.0989	0.0019	0.0000	0.0000	0.0000	0.1033	0.0000	0.0000	0.0000	0.1338	0.2107	0.1073	0.22	0.0963	1.06
OSL2U-0065	0.0000	0.0000	0.0940	0.0000	0.3598	0.0000	0.0000	0.0000	0.0000	0.0000	0.0777	0.0000	0.0467	0.0000	0.0000	0.0000	0.0000	0.0000	0.2786	0.0380	0.1052	0.0000	0.92	-0.0170	1.12
OSL2U-0462	0.0103	0.0000	0.0205	0.1087	0.0000	0.0280	0.0000	0.0000	0.0384	0.0000	0.0000	0.0586	0.0000	0.0000	0.0000	0.1027	0.0281	0.0000	0.5415	0.0000	0.0000	0.0631	0.01	0.2927	1.00
OSL2U-0177	0.0000	0.0215	0.0081	0.0000	0.0000	0.3404	0.0000	0.0000	0.0000	0.0728	0.0000	0.1118	0.0000	0.0002	0.0000	0.1380	0.0221	0.0000	0.2221	0.0036	0.0000	0.0594	0.35	0.0610	1.08
OSL2U-0103	0.0710	0.0000	0.0000	0.0000	0.1646	0.0000	0.0000	0.0000	0.0000	0.0000	0.0525	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5029	0.0808	0.1282	0.0000	0.29	0.0773	1.15
OSL2U-0009	0.0000	0.0358	0.0057	0.0146	0.1268	0.0000	0.0731	0.0744	0.0203	0.1438	0.0000	0.0401	0.0000	0.2782	0.1382	0.0000	0.0000	0.0000	0.0116	0.0000	0.0000	0.0373	0.01	0.2613	0.99
OSL2U-0152	0.0000	0.0000	0.0000	0.0095	0.0000	0.0949	0.0000	0.0000	0.0187	0.0000	0.0000	0.0180	0.0517	0.0538	0.0000	0.3437	0.0000	0.0000	0.0000	0.0887	0.1256	0.1955	0.01	0.2531	1.01
OSL2U-0208	0.0000	0.1676	0.1559	0.0157	0.0331	0.0668	0.0254	0.0000	0.0467	0.1214	0.0000	0.0000	0.0000	0.1227	0.0350	0.1564	0.0000	0.0000	0.0296	0.0000	0.0000	0.0236	0.00	0.4490	0.90
OSL2U-0040	0.0737	0.0000	0.0000	0.0000	0.2300	0.0000	0.0000	0.0000	0.0000	0.0000	0.0993	0.0575	0.2799	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1250	0.1345	0.0000	1.00	-0.0639	1.14
OSL2U-0514	0.0000	0.0643	0.0000	0.0000	0.0512	0.4666	0.0000	0.0000	0.0153	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0921	0.2137	0.0968	0.0000	0.95	-0.0289	1.15
OSL2U-0496	0.0000	0.0000	0.0000	0.0000	0.1340	0.0000	0.0000	0.0000	0.0906	0.0000	0.1709	0.0000	0.0000	0.0110	0.1512	0.0000	0.0724	0.0860	0.1489	0.0000	0.0000	0.1349	0.10	0.1505	1.05
OSL2U-0347	0.0061	0.0000	0.0218	0.0132	0.1772	0.0167	0.0842	0.0194	0.0000	0.2349	0.0000	0.0217	0.0000	0.0200	0.2090	0.0000	0.0186	0.0227	0.1345	0.0000	0.0000	0.0000	0.00	0.5402	0.84
OSL2U-0557	0.0089	0.0000	0.1630	0.0000	0.2390	0.0850	0.0000	0.0000	0.1481	0.0000	0.0676	0.0118	0.0019	0.1721	0.0000	0.0256	0.0000	0.0420	0.0352	0.0000	0.0000	0.0000	0.34	0.0669	1.06
OSL2U-0323	0.0350	0.0000	0.0303	0.2297	0.0711	0.0000	0.1593	0.0000	0.0443	0.1421	0.0000	0.0459	0.0298	0.0627	0.0454	0.0580	0.0097	0.0000	0.0367	0.0000	0.0000	0.0000	0.00	0.4622	0.89
OSL2U-0549	0.0000	0.3406	0.0634	0.0000	0.2170	0.0428	0.0000	0.0010	0.0000	0.1492	0.0000	0.0691	0.0105	0.0000	0.0150	0.0000	0.0000	0.0001	0.0858	0.0000	0.0056	0.0000	0.01	0.3124	0.97
OSL2U-0221	0.0000	0.0000	0.0000	0.1035	0.0000	0.0893	0.0000	0.0000	0.0233	0.0000	0.0000	0.0451	0.0182	0.0000	0.0000	0.0000	0.0489	0.0000	0.6627	0.0000	0.0000	0.0090	0.00	0.4690	0.91
OSL2U-0087	0.0000	0.0000	0.0655	0.0869	0.0260	0.0000	0.0000	0.0000	0.0000	0.1612	0.0000	0.1090	0.0000	0.0000	0.0002	0.0000	0.0080	0.0000	0.5239	0.0000	0.0000	0.0193	0.00	0.4232	0.93
OSL2U-0447	0.1995	0.0340	0.0000	0.0000	0.0027	0.1605	0.1704	0.0000	0.0000	0.0708	0.0000	0.0000	0.0423	0.0000	0.0000	0.0833	0.0000	0.0476	0.1361	0.0000	0.0529	0.0000	0.19	0.1054	1.03
OSL2R-3005	0.0063	0.0000	0.0000	0.2271	0.0000	0.0000	0.0512	0.0000	0.1024	0.0081	0.0000	0.0180	0.0249	0.2776	0.0923	0.0777	0.0289	0.0000	0.0000	0.0085	0.0000	0.0770	0.01	0.2919	0.98
OSL2U-0091	0.1839	0.0000	0.0000	0.0000	0.2454	0.0000	0.1327	0.0015	0.0000	0.0860	0.0000	0.0379	0.0000	0.0397	0.1272	0.0000	0.0000	0.0000	0.1335	0.0000	0.0122	0.0000	0.51	0.0314	1.09
OSL2U-0391	0.0000	0.1875	0.0000	0.1660	0.0000	0.0000	0.0000	0.0000	0.0000	0.1796	0.0000	0.0000	0.0288	0.0490	0.0000	0.0000	0.1016	0.0000	0.0000	0.1599	0.1277	0.0000	1.00	-0.0657	1.13
OSL2U-0123	0.0537	0.0000	0.0319	0.0878	0.0000	0.0128	0.1844	0.0424	0.0002	0.0692	0.0000	0.0185	0.0055	0.0997	0.1128	0.0000	0.0333	0.0017	0.2461	0.0000	0.0000	0.0000	0.00	0.3818	0.92
OSL2R-3043	0.0000	0.0000	0.0000	0.3184	0.0000	0.0000	0.0000	0.0000	0.0502	0.0000	0.0424	0.0000	0.0000	0.0000	0.0000	0.0000	0.0877	0.0000	0.0000	0.2320	0.1787	0.0906	1.00	-0.0676	1.16
OSL2U-0030	0.2078	0.0000	0.0000	0.0000	0.0731	0.0000	0.0000	0.0317	0.1152	0.0000	0.0892	0.0000	0.0000	0.3165	0.0000	0.0000	0.0000	0.0000	0.0000	0.1368	0.0000	0.0296	1.00	-0.0607	1.14
OSL2A-6146	0.0215	0.0000	0.0000	0.2797	0.0036	0.0000	0.0135	0.0674	0.0000	0.2677	0.0000	0.0565	0.0000	0.0512	0.0926	0.0349	0.0000	0.0000	0.0097	0.0000	0.1017	0.0000	0.00	0.4014	0.94
OSL2U-0520	0.0000	0.0000	0.1205	0.0555	0.0000	0.2653	0.0000	0.0000	0.0000	0.0366	0.0000	0.0000	0.0267	0.0000	0.0000	0.0539	0.0000	0.0000	0.2857	0.1558	0.0000	0.0000	0.00	0.4632	0.89
OSL2U-0515	0.0148	0.0000	0.0917	0.0705	0.2079	0.0000	0.0000	0.0000	0.0000	0.0000	0.1671	0.0000	0.0227	0.2900	0.0000	0.0000	0.0368	0.0000	0.0000	0.0664	0.0322	0.0000	0.95	-0.0275	1.12
OSL2R-3013	0.0000	0.0000	0.0000	0.0000	0.0991	0.0000	0.0000	0.0373	0.0884	0.2031	0.0000	0.0000	0.0000	0.4595	0.0517	0.0000	0.0037	0.0000	0.0000	0.0572	0.0000	0.0000	0.00	0.3193	0.98
OSL2R-3028	0.0000	0.1186	0.0000	0.0000	0.1083	0.0000	0.0000	0.0000	0.0367	0.0000	0.2311	0.0040	0.1559	0.1168	0.0000	0.0000	0.0000	0.0000	0.0000	0.2167	0.0117	0.0000	1.00	-0.0832	1.16
OSL2U-0288	0.0000	0.0901	0.0000	0.0000	0.3257	0.0000	0.0000	0.0000	0.1739	0.0000	0.2071	0.0000	0.0655	0.0000	0.0000	0.0000	0.0000	0.0000	0.0695	0.0432	0.0249	0.0000	1.00	-0.1245	1.18
OSL2U-0365	0.3147	0.0000	0.0000	0.0000	0.2556	0.0000	0.0000	0.0000	0.0000	0.0000	0.0437	0.0000	0.0000	0.0000	0.0000	0.0599	0.0000	0.0000	0.1198	0.0000	0.0000	0.2063	1.00	-0.1303	1.17
OSL2U-0190	0.0254	0.1112	0.0000	0.0000	0.2897	0.0961	0.2312	0.0000	0.0000	0.1095	0.0000	0.0000	0.0285	0.0402	0.0507	0.0000	0.0000	0.0167	0.0008	0.0000	0.0000	0.0000	0.00	0.3444	0.95
OSL2U-0485	0.0000	0.0000	0.0477	0.0000	0.0000	0.1272	0.0000	0.0000	0.1365	0.0362	0.0000	0.0598	0.0000	0.0000	0.0000	0.1108	0.0000	0.0000	0.4187	0.0000	0.0000	0.0632	0.05	0.1834	1.04
OSL2U-0450	0.0000	0.0899	0.0000	0.1697	0.0000	0.1000	0.0722	0.0000	0.0155	0.1193	0.0000	0.0861	0.0223	0.0057	0.0323	0.0690	0.1198	0.0297	0.0000	0.0080	0.0356	0.0251	0.00	0.3256	0.96
OSL2U-0145	0.0000	0.0360	0.0350	0.0000	0.0331	0.1661	0.0088	0.0000	0.0000	0.1540	0.0000	0.0465	0.0000	0.0000	0.0396	0.1485	0.0288	0.0000	0.3036	0.0000	0.0000	0.0000	0.23	0.0943	1.06
OSL2U-0026	0.0000	0.0966	0.0000	0.0046	0.0841	0.0000	0.0203	0.0000	0.0121	0.0168	0.0205	0.0000	0.0458	0.4338	0.0000	0.1647	0.0386	0.0000	0.0608	0.0000	0.0000	0.0013	0.05	0.1919	1.05
OSL2U-0222	0.0000	0.1033	0.0000	0.2530	0.0401	0.0000	0.0155	0.0000	0.0302	0.2213	0.0000	0.0058	0.0000	0.0000	0.1466	0.0578	0.0012	0.0000	0.1214	0.0000	0.0000	0.0036	0.00	0.5676	0.82
OSL2U-0217	0.0000	0.0000	0.2286	0.0000	0.2073	0.2210	0.0000	0.0000	0.0000	0.0000	0.1138	0.0000	0.0000	0.0000	0.0000	0.0845	0.0286	0.0000	0.0497	0.0663	0.0003	0.0000	1.00	-0.0842	1.12
OSL2U-0251	0.0000	0.1286	0.0273	0.0000	0.0000	0.0999	0.0000	0.0000	0.0000	0.0000	0.0000	0.0077	0.0744	0.0000	0.0000	0.0000	0.0492	0.0000	0.5650	0.0331	0.0064	0.0084	0.00	0.4408	0.92
OSL2U-0437	0.0000	0.0000	0.0000	0.0000	0.0000	0.0329	0.0000	0.0242	0.1379	0.0000	0.1053	0.0000	0.0164	0.4192	0.0000	0.0000	0.0388	0.0000	0.0000	0.1331	0.0922	0.0000	0.06	0.1668	1.06

Sample ID	B cells naive	B cells memory	Plasma cells	T cells CD8	T cells CD4 naive	T cells CD4 memory resting	T cells CD4 memory activated	T cells follicular helper	T cells regulatory Tregs	T cells gamma delta	NK cells resting	NK cells activated	Monocytes	Macrophages M0	Macrophages M1	Macrophages M2	Dendritic cells resting	Dendritic cells activated	Mast cells resting	Mast cells activated	Eosinophils	Neutrophils	P value	Pearson Correlation	RMSE
OSL2U-0198	0.0000	0.0000	0.2138	0.1461	0.0942	0.1093	0.0000	0.0000	0.0000	0.0000	0.0000	0.0737	0.1316	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1587	0.0726	0.0000	0.64	0.0198	1.09
OSL2U-0457	0.0000	0.1697	0.0000	0.1171	0.2598	0.0000	0.0000	0.0000	0.1450	0.0000	0.0000	0.0205	0.0000	0.1952	0.0000	0.0000	0.0000	0.0000	0.0000	0.0927	0.0000	0.0000	0.56	0.0283	1.10
OSL2U-0226	0.3872	0.0000	0.0210	0.0000	0.2077	0.1620	0.0000	0.0327	0.0000	0.0000	0.0774	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1121	0.0000	0.0000	0.94	-0.0210	1.15
OSL2U-0352	0.0000	0.1951	0.0000	0.2325	0.1889	0.0932	0.0000	0.0000	0.0531	0.0920	0.0000	0.0234	0.0395	0.0000	0.0224	0.0000	0.0000	0.0000	0.0305	0.0000	0.0295	0.0000	0.01	0.2201	1.03
OSL2U-0411	0.0000	0.0000	0.0983	0.0000	0.0308	0.2104	0.0000	0.0000	0.0865	0.0000	0.0000	0.0714	0.0124	0.1006	0.0000	0.0000	0.0077	0.0000	0.3819	0.0000	0.0000	0.0000	0.44	0.0403	1.10
OSL2U-0445	0.0504	0.0482	0.0000	0.0709	0.0034	0.4301	0.1192	0.0000	0.0000	0.0000	0.0000	0.0173	0.1204	0.0084	0.0337	0.0000	0.0000	0.0000	0.0981	0.0000	0.0000	0.0000	0.00	0.3449	0.95
OSL2U-0420	0.0000	0.0134	0.1224	0.0929	0.0000	0.2359	0.0000	0.0000	0.0000	0.1467	0.0000	0.0561	0.1929	0.0000	0.0000	0.0000	0.0000	0.0000	0.1199	0.0182	0.0016	0.0000	0.36	0.0532	1.08
OSL2U-0093	0.0000	0.0000	0.0000	0.0000	0.3831	0.0000	0.0000	0.0052	0.0502	0.0000	0.1764	0.0000	0.0000	0.2979	0.0000	0.0000	0.0000	0.0000	0.0000	0.0479	0.0393	0.0000	0.31	0.0719	1.10
OSL2U-0340	0.0125	0.0529	0.0000	0.0893	0.0052	0.2622	0.0933	0.0000	0.0000	0.2705	0.0000	0.0000	0.0000	0.0000	0.0669	0.0000	0.0000	0.0206	0.1253	0.0000	0.0014	0.0000	0.00	0.5519	0.83
OSL2U-0157	0.0000	0.0243	0.0288	0.0000	0.1303	0.0000	0.1146	0.0723	0.0467	0.0000	0.0664	0.0000	0.0210	0.2887	0.1660	0.0000	0.0000	0.0314	0.0040	0.0054	0.0000	0.0000	0.01	0.2369	1.01
OSL2U-0353	0.0000	0.0000	0.0357	0.2160	0.0842	0.0000	0.0000	0.0000	0.0797	0.0000	0.0000	0.1549	0.0000	0.0701	0.1110	0.0000	0.0000	0.0000	0.2484	0.0000	0.0000	0.0000	0.23	0.0926	1.09
OSL2U-0263	0.0420	0.1166	0.0059	0.0000	0.1650	0.0000	0.0000	0.1303	0.0000	0.0000	0.0000	0.0363	0.0000	0.3484	0.0000	0.0000	0.0000	0.0000	0.0000	0.1555	0.0000	0.0000	0.35	0.0642	1.10
OSL2U-0464	0.0127	0.0000	0.0038	0.0000	0.2089	0.2833	0.0000	0.0000	0.0000	0.0000	0.1044	0.0000	0.0000	0.0000	0.1687	0.0000	0.0000	0.0000	0.1970	0.0000	0.0212	0.0000	0.96	-0.0292	1.12
OSL2U-0043	0.0000	0.2649	0.3166	0.0933	0.0000	0.0000	0.0000	0.0000	0.0711	0.0434	0.0000	0.0129	0.0000	0.0000	0.0522	0.0070	0.0000	0.0026	0.1061	0.0000	0.0000	0.0301	0.00	0.3858	0.92
OSL2U-0220	0.0092	0.3315	0.0000	0.0000	0.1863	0.0755	0.0325	0.0000	0.0000	0.2059	0.0000	0.0198	0.0000	0.0005	0.0586	0.0000	0.0000	0.0294	0.0000	0.0244	0.0262	0.0000	0.00	0.3908	0.93
OSL2U-0552	0.0389	0.2517	0.0000	0.0188	0.1083	0.2243	0.1414	0.0000	0.0000	0.0000	0.0419	0.0000	0.0000	0.0097	0.0000	0.0391	0.0000	0.0179	0.1008	0.0000	0.0000	0.0073	0.06	0.1742	1.02
OSL2U-0319	0.0459	0.1801	0.0000	0.3688	0.0108	0.0000	0.1402	0.0098	0.0325	0.0586	0.0047	0.0300	0.0000	0.0000	0.0992	0.0000	0.0000	0.0000	0.0194	0.0000	0.0000	0.0000	0.06	0.1689	1.05
OSL2U-0390	0.0000	0.0000	0.0000	0.1049	0.0000	0.2316	0.0000	0.1577	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0694	0.0000	0.0000	0.2508	0.0861	0.0290	0.0705	0.80	-0.0001	1.12
OSL2U-0173	0.0000	0.0484	0.0000	0.1066	0.0000	0.1160	0.0136	0.0000	0.0000	0.0768	0.0000	0.0826	0.0670	0.0221	0.0352	0.0398	0.0700	0.0000	0.3220	0.0000	0.0000	0.0000	0.00	0.3469	0.95
OSL2U-0110	0.0869	0.0000	0.0238	0.3371	0.0000	0.0000	0.0000	0.0000	0.1739	0.0000	0.0000	0.0713	0.0000	0.0265	0.0124	0.0708	0.0000	0.0175	0.1797	0.0000	0.0000	0.0000	0.16	0.1126	1.06
OSL2U-0441	0.0000	0.0321	0.1760	0.0000	0.0000	0.1440	0.0000	0.0000	0.0578	0.0446	0.0000	0.0000	0.0000	0.0000	0.0000	0.0004	0.0030	0.0011	0.5410	0.0000	0.0000	0.0000	0.00	0.3971	0.94
OSL2U-0175	0.0248	0.0000	0.0402	0.0000	0.0000	0.0793	0.0256	0.0000	0.0867	0.1759	0.0000	0.0387	0.0274	0.0000	0.0000	0.0203	0.0163	0.0257	0.4391	0.0000	0.0000	0.0000	0.09	0.1530	1.06
OSL2U-0117	0.0000	0.2294	0.0000	0.0000	0.1191	0.0085	0.0000	0.0000	0.1136	0.0000	0.2392	0.0000	0.0476	0.0742	0.0000	0.0000	0.0000	0.0155	0.0000	0.0901	0.0242	0.0386	1.00	-0.1856	1.18
OSL2U-0360	0.1631	0.0000	0.0000	0.0000	0.1335	0.1811	0.0978	0.0000	0.0000	0.0000	0.1922	0.0000	0.0210	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1307	0.0806	0.0000	1.00	-0.0940	1.16
OSL2U-0100	0.0000	0.0500	0.0637	0.0375	0.0000	0.0225	0.0000	0.0000	0.0000	0.0000	0.0000	0.0017	0.0467	0.0000	0.0000	0.0069	0.0150	0.0000	0.6488	0.0901	0.0000	0.0171	0.00	0.5988	0.82
OSL2U-0104	0.0000	0.1824	0.0842	0.1012	0.0000	0.1481	0.0367	0.0205	0.0000	0.1535	0.0000	0.0000	0.0000	0.0000	0.0124	0.0710	0.0636	0.0000	0.0912	0.0000	0.0000	0.0354	0.00	0.3669	0.93
OSL2U-0406	0.0050	0.1141	0.0000	0.2759	0.0778	0.0489	0.0746	0.0000	0.0000	0.1876	0.0000	0.0520	0.0000	0.0009	0.0448	0.0119	0.0000	0.0041	0.1024	0.0000	0.0000	0.0000	0.00	0.4940	0.87
OSL2U-0132	0.0000	0.0000	0.0543	0.0000	0.1077	0.0610	0.0639	0.0000	0.0000	0.0309	0.0000	0.0000	0.0329	0.0000	0.0000	0.0000	0.0171	0.0000	0.6322	0.0000	0.0000	0.0000	0.00	0.4500	0.92
OSL2U-0017	0.0210	0.0000	0.0000	0.0143	0.1298	0.0000	0.0000	0.0000	0.0995	0.0000	0.0678	0.1018	0.1235	0.0000	0.0000	0.0000	0.0000	0.0000	0.2204	0.2038	0.0179	0.0000	0.25	0.0878	1.12
OSL2U-0142	0.0414	0.0000	0.0000	0.0000	0.0000	0.3659	0.0000	0.0000	0.0000	0.0000	0.0000	0.1741	0.0266	0.0197	0.0000	0.0000	0.0000	0.0000	0.1019	0.2463	0.0240	0.0000	0.17	0.1093	1.12
OSL2U-0201	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3079	0.0000	0.0000	0.0000	0.0980	0.0000	0.3339	0.1998	0.0000	0.0000	0.0000	0.0000	0.0283	0.0321	0.0000	0.23	0.0943	1.11
OSL2U-0073	0.0000	0.2078	0.0000	0.1155	0.0615	0.0017	0.1523	0.0358	0.0000	0.1583	0.0000	0.0826	0.0360	0.0073	0.0701	0.0045	0.0291	0.0004	0.0352	0.0000	0.0019	0.0000	0.00	0.3898	0.93
OSL2U-0183	0.0166	0.0000	0.0848	0.0000	0.0000	0.1732	0.0000	0.0000	0.1276	0.0000	0.0560	0.0000	0.0000	0.0000	0.0000	0.0000	0.0313	0.0000	0.4889	0.0216	0.0000	0.0000	0.01	0.2336	1.03
OSL2U-0513	0.1055	0.0000	0.0000	0.0000	0.0114	0.1837	0.0000	0.0000	0.1763	0.0000	0.1578	0.0000	0.1469	0.0767	0.0001	0.0000	0.0057	0.0000	0.1357	0.0000	0.0000	0.0000	0.94	-0.0206	1.10
OSL2U-0461	0.0000	0.0296	0.0000	0.0392	0.0318	0.2531	0.0000	0.0000	0.0000	0.0570	0.0000	0.0230	0.0000	0.0000	0.0000	0.0000	0.0382	0.0000	0.4049	0.0837	0.0395	0.0000	0.10	0.1432	1.08
OSL2U-0148	0.0083	0.0986	0.0062	0.1182	0.0000	0.3492	0.1874	0.0081	0.0000	0.0000	0.0000	0.0000	0.0043	0.0240	0.1067	0.0000	0.0310	0.0074	0.0449	0.0000	0.0000	0.0056	0.00	0.4847	0.87
OSL2U-0431	0.0000	0.1631	0.1088	0.0281	0.0000	0.3124	0.0000	0.0000	0.0000	0.1725	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0252	0.0000	0.1898	0.0000	0.0000	0.0000	0.37	0.0482	1.08
OSL2U-0550	0.0208	0.0000	0.0147	0.0000	0.1858	0.0425	0.0000	0.0084	0.0000	0.0000	0.0751	0.0000	0.1959	0.2688	0.0000	0.1853	0.0000	0.0000	0.0000	0.0027	0.0000	0.0000	0.25	0.0875	1.08
OSL2U-0310	0.0303	0.0488	0.0000	0.0000	0.0094	0.3495	0.0969	0.0000	0.0000	0.0905	0.0000	0.0502	0.0213	0.0000	0.0191	0.0833	0.0000	0.0397	0.1355	0.0000	0.0256	0.0000	0.03	0.2064	1.01
OSL2U-0337	0.0000	0.0276	0.0000	0.0027	0.0000	0.2687	0.0000	0.0000	0.0000	0.0134	0.0000	0.0736	0.0000	0.0000	0.0000	0.1101	0.1020	0.0000	0.4018	0.0000	0.0000	0.0000	0.00	0.4290	0.91

Appendix E: Data generated for survival analysis in METABRIC

Table 9. Overview of the data used in survival analysis in the METABRIC patient cohort. Each patient is annotated with PAM50 subtype, mean expression of the EMT-cluster genes and EMT score.

Sample ID	PAM50 subtype	Mean expression of EMT-cluster genes	EMT score
MB-2901	Luminal B	-2.750	-1.292
MB-4838	Luminal B	-2.633	-1.058
MB-0471	Luminal A	-2.605	-0.977
MB-0472	Luminal B	-2.470	-0.778
MB-0173	Luminal B	-2.454	-0.708
MB-0060	Luminal B	-2.292	-0.428
MB-0570	Luminal B	-2.219	-0.708
MB-5271	Luminal B	-2.248	-0.398
MB-0492	Luminal B	-2.253	-0.850
MB-3487	Luminal B	-2.175	-0.369
MB-3350	Luminal B	-2.181	-0.615
MB-0454	Luminal A	-2.137	-0.743
MB-7216	Luminal B	-2.100	-0.591
MB-0368	Luminal B	-2.082	-0.471
MB-0370	Luminal B	-2.076	-0.443
MB-7188	Luminal B	-2.088	-0.423
MB-0184	Luminal A	-2.105	-0.270
MB-0095	Luminal B	-2.055	-0.356
MB-5628	Luminal B	-2.036	-0.384
MB-7138	Luminal B	-2.042	-0.430
MB-0304	Luminal B	-2.029	-0.584
MB-7086	Luminal B	-2.013	-0.308
MB-7185	Luminal A	-2.007	-0.432
MB-0577	Luminal B	-2.013	-0.347
MB-7236	Luminal B	-1.979	-0.487
MB-5086	Luminal A	-1.924	-0.466
MB-6011	Luminal B	-1.873	-0.271
MB-4374	Luminal A	-1.866	-0.550
MB-4630	Luminal B	-1.815	-0.174
MB-0325	Luminal B	-1.824	0.088
MB-4906	Luminal A	-1.804	-0.560
MB-7157	Luminal B	-1.785	-0.201
MB-4849	Luminal A	-1.781	-0.244
MB-4970	Luminal B	-1.760	-0.153
MB-0125	Luminal B	-1.767	-0.283
MB-0225	Luminal B	-1.719	-0.309
MB-5617	Luminal A	-1.710	-0.509
MB-4091	Luminal A	-1.697	-0.424
MB-4017	Luminal B	-1.648	-0.473
MB-5160	Luminal B	-1.668	-0.133
MB-5370	Luminal B	-1.674	-0.025
MB-5244	Luminal B	-1.632	-0.280
MB-4602	Luminal B	-1.588	-0.098
MB-4912	Luminal B	-1.624	0.178
MB-0321	Luminal A	-1.646	-0.590
MB-5463	Luminal B	-1.627	-0.020
MB-7231	Luminal A	-1.625	-0.258
MB-4339	Luminal B	-1.573	-0.404
MB-0270	Luminal B	-1.602	-0.342
MB-5266	Luminal B	-1.547	-0.058
MB-4802	Luminal B	-1.568	-0.224
MB-4898	Luminal A	-1.565	-0.132
MB-0448	Luminal B	-1.524	0.149
MB-5291	Luminal B	-1.538	0.105
MB-5215	Luminal B	-1.580	-0.187
MB-2730	Luminal B	-1.531	-0.261
MB-4328	Luminal A	-1.519	-0.200
MB-7266	Luminal B	-1.537	-0.057
MB-7199	Luminal B	-1.514	-0.193
MB-3840	Luminal B	-1.500	-0.085
MB-5638	Luminal A	-1.515	-0.006
MB-5062	Luminal B	-1.502	0.044
MB-0427	Luminal A	-1.496	0.339
MB-6075	Luminal B	-1.452	0.026
MB-5636	Luminal B	-1.470	-0.198
MB-6154	Luminal B	-1.441	0.012
MB-5590	Luminal B	-1.419	0.225
MB-5121	Luminal B	-1.460	0.032
MB-0146	Luminal B	-1.463	-0.305
MB-0028	Luminal B	-1.416	0.156
MB-6213	Luminal A	-1.437	-0.061
MB-6183	Luminal B	-1.412	0.167
MB-0053	Luminal B	-1.353	0.176
MB-5514	Luminal A	-1.433	-0.175
MB-7132	Luminal B	-1.379	0.250
MB-3437	Luminal B	-1.379	0.016
MB-4998	Luminal B	-1.371	-0.204
MB-5502	Luminal B	-1.321	0.149
MB-5645	Luminal A	-1.344	0.087
MB-5140	Luminal B	-1.313	-0.216
MB-0119	Luminal B	-1.354	0.303
MB-6008	Luminal B	-1.277	0.250
MB-4675	Luminal A	-1.323	-0.196
MB-5550	Luminal A	-1.295	0.083
MB-2779	Luminal A	-1.285	-0.556
MB-5211	Luminal B	-1.259	0.228
MB-7263	Luminal B	-1.319	0.196
MB-4357	Luminal A	-1.239	0.223
MB-5604	Luminal B	-1.260	0.023
MB-4688	Luminal B	-1.262	-0.001
MB-0526	Luminal B	-1.273	0.369
MB-7226	Luminal B	-1.238	0.067
MB-4969	Luminal A	-1.236	0.279
MB-5122	Luminal A	-1.258	-0.060
MB-6012	Luminal A	-1.179	-0.123
MB-7037	Luminal A	-1.271	-0.112
MB-0630	Luminal B	-1.218	0.174
MB-0877	Luminal B	-1.188	0.149
MB-7229	Luminal A	-1.221	0.050
MB-7297	Luminal B	-1.181	0.118
MB-0440	Luminal B	-1.216	0.122

Sample ID	PAM50 subtype	Mean expression of EMT-cluster genes	EMT score
MB-7186	Luminal B	-1.162	0.270
MB-6145	Luminal B	-1.150	-0.202
MB-0513	Luminal A	-1.222	0.084
MB-6016	Luminal A	-1.167	0.045
MB-0434	Luminal B	-1.192	-0.165
MB-5101	Luminal B	-1.166	-0.178
MB-7253	Luminal B	-1.142	0.227
MB-7130	Luminal B	-1.127	0.177
MB-7234	Luminal B	-1.196	-0.098
MB-4930	Luminal B	-1.133	0.146
MB-4421	Luminal A	-1.116	0.163
MB-0324	Luminal A	-1.111	-0.153
MB-0064	Luminal B	-1.056	0.525
MB-3824	Luminal B	-1.133	0.026
MB-5562	Luminal B	-1.108	0.354
MB-5556	Luminal B	-1.104	0.478
MB-5646	Luminal B	-1.057	0.098
MB-4306	Luminal A	-1.047	0.145
MB-0056	Luminal B	-1.018	0.456
MB-0147	Luminal A	-1.049	0.493
MB-5341	Luminal B	-1.032	0.219
MB-5464	Luminal A	-1.078	0.035
MB-7094	Luminal B	-1.000	0.188
MB-6271	Luminal A	-1.069	0.021
MB-4213	Luminal B	-1.041	0.175
MB-0455	Luminal B	-0.998	0.419
MB-7124	Luminal A	-1.029	0.515
MB-5068	Luminal A	-1.030	-0.033
MB-3781	Luminal A	-1.001	0.224
MB-6026	Luminal B	-0.958	0.361
MB-3272	Luminal B	-0.975	0.162
MB-6124	Luminal A	-1.004	0.131
MB-0880	Luminal A	-1.013	0.392
MB-6149	Luminal A	-0.953	0.179
MB-0385	Luminal B	-0.950	0.055
MB-4996	Luminal A	-0.939	-0.266
MB-3092	Luminal B	-0.942	0.146
MB-3167	Luminal A	-0.938	0.361
MB-5433	Luminal A	-0.926	0.256
MB-0606	Luminal B	-0.949	0.277
MB-0167	Luminal B	-0.898	0.708
MB-7170	Luminal A	-0.921	0.162
MB-5434	Luminal B	-0.851	0.347
MB-4001	Luminal A	-0.849	0.070
MB-0097	Luminal A	-0.881	-0.297
MB-7099	Luminal B	-0.873	0.388
MB-4730	Luminal A	-0.876	0.227
MB-5601	Luminal A	-0.839	0.621
MB-3388	Luminal B	-0.819	0.004
MB-5580	Luminal A	-0.840	0.627
MB-4616	Luminal A	-0.795	0.326
MB-0046	Luminal A	-0.821	0.000
MB-7072	Luminal B	-0.787	0.187
MB-4148	Luminal B	-0.750	0.301
MB-5186	Luminal A	-0.799	-0.145
MB-6079	Luminal B	-0.747	0.337
MB-4965	Luminal A	-0.791	0.350
MB-2803	Luminal B	-0.769	0.609
MB-5251	Luminal A	-0.772	0.253
MB-4956	Luminal B	-0.748	0.164
MB-7171	Luminal B	-0.724	0.330
MB-5167	Luminal B	-0.754	0.468
MB-5525	Luminal B	-0.770	0.366
MB-5475	Luminal B	-0.765	0.677
MB-4642	Luminal B	-0.742	0.093
MB-4767	Luminal B	-0.725	0.551
MB-4749	Luminal A	-0.738	0.432
MB-0541	Luminal A	-0.736	0.553
MB-0646	Luminal B	-0.724	0.468
MB-5152	Luminal A	-0.719	0.825
MB-4744	Luminal B	-0.762	0.379
MB-0134	Luminal B	-0.706	0.680
MB-5396	Luminal B	-0.680	0.614
MB-0412	Luminal B	-0.667	0.567
MB-7095	Luminal A	-0.693	0.402
MB-5592	Luminal B	-0.662	0.527
MB-7276	Luminal A	-0.666	0.494
MB-4735	Luminal A	-0.666	0.270
MB-6001	Luminal A	-0.592	0.397
MB-0642	Luminal A	-0.680	0.639
MB-4323	Luminal A	-0.584	0.377
MB-0360	Luminal B	-0.642	0.676
MB-5197	Luminal A	-0.595	0.451
MB-4649	Luminal B	-0.558	0.315
MB-0328	Luminal B	-0.615	0.424
MB-7193	Luminal B	-0.583	0.575
MB-0258	Luminal A	-0.611	0.664
MB-6179	Luminal A	-0.604	0.477
MB-4737	Luminal B	-0.632	0.324
MB-5163	Luminal A	-0.661	0.802
MB-2990	Luminal A	-0.580	0.136
MB-5305	Luminal A	-0.586	0.512
MB-4233	Luminal B	-0.577	0.617
MB-0654	Luminal A	-0.596	0.599
MB-0501	Luminal A	-0.572	0.725
MB-5399	Luminal A	-0.562	0.718
MB-7011	Luminal B	-0.575	0.935
MB-7197	Luminal B	-0.527	0.582
MB-0637	Luminal B	-0.515	1.006
MB-3365	Luminal A	-0.539	0.686
MB-4977	Luminal A	-0.543	0.636
MB-2999	Luminal A	-0.478	0.523
MB-4829	Luminal B	-0.510	0.603
MB-5410	Luminal A	-0.488	0.392
MB-5290	Luminal A	-0.479	0.341
MB-2613	Luminal B	-0.438	0.613
MB-4564	Luminal A	-0.482	0.488
MB-4908	Luminal B	-0.482	0.573
MB-5647	Luminal A	-0.496	0.574
MB-4313	Luminal A	-0.432	0.425
MB-7162	Luminal A	-0.489	0.517
MB-2984	Luminal B	-0.478	0.475
MB-4333	Luminal A	-0.421	-0.180
MB-3088	Luminal A	-0.510	0.464
MB-5518	Luminal B	-0.464	0.612
MB-3021	Luminal A	-0.508	0.427
MB-3050	Luminal B	-0.446	0.562
MB-4834	Luminal B	-0.426	0.725
MB-4845	Luminal A	-0.514	0.418

Sample ID	PAM50 subtype	Mean expression of EMT-cluster genes	EMT score
MB-5273	Luminal B	-0.457	0.841
MB-4787	Luminal A	-0.456	0.317
MB-0185	Luminal B	-0.450	1.012
MB-5563	Luminal A	-0.462	0.690
MB-3102	Luminal A	-0.430	0.122
MB-4858	Luminal A	-0.452	0.346
MB-4750	Luminal B	-0.441	0.568
MB-7280	Luminal A	-0.394	0.222
MB-4999	Luminal A	-0.424	0.042
MB-5270	Luminal A	-0.430	0.533
MB-4618	Luminal B	-0.434	0.559
MB-7235	Luminal A	-0.447	0.416
MB-0586	Luminal A	-0.434	0.299
MB-5540	Luminal B	-0.408	0.569
MB-6083	Luminal A	-0.365	0.265
MB-5654	Luminal B	-0.427	0.893
MB-2858	Luminal B	-0.385	0.831
MB-5228	Luminal B	-0.423	0.529
MB-6288	Luminal B	-0.393	0.654
MB-5444	Luminal A	-0.407	0.678
MB-2947	Luminal A	-0.373	0.326
MB-7060	Luminal A	-0.318	0.224
MB-5486	Luminal B	-0.360	0.563
MB-4018	Luminal A	-0.331	0.818
MB-0123	Luminal B	-0.301	0.597
MB-4234	Luminal A	-0.326	0.656
MB-2618	Luminal B	-0.241	0.848
MB-0485	Luminal A	-0.328	0.181
MB-4644	Luminal B	-0.306	0.258
MB-2686	Luminal B	-0.279	1.092
MB-5369	Luminal A	-0.317	0.673
MB-7288	Luminal B	-0.324	0.702
MB-3351	Luminal A	-0.320	1.103
MB-5554	Luminal A	-0.310	0.734
MB-3253	Luminal B	-0.286	0.865
MB-0574	Luminal A	-0.319	1.039
MB-5221	Luminal A	-0.281	0.889
MB-7010	Luminal A	-0.310	0.902
MB-4872	Luminal A	-0.305	0.994
MB-0261	Luminal B	-0.327	0.917
MB-4752	Luminal A	-0.295	0.791
MB-0327	Luminal A	-0.281	0.628
MB-7219	Luminal A	-0.251	0.526
MB-4641	Luminal A	-0.267	0.561
MB-7137	Luminal A	-0.243	0.553
MB-6300	Luminal B	-0.298	0.871
MB-5626	Luminal A	-0.254	0.824
MB-5384	Luminal A	-0.191	0.627
MB-5011	Luminal A	-0.239	0.746
MB-0083	Luminal B	-0.140	0.711
MB-5322	Luminal A	-0.261	0.557
MB-0553	Luminal A	-0.217	0.539
MB-4846	Luminal B	-0.209	1.122
MB-4873	Luminal B	-0.204	0.982
MB-0882	Luminal B	-0.170	1.003
MB-0143	Luminal A	-0.183	0.965
MB-0172	Luminal A	-0.138	0.319
MB-5571	Luminal A	-0.156	1.097
MB-4994	Luminal B	-0.211	0.975
MB-3104	Luminal A	-0.138	0.124
MB-5049	Luminal B	-0.207	0.793
MB-2819	Luminal A	-0.144	0.925
MB-0406	Luminal B	-0.155	0.720
MB-0312	Luminal B	-0.149	0.907
MB-5144	Luminal B	-0.171	0.771
MB-3852	Luminal A	-0.076	0.839
MB-5623	Luminal B	-0.070	0.919
MB-0609	Luminal A	-0.097	0.392
MB-0126	Luminal A	-0.175	0.753
MB-0589	Luminal B	-0.136	0.506
MB-5048	Luminal A	-0.122	0.602
MB-5204	Luminal A	-0.182	0.850
MB-2952	Luminal A	-0.076	0.288
MB-5338	Luminal B	-0.070	1.169
MB-4771	Luminal A	-0.121	0.631
MB-0535	Luminal A	-0.088	0.933
MB-7015	Luminal B	-0.091	0.804
MB-2781	Luminal B	-0.099	0.706
MB-0232	Luminal A	-0.093	0.672
MB-4298	Luminal B	-0.046	0.528
MB-5105	Luminal A	-0.088	0.683
MB-0536	Luminal B	-0.041	1.189
MB-7056	Luminal A	-0.108	0.863
MB-5261	Luminal B	-0.082	1.273
MB-3007	Luminal B	-0.041	0.847
MB-0353	Luminal A	-0.049	0.818
MB-5451	Luminal A	-0.047	0.538
MB-3110	Luminal A	-0.038	0.934
MB-5472	Luminal A	-0.077	1.038
MB-6211	Luminal A	-0.033	0.894
MB-0449	Luminal A	0.001	0.323
MB-4691	Luminal A	-0.003	0.986
MB-5454	Luminal B	0.034	1.119
MB-7042	Luminal B	-0.010	1.119
MB-5632	Luminal B	0.024	1.204
MB-4801	Luminal A	0.050	0.963
MB-4716	Luminal A	0.033	1.050
MB-5074	Luminal A	-0.018	0.793
MB-3402	Luminal B	-0.013	0.496
MB-4011	Luminal A	0.044	0.841
MB-2760	Luminal B	-0.002	1.263
MB-0614	Luminal A	0.020	1.243
MB-0309	Luminal A	-0.024	0.918
MB-7299	Luminal B	0.055	1.160
MB-3016	Luminal B	0.028	1.053
MB-5553	Luminal B	0.055	1.093
MB-7150	Luminal B	0.088	0.762
MB-5377	Luminal B	0.035	0.987
MB-5519	Luminal A	0.045	1.121
MB-4230	Luminal A	0.085	0.843
MB-4266	Luminal A	0.042	0.765
MB-5404	Luminal A	0.084	1.092
MB-0529	Luminal A	0.083	0.687
MB-5585	Luminal A	0.062	0.501
MB-0260	Luminal A	0.047	0.814
MB-5206	Luminal A	0.085	0.855
MB-0059	Luminal A	0.059	0.939
MB-6344	Luminal B	0.094	0.981
MB-2960	Luminal A	0.137	0.736

Sample ID	PAM50 subtype	Mean expression of EMT-cluster genes	EMT score
MB-4687	Luminal A	0.136	0.822
MB-5459	Luminal B	0.146	0.814
MB-5402	Luminal A	0.171	1.115
MB-3711	Luminal A	0.138	1.080
MB-5182	Luminal A	0.093	0.670
MB-4173	Luminal A	0.131	1.028
MB-0459	Luminal B	0.143	0.907
MB-5060	Luminal B	0.102	0.823
MB-7164	Luminal A	0.150	0.690
MB-0176	Luminal A	0.176	0.339
MB-5360	Luminal A	0.147	1.073
MB-4623	Luminal A	0.194	0.959
MB-2614	Luminal A	0.207	0.812
MB-2791	Luminal A	0.172	0.954
MB-5040	Luminal B	0.234	1.062
MB-4853	Luminal A	0.209	0.833
MB-0144	Luminal B	0.154	1.082
MB-3026	Luminal A	0.187	1.456
MB-0544	Luminal B	0.162	0.991
MB-4282	Luminal A	0.237	0.660
MB-5189	Luminal A	0.222	0.920
MB-0631	Luminal A	0.208	0.808
MB-0571	Luminal A	0.222	1.279
MB-0336	Luminal B	0.214	0.982
MB-0413	Luminal A	0.200	1.049
MB-7062	Luminal A	0.189	0.896
MB-0591	Luminal B	0.200	0.898
MB-0295	Luminal A	0.216	0.603
MB-5388	Luminal A	0.240	1.005
MB-0585	Luminal B	0.294	1.226
MB-0317	Luminal A	0.191	1.205
MB-4721	Luminal A	0.243	0.833
MB-0162	Luminal A	0.282	1.084
MB-3028	Luminal B	0.217	1.270
MB-7278	Luminal A	0.303	0.609
MB-0550	Luminal B	0.277	0.966
MB-4627	Luminal A	0.283	0.905
MB-5131	Luminal A	0.276	0.981
MB-0301	Luminal A	0.282	1.550
MB-4825	Luminal A	0.257	1.358
MB-6171	Luminal A	0.326	0.921
MB-5603	Luminal A	0.310	0.993
MB-5397	Luminal A	0.318	1.108
MB-5424	Luminal A	0.345	1.020
MB-2610	Luminal A	0.354	0.714
MB-6029	Luminal A	0.333	1.212
MB-5349	Luminal A	0.298	1.114
MB-5576	Luminal B	0.367	1.429
MB-5318	Luminal B	0.309	0.927
MB-2863	Luminal A	0.381	0.794
MB-2747	Luminal A	0.291	1.297
MB-4709	Luminal A	0.322	0.810
MB-0202	Luminal A	0.339	1.096
MB-4666	Luminal A	0.373	0.828
MB-0507	Luminal A	0.383	0.823
MB-7032	Luminal A	0.347	0.955
MB-7254	Luminal A	0.406	0.799
MB-7286	Luminal B	0.335	1.360
MB-4860	Luminal A	0.393	1.326
MB-6071	Luminal A	0.446	1.155
MB-2642	Luminal B	0.433	1.348
MB-5118	Luminal A	0.388	0.886
MB-4805	Luminal A	0.409	1.292
MB-0117	Luminal A	0.433	0.964
MB-5583	Luminal A	0.440	1.241
MB-3033	Luminal A	0.377	1.093
MB-6207	Luminal A	0.382	0.951
MB-7041	Luminal A	0.404	0.993
MB-2801	Luminal A	0.491	1.018
MB-5629	Luminal A	0.495	1.414
MB-0514	Luminal B	0.462	1.391
MB-5382	Luminal A	0.476	1.536
MB-0180	Luminal A	0.403	1.242
MB-3360	Luminal B	0.495	1.221
MB-0122	Luminal A	0.449	0.861
MB-4986	Luminal A	0.448	0.979
MB-5597	Luminal A	0.442	1.012
MB-7006	Luminal A	0.472	0.865
MB-5227	Luminal A	0.503	0.972
MB-4033	Luminal A	0.506	1.185
MB-5284	Luminal A	0.472	1.053
MB-5455	Luminal A	0.538	0.914
MB-3490	Luminal B	0.497	1.302
MB-5195	Luminal A	0.508	1.313
MB-0218	Luminal B	0.515	1.532
MB-4633	Luminal A	0.475	1.161
MB-0006	Luminal B	0.528	1.174
MB-0598	Luminal B	0.512	1.099
MB-5635	Luminal A	0.502	1.254
MB-0359	Luminal A	0.500	0.875
MB-5226	Luminal B	0.553	1.063
MB-0503	Luminal A	0.497	0.953
MB-2711	Luminal A	0.517	1.009
MB-2708	Luminal B	0.524	1.449
MB-6018	Luminal A	0.579	1.196
MB-2750	Luminal A	0.480	1.044
MB-0124	Luminal A	0.546	1.051
MB-0345	Luminal A	0.524	0.748
MB-5591	Luminal A	0.540	1.284
MB-4681	Luminal A	0.551	0.902
MB-5050	Luminal A	0.503	1.036
MB-5158	Luminal A	0.520	1.288
MB-0356	Luminal A	0.555	1.327
MB-3547	Luminal A	0.574	0.984
MB-7005	Luminal B	0.548	1.343
MB-6150	Luminal B	0.602	1.087
MB-5059	Luminal A	0.547	0.972
MB-4862	Luminal B	0.576	1.645
MB-0397	Luminal A	0.628	0.868
MB-0197	Luminal B	0.577	1.505
MB-4961	Luminal A	0.594	1.279
MB-3235	Luminal B	0.651	1.251
MB-5613	Luminal B	0.635	1.571
MB-4698	Luminal A	0.627	1.149
MB-0268	Luminal B	0.628	1.093
MB-0010	Luminal B	0.641	1.155
MB-4705	Luminal A	0.651	0.751
MB-0383	Luminal B	0.648	1.264
MB-2786	Luminal A	0.593	1.223

Sample ID	PAM50 subtype	Mean expression of EMT-cluster genes	EMT score
MB-6232	Luminal A	0.603	1.319
MB-7128	Luminal A	0.665	1.169
MB-3002	Luminal A	0.623	1.164
MB-3344	Luminal A	0.636	0.748
MB-6024	Luminal A	0.689	1.065
MB-0607	Luminal A	0.657	0.942
MB-6225	Luminal A	0.665	1.288
MB-4832	Luminal A	0.675	1.181
MB-2971	Luminal A	0.699	1.311
MB-2838	Luminal A	0.777	1.344
MB-4822	Luminal A	0.702	1.451
MB-3439	Luminal A	0.709	1.319
MB-0166	Luminal A	0.683	0.985
MB-0419	Luminal A	0.661	1.296
MB-3037	Luminal A	0.668	1.256
MB-3008	Luminal A	0.715	1.231
MB-7220	Luminal B	0.759	1.538
MB-6322	Luminal B	0.735	1.236
MB-2953	Luminal B	0.788	1.649
MB-4870	Luminal A	0.741	1.364
MB-0579	Luminal B	0.708	1.500
MB-5499	Luminal A	0.787	1.320
MB-6238	Luminal A	0.731	1.329
MB-3049	Luminal A	0.811	1.266
MB-0099	Luminal B	0.904	1.596
MB-0904	Luminal A	0.777	1.454
MB-6233	Luminal B	0.729	1.443
MB-0899	Luminal A	0.815	1.257
MB-7106	Luminal A	0.796	1.037
MB-3871	Luminal A	0.818	1.100
MB-5330	Luminal A	0.745	1.472
MB-7004	Luminal A	0.757	1.295
MB-2626	Luminal B	0.923	1.443
MB-5053	Luminal A	0.829	1.360
MB-0408	Luminal A	0.897	1.416
MB-0341	Luminal A	0.846	1.141
MB-0226	Luminal A	0.897	1.176
MB-5589	Luminal A	0.884	1.325
MB-6195	Luminal A	0.904	1.481
MB-7093	Luminal A	0.904	1.481
MB-4212	Luminal A	0.921	1.045
MB-4867	Luminal A	0.933	1.520
MB-5395	Luminal A	0.929	1.371
MB-3103	Luminal B	0.962	1.431
MB-3430	Luminal A	1.008	1.067
MB-2931	Luminal B	0.968	1.711
MB-0605	Luminal A	0.973	1.368
MB-5324	Luminal A	0.944	1.187
MB-0236	Luminal A	0.975	1.109
MB-0580	Luminal A	0.990	1.395
MB-0374	Luminal B	0.966	1.536
MB-6297	Luminal A	0.984	1.476
MB-0133	Luminal A	0.970	1.077
MB-2994	Luminal A	1.050	1.328
MB-0398	Luminal B	1.024	1.737
MB-5489	Luminal A	1.072	1.377
MB-5143	Luminal A	1.004	1.192
MB-0170	Luminal A	0.994	1.432
MB-6118	Luminal A	1.149	1.511
MB-0382	Luminal A	1.091	1.480
MB-5428	Luminal A	1.133	1.317
MB-0243	Luminal B	1.087	1.617
MB-0491	Luminal A	1.114	1.555
MB-0386	Luminal A	1.102	1.650
MB-4762	Luminal A	1.088	1.673
MB-3266	Luminal B	1.146	1.958
MB-7091	Luminal A	1.174	1.298
MB-5092	Luminal A	1.193	1.203
MB-4967	Luminal A	1.187	1.626
MB-0266	Luminal A	1.164	1.500
MB-2767	Luminal B	1.197	1.713
MB-4293	Luminal A	1.208	1.428
MB-5614	Luminal A	1.236	1.458
MB-4966	Luminal B	1.208	1.880
MB-0463	Luminal A	1.185	1.707
MB-3403	Luminal A	1.221	1.697
MB-3379	Luminal A	1.220	1.806
MB-7244	Luminal A	1.160	1.746
MB-3754	Luminal B	1.318	1.876
MB-7217	Luminal A	1.299	1.561
MB-6230	Luminal A	1.239	1.640
MB-0425	Luminal A	1.207	1.788
MB-0239	Luminal A	1.202	1.572
MB-6273	Luminal A	1.242	1.347
MB-3301	Luminal B	1.258	1.541
MB-0130	Luminal A	1.260	1.405
MB-2848	Luminal A	1.333	1.851
MB-3252	Luminal A	1.294	1.183
MB-3823	Luminal A	1.320	2.018
MB-5267	Luminal A	1.330	1.239
MB-0891	Luminal A	1.312	1.558
MB-0636	Luminal A	1.285	1.641
MB-6051	Luminal A	1.385	1.456
MB-0649	Luminal A	1.376	1.628
MB-2969	Luminal A	1.449	1.936
MB-5184	Luminal A	1.360	1.442
MB-2916	Luminal A	1.396	1.520
MB-5264	Luminal A	1.446	1.583
MB-0138	Luminal A	1.426	1.441
MB-0256	Luminal A	1.434	1.518
MB-2624	Luminal A	1.507	1.679
MB-0145	Luminal A	1.407	1.366
MB-0511	Luminal A	1.432	1.685
MB-0224	Luminal A	1.477	1.622
MB-2772	Luminal B	1.486	2.023
MB-7077	Luminal A	1.506	1.711
MB-6168	Luminal A	1.555	1.801
MB-4941	Luminal A	1.545	1.709
MB-4779	Luminal A	1.523	1.873
MB-2669	Luminal A	1.537	1.864
MB-6108	Luminal A	1.654	1.603
MB-4883	Luminal A	1.627	1.700
MB-3254	Luminal A	1.636	1.822
MB-4981	Luminal A	1.658	1.827
MB-0111	Luminal A	1.646	1.655
MB-4764	Luminal A	1.632	1.805
MB-0411	Luminal A	1.739	1.673
MB-6138	Luminal B	1.763	2.283
MB-5191	Luminal A	1.723	1.693

Sample ID	PAM50 subtype	Mean expression of EMT-cluster genes	EMT score
MB-0505	Luminal A	1.781	1.695
MB-5412	Luminal A	1.740	1.903
MB-3228	Luminal A	1.770	1.856
MB-5541	Luminal A	1.975	1.966
MB-0247	Luminal A	2.055	2.352
MB-0204	Luminal A	2.028	2.047
MB-0315	Luminal A	2.184	1.962
MB-0599	Luminal A	2.471	2.129

Appendix F: Overview of the data generated for OSL2

Table 10. Overview of the generated OSL2 data. The table provides an overview of all the information generated for the OSL2 patient cohort in this study including the dedicated patient group of the samples, tumor purity estimated by ASCAT, PAM50 subtype, EMT score, Infiltration level from *in silico* nanodissection, proliferation score, stemness score, *ESR1* expression and CpG-cluster A and CpG-cluster B mean methylation.

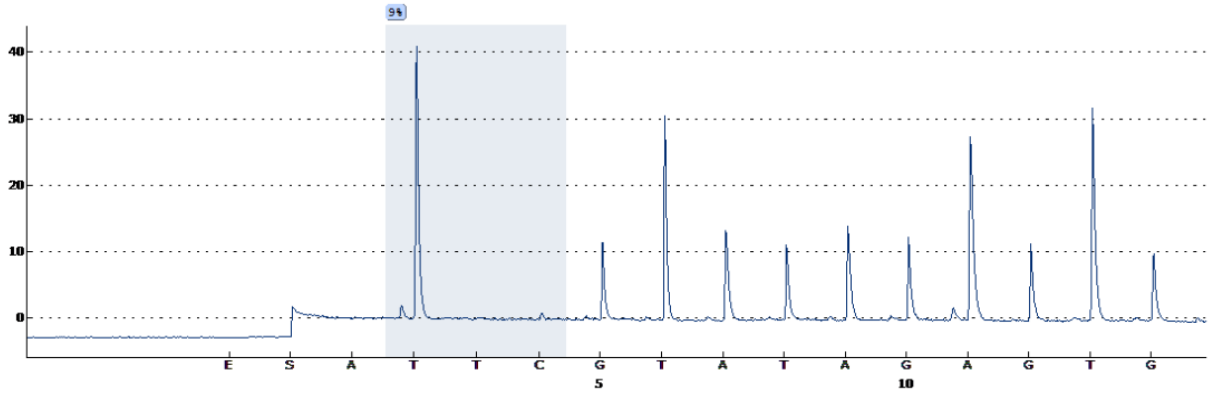
Sample ID	Patient group	Tumor purity	PAM50 subtype	EMT score	Infiltration level	Proliferation score	Stemness score	*ESR1* expression	CpG-cluster A mean methylation	CpG-cluster B mean methylation
OSL2U-0442	A	0.41	Luminal B	-1.5330	-0.1169	1.9883	-0.4709	0.256	0.83	0.24
OSL2U-0090	A	0.48	Luminal B	-1.9322	0.8955	1.3036	-0.2409	-1.190	0.82	0.22
OSL2U-0358	A	0.57	Luminal B	-2.4278	-0.2070	1.4959	-0.8037	2.334	0.82	0.22
OSL2R-3002	A	NA	Luminal B	-1.6927	-0.8339	2.0854	-0.9296	0.754	0.82	0.24
OSL2U-0407	A	0.32	Luminal B	-1.8849	-0.6896	1.7176	-1.9796	0.065	0.82	0.18
OSL2U-0059	A	1.00	Luminal B	-1.6947	-0.9976	1.1435	-1.3043	1.828	0.82	0.24
OSL2U-0534	A	0.67	Luminal B	-0.9391	1.0327	0.5286	-0.0798	1.365	0.82	0.26
OSL2U-0334	A	0.52	Luminal B	-0.0667	1.5553	1.0417	-0.0415	-0.624	0.81	0.31
OSL2U-0262	A	NA	Luminal B	-0.5375	0.5361	0.7641	-0.3153	0.476	0.80	0.34
OSL2U-0289	A	0.78	Luminal B	-0.7743	-0.0539	1.1160	-1.5593	0.391	0.80	0.17
OSL2U-0472	A	0.61	Luminal A	-1.0818	-0.5085	0.0053	0.9277	0.720	0.80	0.24
OSL2U-0486	A	0.71	Luminal B	-1.2018	-1.2158	0.4895	0.4241	0.524	0.79	0.28
OSL2R-3030	A	NA	Luminal A	-0.5992	-0.8375	-1.2076	0.4448	1.153	0.79	0.22
OSL2U-0269	A	0.66	Luminal B	1.0867	0.0045	0.1496	-0.0707	-0.053	0.79	0.36
OSL2U-0020	A	0.54	Luminal B	-1.0926	0.7012	1.6416	-1.2036	1.899	0.78	0.15
OSL2U-0392	A	0.73	Luminal A	-0.8449	0.2650	-2.0041	1.0011	-0.722	0.78	0.22
OSL2U-0370	A	0.74	Luminal A	-1.6056	-0.8494	-0.1125	-1.8553	1.592	0.78	0.22
OSL2U-0383	A	0.46	Luminal A	-0.7667	-0.4166	-1.4580	-1.9952	-0.035	0.78	0.35
OSL2U-0101	A	0.78	Luminal B	-0.3348	-1.1354	0.4687	0.3887	1.900	0.78	0.33
OSL2U-0064	A	0.51	Luminal B	-1.0737	-0.5827	1.5933	-1.2623	-1.668	0.77	0.19
OSL2U-0364	A	0.68	Luminal B	-0.6741	-0.5223	1.1865	-1.7171	0.605	0.77	0.36
OSL2A-6181	A	NA	Luminal B	-1.2020	-0.1587	1.9214	-0.7258	1.606	0.77	0.27
OSL2U-0488	A	0.64	Luminal B	-0.0176	-0.5645	0.8276	-0.4200	-2.202	0.77	0.25
OSL2U-0229	A	NA	Luminal B	0.1003	-1.1245	0.3225	0.7883	-0.363	0.77	0.43
OSL2U-0422	A	0.45	Luminal B	-0.2587	1.3125	0.6899	0.8276	0.520	0.77	0.30
OSL2U-0489	A	0.79	Luminal A	-0.3907	-0.8930	-2.1765	1.6634	-0.046	0.77	0.27
OSL2U-0079	A	0.39	Luminal B	1.1179	1.2650	-0.4163	0.0257	-1.313	0.77	0.35
OSL2U-0179	A	NA	Luminal A	0.7550	-0.5423	-0.3315	-0.5588	-0.749	0.77	0.34
OSL2R-3052	A	NA	Luminal B	0.3704	0.4973	0.3305	-0.6212	-1.737	0.77	0.25
OSL2A-6133	A	NA	Luminal B	-1.6661	1.8586	0.9310	-1.1944	0.352	0.76	0.28
OSL2U-0342	A	0.63	Luminal B	-0.9985	1.3739	0.8244	0.5898	1.124	0.76	0.22
OSL2U-0027	A	NA	Luminal A	0.1100	-0.1821	-0.7509	-0.8755	-0.803	0.76	0.43
OSL2U-0325	A	0.44	Luminal B	0.9513	-0.2519	0.4382	-0.7442	0.951	0.75	0.35
OSL2U-0379	A	0.45	Luminal B	-0.5169	1.2285	0.6858	1.0066	1.820	0.75	0.22
OSL2U-0544	A	0.58	Luminal B	0.0905	-0.6001	1.2526	1.1968	-0.709	0.75	0.25
OSL2U-0242	A	NA	Luminal A	-1.3569	-1.0795	-0.5257	-0.5902	0.508	0.75	0.32
OSL2U-0209	A	0.72	Luminal B	-1.3463	-0.3604	0.2454	-1.8133	1.606	0.75	0.24
OSL2U-0120	A	NA	Luminal A	-0.8570	-0.6308	-1.9881	-1.9588	0.992	0.75	0.33
OSL2U-0519	A	0.69	Luminal A	0.1465	-0.4686	-0.8341	-0.7298	0.499	0.75	0.31
OSL2U-0086	A	0.62	Luminal A	0.4978	-0.4870	-0.3248	-0.3530	-0.611	0.75	0.30
OSL2U-0296	A	0.59	Luminal A	-0.9812	-0.6495	0.9062	-1.2477	0.786	0.75	0.37
OSL2U-0171	A	0.74	Luminal A	-0.2158	-0.7149	-0.1442	-1.0443	0.892	0.75	0.27
OSL2R-3053	A	0.81	Luminal A	0.0596	-0.3737	-0.3986	0.4627	0.544	0.74	0.29
OSL2A-6166	A	NA	Luminal A	-1.5970	-0.9540	0.4484	0.4106	1.399	0.74	0.29
OSL2U-0280	A	0.61	Luminal B	-0.4792	0.7772	2.1096	-0.7303	0.257	0.74	0.23
OSL2U-0250	A	0.62	Luminal B	0.1381	0.3785	0.4791	-2.1167	1.097	0.74	0.29
OSL2U-0285	A	0.50	Luminal B	-0.0508	0.7411	-0.0332	-0.7587	-0.639	0.74	0.27
OSL2U-0318	A	0.60	Luminal A	0.3675	-0.0686	0.2808	0.7452	-1.091	0.74	0.31
OSL2U-0484	A	0.49	Luminal B	-0.7852	-0.5471	0.3271	-1.0146	1.207	0.74	0.22
OSL2U-0232	A	0.69	Luminal A	-0.3639	-1.0955	-0.0748	-1.1896	-0.063	0.74	0.29
OSL2U-0069	A	0.68	Luminal B	0.2745	0.1487	0.8379	0.1442	-0.503	0.74	0.31
OSL2U-0328	A	0.67	Luminal A	0.2630	0.3442	-0.0698	-1.0844	1.460	0.73	0.36
OSL2U-0223	A	0.52	Luminal A	1.0007	-0.6906	-0.1393	-0.9150	-0.462	0.73	0.31
OSL2U-0083	A	0.55	Luminal A	0.3996	-0.5514	-0.3384	-1.3535	0.324	0.73	0.27
OSL2U-0291	A	0.58	Luminal B	-0.2099	-0.4783	0.4654	-0.4806	0.777	0.73	0.25
OSL2U-0260	A	0.65	Luminal A	0.0371	-0.4022	-0.3447	-0.8518	0.810	0.72	0.28
OSL2U-0299	A	0.71	Luminal A	0.1393	-0.5494	-1.8714	-0.6077	0.966	0.72	0.29
OSL2U-0174	A	0.75	Luminal A	-0.0018	-0.9369	0.5534	-1.6769	0.910	0.72	0.31
OSL2U-0477	A	0.53	Luminal B	-0.4109	0.8256	1.0960	0.7877	-0.266	0.72	0.21
OSL2U-0367	B	0.53	Luminal B	0.0933	-0.5728	0.1743	0.1366	0.638	0.72	0.29
OSL2U-0243	B	NA	Luminal A	-0.7927	-0.3625	0.5567	1.1205	-0.567	0.72	0.31
OSL2U-0548	B	0.71	Luminal B	-0.4951	-0.3944	0.6240	-0.8579	-1.229	0.72	0.37
OSL2U-0553	B	0.77	Luminal A	-2.8997	-1.2572	-0.1466	-2.6356	2.231	0.72	0.31
OSL2U-0491	B	0.71	Luminal B	0.1708	-0.0150	-0.1733	-0.1859	0.601	0.72	0.33
OSL2U-0197	B	0.70	Luminal A	1.0963	-0.7314	0.1747	-0.3942	0.574	0.72	0.35
OSL2U-0400	B	0.46	Luminal B	-1.5888	0.4323	1.1875	-1.1523	-0.484	0.71	0.32
OSL2U-0053	B	0.65	Luminal A	1.3537	-0.2461	-0.1171	-0.9474	1.119	0.71	0.29
OSL2U-0096	B	0.43	Luminal A	-0.7335	-0.1310	1.2689	-0.6277	0.804	0.71	0.34
OSL2U-0393	B	0.61	Luminal A	-1.9202	-0.7368	-0.3914	-0.0242	1.350	0.71	0.30
OSL2U-0037	B	0.61	Luminal A	1.0176	0.2554	-0.9979	0.3109	0.095	0.71	0.35
OSL2U-0267	B	0.52	Luminal B	0.5494	-0.2564	1.7886	-0.7282	-0.847	0.71	0.31
OSL2U-0126	B	0.71	Luminal A	1.6318	-0.5771	-1.9358	1.8808	-2.014	0.71	0.32
OSL2U-0324	B	0.55	Luminal B	-0.0084	-0.2716	1.5306	-0.9869	-0.421	0.71	0.43
OSL2U-0225	B	NA	Luminal B	1.6624	0.4726	0.6527	0.7537	-0.365	0.71	0.39
OSL2U-0056	B	NA	Luminal A	0.9664	0.3756	-0.2532	-0.0595	0.410	0.70	0.35
OSL2U-0275	B	0.64	Luminal B	1.4105	0.8671	0.9184	-0.7474	-0.157	0.70	0.28
OSL2U-0130	B	0.66	Luminal B	-0.7687	0.5621	1.6473	0.7628	0.512	0.70	0.31
OSL2A-6134	B	NA	Luminal A	2.0330	0.6223	-0.0318	-0.0103	-2.839	0.70	0.40
OSL2U-0109	B	0.71	Luminal B	-0.1639	0.1247	0.7400	0.2699	0.067	0.70	0.32
OSL2U-0112	B	NA	Luminal A	0.1188	-0.6494	-0.2001	0.1715	1.090	0.70	0.28
OSL2U-0554	B	0.65	Luminal A	-0.3240	-0.2565	-0.8331	-0.0546	-0.419	0.70	0.36
OSL2U-0438	B	0.68	Luminal A	0.9424	0.1194	-2.4952	2.2465	-0.870	0.70	0.35
OSL2U-0128	B	NA	Luminal B	0.6730	0.0241	0.1244	0.6897	0.747	0.70	0.32
OSL2U-0085	B	NA	Luminal A	0.6751	0.2236	-0.5611	0.6489	0.088	0.69	0.36
OSL2U-0249	B	0.60	Luminal A	0.1501	-0.6963	-0.3367	0.1964	0.647	0.69	0.34
OSL2U-0248	B	NA	Luminal A	0.4194	0.5667	-0.1955	-0.7618	-0.735	0.69	0.34
OSL2U-0268	B	0.55	Luminal A	-1.0567	-0.0816	-1.1964	-0.6914	1.362	0.69	0.33
OSL2U-0463	B	0.72	Luminal A	-0.8359	-0.8688	-0.8557	0.6406	0.615	0.69	0.35
OSL2U-0186	B	NA	Luminal A	-0.9390	-0.5320	0.3859	0.9699	0.090	0.69	0.26
OSL2A-6132	B	NA	Luminal A	0.6882	2.1047	1.9501	-0.6163	-2.747	0.69	0.30
OSL2U-0430	B	0.71	Luminal B	0.1525	-0.9268	0.4764	0.5099	0.680	0.69	0.36
OSL2U-0537	B	0.91	Luminal A	-0.3339	-1.2234	-0.1043	-0.4248	0.918	0.69	0.20
OSL2U-0290	B	0.68	Luminal B	-0.0828	-0.8607	1.0603	1.3311	-0.104	0.68	0.26
OSL2U-0227	B	0.69	Luminal A	-0.1443	-1.5629	-1.5554	0.2100	1.533	0.68	0.31
OSL2U-0182	B	0.50	Luminal A	-0.1516	-0.4157	-0.2290	-0.0046	0.889	0.68	0.34
OSL2U-0531	B	0.72	Luminal A	-0.3191	-1.0721	0.4039	0.0565	1.020	0.68	0.29

Sample ID	Patient group	Tumor purity	PAM50 subtype	EMT score	Infiltration level	Proliferation-score	Stemness score	*ESR1* expression	CpG-cluster A mean methylation	CpG-cluster B mean methylation
OSL2U-0159	B	NA	Luminal B	0.6278	1.3791	0.2941	-0.2677	0.090	0.68	0.29
OSL2U-0045	B	NA	Luminal A	0.3803	-0.4036	-1.6638	-0.4889	-0.540	0.68	0.32
OSL2U-0065	B	NA	Luminal A	0.1613	-0.6936	-1.2199	1.6853	-0.469	0.68	0.35
OSL2U-0462	B	0.72	Luminal A	1.6740	-0.0872	-1.3800	0.3309	-1.715	0.68	0.32
OSL2U-0177	B	0.57	Luminal A	0.9459	-0.1111	-1.3159	0.5987	-0.535	0.68	0.39
OSL2U-0103	B	0.69	Luminal B	-0.8681	-1.1253	0.9457	-0.3605	1.414	0.68	0.23
OSL2U-0009	B	0.63	Luminal B	0.8815	0.4345	1.8419	1.0420	-0.413	0.68	0.33
OSL2U-0152	B	NA	Luminal A	0.1595	0.0019	-0.1445	-0.2493	-0.995	0.68	0.30
OSL2U-0208	B	0.47	Luminal B	0.0559	0.8852	0.4211	-1.3604	-0.129	0.67	0.37
OSL2U-0040	B	NA	Luminal A	-0.2930	-1.0997	-2.1598	1.1700	1.080	0.67	0.30
OSL2U-0514	B	0.66	Luminal A	1.0977	-0.7796	-0.9630	0.6711	0.444	0.67	0.36
OSL2U-0496	B	0.63	Luminal A	-0.2715	-0.3704	0.7411	-1.6004	0.205	0.67	0.30
OSL2U-0347	B	0.28	Luminal A	-0.1923	0.5716	0.0332	-0.4392	-0.151	0.67	0.40
OSL2U-0557	B	0.55	Luminal B	-0.6351	-0.4948	1.4868	-0.6088	0.303	0.67	0.29
OSL2U-0323	B	0.33	Luminal A	1.0898	1.0588	-0.5690	1.5779	-0.381	0.67	0.41
OSL2U-0549	B	0.41	Luminal A	-0.1914	0.3484	0.2651	0.2624	-1.221	0.67	0.36
OSL2U-0221	B	0.53	Luminal A	1.2469	-0.4777	-0.6690	-0.5722	-0.696	0.67	0.34
OSL2U-0087	B	0.48	Luminal B	0.9246	-0.1246	-0.1222	0.0726	-1.267	0.66	0.25
OSL2U-0447	B	0.60	Luminal A	-0.7390	0.0283	0.7841	-1.0624	-0.497	0.66	0.29
OSL2R-3005	B	0.67	Luminal A	0.8053	0.3578	-0.0725	-0.9552	-0.734	0.66	0.32
OSL2U-0091	B	0.37	Luminal A	-0.0540	-0.0486	0.1543	1.8780	1.882	0.66	0.37
OSL2U-0391	B	0.82	Luminal A	0.3778	-0.3221	-0.8107	1.0270	-0.466	0.66	0.30
OSL2U-0123	C	0.32	Luminal B	-1.0082	0.7245	2.3623	-0.8152	1.468	0.66	0.39
OSL2R-3043	C	0.72	Luminal B	0.2499	-0.4401	1.1649	-0.9281	-0.395	0.66	0.20
OSL2U-0030	C	0.54	Luminal B	0.7901	-0.8479	1.8100	-0.6702	-0.457	0.66	0.39
OSL2A-6146	C	NA	Luminal B	0.0024	0.5583	0.2815	-0.5618	1.037	0.65	0.37
OSL2U-0520	C	0.51	Luminal A	-0.2522	-0.3118	-0.2089	2.3692	-0.373	0.65	0.41
OSL2U-0515	C	1.00	Luminal A	-0.2666	-0.5757	-1.1587	2.5965	-0.820	0.65	0.33
OSL2R-3013	C	0.55	Luminal B	0.1828	-0.0159	1.0858	0.0286	-0.303	0.65	0.35
OSL2R-3028	C	NA	Luminal A	0.9068	-0.6205	-0.7082	0.5849	0.035	0.65	0.36
OSL2U-0288	C	0.74	Luminal B	-0.4403	-0.7033	1.6604	-0.4710	-0.468	0.65	0.30
OSL2U-0365	C	0.49	Luminal A	0.5895	-0.8633	-0.4405	0.9277	0.883	0.65	0.34
OSL2U-0190	C	NA	Luminal A	-0.2490	1.1073	1.0620	-1.0622	-0.388	0.65	0.42
OSL2U-0485	C	0.63	Luminal A	0.0302	-0.2960	-1.0133	1.9893	-0.628	0.65	0.19
OSL2U-0450	C	0.49	Luminal A	0.8287	0.6059	-0.7531	-0.3970	-1.238	0.65	0.39
OSL2U-0145	C	NA	Luminal A	0.2223	0.1395	-1.1401	0.7761	0.184	0.65	0.44
OSL2U-0026	C	0.39	Luminal A	0.6134	0.5142	-0.2546	-0.0434	0.113	0.65	0.36
OSL2U-0222	C	0.43	Luminal A	0.5447	1.0373	0.4715	1.1579	-0.082	0.64	0.44
OSL2U-0217	C	0.38	Luminal A	-0.0036	-0.2883	0.4082	0.3905	-0.390	0.64	0.38
OSL2U-0251	C	0.57	Luminal A	0.3204	-0.0374	-1.3515	0.6908	-0.649	0.64	0.31
OSL2U-0437	C	0.59	Luminal A	0.7545	-0.5841	-0.4851	-1.4822	-0.188	0.64	0.24
OSL2U-0198	C	0.43	Luminal A	0.9817	-0.3847	-0.2251	-0.1724	0.099	0.64	0.43
OSL2U-0457	C	0.62	Luminal A	0.6038	-0.2684	-0.7278	2.1384	1.419	0.64	0.35
OSL2U-0226	C	0.63	Luminal A	0.3356	-0.5387	-0.9084	0.0277	-0.924	0.64	0.38
OSL2U-0352	C	0.57	Luminal A	-0.2352	-0.0466	0.2019	0.2470	-0.693	0.63	0.29
OSL2U-0411	C	0.94	Luminal A	0.5318	-0.5897	0.2476	0.4780	-1.831	0.63	0.32
OSL2U-0445	C	0.48	Luminal A	0.5074	0.5564	-1.1632	0.6511	-1.648	0.63	0.42
OSL2U-0420	C	0.47	Luminal A	0.0218	-0.1857	-2.1664	0.6276	0.050	0.63	0.42
OSL2U-0093	C	NA	Luminal A	0.7228	-0.8336	-0.6226	0.4182	-0.440	0.63	0.34
OSL2U-0340	C	0.37	Luminal A	-0.0030	0.9669	0.0340	0.5560	-0.137	0.63	0.36
OSL2U-0157	C	0.54	Luminal A	-0.4542	0.5624	-0.2633	0.0125	0.927	0.63	0.26
OSL2U-0353	C	1.00	Luminal A	0.1562	-0.0454	-0.2558	0.0129	-1.363	0.63	0.37
OSL2U-0263	C	0.60	Luminal A	0.0680	-0.6086	-1.0854	1.9539	0.462	0.63	0.43
OSL2U-0464	C	0.52	Luminal A	0.6628	-0.3978	-0.8145	1.2844	-0.044	0.62	0.44
OSL2U-0043	C	0.46	Luminal A	1.2237	0.4178	-0.0772	-0.5970	-1.654	0.62	0.34
OSL2U-0220	C	0.30	Luminal A	-0.0650	0.9873	-0.7130	-0.0947	0.511	0.62	0.44
OSL2U-0552	C	0.45	Luminal A	-0.2082	0.2581	-0.4422	0.5360	-1.940	0.62	0.38
OSL2U-0319	C	0.53	Luminal A	-0.9099	0.3469	-1.1067	1.5254	-0.662	0.62	0.26
OSL2U-0390	C	0.45	Luminal A	-0.2573	-0.2391	0.1746	-0.7908	1.337	0.62	0.34
OSL2U-0173	C	NA	Luminal A	2.0959	0.3975	-0.1446	1.6972	-0.552	0.60	0.39
OSL2U-0110	C	NA	Luminal A	0.8903	-0.0127	-0.5014	0.4335	-0.255	0.60	0.32
OSL2U-0441	C	0.43	Luminal A	0.8714	-0.0621	-0.3152	-0.0389	-0.905	0.60	0.39
OSL2U-0175	C	0.51	Luminal A	-1.8088	0.1704	-1.1824	0.3657	-0.119	0.60	0.27
OSL2U-0117	C	0.57	Luminal B	0.2039	-0.4948	0.2981	0.2121	0.908	0.60	0.24
OSL2U-0360	C	0.54	Luminal A	0.1023	-0.5052	0.5686	1.1982	-0.157	0.60	0.38
OSL2U-0100	C	0.28	Luminal A	1.7846	-0.2373	-0.5351	1.5442	-2.228	0.59	0.39
OSL2U-0104	C	0.31	Luminal A	0.0074	0.9523	0.1000	-0.4470	-0.346	0.59	0.44
OSL2U-0406	C	0.38	Luminal A	0.5342	0.9579	-1.3551	0.1450	-1.792	0.59	0.44
OSL2U-0132	C	NA	Luminal A	0.8002	-0.3421	-0.7557	0.8504	0.324	0.59	0.43
OSL2U-0017	C	0.47	Luminal A	1.1531	-0.8620	-1.7697	0.5228	-0.832	0.58	0.41
OSL2U-0142	C	0.48	Luminal A	0.7311	-0.3421	0.3641	0.2694	-0.442	0.57	0.32
OSL2U-0201	C	0.45	Luminal B	1.1869	-0.1980	0.4395	-0.8098	-0.317	0.57	0.43
OSL2U-0073	C	0.37	Luminal A	0.1799	1.3934	0.6913	1.0553	-0.719	0.57	0.47
OSL2U-0183	C	NA	Luminal A	1.3545	-0.5888	-1.1180	1.5615	-1.264	0.57	0.46
OSL2U-0513	C	0.34	Luminal A	0.4056	-0.1349	-0.8998	0.5861	-0.458	0.56	0.48
OSL2U-0461	C	0.37	Luminal A	0.0990	-0.4877	-1.1052	0.2969	0.378	0.55	0.45
OSL2U-0148	C	1.00	Luminal A	-0.1011	0.7253	-0.1095	-0.5126	-0.204	0.55	0.50
OSL2U-0431	C	0.32	Luminal A	0.6979	0.0235	-0.8791	1.8836	-1.062	0.54	0.49
OSL2U-0550	C	0.31	Luminal A	0.5402	-0.4149	-1.1021	-0.2116	-1.603	0.53	0.52
OSL2U-0310	C	0.81	Luminal A	1.4782	0.4552	-0.5122	1.0644	0.343	0.53	0.54
OSL2U-0337	C	0.31	Luminal A	1.2864	-0.2385	-1.1360	0.7441	-0.424	0.53	0.48

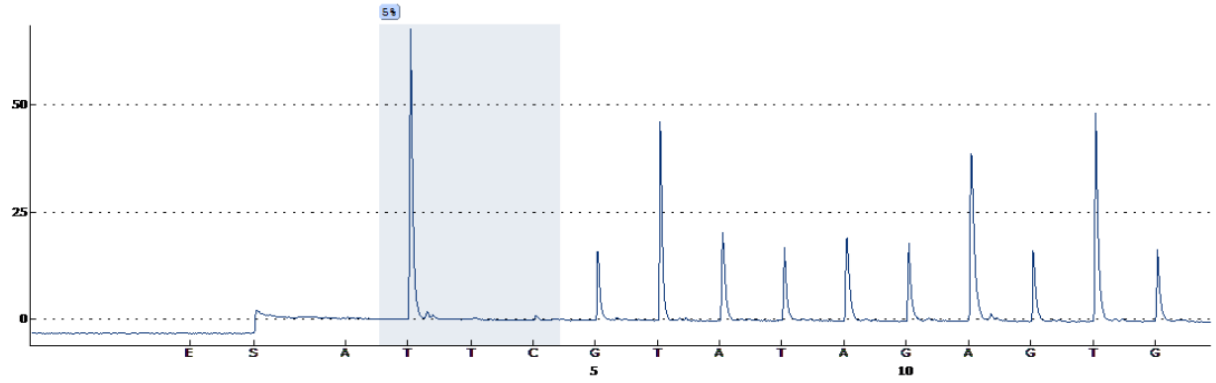
Appendix G: Pyrograms of the local DNA sequence around the target CpGs for D492 and D492M

Table 11. Pyrograms of the local DNA sequence of the target CpGs for the D492 and D492M cell lines. The value in the blue boxes represents the percent of the DNA fragments in the sample that were methylated.

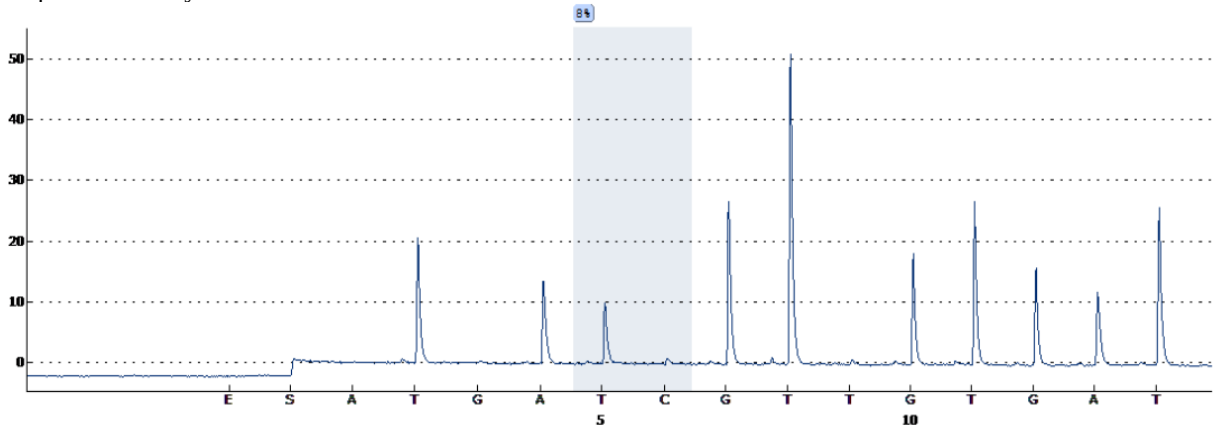
Assay: cg06947286
Sample: D492
Sequence to analyze: TTTYGTTTATAGAAGTTTGAATGTATTTT



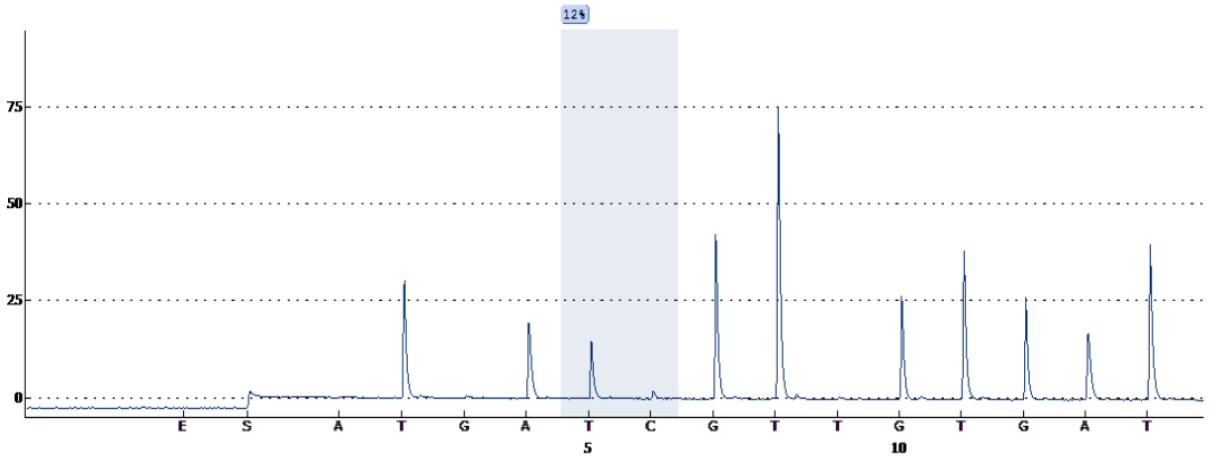
Assay: cg06947286
Sample: D492M
Sequence to analyze: TTTYGTTTATAGAAGTTTGAATGTATTTT



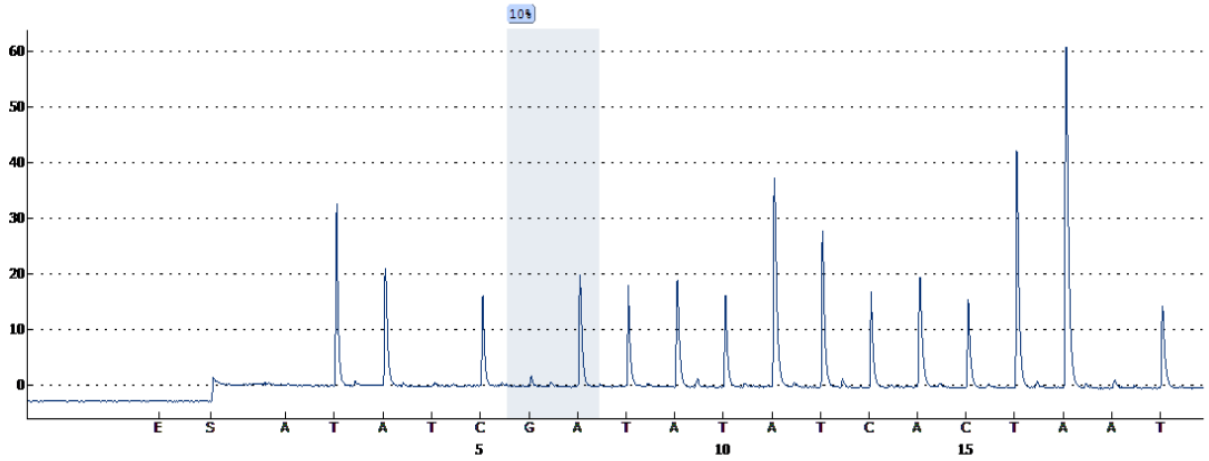
Assay: cg05223441
Sample: D492
Sequence to analyze: TTAYGGGTTTTTTGGTTTGGATTTA



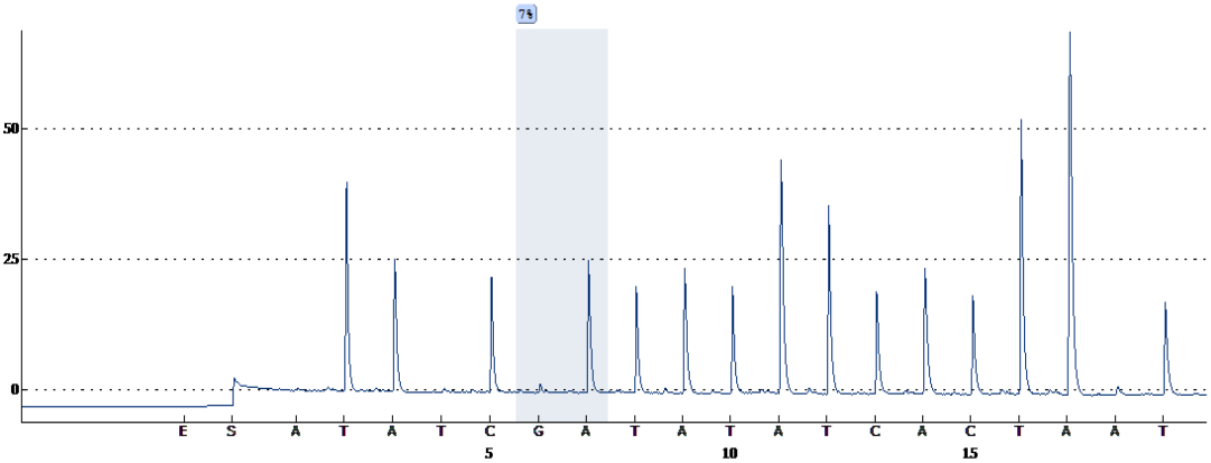
Assay: cg05223441
Sample: D492M
Sequence to analyze: TTAYGGGTTTTTTGGTTTGGATTTA



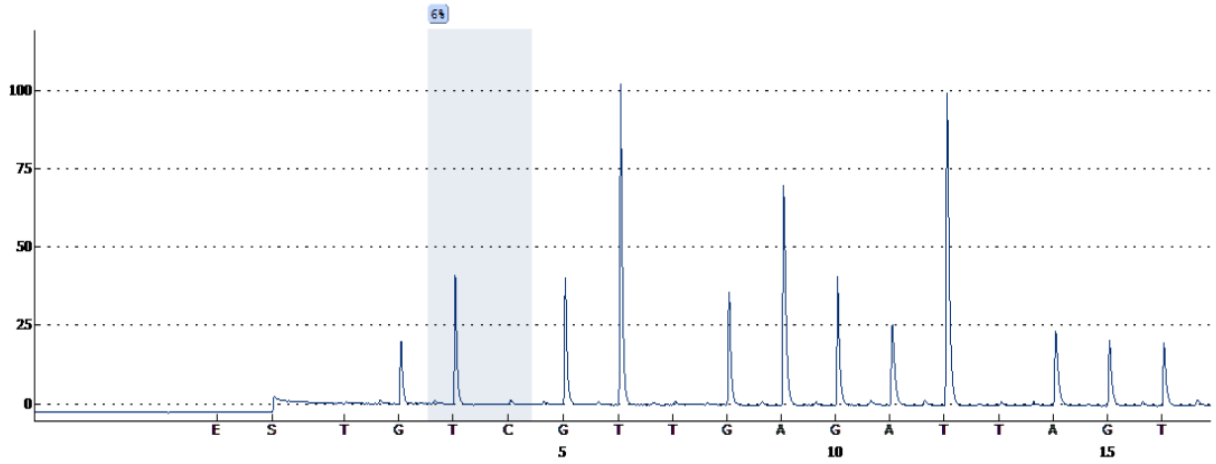
Assay: cg12232146
Sample: D492
Sequence to analyze: TTACRTATAATTCACTTTAAAATAAATTTC



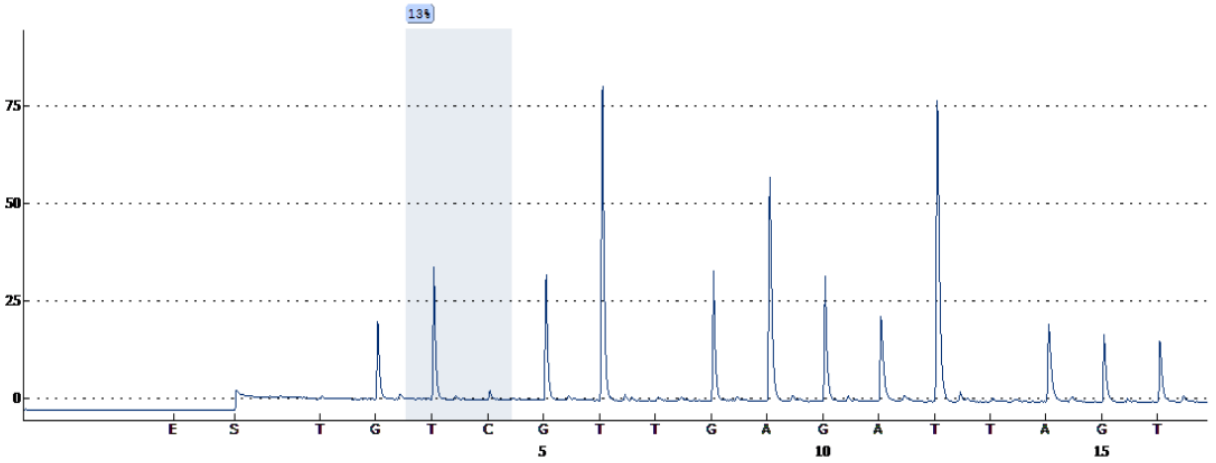
Assay: cg12232146
Sample: D492M
Sequence to analyze: TTACRTATAATTCACTTTAAAATAAATTTC



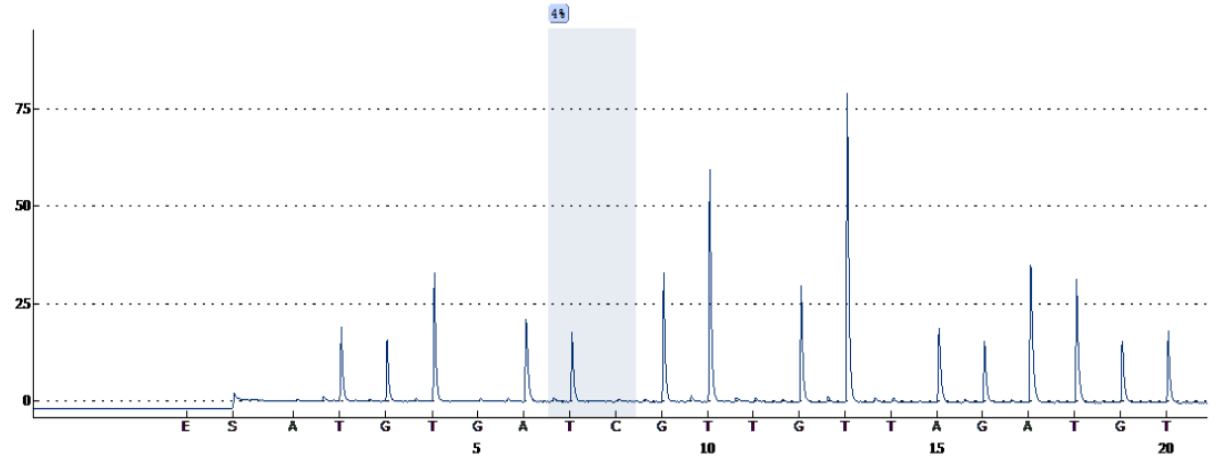
Assay: cg16888565
Sample: D492
Sequence to analyze: GTYGGTTTTTTGGAAAGGATTTTTTAGT



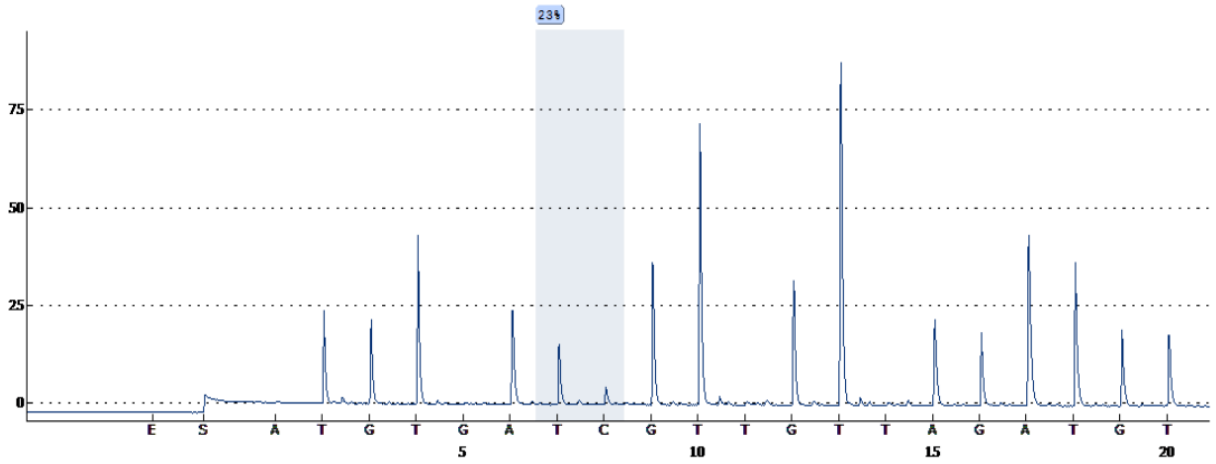
Assay: cg16888565
Sample: D492M
Sequence to analyze: GTYGGTTTTTTGGAAAGGATTTTTTAGT



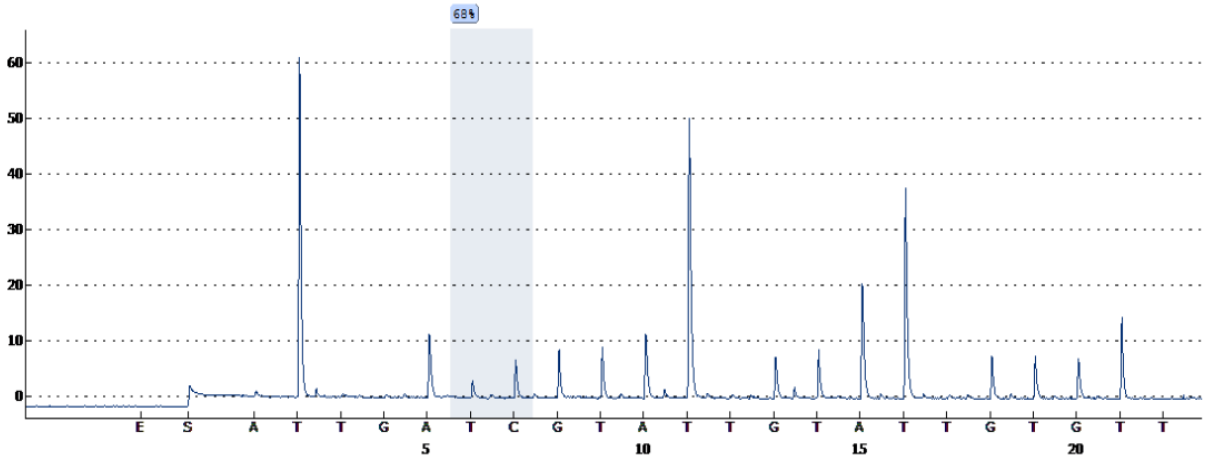
Assay: cg20909017
Sample: D492
Sequence to analyze: TGTTAYGGTTTTGGTTTTTTAGAATTGTGGG



Assay: cg20909017
Sample: D492M
Sequence to analyze: TGTTAYGGTTTTGGTTTTTTAGAATTGTGGG



Assay: cg10233454
Sample: D492
Sequence to analyze: TTTTTTTAYGTATTTTTTTGTAATTTTTGTGTTG



Assay: cg10233454
Sample: D492M
Sequence to analyze: TTTTTTTAYGTATTTTTTTGTAATTTTTGTGTTG