

# Deciphering molecular heterogeneity and relevance of subtypes in breast cancer progression

**Helga Bergholtz**

Department of Cancer Genetics  
Institute for Cancer Research  
Division of Cancer Medicine  
Oslo University Hospital

Faculty of Medicine  
University of Oslo



UiO : **University of Oslo**



© **Helga Bergholtz, 2020**

*Series of dissertations submitted to the  
Faculty of Medicine, University of Oslo*

ISBN 978-82-8377-602-7

All rights reserved. No part of this publication may be reproduced or transmitted, in any form or by any means, without permission.

Cover: Hanne Baadsgaard Utigard.  
Print production: Reprintentralen, University of Oslo.

## ACKNOWLEDGMENTS

The work presented in this thesis has been carried out at the Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital (the Norwegian Radium Hospital) during the period of 2012-2019. I want to thank the Faculty of Medicine at the University of Oslo for admitting me into the PhD program and the South-Eastern Norway Regional Health Authority for financial support.

The road was long, but it has been a great journey thanks to all the fantastic people I have been surrounded by during my period as a PhD-student. First of all, I want to thank my main supervisor Therese for giving me the opportunity to join your group and for the invaluable support you provide, both academically and personally. You have a seemingly endless positivity and your open office door is highly appreciated. Jens Henrik, I want to thank you for accepting the dubious task of becoming my supervisor. Your fountain of endless ideas and suggestions has been both a blessing and a challenge for me. I highly appreciate your mentoring in the animal department and in the lab. Thank you Anne-Lise for being a great inspiration, I am proud to have had the opportunity to work with you.

To all current and former members of the *Breast Tumor Initiation and Progression* group: Thank you so much! We are a great team, and we should consider patenting our unique combination of fun and academic discussion. A special thanks to Tonje who has provided vital psycho-statistical support. I also want to thank all my other wonderful colleagues in the department of cancer genetics, especially Laxmi and Inger for making bad VY-days so much better, and to all my different office-mates for important coffee-breaks and extracurricular laughs. A special recognition goes to the exceptional engineering staff in the department. You are the backbone of the lab, and you do an amazing job. Thank you Eldri, Phuong and Veronica for patiently supervising a lab-novice like me and Gry for administering it all. I also want to thank all co-authors and external collaborators for interesting discussions and great input.

Mamma, Pappa, Tore and Wenche, I appreciate your unconditional support and invaluable help in taking care of kids and dogs. To all my friends outside of work who have encouraged me and patiently listened to my frustrations: I am finally done! Johan and Alfred, thank you for helping me see that there are more important things to life than work. And to my loyal furry therapists Diva and Klara, thank you for getting me out in fresh air and for never *ever* asking me when I planned to finish my thesis. Finally, I want to express my deep gratitude to Raymond who has been endlessly supporting, patient and caring throughout the whole journey.



Ski, October 2019



## Table of Contents

LIST OF PAPERS .....	I
ABBREVIATIONS .....	II
INTRODUCTION .....	1
Cancer genomics .....	1
Normal breast anatomy and physiology .....	4
Ductal carcinoma in situ .....	6
Breast tumor progression .....	10
Breast cancer invasion .....	12
The role of the tumor microenvironment in breast cancer progression .....	14
Molecular subtyping of breast tumors .....	18
Comparative breast cancer biology .....	20
AIMS .....	24
RESULTS IN BRIEF .....	25
Paper I: A longitudinal study of the association between mammographic density and gene expression in normal breast tissue .....	25
Paper II: Claudin-low-like mouse mammary tumors show distinct transcriptomic patterns uncoupled from genomic drivers .....	26
Paper III: Contrasting DCIS and invasive breast cancer by subtype suggests basal-like DCIS as distinct lesions .....	27
Paper IV: Comparable cancer–relevant mutation profiles in synchronous ductal carcinoma in situ and invasive breast cancer .....	28
METHODOLOGICAL CONSIDERATIONS .....	29
Material .....	29
Material in paper I .....	29
Material in paper II .....	29
Material in paper III .....	30
Material in paper IV .....	31
Gene expression microarrays .....	32
Gene expression data analyses .....	33
Data preprocessing .....	33
Subtyping of mammary gland tumors .....	33

Creating a DCIS score using multivariate logistic regression .....	35
DNA copy number analyses .....	37
DNA methylation analyses .....	38
Targeted DNA sequencing .....	39
Statistical considerations .....	41
ETHICAL CONSIDERATIONS .....	43
DISCUSSION.....	45
Molecular subtyping of breast tumors .....	45
Subtype specific breast tumor progression .....	46
The role of the microenvironment in breast tumor progression .....	50
CONCLUSIONS AND FUTURE PERSPECTIVES.....	53
REFERENCES .....	55

## LIST OF PAPERS

- I. **A longitudinal study of the association between mammographic density and gene expression in normal breast tissue**  
*Helga Bergholtz, Tonje Gulbrandsen Lien, Giske Ursin, Marit Muri Holmen, Åslaug Helland, Therese Sørлие and Vilde Drageset Haakensen.*  
Journal of Mammary Gland Biology and Neoplasia 2019, **24**, 163–175.
  
- II. **Claudin-low-like mouse mammary tumors show distinct transcriptomic patterns uncoupled from genomic drivers**  
*Christian Fougner, Helga Bergholtz, Raoul Kuiper, Jens Henrik Norum and Therese Sørлие.*  
Breast Cancer Research 2019. **21**, 85.
  
- III. **Contrasting DCIS and invasive breast cancer by subtype suggests basal-like DCIS as distinct lesions**  
*Helga Bergholtz, Tonje Gulbrandsen Lien, David M. Swanson, Arnaldo Frigessi, Oslo Breast Cancer Research Consortium (OSBREAC), Jörg Tost, Maria Grazia Daidone, Fredrik Wärnberg, and Therese Sørлие.*  
Manuscript
  
- IV. **Comparable cancer-relevant mutation profiles in synchronous ductal carcinoma in situ and invasive breast cancer**  
*Helga Bergholtz, Surendra Kumar, Fredrik Wärnberg, Torben Lüders, Vessela Kristensen and Therese Sørлие.*  
Manuscript

## ABBREVIATIONS

<b>ADH</b>	Atypical ductal hyperplasia	<b>MIN</b>	Mammary intraepithelial neoplasia
<b>BCS</b>	Breast conserving surgery	<b>MIND</b>	Mouse mammary intraductal method
<b>BM</b>	Basement membrane	<b>miRNA</b>	Micro-ribonucleic acid
<b>CAF</b>	Cancer associated fibroblast	<b>MMP</b>	Matrix metalloproteinase
<b>COSMIC</b>	Catalogue of somatic mutations in cancer	<b>MMTV</b>	Mouse mammary tumor virus
<b>cPCDH</b>	Clustered protocadherins	<b>MPA</b>	Medroxyprogesterone acetate
<b>DCIS</b>	Ductal carcinoma in situ	<b>PCR</b>	Polymerase chain reaction
<b>dNTP</b>	Deoxyribonucleotide triphosphate	<b>PDX</b>	Patient derived xenograft
<b>ddPCR</b>	Digital droplet PCR	<b>PLC</b>	Pregnancy-lactation cycle
<b>DMBA</b>	7,12-dimethylbenz[a]anthracene	<b>PR</b>	Progesterone receptor
<b>ECM</b>	Extracellular matrix	<b>RNA</b>	Ribonucleic acid
<b>EMT</b>	Epithelial to mesenchymal transition	<b>RNAseq</b>	RNA sequencing
<b>ER</b>	Estrogen receptor	<b>RT</b>	Radiation therapy
<b>FDR</b>	False discovery rate	<b>RT-qPCR</b>	Reverse transcription quantitative PCR
<b>FEA</b>	Flat epithelial atypia	<b>SMA</b>	Smooth muscle actin
<b>FFPE</b>	Formalin fixed paraffin embedded	<b>SMMHC</b>	Smooth muscle myosin heavy chain
<b>GEMM</b>	Genetically engineered mouse model	<b>TDLU</b>	Terminal ductal lobular unit
<b>HER2</b>	Human epidermal growth factor receptor 2	<b>TEB</b>	Terminal end bud
<b>IBC</b>	Invasive breast cancer	<b>TGF<math>\beta</math></b>	Transforming growth factor $\beta$
<b>IHC</b>	Immunohistochemistry	<b>TIL</b>	Tumor infiltrating lymphocyte
<b>MaSC</b>	Mammary stem cell	<b>TME</b>	Tumor microenvironment
<b>MD</b>	Mammographic density	<b>TP53</b>	Tumor protein 53
<b>MEC</b>	Myoepithelial cells		

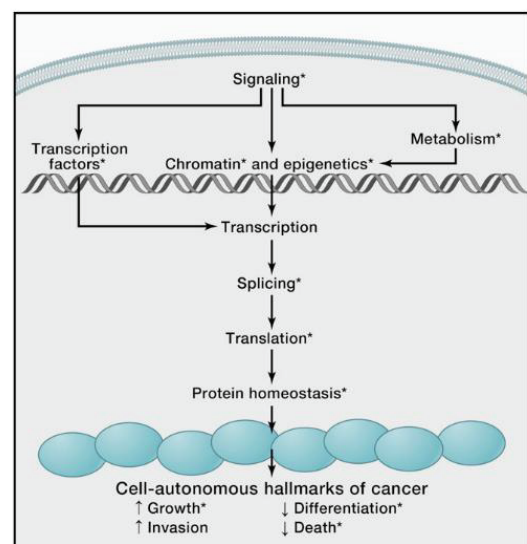


## INTRODUCTION

Cancer is a group of diseases where normal cells in the body change and gain the ability to multiply out of normal control with potential for invasion of surrounding tissue and dissemination to distant organs<sup>1</sup>. Cancer is an ancient disease and has been described in papyrus scripts from Egypt dated to year 1500 B.C. and even back then, cancerous tumors were surgically removed. The word cancer was presumably first used by Hippocrates around 400 B.C. to explain the disease's resemblance to a crab with legs that spread out into the surrounding tissue<sup>2</sup>. Cancer occurring in the mammary glands is called breast cancer. Breast cancer has been diagnosed in excavated mummies and it is believed that also Renaissance paintings show women suffering from breast cancer<sup>3,4</sup>. Today, breast cancer is the most common cancer in women across the world<sup>5</sup>. In 2018, the number of new breast cancer cases worldwide was estimated to 2.08 million and there were around 626.000 breast cancer related deaths<sup>6</sup>. In Norway in 2017, 3589 women were diagnosed with breast cancer and 623 women died of the disease. Compared to many other cancers, the long term survival of breast cancer is good, and in Norway from 2013-2017, the relative 5 year survival was 90.4%<sup>7</sup>.

## Cancer genomics

The central dogma of molecular biology is the theory of how information flows in the cell from DNA via mRNA to proteins (Figure 1)<sup>8,9</sup>. The genome of a mammalian cell consists of all its genetic material, i.e. the DNA. It may be subject to a multitude of different aberrations, for instance single base substitutions (point mutations), insertions or deletions of small or large segments of DNA, rearrangements and copy number changes (amplifications and deletions)<sup>10</sup>. Genes are transcribed into mRNA and these molecules constitute the *transcriptome* of the cell. The mRNA then serves as a template for translation into proteins<sup>11</sup>. mRNA molecules may be subject to splicing before translation, yielding different isoforms of proteins based on the same original



**Figure 1. Genetic alterations may disrupt cellular processes at multiple different levels and contribute to cancer formation. Reproduced from Garraway et al. with permission from Elsevier<sup>13</sup>.**

## INTRODUCTION

---

mRNA molecule. The set of all expressed proteins is called the *proteome*. Proteins are the executive part in a cell, they provide structure, serve as enzymes, and are involved in transport and many other tasks. Proteins are the end result of the transcription and translation, and proteomic investigation is undoubtedly very informative. However, studying the proteome is more challenging than studying DNA and mRNA; especially since there exist >100.000 different proteins due to post translational modifications. The correlation between gene and protein expression is, in many cases, low. This may be due to differences in translational efficiency, splicing events, different rate of degradation of proteins and mRNA molecules, or technical issues<sup>12</sup>. *Genomics* is a general term for the systematic studies of (some or all of) the genome or its products, e.g. DNA, mRNA or proteins<sup>11,13</sup>.

Already in 1902 did the German zoologist Theodor Boveri postulate a genetic basis for cancer: “A malignant tumor cell is (...) a cell with a specific abnormal chromosome constitution.”<sup>14,15</sup> He also presented ideas of inhibitory and stimulatory chromosomes which when perturbed could drive the cell into abnormal cell division. This harmonizes well with what today is known about tumor suppressor genes and proto-oncogenes<sup>1</sup>. Tumor suppressor genes are genes that in normal cells work to slow down cell division, repair DNA damage or induce apoptosis (programmed cell death), i.e. they act as cell-proliferation “brakes”. In contrast, proto-oncogenes are genes that usually stimulate cell division. If a tumor suppressor gene is inactivated (e.g. by mutation or deletion), the brakes of the cell proliferation

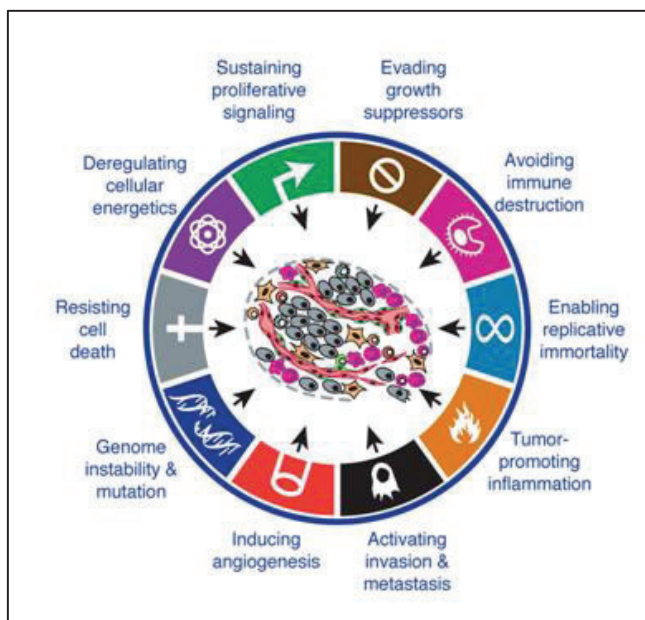


Figure 2. The hallmarks of cancer. Adapted from Hanahan et al.<sup>18</sup> with permission from Elsevier.

machinery may fail to inhibit proliferation. Likewise, if a proto-oncogene is activated (e.g. by mutation or amplification), it may turn into an oncogene that fuels proliferation in an abnormal way. Both these situations may lead to increased cell proliferation, a fundamental characteristic of cancer cells<sup>16</sup>. However, cell proliferation alone is not sufficient for development of a tumor. In two seminal papers, Hanahan and Weinberg proposed several additional key features that cancer cells need to overcome to “succeed” as a tumor, the so-called *hallmarks of cancer* that are illustrated in Figure 2<sup>17,18</sup>.

The discovery that mutations in normal genes may cause cells to transform to cancer cells, motivated researchers to start the formidable work of deriving the sequence of the whole human genome<sup>19</sup>. This work was initiated in 1990 and completed in 2003<sup>20</sup>. During this time, the sequencing technology improved significantly and with the development of massive parallel sequencing technology and the availability of a complete human reference genome, cancer research was led into the *genomic era*<sup>21</sup>. With the human genome as reference, somatic mutations (those that arise in the tumor), could be identified and catalogued<sup>22</sup>.

A common characteristic of cancer is genomic instability, which generates genetic diversity, enabling the cell to acquire features necessary for growth and progression<sup>18</sup>. Most human cancers arise due to only a few (two to eight) mutations/alterations that occur in a cell sequentially over time. Each of these *driver* mutations causes a selective growth advantage to the cell, in contrast to mutations that give no advantage to the cell, so-called *passenger* mutations<sup>13</sup>. The exact number of genes harboring driver mutations (driver genes) is not known. Importantly, not all mutations in a driver gene are driver mutations, as the specific effect of the mutation plays a role. In addition may other genomic aberrations such as copy number changes (amplifications or deletions) and epigenetic changes (DNA methylation, histone modifications etc.), play a role in carcinogenesis<sup>23</sup>. Hypermethylation of certain cytosine residues in gene promoters may lead to reduced gene transcription. When this occurs in tumor suppressor genes, it may promote cancer. Methylation also plays an important role in cellular differentiation<sup>24</sup>.

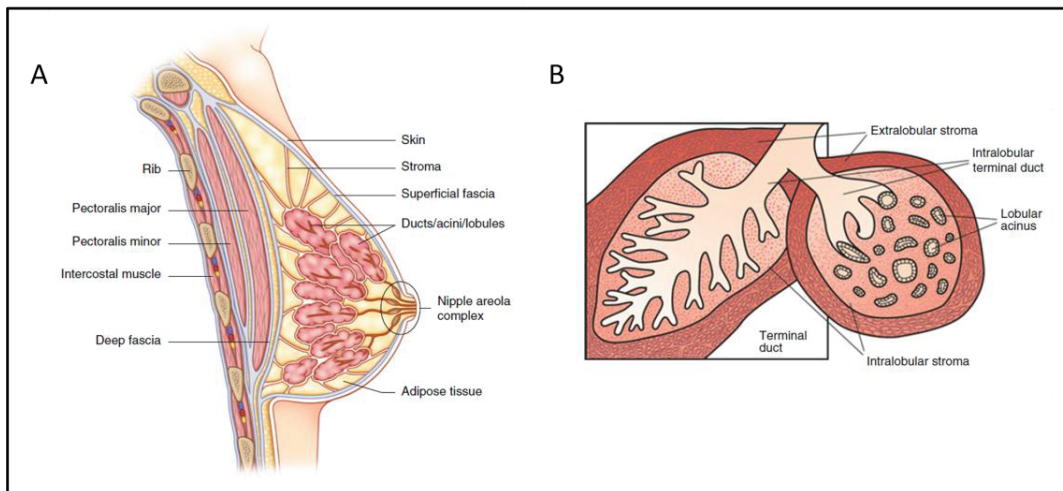
The effect of genomic aberrations in cancer cells is elicited through only a dozen signaling pathway that regulate three core cellular processes: *cell fate determination*, *cell survival* and *genome maintenance*. Because of genomic instability and the random occurrence of aberrations in a cancer cell, each individual tumor exhibits its own distinct fingerprint of genetic alterations, even distinct from tumors of the same histopathologic or molecular subtype. This is called *intertumoral* heterogeneity. Also, the cells that make up one individual tumor may have different genetic aberrations, causing *intratumoral* heterogeneity. Tumors with different genetic aberrations may still have strikingly similar phenotypes as similar pathways may be affected<sup>25</sup>.

In cancer research, the basic scientific hypothesis-testing paradigm has been a preferred method for a long time, presumably due to lack of high throughput assays, entailing a need for studies with very narrow focus and defined hypotheses<sup>26</sup>. However, the constantly increasing pool of data from *omics* high-throughput technologies, has led to a higher perception of the complexity of cancers, and insight

into why therapies targeting specific cancer genes are not always effective. It enables us to have a more holistic view of how aberrations in cancer cells alter the homeostasis of signaling networks, within the cancer cell and between cancer cells and the microenvironment. This approach is called *systems biology* and is a more *hypothesis-generating* method than traditional methods which are more *hypothesis-driven*. By studying the behavior of many molecules simultaneously, systems biology may help elucidate the complexity of perturbations in cancer cells. Omic approaches are undoubtedly valuable for research purposes, however in the clinic, acquisition of whole omic data is usually not feasible since it is costly, time consuming and large amount of data makes interpretation difficult. Nevertheless, exploration of such data in a research setting is valuable to enhance the understanding of tumor biology which may contribute to improving personalized therapy and aid the discovery of new predictive and prognostic biomarkers and subtype classifications for use pre-clinically and clinically<sup>13,27</sup>.

### **Normal breast anatomy and physiology**

Mammary glands are unique to mammals and have the important role of synthesizing, secreting and delivering milk to the newborn baby. The mammary gland is an organ of simple function, but has complex biology, and to understand breast tumor development one needs to have an understanding of the normal gland function, development, structure and regulation<sup>28</sup>. The anatomy of the human breast was thoroughly described by Sir Astley Paston Cooper already in 1840 (Figure 3A)<sup>29</sup>. The breasts rest on the pectoralis major muscle of the chest. They are composed of varying amounts of adipose tissue whose role is to support the parenchyma (the functional tissue), which is made up of tubuloalveolar glands. This is in turn surrounded by a loose framework of fibrous connective tissue called Cooper's ligaments. The parenchyma is divided into lobes which are made up of 4-18 lobules containing 10-100 alveoli; small sacs of lactocytes responsible for producing milk. The glands are drained by ductules that converge into lactiferous ducts which dilate to form the lactiferous sinuses where milk is stored temporarily during feeding. The ducts pass through the nipple and opens up onto its surface<sup>30</sup>.



**Figure 3. Normal breast anatomy. A: Cross-section of a normal breast. B: Cross-section of a fully developed terminal ductal lobular unit (TDLU). Reproduced from McGuire with permission from Springer Nature<sup>31</sup>.**

The basic functional unit of the breast is the terminal ductal lobular unit (TDLU) (Figure 3B)<sup>31</sup>. This is comprised of an extralobular duct, an intralobular duct and the lobule itself<sup>32</sup>. The walls of the ducts are lined with two layers of epithelial cells: an inner (luminal) layer of cuboidal cells surrounding the lumen and an outer (basal) layer of myoepithelial cells (MEC). The luminal cells of the lobules are secretory cells that have the ability to transform to milk-producing lactocytes during lactation, while MECs are contractile, i.e. they resemble smooth muscle cells and serve to expulse the milk when the child is suckling. The basal layer lies on the basement membrane (BM) separating the epithelium from the surrounding stroma<sup>30</sup>. In the mammary gland, a common progenitor is thought to be the ancestor of both luminal and myoepithelial cells<sup>33</sup>.

A remarkable feature of the breasts is the drastic changes in structure and function that the organ goes through during a woman's lifetime. The purpose of all these changes is to prepare for and perform as milk-producing organ. Most organs develop to a relatively mature state in the embryonic stage. The mammary gland, however, is very immature at birth and reaches its mature state only during the pregnancy-lactation cycle (PLC). The mammary epithelium is during the PLC able to undergo cycles of proliferation, differentiation and apoptosis due to self-renewing multipotent mammary stem cells (MaSC) that are capable of generating the entire epithelial architecture<sup>34,35</sup>. There is a clear link between the physiological processes occurring in the breasts and breast cancer risk. Lactation decreases the risk of developing breast cancer, especially if the first PLC takes place before the age of 30. Increased total duration of lactation and multiple pregnancies also decreases the risk<sup>36</sup>. It is possible that the decreased risk following an early pregnancy is caused by a reduction of the number of MaSC available for tumor

transformation, since it is believed that the most aggressive breast cancer subtypes originate from MaSCs<sup>34</sup>. Interestingly, during the PLC, a mammary gland displays transient characteristics that are also involved in breast cancer initiation and progression such as *epithelial-to-mesenchymal transition* (EMT)<sup>37</sup>.

### Ductal carcinoma in situ

Carcinomas are tumors originating from epithelial tissues<sup>1</sup>. Mammary gland tumors are mostly adenocarcinomas originating from glandular epithelial cells in the TDLU, either in cells lining the ducts (resulting in ductal carcinomas) or in cells of the lobules (resulting in lobular carcinomas). Around 80% of mammary gland tumors are of ductal origin<sup>38</sup>. Ductal carcinoma in situ (DCIS) are tumors made up of carcinoma cells proliferating inside the ducts of the mammary gland with no evidence of cancer cell invasion into the surrounding stroma<sup>39</sup>. *In situ* means *in place*, which refers to the tumor cells being confined inside the ducts of the mammary gland without breaching the basement membrane. Similarly,

“My plea in regard to neoplasms of the breast is, that they should all be held to be malignant until their innocence is proved; and the complement is, let no guilty tumor escape.”

– Chas. Langdon Gibson, *Ann Surg* (1909)<sup>42</sup>

lobular carcinomas in situ refer to carcinomas confined to lobuli. DCIS was first comprehensively described in the 1930s<sup>40,41</sup>, but management of pre-malignant or border-line breast tumors had already been a topic for discussion many years earlier when Dr. Gibson expressed his concerns about under-treatment of such tumors<sup>42</sup>.

Most DCIS tumors are detected through screening mammography (low-dose X-ray imaging used to detect breast cancer in healthy (non-symptomatic) women)<sup>43</sup>. During the last decades, the incidence of DCIS has increased dramatically, predominantly due to introduction of screening. In the US, the incidence of DCIS in 1975 was 5.8/100.000 while in 2004 it had increased to 32.5/100.000. Now, DCIS constitutes around 25% of all breast cancer cases in the US<sup>44</sup>. Similar numbers are seen in Norway: from 2006 to 2016, 17% of all screen-detected breast cancers were DCIS, while only 6% of the so-called interval cancers (those diagnosed between two routine screenings) are DCIS<sup>45</sup>. Notably, the incidence of aggressive DCIS has not risen as much as non-aggressive DCIS types after the introduction of screening<sup>44</sup>. The benefits of breast cancer screening programs has been debated<sup>46,47</sup>, however meta-analyses show protective effects of screening with 20 to 35 percent reduction in mortality from breast cancer<sup>48,49</sup>. As for all screening tests, there is a risk of false positive findings on mammograms that may cause anxiety

and unnecessary interventions and costs due to biopsy-taking. The rate of false positive findings has been reported to be around 10%. In addition comes the risk of so-called overdiagnosis, where screening detects small lesions (often DCIS) that may never have constituted a risk for the patient during the rest of her lifetime if it hadn't been detected<sup>50,51</sup>. Autopsy studies confirm that there is a certain "reservoir" of DCIS in the population that would probably never progress to invasive cancer<sup>52-55</sup>.

DCIS tumors by themselves do not pose any danger for patients as long as the tumors remain intraductal; thus therapy for DCIS is initiated in order to prevent subsequent invasive breast cancer. Standard therapy for DCIS is surgery, either mastectomy alone or breast conserving surgery (BCS) combined with whole breast radiation therapy (RT)<sup>56,57</sup>. A large scale study performed by Sagara et al. demonstrated no additional benefit of surgery on low-grade DCIS, while for DCIS of intermediate or high grade, surgery increased breast cancer specific survival<sup>58</sup>. Mastectomy is performed when there is multicentric disease, large lesions or in case of personal preference, and reduces mortality to around 1% while BCS reduces mortality slightly less<sup>58-60</sup>. Observational studies investigating the effect of RT following BCS in DCIS showed that RT reduced local recurrence (both in situ and invasive) by approximately 50% compared to those that did not receive RT, however no effect was seen on overall or breast cancer specific survival<sup>61-63</sup>. Endocrine therapy such as tamoxifen may also be initiated in estrogen receptor (ER) positive DCIS cases<sup>64,65</sup>. This is more common in the US, and is not standard therapy in Norway, however it has been shown to reduce the risk of recurrence and also the risk of contralateral disease<sup>66</sup>. At the current time, there is no reliable way of predicting which DCIS lesions are low-risk, i.e. which lesions have low probability of progressing into invasive and potentially harmful disease. As a consequence, women with low-risk DCIS are at risk of experiencing side-effects from unnecessary treatment. In addition, there has not been identified a clear reduction in mortality by DCIS treatment<sup>67</sup>. This suggests that there is substantial overdiagnosis and also overtreatment of DCIS. On the other hand, there is a risk of missing invasive foci in routine diagnostics. A major challenge in DCIS management is therefore to identify the potentially hazardous DCIS to initiate appropriate treatment, while leaving the indolent ones.

The risk of recurrence after being treated for DCIS (i.e. the risk of being diagnosed with another breast malignancy) is reported to be between 10% and 24%. Even though the mortality rate for patients diagnosed with DCIS is only around 2%, there is still a four times higher risk of dying of breast cancer after a DCIS-diagnosis than for a woman in the general population<sup>44,68</sup>. Death following a DCIS diagnosis is either caused by recurrence of invasive disease, or undetected invasive foci at original diagnosis<sup>69</sup>.

Recurrences may be seen many years after an initial DCIS diagnosis, and in situ recurrences generally happen earlier than recurrences that are invasive<sup>70</sup>. The long time span from a DCIS-diagnosis to a possible recurrence complicates the study of DCIS, as studies with too short follow-up time will fail to identify late recurrences. Since most DCIS patients undergo treatment, there are few studies that explore the natural history of DCIS. However, identifying DCIS that have initially been misdiagnosed as benign tumors has enabled studies of untreated DCIS. These studies have estimated that only 14-53% of all DCIS will progress to invasive breast cancer (IBC)<sup>71-74</sup>.

DCIS constitute a heterogeneous group of tumors that display a high degree of diversity from well-differentiated, slow growing tumors to lower differentiated, rapidly proliferating tumors (and everything in between). There is therefore a need for classification systems that can help stratify tumors into meaningful groups with prognostic relevance<sup>75</sup>. Classification of DCIS has traditionally been performed based on architectural features and growth patterns (Table 1)<sup>76,77</sup>. Multiple architectural patterns may be present in one lesion, and the prognostic value of these features are limited, although comedo-type DCIS is often associated with high-grade tumors and poorer breast-cancer specific survival<sup>43</sup>.

Since architectural pattern has shown to be insufficient as a prognostic factor, several alternative classification systems for histopathological assessment that correlate better with other prognostic markers and more precisely predict recurrence have been proposed<sup>78,79</sup>. The different systems classify tumors based on cellular and nuclear appearance, growth pattern, cellular differentiation and polarization or presence of necrosis, and usually separate DCIS lesions into three categories: low, intermediate and high grade (although using different terminology and definitions)<sup>80-82</sup>. There is currently no universal agreement on grading of DCIS, which is a source for confusion<sup>83</sup>. The method used

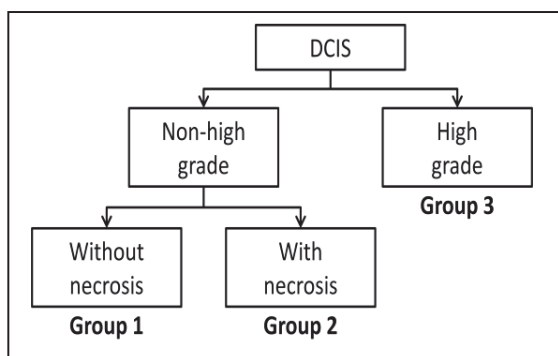
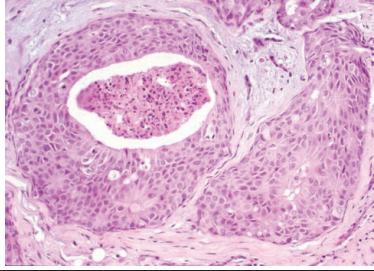
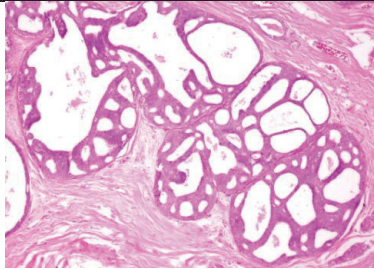
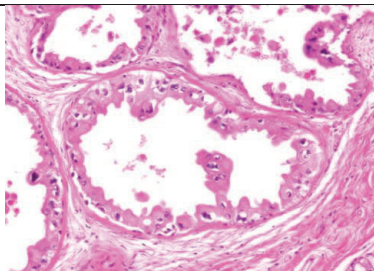
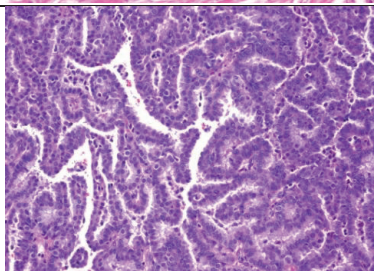
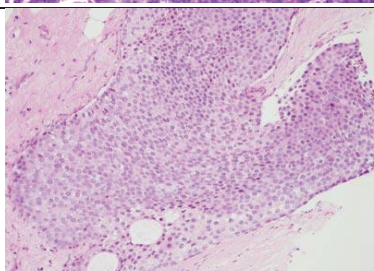


Figure 4. Van Nuys DCIS histopathological classification system.

in Norway is the Van Nuys's classification where DCIS tumors are classified into three groups according to nuclear grade and presence of necrosis (Figure 4)<sup>57,81</sup>. A modified system of the Van Nuys classification (the USC/Van Nuys prognostic index) also includes information about tumor size, margin width and patient age and aims to predict which patients with DCIS could benefit from radiation therapy<sup>84</sup>.



Table 1. Architectural patterns of DCIS. H&E stained sections<sup>76,77</sup>. Images reproduced with permission from [www.webpathology.com](http://www.webpathology.com)<sup>85</sup>

<b>Type</b>	<b>Characteristics</b>	<b>Histology</b>
<b>Comedo-type</b>	Prominent necrosis in the center of the lesion. Necrotic material frequently calcified (may be detected mammographically). Often large tumor cells with nuclear pleomorphism and prominent mitotic activity. More often associated with invasion and the degree of comedo necrosis is a strong predictor for recurrence after treatment	
<b>Cribriform</b>	Tumor cells that grow in a glandular pattern without intercellular stroma. The cells are small to medium of size and have uniform hyperchromatic nuclei with infrequent mitosis. Any necrosis is limited to single cells or small cell clusters.	
<b>Micropapillary</b>	Small club-like protrusions of cells without a fibrovascular core are oriented perpendicular to the basement membrane and project into the lumen. Tumor cells are usually small to medium of size, nuclei show diffuse hyperchromasia and mitoses are infrequent.	
<b>Papillary</b>	Intraluminal protrusions of tumor cells that have fibrovascular cores, i.e. true papillations.	
<b>Solid</b>	Tumor cells fill and distend the lumen of the ducts without necrosis, fenestrations or papillations. Tumor cells may be of various sizes.	
<b>Other</b>	Rare DCIS variants: <ul style="list-style-type: none"> <li>• Clinging carcinoma<sup>86</sup></li> <li>• Intraductal signet ring cell carcinoma<sup>87</sup></li> <li>• Cystic hypersecretory duct carcinoma<sup>88</sup></li> </ul>	

## Breast tumor progression

The most commonly proposed model for tumor progression of ductal carcinomas is the linear model put forth by Welling and colleagues which proposes that IBCs evolve through a non-obligate series of increasingly abnormal stages over a period of time: Changes occur in normal ductal epithelium leading to flat epithelial atypia (FEA), followed by atypical ductal hyperplasia (ADH), and ductal carcinoma in situ (DCIS), which may develop into invasive ductal carcinoma if the tumor cells break out of the ducts into surrounding stromal tissue (Figure 5)<sup>89,90</sup>. Before the DCIS-stage, lesions are not considered cancers; however, the progression should be seen as a continuum and there are no clear cut boundaries between the different stages. In particular, the distinction between ADH and low-grade DCIS is difficult to determine as the histomorphological diagnostic criteria separating ADH and DCIS are predominantly quantitative rather than qualitative<sup>77,91</sup>. All the described progression stages are not obligate prior to development of invasive ductal carcinoma, but since most breast tumor arise inside the ducts, intraductal cancer cells will in most cases be present at some point (for longer or shorter time) before any invasion occurs.

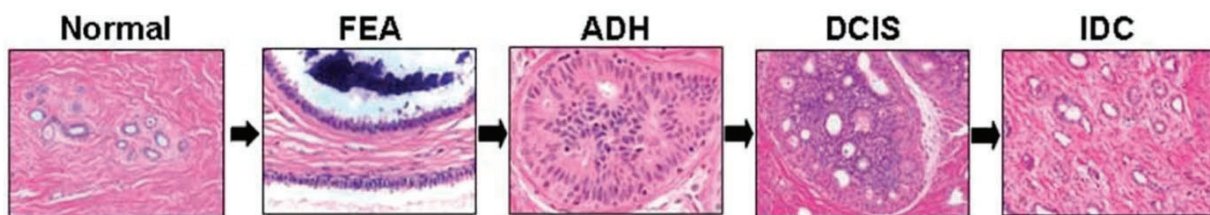


Figure 5. The classic model of breast cancer progression. Neoplastic evolution initiates in normal epithelium and progresses to flat epithelial atypia (FEA), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC). Adapted from Bombonati & Sgroi, with permission from John Wiley and Sons<sup>89</sup>.

Currently there exist no convincing molecular markers that can predict whether a DCIS will become invasive if left untreated. As immunohistochemistry (IHC) is a readily available method, attempts have been made to identify IHC markers capable of differentiating between subgroups of DCIS, especially related to aggressiveness. The hormone receptors estrogen receptor (ER) and progesterone receptor (PR) are in DCIS as in IBC, related to low aggressiveness. ER and PR positivity has been shown to negatively correlate with increasing tumor grade<sup>92</sup>. Over-expression of human epidermal growth factor receptor 2 (HER2) has shown to predict local recurrence, however there is a higher frequency of HER2 positive tumors among DCIS compared to IBC, indicating that HER2 alone is not able to drive the process of invasion of tumor cells into the surrounding stroma<sup>93,94</sup>. Increased expression of *K167* (a proliferation marker) and *TP53* (a tumor suppressor) is also correlated to higher tumor grade in DCIS<sup>93</sup>. Other markers

have also shown to have prognostic relevance; for instance, in a study by Kerlikowske et al., COX2<sup>+</sup>/p16<sup>+</sup>/Ki67<sup>+</sup> DCIS tumors were significantly associated with subsequent invasive cancer<sup>95</sup>.

Studies of genomic differences between DCIS and IBC show that DCIS often harbor similar chromosomal aberrations as IBC tumors. Copy number aberrations appear to arise early in breast tumor progression as many of these are found already at the ADH stage, supporting the belief that ADH is a precursor to DCIS and IBC<sup>96</sup>. No consistent differences in quantity or quality of copy number aberrations have been identified between DCIS and IBC, indicating that DCIS may be a precursor to IBC and also that copy number aberrations are not driving invasion<sup>97</sup>. Likewise has the mutational profile of DCIS shown to be highly similar to IBC<sup>98-100</sup>. A number of studies have compared transcriptomic profiles of DCIS and IBC, resulting in a multitude of gene lists attempting to differentiate between the two tumor stages. However, the overall results from these studies show that there are very few differences between DCIS and IBC also on the transcriptomic level<sup>101-104</sup>. Importantly, many of the studies comparing genomic and transcriptomic differences between DCIS and IBC have not performed analyses stratified by molecular subtype. This may have obscured important findings. In Lesurf et al., differential expression analyses (mRNA and miRNA) and copy number analyses comparing DCIS and IBC was carried out separately for each molecular subtype<sup>105</sup>. This study revealed that there exist molecular features associated with breast cancer progression unique to each intrinsic subtype and this opens new possibilities for studying breast cancer progression.

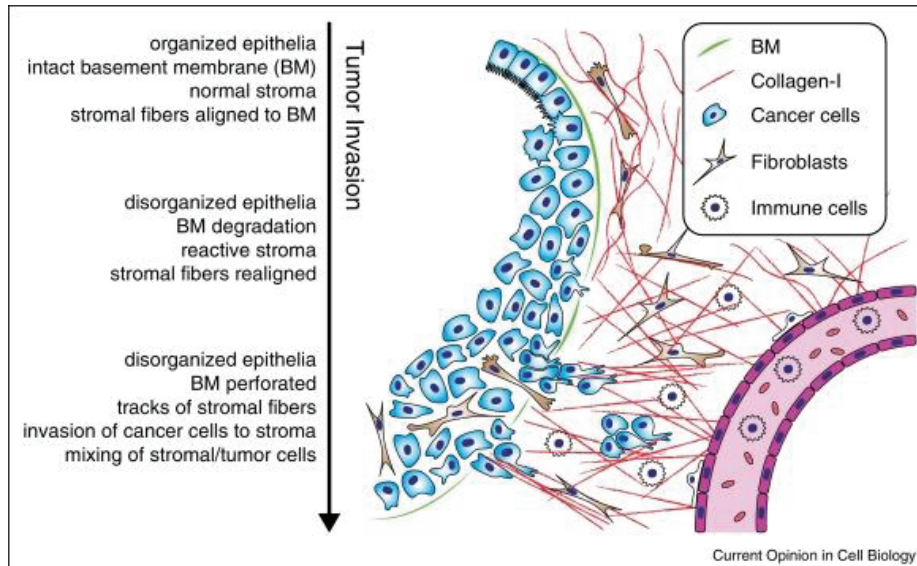
Since histopathological evaluation provides limited prognostic information, there is a need for better stratification tools by integrating clinical, morphological and molecular data. To explore this, several studies, such as LORD, LORIS and COMET have been initiated<sup>106-108</sup>. These are all prospective open-label non-inferiority studies that randomize patients with low-grade DCIS into a standard treatment arm (surgery with or without RT) or an active surveillance arm (where patients are not treated, but are monitored for signs of progression). Active surveillance has been shown to be an appropriate management strategy for DCIS in selected patients<sup>109,110</sup>. A multigene gene expression based assay, Oncotype DX DCIS (Genomic Health, Redwood City, California), is a commercially available assay that may be used to stratify patients into low-risk and high-risk groups<sup>111,112</sup>. It estimates the 10 year risk of recurrence, resulting in a personalized score for each patient; however, the assay is costly and has not yet proven cost-beneficial<sup>113,114</sup>. For research purposes, to learn more about the biology of DCIS, it is important to integrate molecular data with clinical, pathological, radiological and cancer registry data<sup>115</sup>.

### **Breast cancer invasion**

Only when invasion occurs, can a ductal carcinoma be considered to have a truly malignant phenotype. The invasion process is therefore a crucial point in tumor progression and it is defined as one of the hallmarks of cancer (Figure 2)<sup>17</sup>. Invasion is defined as “cancer that has spread beyond the layer of tissue in which it developed and is growing into surrounding, healthy tissues”<sup>116</sup>. Cell migration and invasion are essential processes of normal embryonic development and organogenesis, and is also important in inflammation and wound healing in the adult body<sup>117</sup>. These properties are exploited by cancer cells, enabling them to invade surrounding tissue and metastasize<sup>118,119</sup>. In breast cancer, invasion is the process where tumor cells break through the myoepithelial cell layer and the basement membrane of the ducts and invade surrounding stromal tissue<sup>119</sup>. Invasion is diagnosed by histopathology, but radiological examination prior to surgery may also give indications of invasive disease. MRI may be used to predict invasion, however this is not used routinely<sup>120,121</sup>. Differentiation of in situ and invasive lesions by histopathological diagnosis is based on the presence of an intact barrier of MECs and BM between malignant cells and the stroma; however this task is not always straightforward. Identifying the BM by IHC may be challenging, so disruption of the MEC layer is often used as a surrogate marker for BM destruction. The immunohistochemical markers most commonly used to identify BM is Laminin and Collagen IV, while for MEC, smooth muscle actin (SMA), smooth muscle myosin heavy chain (SMMHC) or p63 are useful markers<sup>122</sup>. DCIS lesions may be associated with so-called micro-invasion which is defined as invasive carcinoma foci that measure less than 1mm in greatest extent<sup>59,123</sup>. DCIS with micro-invasion is more likely to be found in DCIS that are large, those that have been detected due to clinical symptoms (as opposed to those detected through mammography) and in tumors with aggressive features such as high grade, comedo-necrosis and ER-negativity. Furthermore, there has been identified differences in breast-cancer specific and overall survival between DCIS tumors with and without micro-invasion<sup>124</sup>. Because of the small size, studies of genomic changes in the cells at these micro-invasive foci are limited, but there are indications of involvement of the microenvironment at the site of micro-invasion<sup>43</sup>.

Invasion is a dynamic process and may be influenced by both underlying genetics, signaling from neighboring cells and the structure of the microenvironment<sup>125</sup>. During invasion, the proliferating epithelium loses its two-layered arrangement, and becomes increasingly disorganized as epithelial cell polarization and cell-cell-adhesion is gradually lost. The fibers of the stroma become increasingly aligned perpendicular to the BM and there is an increase in immune cell infiltration and fibroblasts in the stroma

(so-called reactive stroma). After breaching the BM, tumor cells move into the stroma and intermix with stromal cells (Figure 6)<sup>126</sup>.



**Figure 6. Summary of processes in mammary gland epithelium and the immediate tumor microenvironment during tumor progression and invasion. Reproduced from Clark et al. licensed under CC BY-NC-ND 4.0<sup>126</sup>.**

Cancer cells need to acquire a motile phenotype to be able to move into surrounding tissue. Such cancer cell migration may happen through several different modes: *Single-cell migration* (single cells migrating separately either in an amoeboid-like or mesenchymal-like fashion), *multicellular streaming* (non-adherent cells following each other rapidly along the same path) and *collective migration* (several adherent cells migrate together either as strands or sheets of cells)<sup>126,127</sup>. Breast cancer invasion typically happens through collective migration<sup>128</sup>. Lower differentiated tumors with less marked epithelial phenotype and lower cell-cell-adhesion may show a higher tendency of single-cell migration compared to higher differentiated tumors, and may be regarded more aggressive<sup>119</sup>. The process where epithelial tumor cells obtain a mesenchymal phenotype with reduced cellular adhesion and polarization, separation into individual cells and increased cell motility is called epithelial-to-mesenchymal transition (EMT)<sup>119</sup>. In addition to increasing the invasive abilities of tumor cells, EMT also contributes to loss of contact inhibition and altered growth control which are hallmark features of carcinomas.

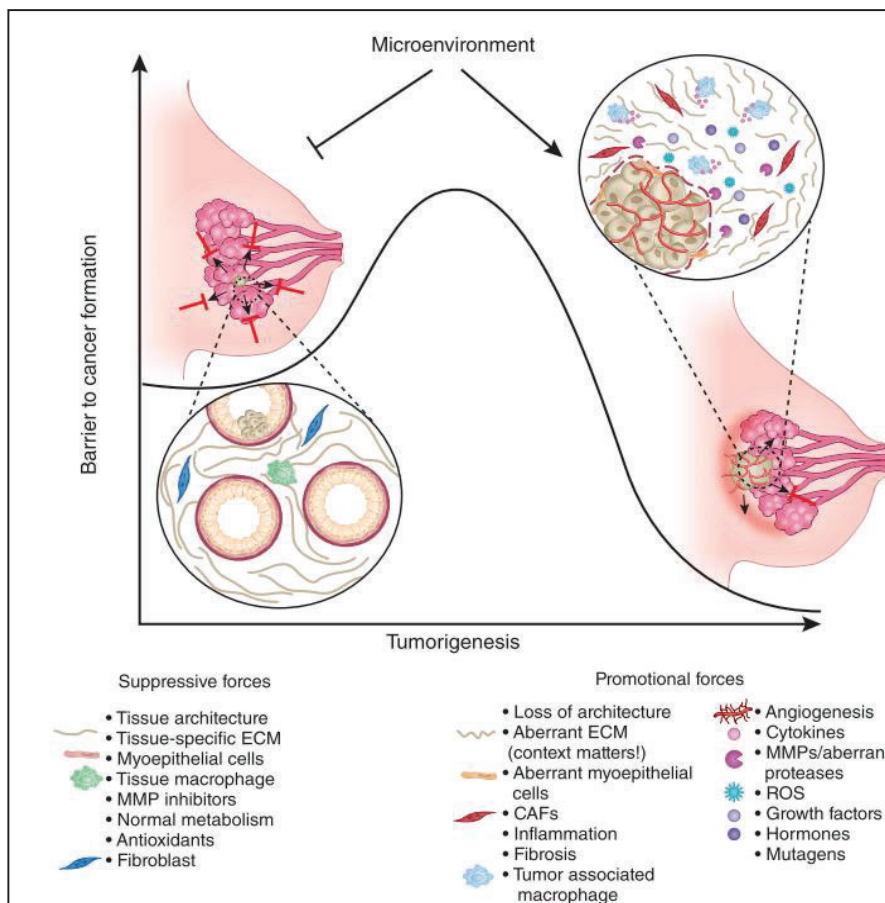
Following invasion, tumor cells may continue to disseminate throughout the body i.e. metastasize. Metastasis is defined as “The spread of cancer cells from the place where they first formed to another part of the body”<sup>129</sup>. The cells in the metastasis are of the same type as the primary tumor. Cancer deaths are usually caused by the metastatic lesions, not by the primary tumor itself. The metastatic

process is initiated from a local (invasive) tumor from which cells detach into blood or lymph vessels<sup>118</sup>. Metastasis is dependent on properties of tumor and host cells and certain tumor types tend to metastasize to specific organs. Breast cancers have a tendency to spread to bone, lung and liver<sup>130,131</sup>. This non-random pattern of metastasis has been known for more than a century since Stephen Paget proposed his “seed and soil” hypothesis in 1889, stating that different *seeds* (i.e. tumor cells) have different requirements of the *soil* (i.e. the target organ for the metastasis)<sup>131,132</sup>. In DCIS, tumor cells are confined to a limited space, and breaching the BM has been regarded as obligate for a DCIS to develop into a cancer with true malignant potential. Narod et al. however, proposes the possibility that DCIS cells may metastasize without a preceding invasion of the stromal tissue surrounding the ducts<sup>67</sup>. They claim that DCIS tumor cells may spread through neovascularization (small vessels invading the ducts enabling hematogenous dissemination of tumor cells) and they justify this by referring to the finding of circulating tumor cells and tumor cells in lymph nodes in patients with DCIS without known invasion. However, this theory is rejected by most researchers who explain these findings by *occult* micro-invasion, i.e. invasion not detected in histopathological sections<sup>67,133–135</sup>.

### **The role of the tumor microenvironment in breast cancer progression**

When studying invasion, much emphasis has been put on the changes happening in the tumor cells themselves assuming that invasion is mainly a tumor cell-driven process. However, more and more attention has been given to the role of the microenvironment in determining cancer cell invasion, migration and metastasis. In our bodies, aberrations arise in cells constantly, but under physiological conditions, the normal microenvironment exerts suppressive forces to keep these cells from turning into tumors. However, sometimes, this suppression is corrupted and the microenvironment becomes permissive for tumor growth (Figure 7)<sup>136</sup>.

The tumor microenvironment (TME) is the tissue immediately surrounding a tumor and consists of cells (e.g. fibroblasts, endothelial cells, macrophages, lymphocytes), extracellular matrix (ECM) and soluble molecules<sup>97,137</sup>. Several of the hallmarks of cancer include processes where TME plays a significant role: *evading growth suppressors, avoiding immune destruction, inducing angiogenesis and tumor promoting inflammation* (Figure 2)<sup>18</sup>. The interplay between TME and tumor cells may affect all stages of tumor progression from tumor initiation and invasion to metastasis<sup>119</sup>. The TME influences the tumor cells through paracrine signaling from normal cells and interaction between constituents of the ECM and cellular adhesion molecules on the surface of tumor cells. Importantly, the interplay is bidirectional, as



**Figure 7. The role of the microenvironment during breast tumorigenesis. In normal conditions the microenvironment exerts suppressive forces and acts as a barrier to tumorigenesis, but it may at some point change and become permissive to tumor growth. Reproduced from Bissell & Hines with permission from Springer nature<sup>136</sup>.**

tumor cells also influence the microenvironment<sup>119,138</sup>. Tumor-adjacent normal tissue may therefore never be considered completely normal. Already at the DCIS stage is stroma surrounding the lesions altered compared to normal stroma, suggesting co-evolution of stroma and tumor cells before invasion occurs<sup>139</sup>. The morphology of the microenvironment is highly variable, both regarding the extent of collagen deposition and the number and distribution of cells<sup>140</sup>. The tumor microenvironment may be different from patient to patient, not only because of features of the tumor itself, but also due to different genetic predispositions. For instance, some women display an innate high breast tissue density which is a strong independent risk factor for breast cancer<sup>141,142</sup>.

The ECM is the non-cellular component of tissues and provides a scaffold for the cells and also plays an important role in eliciting biochemical and biophysical signaling important for tissue morphogenesis, differentiation and homeostasis<sup>143</sup>. Fibroblasts are stromal cells responsible for producing molecules constituting the ECM such as collagen and fibronectin and they also secrete proteases such as matrix metalloproteinases (MMP) which contribute to stromal reorganization during branching morphogenesis

## INTRODUCTION

---

in the developing gland<sup>140,144</sup>. Fibroblasts are also able to enhance mammary stem- and progenitor cell function through paracrine factors<sup>145</sup>. During wound healing, fibroblasts are in an activated state in order to remodel tissue and restore normal function. Activated fibroblasts resemble so-called myofibroblasts with a phenotype similar of both fibroblasts and smooth muscle cells. When wound healing is completed, activation is turned off. However, in tumor tissues, a high proportion of the fibroblasts are in an activated state, supporting the claim by Dvorak in 1986 that tumors are “wounds that never heal”<sup>144,146</sup>. In tumors, cancer associated fibroblasts (CAFs) are situated in proximity of the tumor. They constitute a heterogeneous group of cells, originating from several cell types such as resident fibroblasts, epithelial cells, endothelial cells and adipocytes<sup>147</sup>. CAFs may contribute to tumor proliferation, invasion and metastasis through secretion of growth factors and cytokines and through degradation of proteins in the ECM<sup>140</sup>. Specific changes in the collagen matrix organization surrounding tumors are associated with increased invasiveness. In early lesions, there is increased collagen density around the tumor, while at later stages, the collagen fibers are oriented more perpendicular to the lesion along which tumor cells may migrate. This type of collagen alignment is associated with more aggressive tumors, and is also seen in DCIS of high grade<sup>148</sup>. CAFs have also been implicated as modulators of hormone responsiveness in breast tumor cells, enhancing response to estrogen and promoting proliferation of ER+ tumor cells as well as modulators of angiogenesis<sup>149,150</sup>.

The immune environment in breast tumors varies considerably. Compared to normal tissue, it has been demonstrated an increase in the leukocyte population in DCIS and IBC involving both innate and adaptive immune cells, although some tumors may completely lack an immune response<sup>151</sup>. Neoplastic cells are usually recognized as foreign, eliciting an antitumor immune response characterized by infiltration of type 1 macrophages, dendritic cells, natural killer cells, CD8+ cytotoxic T-cells and CD4+ Th1 cells that prevents further tumor growth. Through a process called immune-editing, tumor cells may acquire the ability to escape immune control through expression of immune checkpoint proteins. Also, the immune cell composition may shift towards more immunosuppressive cell types such as myeloid-derived suppressor cells, CD4+FOXP3+ regulatory T cells and type 2 macrophages. Tumor-associated immune cells may therefore have both positive and negative effects on cancer progression. There is a gradual increase in immune cell density throughout breast cancer progression, with the highest number in invasive tumors. However, already at the in situ stage there may be an extensive immune response surrounding the lesions<sup>152</sup>.



Tumor infiltrating lymphocytes (TILs) is a prognostic factor in invasive breast cancer and is associated with improved survival<sup>97</sup>. In DCIS, high level of TILs is also seen in high grade, ER- or HER2+ tumors or in tumors with a high degree of genomic imbalance. Recurring lesions tend to have lower TIL-infiltration than the primary tumor, suggesting that suppression of anti-tumor immune responses may be involved in recurrence<sup>153</sup>. TILs consist of several subsets. The CD8+ lymphocytes have cytotoxic properties and are associated with anti-tumor effects<sup>140</sup>. CD4+ TILs may have tumor suppressive effects when they are of the Th1 type (expressing tumor suppressing INF $\gamma$ ), while the Th2 type (expressing IL-4) may have a tumor promoting role through differentiating macrophages towards a pro-tumorigenic phenotype<sup>154,155</sup>. Regulatory T-cells (Tregs) express FOXP3 and are involved in suppression of cytotoxic T-cells. A low CD8+/Treg ratio is more commonly found in ER-/high grade DCIS lesions and may be an indication of an immunosuppressive environment so even though the total number of TILs may be high, the effect of the TILs may be in favor of the tumor<sup>156,157</sup>.

Since DCIS tumor cells are protected from their surroundings by the MEC layer and the BM, there is little direct contact between tumor cells and immune cells. However, immune cells still increase in numbers and migrate to the tumor in DCIS. T-cells surrounding DCIS foci have been shown to be of an active phenotype, expressing GZMB and MKI67, and the frequency of activated T-cells decrease in invasive disease. There is in general higher expression of checkpoint proteins CTLA4 and PD-L1 in immune cells invading invasive tumors compared to DCIS, and PD-L1 positive immune cells in DCIS are mainly found in ER-negative tumors<sup>151</sup>. In addition, the expression of PD-L1 in the DCIS tumor cells themselves has shown to be low compared to IBC<sup>157</sup>. These findings indicate an increasing suppressive immune microenvironment during progression from DCIS to IBC and further studies could reveal useful immune response biomarkers to predict invasion. Immunotherapy has been discussed as a possible approach to treat DCIS, either by checkpoint blockade or dendritic-cell-based vaccines, however it is still unclear whether this would be effective and advantageous in DCIS patients since especially checkpoint inhibitors are associated with many side effects<sup>152,157,158</sup>.

The myoepithelial cell layer plays an important role in the transition from DCIS to IBC. In addition to their role during feeding, MECs in a normal breast affect differentiation, proliferation and polarity of luminal cells. The MECs also contribute to synthesis and maintenance of the BM<sup>140,159</sup>. MECs are believed to have a tumor suppressive role, and may act as a natural barrier against invasion exhibiting anti-angiogenic, anti-proliferative and anti-invasive properties through for instance production of protease

inhibitors and paracrine downregulation of MMPs in both tumor cells and fibroblasts<sup>160–162</sup>. However, MECs may also promote invasion in breast cancer<sup>163</sup>.

Two non-exclusive theories have been proposed on the mechanisms of the transition from DCIS to IBC: The *proteolytic* (“escape”) theory states that tumor cells themselves secrete proteases that degrade BM and ECM to be able to invade the stroma, while the *focal myoepithelial cell layer disruption* (“release”) theory suggests that tumor invasion begins with disruption of the MEC layer due to genetic changes, inflammation, localized trauma or other mechanisms in the MEC layer itself. The death of the MECs then leads to localized loss of tumor suppressors which causes a focal change in the microenvironment that promotes tumor invasion through subsequent destruction of BM, further destruction of MECs and finally invasion of tumor cells into the stroma surrounding the duct<sup>159,164–166</sup>. The mechanisms involved in DCIS to IBC invasion are complex and heterogeneous, however there is no doubt that permissive changes in the tumor microenvironment play an important role in breast cancer progression.

### **Molecular subtyping of breast tumors**

Breast cancer is not a single uniform disease, but varies extensively in biological properties, clinical behavior and histological features. In fact, all tumors are essentially different; however, some features are shared across several tumors. To be able to select appropriate treatment and estimate prognosis for the patients, there is a need for a robust and objective classification system. Such stratification is also important when performing clinical trials and when studying the underlying biology of tumor evolution<sup>167</sup>.

In 2011, the consensus meeting in St. Gallen recommended using a more comprehensive multigene test for characterization of breast cancer tumors when feasible<sup>168,169</sup>. The background for this recommendation was the work performed by Perou, Sørli and colleagues where the intrinsic subtypes of breast cancer were discovered<sup>170,171</sup>. They performed gene expression analyses using cDNA microarrays and compared transcriptomic profiles of breast cancer tumors before and after treatment. An *intrinsic* gene list was determined by identifying genes whose expression varied less between samples from the same patient compared to samples from different patients. Hierarchical clustering of the tumors based on the intrinsic gene list, revealed two main clusters (mainly separated according to ER status) and five sub-clusters which represent the intrinsic subtypes: luminal A, luminal B, HER2-enriched, basal-like and normal-like. The luminal A subtype is characterized by ER positive tumors that

express genes normally expressed by luminal breast epithelial cells. Luminal B is similar to luminal A, but has higher expression of genes involved in proliferation. The cluster enriched for ER negative tumors consists of basal-like tumors that are mainly ER/PR/HER2-negative tumors with a gene expression pattern resembling myoepithelial/basal cells, HER2-enriched tumors that often show high expression of HER2 (ERBB2) and the normal-like with an expression pattern resembling normal breast tissue<sup>167,171</sup>. The intrinsic subtypes also have prognostic value as they showed statistically significant different outcomes, most notably poor prognosis for the basal-like subtype and significantly different outcomes between the two luminal subtypes<sup>170,172</sup>. Furthermore, the different subtypes show characteristics that to a certain degree may reflect different cells of origin<sup>173-175</sup>.

The original intrinsic gene list consisted of several hundred genes, complicating implementation of intrinsic subtyping in a clinical setting. Therefore, Parker et al. aimed for making an assay to subtype breast tumors for use in the clinic. They used several datasets and reduced the intrinsic gene list down to 50 genes (PAM50)<sup>176</sup>. This subtyping method is now approved by the US Food and Drug Administration as the Prosigna assay for use in patients with ER-positive tumors<sup>177-179</sup>. Gene expression analyses are costly and not accessible for all patients. Therefore, IHC markers for ER, PR, HER2 and Ki67 are commonly used as surrogate markers as a convenient approximation to the molecular subtypes (Table 2), however the overlap between the two methods is far from complete.<sup>168,180</sup>

**Table 2. Surrogate definitions of intrinsic breast cancer subtypes. Adapted from St. Gallen recommendations, 2011<sup>168</sup>.**

<b>Intrinsic subtype</b>	<b>Surrogate subtype</b>	<b>IHC surrogate markers</b>
<b>Luminal A</b>	Luminal A	ER and/or PR positive HER2 negative Ki67 low
	Luminal B (HER2 negative)	ER and/or PR positive HER2 negative Ki67 high
<b>Luminal B</b>	Luminal B (HER2 positive)	ER and/or PR positive HER2 overexpressed/amplified Any Ki67
	HER2 positive (non-luminal)	ER and PR negative HER2 overexpressed/amplified
<b>Basal-like</b>	Triple-negative	ER and PR negative HER2 negative

Even though the PAM50 subtypes are inherently different, there is substantial heterogeneity between tumors of the same subtype. For instance, is there evidence that luminal A tumors may be further divided into two different subtypes with different prognosis<sup>181</sup>. Also, a group of *core basal* tumors with worse prognosis than other basal-like tumor has been identified. These are, in addition to being ER/PR/HER2 negative, positive for either EGFR or CK5/6<sup>182-184</sup>. In a study by Herschkowitz et al. breast cancer tumors that did not fit well with any of the PAM50 subtypes were identified both in human and in mice. These tumors had low expression of tight junction and cell-cell adhesion proteins such as claudins and E-cadherin, low expression of luminal genes, low degree of differentiation, a mesenchymal phenotype and high immune infiltration. They proposed these *claudin-low* tumors as a separate subtype, and studies have shown that 7-14% of breast cancers may have claudin-low properties, however, the biology and clinical significance of this proposed subtype remains to be elucidated<sup>174,185-187</sup>.

Ductal carcinoma in situ has not been nearly as rigorously characterized as invasive breast cancer and even though many studies have investigated the transition from DCIS to IBC, few have taken molecular subtype into consideration. However, it is not reasonable to believe that the subtypes suddenly emerge during invasion. Analyses of DCIS using both IHC<sup>188-190</sup> and gene expression analyses<sup>101,191,192</sup> confirm that subtypes are present also at the DCIS stage. Importantly, the molecular subtypes of IBC have vastly different characteristics so it could be presumed that progression from DCIS to IBC is distinct for the different subtypes. This has not been investigated thoroughly, however, in the study by Lesurf et al., they found evidence of unique molecular features associated with disease progression between the different subtypes<sup>105</sup>.

### **Comparative breast cancer biology**

In cancer research, *in vitro* methods are invaluable tools. However, there is a huge need for *in vivo* experiments to close the translational gap between the lab bench and the patient. Research animals are imperative preclinical models for the development of new therapies (to test both efficacy and pharmacokinetics of new drugs) and to study mechanisms of diseases. Ideally, the perfect *in vivo* model should recapitulate all relevant clinical features of the human disease; however, this is generally not the case. It is therefore important to be aware of the limitations as well as the strengths of the different

relevant models, and also understand the fundamental differences between the research animal and humans.

The most commonly used laboratory animal in cancer research is the house mouse (*Mus musculus*). It is of small size, easily handled, inexpensive to house, breeds rapidly and has a fairly long lifespan (up to 3 years). The mouse genome is thoroughly characterized and may be manipulated relatively easily. It also shares many physiological attributes with humans<sup>193</sup>. There are several anatomical and physiological differences between the murine and human mammary gland that needs to be taken into consideration when using mouse models to study breast cancer. The human breast has a complex structure of lobes and lobuli with the TDLUs as the functional unit<sup>194,195</sup>. The mouse mammary gland has a simpler structure with a less branched network of ducts ending in stem cell enriched terminal end buds (TEBs) which are the functional (milk producing) units in the mouse and also responsible for driving ductal branching and elongation. TDLUs (in humans) and TEBs (in mouse) are built up similarly with luminal epithelial cells and myoepithelium and are both hormone responsive and dynamically active through the reproductive cycle<sup>196,197</sup>. In the mouse, the stromal tissue surrounding the ducts consists mostly of adipose tissue, with little ECM, while in the human, connective tissue is much more abundant with intra-lobular stroma made up of a loose collagen matrix with many specialized stromal cells that exert paracrine effects upon the mammary epithelium<sup>194</sup>.

Mice produce multiple offspring and have a correspondingly higher regenerative capacity of glandular tissue compared to humans. This may possibly be paralleled by different mammary stem cell composition and efficacy compared to human. The degree of involution after lactation is also much greater in mouse than in women<sup>194</sup>. Also, of relevance for the use of murine breast cancer models, is the endogenous plasma estrogen levels which in mice is up to ten-fold higher than in women<sup>198</sup>.

There is an extensive repertoire of mouse models available to study breast cancer utilizing different approaches for tumor formation e.g. xenografts, genetically engineered mouse models (GEMM) and chemically induced mouse tumors<sup>199</sup>. Both in mice and humans, mammary tumors may go through comparable tumor progression stages. Intraductal lesions with invasive potential are comparable to human DCIS and are referred to as mammary intraepithelial neoplasia (MIN)<sup>200</sup>.

Xenografts are made by transplanting human cells or tumor tissue into mice. There exist numerous human cell lines that represent the different subtypes of breast cancer. Cell lines also have the advantage that they may be genetically manipulated prior to engraftment and they grow reasonably fast

## INTRODUCTION

---

in mice. The drawback of immortalized cell lines is that they are heavily selected for characteristics that make them able to grow in an *in vitro* environment and they are unnaturally homogenous. An alternative to cell lines is patient derived xenografts (PDXs) where tumor tissue from patients is directly transplanted to mice. In PDXs the heterogeneous nature of breast tumors may be more conserved compared to cell line xenografts. PDXs have been shown to maintain genetic and molecular characteristics through several passages, and they retain clinical responses to drug treatment<sup>97,193</sup>. The major drawback of xenografts is that they require immune-compromised mice. Different humanization techniques (injection of human immune cells) may overcome this problem, however such techniques are expensive and labor-intensive and are currently not able to recapitulate a complete immune response<sup>201</sup>. The differences between human and murine stroma is also a limitation of xenografts models that need to be taken into account. Engraftment of breast tumor xenografts may be done subcutaneously, orthotopic (within the mammary fatpad) or intraductally<sup>97,202,203</sup>. The route of transplantation may be relevant for tumor development; for instance, Sflomos et al. showed that MCF7 cells (a luminal cell line) maintained luminal characteristics when transplanted intraductally while when transplanted into the fat-pad, the tumors obtained a basal-like phenotype<sup>204</sup>. There exist several human DCIS cell lines that may be used to study the transition of DCIS, e.g. MCF10ADCIS.COM, 12NTci and SUM-225<sup>193,205</sup>. Another relevant model for studying DCIS is the mammary intraepithelial neoplasia outgrowth (MINO) model, which are murine cell lines derived from MMTV-PyV-MT (mouse mammary tumor virus, polyomavirus middle T) lesions. This model enables studying MIN to invasive transition in immune-competent mice<sup>206</sup>.

GEMMs are usually generated using promoters such as mouse mammary tumor virus (MMTV) to over-express genes that promote tumorigenesis in a targeted fashion to the mammary gland. In contrast to xenografts, GEMMs are able to model tumor initiation and stepwise tumor progression (including MIN stage) and the invasion process, which makes them valuable tools for studying DCIS. They also have the advantage that both the microenvironment and the immune system are native, which may reflect human disease more precisely than xenografts, however the behavior of the tumor is less random, since tumor initiation is caused by one specific mutation. Few mammary gland tumors from GEMMs truly express ER; however, they may display luminal characteristics independently of ER. GEMM models especially relevant for studying DCIS include MMTV-PyV-MT, WAP-T and MMTV-neu<sup>97,193,205</sup>.

Chemical induction of tumors using carcinogens is another approach of inducing mammary tumors. One example of such is the MPA/DMBA model. 7,12-dimethylbenz[a]anthracene (DMBA) acts as a *tumor*

*initiator* by inducing DNA damage while medroxyprogesterone acetate (MPA) acts as a *promoter*, which targets tumor initiation to the mammary epithelium and drives proliferation. The DNA damage inflicted by DMBA is random, making this model highly heterogeneous and yielding tumors of different subtypes<sup>207–210</sup>.

As seen in human mammary tumors, murine tumors also display a high degree of intertumoral heterogeneity, and it is therefore important to be aware of the clinical and molecular features of each specific model and how well they represent human disease. Human xenografts in mice may be characterized similarly to human tumors with regards to subtype etc, since many features are preserved when they are transplanted to mice. Herschkowitz and Pfefferle and colleagues have explored transcriptomic characteristics of multiple different murine mammary tumor models and discovered as many as 17 different murine subtypes across 27 models. All human subtypes were represented by a murine model, however many of the models were heterogeneous, i.e. the model may yield tumors of different subtypes. Herschkowitz and Pfefferle also explored the role of ER in murine mammary tumors and found that some of the models resembled luminal mammary tumors by gene expression despite of being negative for ER, suggesting that luminal signatures in murine mammary tumors may be driven by GATA3<sup>185,211</sup>. These studies have illustrated the importance of being aware of the different mammary tumor models characteristics and to bear in mind the potential caveats of comparative oncology in general and murine models for breast cancer specifically.

## **AIMS**

The overall aim of this thesis was to explore the heterogeneity of breast cancer progression by performing a comprehensive genomic and transcriptomic mapping of breast tissue through different stages of tumor development. Through this, we aspired to obtain further insight into breast tumor invasion mechanisms and contribute to the understanding of why many early breast tumors never acquire invasive properties.

We have addressed this subject through five specific aims by:

- Identifying transcriptomic changes in normal breast tissue over time in relation to mammographic density.
- Exploring the heterogeneity of human and murine mammary tumors through molecular subtyping.
- Investigating the involvement of the microenvironment in breast cancer initiation and progression.
- Exploring subtype specific differences between DCIS and invasive human breast tumors.
- Mapping molecular heterogeneity in DCIS through mutational analyses.



## RESULTS IN BRIEF

### **Paper I: A longitudinal study of the association between mammographic density and gene expression in normal breast tissue**

*Journal of Mammary Gland Biology and Neoplasia*, 2019<sup>212</sup>, doi: 10.1007/s10911-018-09423-x

This paper describes a follow-up-study of the work by Haakensen et al. from 2010 (MDG1) where normal breast biopsies from 65 women were subjected to gene expression analyses, and associations between gene expression and mammographic density (MD) were explored<sup>213</sup>. High MD is associated with an increased risk of developing breast cancer, however, the underlying biological mechanisms are not elucidated, neither is the change in gene expression of normal breast tissue over time.

In the current study (MDG2), a selection of the women who participated in the first study were asked to donate new tissue biopsies and have new mammograms taken. The time between first and second biopsies ranged from 5 to 8 years. We calculated MD based on new mammograms, extracted mRNA from biopsies and performed microarray gene expression analyses on biopsies from 17 women. Eleven of these also had gene expression data and MD calculations available from the first study.

We first explored those genes that in the previous study were shown to be correlated to MD. Despite of low number of samples, impeding the discovery of significant findings, we validated an inverse correlation between RBL1 gene expression and MD, indicating involvement of the transforming growth factor  $\beta$  (TGF $\beta$ ) pathway. To study longitudinal changes in gene expression, we used a rank-based approach and correlated change in gene expression with change in MD for the 11 cases with data from MDG1 and 2. This revealed that breast tissue samples with a large decrease in MD from the first to the second time point sustained a high expression of several genes of the histone H4 family. In both cohorts, we assigned microenvironment subtypes and found that the active subtype had characteristics similar to the claudin-low breast cancer subtype. This corresponds well with a previous study performed on the MDG1 tissue samples<sup>214</sup>. We did not find any association between microenvironment subtypes and MD or RBL1 gene expression.

This study showed that inverse correlation between mammographic density and *RBL1* gene expression in normal breast tissue is consistent over time, and that the microenvironment subtypes are biologically meaningful also in normal tissue.

## **Paper II: Claudin-low-like mouse mammary tumors show distinct transcriptomic patterns uncoupled from genomic drivers**

*Breast Cancer Research, 2019<sup>210</sup>, doi: 10.1186/s13058-019-1170-8*

In this study we performed exome sequencing and gene expression microarrays to characterize chemically induced mouse mammary gland tumors. We induced tumors applying a MPA/DMBA protocol in 14 mice, and eighteen tumors and five normal mammary glands were included in the study.

We classified tumors and normal tissue by applying a centroid based murine subtyping method using gene expression data. This revealed high degree of intertumoral heterogeneity with nine subtypes represented. Clustering based on the murine intrinsic gene list showed that the tumors split into two clusters: one cluster showed a mostly homogeneous gene expression profile and consisted of tumors of either Claudin-low-like<sup>Ex</sup> or Squamous-like<sup>Ex</sup> subtype, the other cluster was more heterogeneous and consisted of tumors of seven different subtypes. The tumors in the homogeneous cluster resembled human claudin-low (CL) tumors with low degree of differentiation, low expression of genes involved in adhesion and high expression of genes involved in EMT and neoangiogenesis compared to the other cluster. The CL-like tumors also showed high immune scores and high expression of interferons and activation of immunosuppressive mechanisms. Of special note were *Ptgs2* (encoding COX-2) and *Cd274* (encoding PD-L1) highly expressed in both the murine CL-like tumors and in CL breast tumors in a human cohort (Metabric). Both these genes are clinically targetable and the results from our study suggest immune checkpoints as potential therapeutic targets in CL breast cancer.

Exome sequencing of the MPA/DMBA-induced tumors also revealed high intertumoral heterogeneity and marked higher mutation-rate compared to human breast cancer. We found mutations in known driver genes in all tumors, however we did not identify any association between specific mutations and murine subtype, indicating that cell of origin may play a larger role in explaining the observed gene expression phenotype than the specific driver mutation. We found that all tumors carried a characteristic mutational signature with an overweight of T>A transversion in TG dinucleotides. There was a tendency of lower copy number aberration burden in the CL-like tumors compared to all others.

Even though the MPA/DMBA model yields tumors of many different subtypes, and the mutational signature is different from human breast cancer, the transcriptomic phenotype of the resulting CL-like tumors indicates that these tumors may be appropriate as a model for studying human CL breast cancer.

### **Paper III: Contrasting DCIS and invasive breast cancer by subtype suggests basal-like DCIS as distinct lesions**

#### *Manuscript*

In this study, we explored differences in gene expression, DNA copy number and DNA methylation between ductal carcinoma in situ (DCIS, n=57) and invasive breast cancer (IBC, n = 313) in a subtype stratified manner. The distribution of PAM50 subtypes and ESR1 gene expression differed between DCIS and IBC with a higher proportion of tumors of HER2-enriched subtype and higher frequency of ESR1 negative tumors in DCIS compared to IBC. In general, DCIS showed lower correlation to the PAM50 subtype centroids compared to IBC. This was particularly evident for basal-like tumors where *core basal* tumors were found only in IBC and not in DCIS.

Subtype stratified comparison of DCIS and IBC showed marked differences between the subtypes. There was high resemblance between Luminal A DCIS and Luminal A IBC, while for the basal-like subtype, the DCIS were markedly different from IBC at all genomic levels. The basal-like DCIS showed lower proliferation and higher differentiation characteristics compared to basal-like IBC while we did not find significant differences in immune, stromal or EMT scores between basal-like DCIS and IBC. Copy-number data revealed that basal-like DCIS showed some basal-like features (as defined for IBC), but had overall much fewer copy number aberrations than basal-like IBC. Genes differentially methylated between DCIS and IBC were identified for each subtype separately and no differentially methylated genes were common between the subtypes. The basal-like subtype had notably more differentially methylated genes between DCIS and IBC compared to the other subtypes. Most notable was hypermethylation of 19 clustered protocadherin (cPCDH) genes in basal-like IBC compared to basal-like DCIS. These genes are located on chromosome 5q which is commonly deleted in core basal tumors. Hypermethylation of cPCDH genes has been reported to occur in breast cancer and other cancer types and is most likely the result of long range epigenetic silencing in which hypermethylation occurs over longer stretches of the genome.

This study affirms that subtype stratification is important when studying progression from DCIS to IBC, and we suggest that basal-like DCIS and basal-like IBC may represent different entities.

## **Paper IV: Comparable cancer-relevant mutation profiles in synchronous ductal carcinoma in situ and invasive breast cancer**

### *Manuscript*

In this study we performed targeted sequencing of 26 mixed tumors (invasive tumors with synchronous DCIS) and 10 pure DCIS tumors. The mixed tumors were microdissected using laser capture microdissection in up to three cellular compartments per tumor: Normal epithelium, DCIS and invasive tumor. After DNA isolation, we performed Ion Torrent sequencing of hotspot regions of 50 known cancer driver genes. In total were 44 cellular compartments from mixed tumors sequenced in addition to the 10 pure DCIS.

We identified 22 hotspot variants in eight different genes across the microdissected mixed tumors. The genes harboring most variants were *PIK3CA* (four different variants in ten tumors) and *TP53* (nine different variants in eight tumors). The most common variant was *PIK3CA*:p.H1047R. We sequenced thirteen pairs of DCIS and invasive compartments and in six of these, DCIS and invasive compartments harbored identical variants. In the remaining cases, some variants were found in the invasive compartment only, while in other cases, variants were found in the DCIS compartment while not in the invasive. In eight out of nine normal samples no hotspot variants were identified. One normal sample harbored the same variant as found in the corresponding tumor. Digital droplet polymerase chain reaction (PCR) was used to validate sequencing of one variant (*PIK3CA*:p.H1047R). We were able to detect this variant in comparable frequencies as identified by sequencing. Progesterone receptor (PR) positivity was significantly correlated with presence of *PIK3CA* variants. For comparison, ten pure DCIS were included in the study. Only three hotspot variants in three different tumors (one in *TP53* and two in *PIK3CA*) were identified in total across all pure DCIS tumors. The frequency of tumors carrying variants was significantly different between pure DCIS and synchronous DCIS from mixed tumors

We found few mutational differences between synchronous DCIS and IBC which is in accordance with previous studies. The number of samples in this study is too small to draw firm conclusions, but our results may indicate fewer potential driver mutations in pure DCIS compared to DCIS concurrent with IBC.

## METHODOLOGICAL CONSIDERATIONS

### Material

A major challenge in biomedical research is limited access to patient material. In all studies including human tissues, there is a possibility of introducing bias due to low amount of available tissue. Especially DCIS tumors are difficult to obtain since these lesions often are small and a large part of the tumor is used for routine histopathological evaluation. The tumors available for research may therefore be larger tumors, possibly skewed towards late stage or fast growing tumors. Tumor cell lines and murine tumor models are useful supplements to studies on human material as there are limited possibilities to perform interventions in human tumors; however such models may have artificially low variance and do not recapitulate the large degree of intertumoral heterogeneity observed in human breast tumors. The biological differences between the species also need to be taken into consideration.

### Material in paper I

This study (Mammographic Density and Genetics 2 - MDG2) included material from seventeen normal breast tissue biopsies obtained from healthy women. These women had previously donated breast biopsies for the preceding study (MDG1)<sup>213,215</sup>. The women included in the first study were asked to participate because of suspicious findings on routine mammograms. Therefore, there is a possibility that these women had an overall higher MD than the average populations since mammograms of high density breasts may be more difficult to interpret<sup>50</sup>. Nevertheless, the variation of MD in the MDG1 population was high, and there were no indications of a bias towards high MD subjects. Normal breast biopsies from relatively young women are hard to obtain and it would have been beneficial to include more subjects to improve statistical power, however this was not feasible. Even though biopsies were taken in an area of dense mammary tissue, there was not enough material for histological evaluation of the biopsies in addition to the molecular analyses. We therefore do not know whether the cellular composition differed between the samples and thus impacted the results. Samples with low RNA yield could not be included in the gene expression analysis, but since we did not observe any association between RNA yield and MD, excluding these samples did most likely not influence the results.

### Material in paper II

In this study, mammary tumors from transgenic (Lgr5-creERT2-EGFP;R26R-Confetti) mice on a FVB/N background were studied. The transgenes themselves were not relevant for this study and are inert, so the specific genotype did not affect the results. Tumors were induced in mice according to an

established protocol including medroxyprogesterone acetate (MPA) and the carcinogen 7,12-dimethylbenzanthracene (DMBA)<sup>207</sup>. In total, 18 mammary gland tumors from 14 mice and 6 normal mammary glands (from non-treated mice) were subject to genomic analyses. Histological assessment was performed by a trained veterinary pathologist.

We explored the genomic and transcriptomic features of MPA/DMBA induced tumors. Previous transcriptomic analyses of tumors generated using the DMBA/MPA model system has shown that this model is heterogeneous, yielding very diverse tumors<sup>211</sup>. This was also the case in our experiment; even two tumors from the same mouse were highly different. In contrast to homogeneous mouse models where the tumors' phenotypes are quite predictable, we needed to account for large intertumoral heterogeneity by increasing the number of animals to ensure statistically relevant analyses. However, the need for high N had to be balanced against ethical considerations of the included number of animals. Tumor growth in each mouse was monitored and the mice were euthanized either at a pre-defined time point or when the maximum allowed tumor volume was reached.

### Material in paper III

This study includes tumors from three different cohorts: "OSLO2"<sup>216</sup>, "Uppsala"<sup>192</sup> and "Milano" (previously unpublished). We decided to include tumors from different cohorts since DCIS tumors are not readily available for research purposes and DCIS cohorts often are quite small. We also included invasive tumors from the same cohorts. An overview of the samples is shown in Table 3.

**Table 3. Overview of samples included in paper III**

<b>Cohort</b>	<b>DCIS</b>	<b>IBC</b>
OSLO2	7	302
Uppsala	22	6
Milano	28	5
Total	57	313

All tumors were subjected to RNA and DNA isolation in our laboratory, and were analyzed using the same microarray platform to avoid merging of data from different array types. The cohorts were run on microarrays at different time points, so randomization across cohorts was not possible. In total, 57 DCIS and 313 IBC tumors were included. All tumors were evaluated by a breast pathologist ensuring that no DCIS tumors had invasive foci larger than 1mm (i.e. DCIS with micro-invasion was still considered pure DCIS). DCIS tumors were classified according to the European Organization for Research and Treatment

of Cancer (EORTC) system<sup>80</sup> while the IBC tumors were classified according to the Nottingham (Elston & Ellis) system<sup>217</sup>. Tumors originated from different countries, and for many of the patients clinical information was sparse and we lacked reliable recurrence and survival data. We also lacked IHC results for ER, PR and HER2 for some samples and different threshold values were used for evaluation. To be able to interpret these across cohorts, we therefore used genomic data (*ESR1* and *PGR* expression and *HER2* copy number) to determine the status of these biomarkers.

The DCIS tumors in this study were all pure DCIS with no signs of invasion. This means that DCIS and IBC tumors were compared as groups, not as DCIS-IBC pairs from the same (mixed) tumor. Several studies have been performed on tumors with both DCIS and IBC components present, enabling comparison between different tumor compartments (using micro-dissection)<sup>43,218,219</sup>. In such studies, DCIS and IBC have been shown to be very similar. However, DCIS from mixed tumors and pure DCIS (from lesions without invasion) may have very different biology and it may be advantageous to consider these tumors as distinct groups. All analyses were performed on bulk tumor, i.e. without micro-dissecting into cellular compartments (except trimming of excess adipose tissue during tumor preparation). This was done for two reasons: First, we wanted all tumors in the cohort to be prepared the same way, and second, we wanted to also include cells from the immediate microenvironment surrounding the tumor. The drawback of studying bulk tumor rather than micro-dissected cellular compartments is that we do not obtain information about the specific contribution from the different cellular compartments.

#### **Material in paper IV**

In this paper, we have included material from mixed tumors (i.e. synchronous DCIS and IBC) and pure DCIS from the Uppsala cohort (described in paper III). The mixed tumors were micro-dissected using laser capture micro-dissection (LCM) into three different cellular compartments: normal, DCIS and IBC. Initially, 76 samples from 33 patients with mixed tumors were selected. However, due to difficulties during DNA extraction mainly because of sample storage in Trizol, many samples were discarded. In total, 44 samples (19 IBC, 16 DCIS and 9 normal) from 26 different patients were successfully sequenced. These included three triplets (IBC, DCIS and normal from the same patient) and 10 IBC/DCIS pairs. For comparison, 10 pure DCIS tumors were included to explore whether there were any consistent distinctions in mutational profiles between DCIS from mixed tumors and pure DCIS.

### **Gene expression microarrays**

Gene expression microarrays enable detection and relative quantification of RNA levels of genes expressed in a sample. In this thesis, I have used gene expression microarrays in paper I, II and III. In all three papers, we used Agilent SureprintG3 Gene Expression 8x60K arrays with the Low Input Quick Amp Labeling protocol<sup>220</sup>. The human version of the array was used in paper I and III and the murine version in paper II. In microarrays, gene expression is measured by hybridization of RNA from a sample to DNA probes immobilized on a glass surface. The measurement of gene expression by microarrays is highly indirect and due to the kinetics of hybridization, the fluorescence signal that is detected is not proportional to RNA content for all RNA concentrations, i.e. the dynamic range of detection is limited. A second challenge of microarrays is the fact that probes are not 100% specific, and some cross-hybridization with RNA molecules with similar sequence may occur. Finally, microarrays only detect known mRNAs, thus novel or un-annotated transcripts are not covered in this type of analyses<sup>221,222</sup>.

At the time our study was initiated, microarray was the preferred method. Since then, RNA sequencing (RNAseq) has become increasingly common. Compared to microarrays, RNA sequencing is a more direct method of analyzing the transcriptome with a wider dynamic range than microarrays and no upper limit of quantification. It is also more sensitive at low expression levels and has the ability to detect short reads and base-level resolution. The largest drawback of RNAseq is the high cost<sup>222</sup>. Another recently developed technology is a digital molecular barcode counting system (Nanostring). It is limited to ~800 different transcripts, but requires no amplification, cDNA conversion or library prep resulting in clean and reliable data also from formalin fixed paraffin embedded (FFPE) material<sup>223,224</sup>.

In paper I, gene expression analysis was performed on normal breast samples. We used Agilent 60K microarrays, while the previous project was performed using Agilent 44K arrays. It was not feasible to merge the two datasets without losing biological signal; therefore the datasets were analyzed separately. The basic question of whether or not we could validate the association between MD and expression of specific genes was not affected by this, however direct comparison of gene expression between the two time-points was more challenging. To circumvent this problem, we used a rank-based approach where gene-ranks were used as a proxy for gene expression and the change in gene rank from time point 1 to time point 2 would represent change in gene expression. In paper III, all three cohorts were run on the same type of array, but at different time points, by different operators and using RNA isolated by different methods. This could potentially create problems, hence we investigated whether these issues was causing batch-effects.



## Gene expression data analyses

### Data preprocessing

In all papers including gene expression data, preprocessing was performed similarly. Quality control was performed using the Agilent Feature Extraction Software and samples that failed were rerun. Since the distribution of raw intensity values in gene expression analyses is highly skewed, with some extremely high values, we log-transformed the intensity values.

A frequent objective of gene expression microarray studies is to identify biological differences between samples. However differences due to technical issues may occur along virtually all steps of the procedure. To compensate for systematic differences between samples run on microarrays, the data are normalized. Normalization of arrays is made possible based on the assumption that most genes on the array are expressed at approximately the same level. The relatively few genes that are differently expressed between samples should not influence the normalization substantially, however, if the dataset consists of samples that are profoundly different (such as samples from normal tissue and tumor samples) this assumption is no longer valid and normalization of such datasets using standard methods may prove difficult<sup>225</sup>. The most common way of normalizing microarray gene expression data is through quantile normalization. This method forces the data from each sample into the same distribution making it possible to compare gene expression between samples. It may however also reduce true biological differences since extreme values are artificially reduced<sup>226,227</sup>. To normalize the data for paper III, including compiled data from three different cohorts, we used quantile normalization across all tumors in all three cohorts.

In the gene expression arrays used in our studies, multiple probes may represent one gene. In paper II and III we wanted the gene expression data to be on gene level. For this, we calculated the mean of all probes representing the same gene. In paper I (MDG2), we wanted to compare the data to the results from the previous study (MDG1), and since the former study was analyzed on probe-level, we used the same approach on MDG2. In all papers, we performed principal component analyses (PCA) on the final dataset to identify outliers and inspect batch effects. This was especially important in paper III where gene expression arrays from three cohorts were run at different times.

### Subtyping of mammary gland tumors

A major aim of this thesis was to identify and explore the properties and relevance of different molecular subtypes in breast cancer progression. Subtyping enables classification of tissue or tumors into groups that share molecular characteristics. This information may be valuable for e.g. treatment

decisions or for studying underlying biological processes. In three of the papers in this thesis (papers I, II and III), different subtyping methods have been performed.

In paper I, we used microenvironment subtyping to characterize normal mammary gland tissue. The method is based on publications from Román-Pérez et al. and Sun et al. who analyzed tissue samples adjacent to breast tumors and characterized the tissue as *active* or *inactive*<sup>228,229</sup>. These subtypes had not previously been investigated in normal breast tissue. Microenvironment subtypes were assigned by calculating Pearson correlation between a vector of weights obtained from the original paper and gene expression values of genes in a signature gene list<sup>228</sup>. The sample was defined as *active* subtype if the correlation coefficient was positive, and *inactive* if negative. Neither the correlation coefficient, nor the P-value was taken into consideration. In retrospect, including these measures in the analyses to determine the strength of the association to the assigned microenvironment subtype could have been advantageous.

The MPA/DMBA-induced mammary gland tumors and normal mammary gland tissue in paper II were subtyped using a method described by Pfefferle et al.<sup>211</sup>. This method is an example of a nearest centroid classifier, where a tumor's subtype is derived by determining how well the sample fits with several predefined centroids representing the different subtypes<sup>230</sup>. The subtyping of our cohort revealed that 8/17 tumors were of the Claudin-low<sup>Ex</sup> or Squamous-like<sup>Ex</sup> subtype. Pfefferle et al. proposed that these subtypes resembled human claudin-low tumors and samples of these subtypes were the focus of our study. The remaining nine tumors were of seven different subtypes and this mixed group of tumors served as a basis of comparison. The high number of different subtypes identified illustrates what has been shown before, that the MPA/DMBA-model is a heterogeneous model yielding tumors of several different subtypes. Since our cohort was relatively small, we decided to compare the two main groups that emerged after hierarchical clustering. This approach inevitably dilutes the specific signal from individual tumors and subtypes, however we considered it necessary to obtain sufficient statistical power. We used the SigClust tool to confirm that the two tumor clusters were significantly different<sup>231</sup>.

In paper III, on the cohort including human DCIS and IBC tumors we performed PAM50 subtyping: A simplified 50-gene subtype predictor designed to capture the original human intrinsic subtypes<sup>171,176</sup>. This method is also a nearest centroid classifier. All DCIS and IBC were subtyped concurrently using the normalized dataset. We obtained the subtype centroids from the original publication on PAM50 subtyping by Parker et al.<sup>176</sup>. In this dataset, the fraction of ER-positive tumors was ~60%, while the

fraction was ~80% in our dataset. This may influence the centering of the data that is performed prior to subtyping, and consequently the subtyping results. To account for this difference, we calculated the centering of each gene separately for ER+ and ER- patients (C+ and C-). We then found the gene's centering factor (C) by multiplying C+ and C- by the original fraction of ER+ and ER- tumors in the training set using this formula:

$$C = 0.6 C^+ + 0.4 C^-$$

After centering, we calculated spearman correlation between gene expression of the 50 genes in each sample and the subtype centroids, obtaining four correlation coefficients for each sample (basal-like, HER2-enriched, luminal A and luminal B). A sample was assigned to the subtype to which it showed highest correlation. Importantly, the four correlation coefficients serve as a continuous measure that may provide additional information about the tumor's characteristics.

In all centroid-based subtyping, the centering of the data is a critical step. Data sets with very few samples or a cohort consisting of samples with highly different biology (such as tumors and normal tissue) will skew the centering of the data and affect subtyping. It is also important that a dataset represents all subtypes. Importantly, PAM50 subtyping is a tool for subtyping of human tumors. Subtyping of murine tumors using this method would yield unreliable results.

### **Creating a DCIS score using multivariate logistic regression**

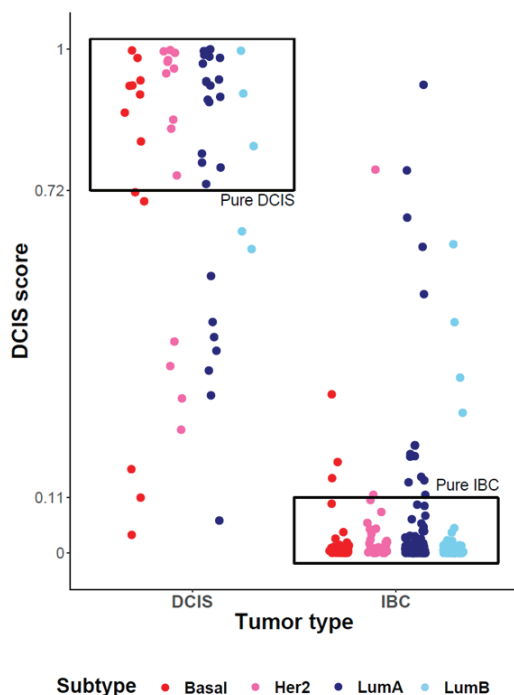
In paper III, we aimed to create a score that could be used to predict the "DCISness" of a tumor, i.e. how much the tumor resembled a DCIS (and simultaneously differed from IBC). We postulated that this approach could be used to predict which DCIS patients could be spared treatment. Although we decided to not include this part in the current manuscript, the DCIS-score is discussed in this thesis since the concept is highly relevant for studying DCIS biology.

For each of the three genomic levels (gene expression, copy number and methylation), we fitted a high dimensional logistic regression model with tumor stage (DCIS or IBC) as independent variable (response variable) and genes as dependent variables (covariates). To handle the much larger number of covariates (>20 000 genes) than number of samples, we used LASSO (least absolute shrinkage and selection operator) which is a regression analysis method that includes both variable selection and penalization<sup>232</sup>. It assumes that many genes are of limited importance and adds penalty to the

coefficients. Many of the coefficients will be set as 0 and thus eliminated from the model. This results in a sparse model with few covariates. To attempt to account for effects caused by molecular subtype, we included correlation coefficient to the basal-like centroid and its interactions with each gene in the model.

To obtain DCIS scores, we used LASSO to predict tumor stage for each sample separately. To avoid overfitting, we used a leave-one-out approach. The result from this method is each sample's probability of being a DCIS (the DCIS score). The range of the score is 0 to 1, where 1 means DCIS-like, while 0 means invasive-like. The overall performance of the models was measured by the predicted mean square error pMSE, i.e. the mean squared difference between the DCIS score and the known tumor state (the histological diagnosis set by the pathologist)<sup>233</sup>.

The model using gene expression data excelled the models using DNA copy number or DNA methylation data. Figure 8 shows gene expression based DCIS scores for all samples and illustrates that most IBC were very invasive-like, while the variance was much greater for the DCIS tumors. To uncover whether any specific biology was associated with the DCIS score, we defined two groups of "pure" tumors:



Standard deviation was calculated separately for DCIS and IBC ( $sd_{DCIS}$  and  $sd_{IBC}$ ). Pure DCIS was defined as DCIS tumors with DCIS-score  $> (1 - sd_{DCIS})$  while pure IBC was defined as IBC tumors with DCIS-score  $< sd_{IBC}$ . We analyzed pure DCIS and pure IBC in a subtype-specific manner. However, when comparing these analyses to the analysis that included all tumors, we found only minor differences. To further develop the score to predict which DCIS tumors are indolent, we would need to include relapse or survival data for validation which unfortunately was not complete for this cohort. Based on the other results in the manuscript of paper III, we are convinced that studying DCIS to IBC progression needs to be carried out in a subtype specific manner. It is therefore reasonable to think that a potential DCIS score should be derived for each subtype separately, or at least separate for basal-like and non-basal-like.

**Figure 8. Gene expression based DCIS scores for DCIS and IBC tumors from all cohorts in paper III. Boxes illustrate pure tumors (determined by standard deviation).**

However, this approach will inevitably reduce statistical power and we would need a larger dataset (most importantly, many more DCIS) to be able to perform such analyses with a sufficiently high statistical power.

### **DNA copy number analyses**

Some cancer types are predominantly driven by mutations, while breast cancer is characterized by recurring copy number aberrations<sup>23</sup>. We have used DNA copy number analyses in paper II and III. In paper II, copy numbers were estimated based on exome sequencing data using EXCAVATOR 2<sup>234</sup>. The copy number analyses in paper III were performed using Affymetrix™ Genome-Wide Human SNP Array 6.0 (SNP6). This array includes 1.8 million markers covering large parts of the human genome, including both SNP-probes and non-polymorphic probes (to increase coverage across the whole genome). The principle behind SNP6 arrays is the same as for gene expression arrays: Oligonucleotides immobilized on a glass slide bind fragmented genomic DNA in an allele specific fashion. After hybridization, a scanner detects emission of fluorescent signals. The data from such arrays may be exploited to calculate copy number of alleles at all loci represented on the array as signal will increase with increased number of DNA copies. The output from SNP6 arrays consist of two intensity measures, one for the major allele (A - the allele with highest frequency) and one for the minor allele (B). After preprocessing, the total intensity (logR) may be calculated as  $\log(A+B)$  and the B-allele-frequency (BAF) as  $B/(A+B)$ .

Cancer genomes are commonly highly aberrant which complicates copy number analyses of tumors. Examples of such aberrations are aneuploidy (tumor lacks the normal diploid state of two copies at each locus), normal cell infiltration (the signal that comes from a tumor is altered due to contribution from normal cells) and tumor heterogeneity (different parts of the tumor may harbor different aberrations). An algorithm termed ASCAT (allele-specific copy-number analysis of tumors) may in part handle these complicating factors. This algorithm takes advantage of the allele specific information from the SNP6 arrays and returns an estimate of the tumor's aberrant cell fraction (tumor cell percentage) and tumor ploidy in addition to calculating adjusted allele specific copy number<sup>235,236</sup>.

In paper III we analyzed copy number data from both DCIS and IBC tumors. Data was preprocessed using the PennCNV-Affy library with the HapMap samples as reference set and corrected for GC content<sup>237-239</sup>. GC correction reduces the impact of GC content on the results, since this may create a wavy artifact. Copy number data may be noisy, which complicates interpretation. To attenuate the noise, a

segmentation algorithm was applied to partition the intensities into regions of homogeneous mean intensity levels. For segmentation, we used the piecewise constant fit (PCF) algorithm in the R “copynumber” package<sup>240</sup>. To obtain one copy number value for each gene, we selected the segment with highest overlap. Finally, we applied the ASCAT algorithm to the data and obtained aberrant cell fraction and ploidy for all tumors. We also calculated the genomic instability index (GII) which gives a measure of the overall instability of a tumor’s genome.

An alternative approach of obtaining copy number data is by whole genome DNA sequencing, which provides not only copy number data and SNPs, but also mutations, structural variants etc. Deriving copy number data from exome sequencing is also possible, but yields lower resolution since large parts of the genome are not covered.

### **DNA methylation analyses**

DNA methylation is an epigenetic regulator of gene transcription and plays an important role coordinating biological processes in physiological conditions. DNA methylation is the covalent addition of a methyl group (-CH<sub>3</sub>) to the 5-carbon of the cytosine ring within a CpG dinucleotide context (a CpG is a cytosine immediately followed by a guanine in 5’->3’ direction). Around 60% of all mammalian genes harbor CpG islands (clusters of CpGs) in their promoter regions. Hypermethylation of such promoters often lead to decreased expression of the gene. In tumors, aberrations such as hypermethylation of tumor suppressor genes or hypomethylation of oncogenes may play a part in deregulating gene expression of these genes thus contributing to cancer development<sup>241</sup>.

DNA methylation analyses were performed in paper III, to compare DCIS and IBC in a subtype specific manner. The analyses were performed using the Illumina Infinium HumanMethylation450 microarray. This array uses the Infinium “BeadChip” technology that quantifies methylation level at >450.000 methylation sites across the genome. Briefly, genomic DNA was treated with bisulfite to convert unmethylated cytosine to uracil. This enables detection of methylated and un-methylated states of CpGs by measuring fluorescence of beads representing the two methylation states. The intensity ratio between the two beads ( $\beta$ ) represent the methylation values ( $\beta=0$ : no methylation,  $\beta=1$ : total methylation). Importantly, when interpreting tumor DNA methylation, normal cell infiltration, ploidy and tumor heterogeneity need to be taken into consideration<sup>241,242</sup>. The BeadChip provides genome-wide coverage of DNA methylation and is a widely used technology. Today, whole genome bisulfite

sequencing using next generation sequencing is considered gold standard for the most comprehensive and quantitative measurements of DNA methylation, but this is a costly technology<sup>243</sup>.

Normalization of methylation data in paper III was performed using subset quantile normalization<sup>244</sup>. The number of CpGs associated with each gene is highly variable, from 1 up to ~1200. To obtain only one value for each gene per sample, we collapsed the  $\beta$ -values. By doing this, we lost specific information about e.g. promoter or gene body methylation, but obtained a substantial dimension reduction (necessary for doing e.g. differential methylation analyses on gene level) and were also able to more easily compare the methylation data with other gene level data (such as gene expression and copy number data). One way of collapsing  $\beta$ -values to gene level would simply be by calculating the arithmetic mean across the  $\beta$ -values in each gene, however, using this approach would weight all  $\beta$ -values equally. Instead, we used PCA including all CpGs within each gene and 50kB upstream or downstream from the gene. We defined the first principal component value as the gene's methylation *profile value*. This is created so that CpGs with highest variance in  $\beta$ -values across samples contribute more than CpGs with low variance. With the methylation profile values, we performed differential methylation analyses separately for each subtype. The p-values were corrected for multiple testing using false discovery rate (FDR). To obtain gene lists for functional enrichment analyses, thresholds including both FDR (<0.05) and effect-size was set. Effect-size threshold was included to increase probability of identifying not only statistically significant differences between DCIS and IBC, but also genes with biological relevant differences.

### **Targeted DNA sequencing**

Although copy number aberrations are important oncogenic drivers in breast cancer, somatic mutations also contribute to breast cancer formation<sup>245</sup>. Next generation DNA sequencing is a valuable tool for assessment of the mutational status of tumors. However, whole genome sequencing or even whole exome sequencing is expensive and produce large amounts of data that requires complex and resource-intensive data analysis. Using a more targeted approach, with a limited panel of genes focused towards known driver genes may, in many cases, be more appropriate. Targeted sequencing allows for increased depth of coverage which enhances sensitivity, thus increasing the chance of discovering low frequency variants. In addition, providers of targeted sequencing panels often offer streamlined data handling and analysis tools that facilitates data analyses for non-bioinformaticians.

In paper IV, we used the Ion Torrent sequencing platform with the Ion AmpliSeq™ Cancer Hotspot Panel v2 to sequence microdissected samples from mixed DCIS/IBC tumors. The Cancer Hotspot Panel includes primers for amplification of 207 amplicons covering ~2800 COSMIC mutations (Catalogue Of Somatic Mutations In Cancer<sup>22</sup>) in 50 oncogenes and tumor suppressor genes relevant for solid tumors<sup>246</sup>. We used 100pg DNA as input, lower than described in the standard protocol (10ng). Ion Torrent uses the ion semiconductor sequencing technology, which detects hydrogen ions released during polymerization of DNA. The template DNA to be sequenced was flooded with deoxyribonucleotide triphosphate (dNTP), one species at a time, and hydrogen ions were detected as the complementary strand is built onto the template<sup>247</sup>. Data was analyzed using the AmpliSeq™ Variant Caller plug-in within the Ion Torrent Suite software. Three samples failed quality control and these were excluded from further analyses. In the remaining samples, the Ion Torrent Suite called a high number of variants; many with low frequency and low quality score. To avoid false positive variants, a strict threshold was applied and the remaining variants were assessed manually in Integrative Genomics Viewer<sup>248</sup> to evaluate strand bias and possible technical errors. Variants demonstrating obvious PCR duplication errors were also excluded. To only include variants that may be of clinical significance, we included only variants present in the COSMIC database and since only somatic mutations were of interest, we excluded SNPs present in the variant database in the 1000 Genomes Project<sup>249</sup>.

To validate the sequencing results, we used digital droplet PCR (ddPCR) on the RainDrop system (RainDance technologies) on one selected variant (*PIK3CA*:p:H1047R) in nine samples. We identified the variant at similar frequencies as was seen in the sequencing experiment. In a previous paper, three different *TP53* mutations were identified using Sanger sequencing of DNA from three tumors that were also included in our study<sup>250</sup>. Two of these mutations were identified in our variant calling pipeline. The third mutation, a 10bp deletion, was not called, however; the deletion could easily be identified by inspecting the data in IGV. The reason that this mutation has not been called in our pipeline may be because of a known issue in the Ion Torrent system, where deletions in homopolymer regions (several equal nucleotides in a row) are systematically under-called<sup>251</sup>. The identification of called variants by alternative methods confirms the validity of our findings, however; the stringent filtering that we applied may have caused potentially important low frequency variants to pass undetected. Due to noisy data these would be indistinguishable from false positive calls. The Ion Torrent technology uses small sized amplicons (~150 nucleotides) which enables the system to tackle highly degraded DNA (e.g. DNA from FFPE tissues). In our study, despite using low quality DNA that had been stored in sub-optimal conditions and a lower amount of DNA input than recommended, we still obtained reliable variant



callings, however we might be approaching the limits for what the Ion Torrent platform may be able to handle.

In retrospect we see that the choice of gene panel for this study might not have been optimal, as the Ion AmpliSeq™ Cancer Hotspot panel is not specific for breast cancer. For instance are genes commonly mutated in breast tumors such as GATA3 and KMT2C, not included in the panel, while mutations in e.g. VHL and SMARCB1 (genes that are included in the Cancer Hotspot panel) are very rare in breast cancer according to COSMIC<sup>22</sup>. An alternative gene panel for this study would have been the Ion AmpliSeq™ Comprehensive Cancer Panel. This encompasses >400 genes implicated in cancer. However, this panel requires substantially more DNA as input than was available in our case. It also would have been beneficial to include more cases especially when considering the results from paper III where we found that molecular subtype should be taken into consideration when studying DCIS. Since RNA for these tumors was not available, we were unfortunately unable to compute PAM50 subtypes in this study.

### Statistical considerations

A recurring issue in all the papers included in this thesis and many similar studies is the limited number of samples available. Especially in *omic* studies where the number of variables ( $p$ ) is high compared to the number of samples ( $n$ ), obtaining sufficient statistical power is a challenge. In univariate analyses of e.g. gene expression data, a statistical test is performed for each gene separately. In every test, there is a possibility that the difference identified is not a result of true biological effects, but has just occurred by chance, and for each additional test performed there is an increased chance that one of the tests might be a false positive. The P-value represents the probability of observing something more extreme than our data show given that the null hypothesis is true. Using the threshold  $\alpha=0.05$ , there is a 5% chance that a significant difference is actually not true (i.e. the result is a false positive). When performing multiple tests (e.g. when analyzing whole genome expression microarrays with >20.000 features), >1000 tests will on average be false positive at a threshold of 0.05. Hence, there is a need for a stricter definition of significance. Several methods are available to adjust for multiple testing<sup>252,253</sup>. In our papers, we have used *false discovery rate* (FDR). FDR controls the expected number of tests where the null hypothesis has been rejected falsely (false positives). An adjusted P-value (representing FDR) is calculated for each gene and threshold is set at the preferred level of false positives<sup>254</sup>. There is a tradeoff between the consequences associated with false positive results versus the benefit of identifying true positive results; therefore FDR thresholds must be adapted to the specific research

question. In addition, FDR is dependent on the number of tests performed. It could therefore be of benefit to reduce the number of genes to be tested i.e. by filtering out genes with low variance across the dataset prior to analysis, however filtering may also introduce bias<sup>255</sup>. Several statistical methods have been developed to deal with the  $p > n$  problem in microarrays. Examples of such are significance testing of microarrays (SAM)<sup>256</sup> and *limma*<sup>257</sup>.

When performing statistical tests, the goal is to assess statistical significance. However, if the result of a test is not significant, we cannot claim that no difference exists. There might actually be a true difference, however too few samples, random variation or noisy data may have interfered and obscured the results. In line with this follows that P-value thresholds should not be used for uncritically dichotomizing results into *significant* and *non-significant*. There is no fundamental difference between two tests that have P-values just below or just above the threshold, and important biological findings may be lost due to too rigid interpretations of P-values. Several scientists have been speaking up against uncritical use of statistical significance lately, promoting scrutiny of the data behind the P-values and this is important food for thought for everyone using statistical tests in data analyses<sup>258,259</sup>.

## ETHICAL CONSIDERATIONS

The studies in this thesis are based on material obtained from human breast cancer patients and mice. In all studies comprising human material, approval from regional ethics committees (REC) and patient consent have been obtained. The relevant REC approval numbers are listed in Table 4.

**Table 4. REC approval numbers for human cohorts included in this thesis.**

<b>Cohort</b>	<b>Paper number</b>	<b>Approval number</b>	<b>Location</b>
MDG2	I	2009/1898	Oslo, Norway
Metabric	II	07/H0308/161 12/EE/0484 07/Q0106/63	Cambridge, UK
OSLO2	III	2016/433	Oslo, Norway
Milano	III	PG/U-25/01/2012-00001497	Milano, Italy
Uppsala	III & IV	2005/118	Uppsala, Sweden

Animal experiments in paper II were performed according to the regulation on the use of animals in research, and approval was obtained from the Norwegian Food Safety Authority (approval number: FOTS 4385). All mouse experiments were designed and performed according to the three R's: *Replace, Reduce, Refine*. Replacement was not applicable as the aim of the study was to characterize the MPA/DMBA tumors as a model for human breast cancer. We obtained reduction in the number of animals by using our mouse cohort in several projects. Also, when more than one tumor arose in the same animal, but in different mammary glands, they were considered as independent, and this contributed to reducing the number of animals considerably. However, the MPA/DMBA model is heterogeneous and to obtain sufficient statistical power, more animals needed to be included than would have been necessary when studying a homogeneous tumor model. Refinement was obtained by conscientiously following the animal welfare guidelines provided by the animal facility and national regulations. All experiments were performed by trained personnel. Mice were inspected daily and were euthanized when the volume of a single tumor exceeded 1000mm<sup>3</sup> or the total tumor volume exceeded 2000mm<sup>3</sup>. Likewise, mice were euthanized if they showed signs of ill health.

Data created in these studies are or will be made publicly available through ArrayExpress<sup>260</sup> and the European Genome-Phenome Archive (EGA)<sup>261</sup> following Minimum information about a microarray experiment (MIAME) guidelines<sup>262</sup>. All data storage and handling is performed in compliance with the EU General Data Protection Regulation (GDPR).



## DISCUSSION

This thesis encompasses studies concerning different stages along breast tumor progression, from normal mammary epithelium to invasive breast cancer. In paper I, we studied processes in the normal mammary gland with relevance for breast tumorigenesis. Paper II describes a carcinogen-induced mouse breast tumor model that may be used to study tumor initiation and progression of a specific subtype, the claudin-low breast cancer, and paper III and IV address ductal carcinoma in situ and invasive breast cancer. Here, I will discuss molecular subtyping of mammary gland tumors and the importance of subtype stratification in tumor progression studies. I will also discuss the role of the microenvironment in breast tumor progression.

### Molecular subtyping of breast tumors

Cancer is an “N of 1” disease; all tumors are essentially different. Ideally, all cancer therapy should be tailored to each individual tumor, however this is a complex task, and requires resources and detailed expertise that is generally not available in a clinical setting. For many cancer types, there exist various predictive and prognostic biomarkers. Predictive biomarkers provide information about the effect of a therapeutic intervention, while prognostic biomarkers indicate the likelihood of patient outcome<sup>263</sup>. In some cases, single aberrations (e.g. EGFR-mutations in lung cancer<sup>264</sup> or PML-RAR $\alpha$  translocation in acute promyelocytic leukemia<sup>265</sup>) may provide sufficient information about the tumor to initiate specific treatments. However, in many tumor types, single biomarkers are not sufficient. Molecular subtyping, based on multiple features in combination, provides comprehensive characterization and classification of tumors and may have both predictive and prognostic value. In breast cancer, the intrinsic subtypes (later condensed to the PAM50 predictor) are thoroughly documented to have clinical importance and are currently being implemented as a tool for therapeutic decisions<sup>177,179</sup>. Subtyping is also valuable in explorative studies where it enables grouping of tumors with similar genomic characteristics to increase statistical power. It is however, important to keep in mind that subtyping also may lead to unwanted bias<sup>266</sup>.

Mouse models are valuable tools for studying human breast cancer, however, due to the substantial differences among human breast cancer subtypes and the large heterogeneity of mouse models, it is important to select a mouse model that best represents the human subtype in question<sup>199</sup>. Using human subtyping tools on murine tumors would lead to erroneous results, especially since the role of ER in mammary tumors differs substantially between the two species. Therefore, a separate method for subtyping murine mammary tumors has been developed. Seventeen murine mammary tumor subtypes

have been characterized, which is noticeably more than the 4-6 subtypes seen in human breast cancer<sup>211</sup>. In paper II, we used the murine subtyping method on MPA/DMBA-induced invasive mammary tumors that arose in mice. The tumor cohort was heterogeneous and approximately half of the tumors demonstrated a claudin-low-like phenotype. Since human breast cancer is a heterogeneous disease, heterogeneous mouse models could be believed to better represent human breast cancer than homogeneous models. However, the chemical induction of the MPA/DMBA-model leads to much higher mutation frequency rate, lack of luminal-like tumors and overrepresentation of claudin-low-like tumors compared to human tumors. This model could therefore not be regarded as representative for the whole spectrum of human breast cancer, but is a valuable model for studying human claudin-low breast cancer.

It is generally accepted that the molecular breast tumor subtypes are present also at the DCIS stage<sup>101,191,192</sup>, yet few have compared subtype characteristics of DCIS and IBC. The highly heterogeneous nature of breast cancer makes a subtype specific approach valuable when exploring breast tumor progression. In paper III we explored the relevance of molecular subtypes in DCIS. We found that DCIS exhibit subtype specific characteristics similar to those in IBC, but at a more moderate level. Notably, we did not find any significant difference in tumor cell percentage between DCIS and IBC. We appreciated in this study that even though tumors are categorized into different subtypes, the association to each subtype should be interpreted as a continuum. Furthermore, when performing PAM50 subtyping it would be appropriate to define tumors that correlate poorly to any of the subtype centroids as *indeterminable* and classify these as a separate group. This could reduce confounding noise and facilitate discovery of important biological differences between subtypes. However, in a clinical setting, an indeterminable category would probably not be advantageous.

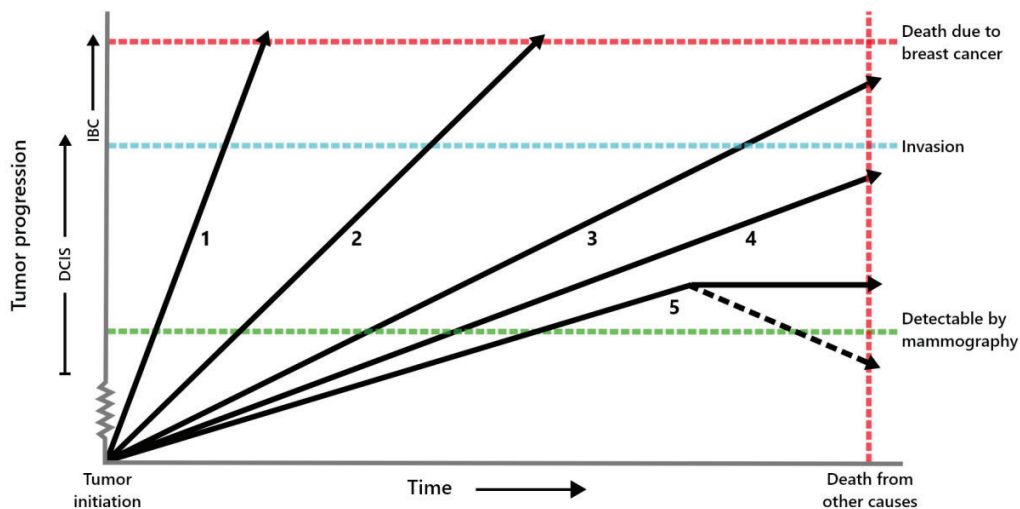
### **Subtype specific breast tumor progression**

During breast tumor progression, a tumor may go through different stages of increasing malignity<sup>89</sup>. DCIS is an important stage along this progression since this is usually the earliest stage where breast tumors are diagnosed either through mammographic screening or because of clinical symptoms. Furthermore, the transition of a tumor from an intraductal to an invasive state is a critical event in tumor progression since only invasive tumors are considered to have metastatic potential<sup>69</sup>. Breast cancer is highly heterogeneous and it is reasonable to believe that the mechanism and rate of progression from DCIS to IBC may differ considerably from tumor to tumor. Several previous studies

have concluded that there are few genomic differences between DCIS and IBC<sup>96,101,102</sup>. However, as we discuss in paper IV, pure DCIS (i.e. DCIS with no associated invasive foci) is potentially a distinct entity compared to DCIS from a mixed lesion (with synchronous DCIS and IBC). In paper III, we aimed to explore the subtype specific difference between DCIS and IBC to elucidate mechanisms of invasion. Here, only pure DCIS were included in the DCIS group. DCIS in mixed lesions has already proven to have invasive potential and considering these DCIS as representative also for pure DCIS would be misleading. This is supported by a study by Knudsen et al. where genes differentially expressed between pure DCIS and pure IBC were shown to be already deregulated in DCIS cells in mixed lesions<sup>267</sup>.

The molecular subtypes of invasive breast tumors are highly disparate in biology, prognosis and response to treatment<sup>170-172</sup>. The subtypes are linked to different cells of origin, which implies that a tumor's subtype is determined before the tumor becomes invasive and one can assume that tumor progression and tumor invasion also may differ between the subtypes. In paper III, we stratified the tumors by subtype before comparing DCIS and IBC. We found extensive differences between DCIS and IBC, especially in the basal-like subtype, while luminal A DCIS and IBC were relatively similar. This is in accordance with results in other studies<sup>105,268</sup> and strongly suggests that subtypes should be taken into consideration when studying breast cancer progression. Previous studies concluding high resemblance between DCIS and IBC may have been confounded by subtype: Since there is overabundance of luminal A tumors of both DCIS and IBC in most breast tumor cohorts, the signal from luminal A tumors would dominate and could confound the results, leading to the erroneous conclusion that DCIS and IBC in general are very similar.

From the point of breast tumor initiation, a lesion may follow one of several different progression paths. Figure 9 (inspired by a figure in Groen et al.<sup>269</sup>) illustrates this heterogeneity and depicts a model of five possible natural breast cancer progression paths. Path number 1 includes rapidly growing tumors that are in the DCIS stage only transiently and can quickly lead to metastasis and death. Invasive tumors of medullary type and those associated with BRCA1-mutations have been shown to lack an in situ stage, and may be examples of tumors following path 1<sup>54</sup>. The interval cancers (those that are diagnosed between two regular mammographic screenings) are also likely to populate this group, as time from detection by mammography to aggressive disease is short. Path number 2 represents the tumors that have slow progression with longer time spent in the DCIS stage. They still carry potential to become lethal. Path 3 consists of even slower developing tumors with invasive potential, but these tumors are developing so slowly that even though they become invasive, they will not lead death caused by breast



**Figure 9. Heterogeneity of breast cancer progression.** The black arrows represent different progression paths after tumor initiation. The green line represents the point where a tumor may be detected by mammography and the blue line represents the point of invasion. The red lines represent death, either due to breast cancer (horizontal line) or other causes (vertical line). Figure modified from Groen et al.<sup>269</sup>.

cancer. Path number 4 and 5 are tumors that never become invasive. The tumors of path 4 lack invasive ability and can be thought to grow by dilatation of the ducts or along the ducts, while the tumors in path 5 stop growing completely. There are even reports of DCIS tumors regressing spontaneously (path 5, stippled line)<sup>270</sup>. Patients carrying tumors in path 1 and 2 will be in need of treatment, while patients carrying tumors in the other paths may benefit from less extensive treatment or active surveillance. Most likely, the distribution of the molecular subtypes differs between the paths; it is reasonable to assume that basal-like and high-grade tumors are overrepresented in path 1 and 2, while low-grade tumors and luminal subtypes are more likely to be found in path 3, 4 or 5. This would be in accordance with other studies that have shown that DCIS tumors follow one of two major courses of development: A low grade course with overrepresentation of ER-positive tumors and a high-grade course, with ER-negative and HER2-positive tumors<sup>54,89,271</sup>. A tumor diagnosed as DCIS can belong to any of the five developmental paths. Tumors of path 1 are only transiently in an intraductal stage, so the time frame for detecting a DCIS in this path is short. Since several studies show that many DCIS, if left alone, never progress to invasive disease, many DCIS probably grow according to path 4 or 5. These would not be in need for treatment. Not even all invasive breast tumors will lead to death of the patient. The ultralow-risk invasive tumors identified in a study by Esserman et al.<sup>272</sup> may be examples of tumors belonging to developmental path 3 and may not require standard treatment.



This model gives a simplified picture of breast tumor progression and there may exist progression paths that are more complicated than portrayed here. For instance, some tumors may develop slowly to begin with and later gain aggressive features accelerating progression. HER2-enriched tumors, for instance, are overrepresented at the DCIS stage indicating that HER2 does not facilitate invasion. However, once HER2-enriched tumors have become invasive they are known to be quite aggressive which would “kink” the tumor progression path. Other events may also divert the path of progression, such as enhanced immune response (inhibiting tumor growth) or increased vascularization (promoting tumor progression). It is also important to acknowledge that tumors may become invasive prior to being detectable by mammography.

During breast tumor progression, if the assumption that cell of origin dictates the subtype is correct, one would expect that the subtype to a certain extent is conserved throughout tumor progression, which supports a subtype specific approach when studying DCIS. In paper III, we saw that luminal A DCIS and IBC were largely similar, suggesting that luminal A DCIS are direct precursors to luminal A IBC. However, this does not imply that all luminal A DCIS will become invasive. The observed differences between basal-like DCIS and IBC might suggest that basal-like invasive tumors (especially those that are *core basal*) develop so rapidly that most tumors have become invasive before they are detected and surgically removed (i.e. they follow progression path 1 (Figure 9)). This is supported by the observation that basal-like IBCs are less likely to have synchronous DCIS compared to luminal IBCs<sup>273</sup>. Other studies have also shown that the incidence of DCIS tumors with core basal features is low<sup>101,274</sup>. Interestingly, microglandular adenosis, a rare breast lesion considered as non-malignant, has been proposed as a precursor to high-grade IBC, possibly corresponding to the core basal tumors in our cohort<sup>275,276</sup>. The basal-like DCIS in our study resembled to a larger extent the *non-core-basal* invasive tumors, indicating that DCIS of basal-like subtype may belong to the same developmental path as non-core basal invasive tumors.

In paper III, we found hypermethylation of clustered protocadherin genes (cPCDH) in basal-like IBC compared to basal-like DCIS. cPCHDs are cell-cell adhesion molecules especially important for self-avoidance in neuronal dendrites<sup>277</sup>. During EMT, loss of the usual intraepithelial cell-cell adhesion is one of the changes that enable epithelial tumor cells to invade surrounding tissue<sup>278,279</sup>. Not much is known about the role of cPCDHs in cancer, but striking similarities have been observed between neuronal dendrites and the invadopodia of cancerous cells<sup>280</sup>. Repression of cPCDH expression through hypermethylation has been shown to occur in breast cancer and other cancer types<sup>279</sup>. It could be

hypothesized that silencing of protocadherin expression in basal-like IBC enables tumor cells to detach and migrate through either single-cell or multicellular streaming so that the DCIS architecture is rapidly lost after BM has been breached. This may explain why basal-like IBC less frequently exhibit synchronous DCIS compared to other subtypes.

Low expression of adhesion molecules is a characteristic feature of the proposed claudin-low breast cancer subtype. These tumors have low expression of several claudin genes involved in cell-cell adhesion, and high expression of genes involved in EMT, resulting in a mesenchymal-like phenotype<sup>186</sup>. In paper II, half of the murine mammary tumors that occurred, were of subtypes that resemble the human claudin-low subtype<sup>211</sup>. Even though loss of cell-cell adhesion contributes to tumor cell migration, it does not explain the mechanism of BM degradation. In paper II, due to the highly heterogeneous nature of the tumor model and the very rapid tumor development, we were not able to sample tumors at the MIN stage (the stage corresponding to human DCIS) so all tumors in this cohort were invasive at the time of sampling. However, it would be very interesting to characterize the claudin-low-like tumors from this model also at the MIN stage and throughout the invasion process. For this, we would need a predictable manner of obtaining claudin-low-like tumors. Currently, no homogeneous claudin-low mouse breast tumor models exist, however serially transplanting known claudin-low-like tumors from the MPA/DMBA model as allografts could be an option. Using the mouse mammary intraductal method (MIND)<sup>202</sup> (intraductal injection of tumor cells) while sampling at several time points throughout tumor development, we would be able to study progression from the in situ to invasive stage. It would also be of interest to explore whether the claudin-low phenotype is maintained throughout tumor progression and through different transplant generations. We have previously shown that GLI1-induced transgenic mammary gland tumors maintain molecular features through serial transplantation<sup>281</sup>. Importantly, MPA/DMBA mammary gland tumor induction and serial transplantation could (and should) be performed in immunocompetent mice, since immune processes are important in the claudin-low subtype<sup>174</sup>.

### **The role of the microenvironment in breast tumor progression**

During breast tumor progression, the processes in the microenvironment surrounding the tumor may play an equally important role to those occurring in the tumor cells<sup>140</sup>. In paper I, II and III, we have explored aspects of the microenvironment at different stages of tumor progression. In paper I, we looked at the association between mammographic density and gene expression in normal breasts over

time and we explored the microenvironment of normal breast tissue samples using a molecular subtyping method assigning each sample to an *active* or *inactive* subtype. These microenvironment subtypes were developed for characterization of tumor-adjacent normal tissue, and has previously not been explored in normal breast tissue without malignant disease<sup>228,229</sup>. The normal breast samples of the active subtype showed features similar to those of claudin-low breast cancer, such as low expression of genes involved in cell-cell adhesion and high expression of EMT-related genes. We also found that samples of the active subtype showed a wound-healing phenotype with higher expression of fibrosis-related genes and higher activation of the TGF $\beta$  pathway compared to the samples of the inactive subtype. The significance of these findings is unclear as we did not find any association between mammographic density and microenvironment subtypes. Also, microenvironment subtypes were not consistent between time points. This may be explained by intra-mammary heterogeneity, i.e. that the specific location of a biopsy plays a role, or that there are dynamic changes between the subtypes over time, for instance due to hormonal influences (such as menopause). We could not, however, identify any association between microenvironment subtype and menopause status.

The immune microenvironment is a highly important part of the tumor surroundings<sup>136</sup>, and in paper II, we found that the claudin-low-like tumors of both the murine and human cohort (Metabric), showed high degree of immune infiltration compared to other subtypes. These tumors also showed an immunosuppressive phenotype with high expression of genes such as Cd274 (encoding PD-L1) and Ptgs2 (encoding COX-2) compared to the other tumors. High immune infiltration in tumors may affect tumor subtyping since gene expression signal from the immune cells will be mixed with the signal from the tumor cells themselves. This may be particularly true for the claudin-low subtype because immune cell contribution affects the subtyping results substantially, possibly masking the “true” claudin-low phenotype. Since there is high correlation between claudin-low features and immune infiltration, it could be of benefit to improve the claudin-low subtyping methods by disentangling the immune signature from the claudin-low signature. This could lead to a more accurate classification of claudin-low tumors and improve understanding of how this subtype relates to the PAM50 subtypes.

In paper III, we found subtype specific differences in the immune response between DCIS and IBC. Our findings indicate that immune cell infiltration is similar in DCIS and IBC in ER-negative tumors, while in luminal A tumors, the DCIS lesions are less immunogenic than IBC. The immune scoring method used in this study is a crude estimate of the total immune infiltration in the tumor and does not discriminate between different immune responses. Immune cell composition has been shown to be different

## DISCUSSION

---

between DCIS and IBC<sup>151</sup>, and future studies should be instigated to explore further the subtype specific differences in immune cell composition in DCIS tumors, including the dual roles of the immune system as both pro- and antitumorigenic.

## CONCLUSIONS AND FUTURE PERSPECTIVES

In a strict histopathological sense, ductal carcinoma in situ should be considered malignant; the tumor cells have the appearance of carcinoma cells, and they even share many genomic aberrations with invasive breast cancer. However, the enclosed location of the tumor cells inside the mammary ducts restricts the tumor cells from performing mischief until invasion occurs, so in that sense, DCIS should be considered a benign disease. For this reason, there is an ongoing debate whether or not DCIS should be called a cancer<sup>282,283</sup>. Patients diagnosed with DCIS certainly have an increased risk of developing invasive breast cancer, but this risk is low. There is an unmet need for more personalized treatment of DCIS, and for low-risk lesions, this would entail reducing treatment or initiating active surveillance instead of standard treatment<sup>109,110,284</sup>. Currently, treatment of DCIS is following a *just-in-case* philosophy, which most likely leads to overtreatment of low-risk lesions. The ultimate goal for treatment of DCIS would be to identify those lesions that are low risk, and by active surveillance be able to detect any changes that instigate treatment. Decreased treatment of low-risk DCIS would have multiple benefits; physical, psychological and economical. In this context, semantics is important, and for a physician, it would be easier to convince a patient that active surveillance is the best treatment option if the word cancer was avoided. Even *pre-invasive* and *precursor* lesions, commonly used terms for DCIS, hold a promise of an imminent invasion and could be misleading.

“Overtreating people who are not at risk of death does not improve the lives of those at highest risk”

– Laura Esserman, *BMJ*, 2019<sup>283</sup>

The main challenge when managing DCIS is to identify the low-risk lesions from high risk. Currently, no reliable low- or high-risk biomarkers exist. The reason for this may lie in the lack of subtype stratification in previous studies. In this thesis, I have explored molecular subtyping in murine and human tumors and explored the heterogeneity of normal and tumor tissues throughout breast tumor progression. Particularly intriguing was the large difference that I identified between DCIS and IBC of basal-like subtype in gene expression, copy number and methylation data. These findings may have significance for how basal-like DCIS should be interpreted and handled in the clinic.

Future studies of subtype specific breast tumor progression should be performed in cohorts where clinical follow-up data such as recurrences and mortality is available. The number of tumors in the cohort needs to be sufficient to account for heterogeneity in tumor progression. Since survival is excellent and recurrence rate is low for DCIS there is a need for long-term follow-up data. New

## CONCLUSIONS AND FUTURE PERSPECTIVES

---

technology has made it possible to extract DNA and RNA from formalin fixed tumor tissue to generate genomic data. FFPE from breast cancer tumors is commonly collected and stored after routine histopathological assessment, and such cohorts may be very valuable for studying DCIS. In Norway, a well organized Cancer Registry makes such studies possible. Prospective studies examining the effect of active surveillance in low-risk DCIS patients would also most likely result in very valuable data as this would contribute with increased knowledge of the natural development of DCIS without compromising patient health.

## REFERENCES

1. Weinberg, R. A. *The biology of cancer*. (Garland Science, 2013).
2. Sudhakar, A. History of Cancer, Ancient and Modern Treatment Methods. *J. Cancer Sci. Ther.* **01**, i–iv (2009).
3. CT scans of Egyptian mummies reveal oldest known cases of breast cancer and multiple myeloma. *ScienceDaily* (2017). Available at: <https://www.sciencedaily.com/releases/2017/12/171214101215.htm>. (Accessed: 29th July 2019)
4. Bianucci, R., Perciaccante, A., Charlier, P., Appenzeller, O. & Lippi, D. Earliest evidence of malignant breast cancer in Renaissance paintings. *Lancet. Oncol.* **19**, 166–167 (2018).
5. Ferlay, J. *et al.* Global Cancer Observatory: Cancer Today. *International Agency for Research on Cancer* (2018). Available at: <https://gco.iarc.fr/today>. (Accessed: 17th August 2019)
6. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
7. Cancer Registry of Norway. *Cancer incidence, mortality, survival and prevalence in Norway Cancer in Norway 2017*. (2018).
8. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
9. Beadle, G. W. & Tatum, E. L. Genetic Control of Biochemical Reactions in *Neurospora*. *Proc. Natl. Acad. Sci. U. S. A.* **27**, 499–506 (1941).
10. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
11. Horgan, R. P. & Kenny, L. C. 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *Obstet. Gynaecol.* **13**, 189–195 (2011).
12. Maier, T., Güell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* **583**, 3966–3973 (2009).
13. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
14. Boveri, T. Concerning the Origin of Malignant Tumours. Translated and annotated by Henry Harris. *J. Cell Sci.* **121**, 1–84 (2008).
15. Boveri, T. *Zur Frage der Entstehung maligner Tumoren, 1914*. (Gustav Fischer Verlag, 1929).
16. American Cancer Society. Oncogenes and tumor suppressor genes. (2014). Available at: <https://www.cancer.org/cancer/cancer-causes/genetics/genes-and-cancer/oncogenes-tumor-suppressor-genes.html>. (Accessed: 29th July 2019)
17. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).
18. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
19. Dulbecco, R. A turning point in cancer research: sequencing the human genome. *Science* **231**, 1055–1056 (1986).
20. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
21. Wheeler, D. A. & Wang, L. From human genome to cancer genome: The first decade. *Genome Research* **23**, 1054–1062 (2013).
22. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2014).
23. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–33 (2013).
24. Berman, B. P. *et al.* Exploring the cancer methylome. *BMC Proc.* **6**, O24 (2012).
25. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **340**, 1546–1558 (2013).
26. Biesecker, L. G. Hypothesis-generating research and predictive medicine. *Genome Res.* **23**, 1051 (2013).
27. Werner, H. M. J., Mills, G. B. & Ram, P. T. Cancer systems biology: A peek into the future of patient care? *Nat. Rev. Clin. Oncol.* **11**, 167–176 (2014).
28. Medina, D. The Mammary Gland: A Unique Organ for the Study of Development and Tumorigenesis. *J. Mammary Gland Biol. Neoplasia* **1**, (1996).
29. Cooper, A. P. *On the anatomy of the breast, volume I*. (1840).
30. Hassiotou, F. & Geddes, D. Anatomy of the human mammary gland: Current status of knowledge. *Clin. Anat.* **26**, 29–48 (2013).
31. McGuire, K. P. Breast Anatomy and Physiology. in *Breast Disease: Diagnosis and Pathology* **1**, 1–14 (Springer International Publishing, 2015).
32. Di Muzio, B. & Pacifici, S. Terminal ductal lobular unit. *Radiopedia.org* Available at: <https://radiopaedia.org/articles/terminal-ductal-lobular-unit>. (Accessed: 14th January 2019)
33. Gudjonsson, T. *et al.* Isolation, immortalization, and characterization of a human breast epithelial cell line with stem cell properties. *Genes Dev.* **16**, 693–706 (2002).
34. Visvader, J. E. & Stingl, J. Mammary stem cells and the differentiation hierarchy: Current status and perspectives. *Genes and Development* **28**, 1143–

## REFERENCES

- 1158 (2014).
35. Macias, H. & Hinck, L. Mammary gland development. *Wiley Interdisciplinary Reviews: Developmental Biology* **1**, 533–557 (2012).
  36. Key, T. J., Verkasalo, P. K. & Banks, E. Epidemiology of breast cancer. *Lancet Oncol.* **2**, 133–140 (2001).
  37. May, C. D. *et al.* Epithelial-mesenchymal transition and cancer stem cells: a dangerously dynamic duo in breast cancer progression. *Breast Cancer Res.* **13**, 202 (2011).
  38. Li, C. I., Anderson, B. O., Daling, J. R. & Moe, R. E. Trends in incidence rates of invasive lobular and ductal breast carcinoma. *JAMA* **289**, 1421–1424 (2003).
  39. Allred, D. C. & Mohsin, S. K. Biological Features of Premalignant Disease in the Human Breast. *J. Mammary Gland Biol. Neoplasia* **5**, 351–364 (2000).
  40. Broders, A. C. Carcinoma in situ contrasted with benign penetrating epithelium. *J. Am. Med. Assoc.* **99**, 1670–1674 (1932).
  41. Bloodgood, J. C. Border-line breast tumors. *Ann. Surg.* **93**, 235–249 (1931).
  42. Gibson, C. L. II. The Rational Treatment of Non-malignant and Border-line Tumors of the Breast. *Ann. Surg.* **49**, 478–486 (1909).
  43. Gorringer, K. L. & Fox, S. B. Ductal Carcinoma In Situ Biology, Biomarkers, and Diagnosis. *Front. Oncol.* **7**, 248 (2017).
  44. Virnig, B. A., Tuttle, T. M., Shamliyan, T. & Kane, R. L. Ductal Carcinoma In Situ of the Breast: A Systematic Review of Incidence, Treatment, and Outcomes. *JNCI J. Natl. Cancer Inst.* **102**, 170–178 (2010).
  45. Cancer Registry of Norway. *Cancer incidence, mortality, survival and prevalence in Norway; Celebrating 20 years of organised mammographic screening.* (2017).
  46. Olsen, O. & Gøtzsche, P. C. Cochrane review on screening for breast cancer with mammography. *Lancet* **358**, 1340–1342 (2001).
  47. Zahl, P. H., Holme, O. & Loberg, M. Norwegian mammography screening - numerous self-contradictions in the evaluation. **136**, 1616–1618 (2016).
  48. U.S. Preventive Services Task Force. Screening for breast cancer: recommendations and rationale. *Ann. Intern. Med.* **137**, 344–346 (2002).
  49. Marmot, M. G. *et al.* The benefits and harms of breast cancer screening: an independent review. *Br. J. Cancer* **108**, 2205–2240 (2013).
  50. Fletcher, S. W. & Elmore, J. G. Mammographic screening for breast cancer. *N. Engl. J. Med.* **348**, 1672–1680 (2003).
  51. Welch, H. G. & Black, W. C. Overdiagnosis in Cancer. *JNCI J. Natl. Cancer Inst.* **102**, 605–613 (2010).
  52. Welch, H. G. & Black, W. C. Using autopsy series to estimate the disease ‘reservoir’ for ductal carcinoma in situ of the breast: how much more breast cancer can we find? *Ann. Intern. Med.* **127**, 1023–1028 (1997).
  53. Nielsen, M., Thomsen, J. L., Primdahl, S., Dyreborg, U. & Andersen, J. a. Breast cancer and atypia among young and middle-aged women: a study of 110 medicolegal autopsies. *Br. J. Cancer* **56**, 814–819 (1987).
  54. Erbas, B., Provenzano, E., Armes, J. & Gertig, D. The natural history of ductal carcinoma in situ of the breast: A review. *Breast Cancer Research and Treatment* **97**, 135–144 (2006).
  55. Thomas, E. T. *et al.* Prevalence of incidental breast cancer and precursor lesions in autopsy studies: a systematic review and meta-analysis. *BMC Cancer* **17**, 808 (2017).
  56. Punglia, R. S. *et al.* Epidemiology, Biology, Treatment, and Prevention of Ductal Carcinoma In Situ (DCIS). *JNCI Cancer Spectr.* **2**, (2018).
  57. Norsk Bryst Cancer Gruppe (NBCG). *Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft.* (2016).
  58. Sagara, Y. *et al.* Survival Benefit of Breast Surgery for Low-Grade Ductal Carcinoma In Situ. *JAMA Surg.* **150**, 739 (2015).
  59. Leonard, G. D. & Swain, S. M. Ductal carcinoma in situ, complexities and challenges. *JNCI J. Natl. Cancer Inst.* **96**, 906–920 (2004).
  60. Lee, L. A. *et al.* Breast cancer-specific mortality after invasive local recurrence in patients with ductal carcinoma-in-situ of the breast. *Am. J. Surg.* **192**, 416–419 (2006).
  61. Early Breast Cancer Trialists’ Collaborative Group (EBCTCG) *et al.* Overview of the Randomized Trials of Radiotherapy in Ductal Carcinoma In Situ of the Breast. *JNCI Monogr.* **2010**, 162–177 (2010).
  62. Bijker, N. *et al.* Breast-Conserving Treatment With or Without Radiotherapy in Ductal Carcinoma-In-Situ: Ten-Year Results of European Organisation for Research and Treatment of Cancer Randomized Phase III Trial 10853. *J. Clin. Oncol.* **24**, 3381–3387 (2006).
  63. Donker, M. *et al.* Breast-Conserving Treatment With or Without Radiotherapy in Ductal Carcinoma In Situ: 15-Year Recurrence Rates and Outcome After a Recurrence, From the EORTC 10853 Randomized Phase III Trial. *J. Clin. Oncol.* **31**, 4054–4059 (2013).



64. Hanna, W. M. *et al.* Ductal carcinoma in situ of the breast: an update for the pathologist in the era of individualized risk assessment and tailored therapies. *Mod. Pathol.* **32**, 896–915 (2019).
65. Nichols, H. B. *et al.* Tamoxifen Initiation After Ductal Carcinoma In Situ. *Oncologist* **21**, 134–40 (2016).
66. Fisher, B. *et al.* Prevention of invasive breast cancer in women with ductal carcinoma in situ: an update of the National Surgical Adjuvant Breast and Bowel Project experience. *Semin. Oncol.* **28**, 400–418 (2001).
67. Narod, S. A., Iqbal, J., Giannakeas, V., Sopik, V. & Sun, P. Breast Cancer Mortality After a Diagnosis of Ductal Carcinoma In Situ. *JAMA Oncol.* **1**, 888–896 (2015).
68. Falk, R. S., Hofvind, S., Skaane, P. & Haldorsen, T. Second events following ductal carcinoma in situ of the breast: a register-based cohort study. *Breast Cancer Res. Treat.* **129**, 929–938 (2011).
69. Makki, J. Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. *Clin. Med. Insights. Pathol.* **8**, 23–31 (2015).
70. Wallis, M. G. *et al.* The effect of DCIS grade on rate, type and time to recurrence after 15 years of follow-up of screen-detected DCIS. *Br. J. Cancer* **106**, 1611–1617 (2012).
71. Sanders, M. E., Schuyler, P. A., Dupont, W. D. & Page, D. L. The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. *Cancer* **103**, 2481–2484 (2005).
72. Collins, L. C. *et al.* Outcome of patients with ductal carcinoma in situ untreated after diagnostic biopsy: results from the Nurses' Health Study. *Cancer* **103**, 1778–1784 (2005).
73. Eusebi, V. *et al.* Long-term follow-up of in situ carcinoma of the breast. *Semin. Diagn. Pathol.* **11**, 223–235 (1994).
74. Rosen, P., Snyder, R. E., Foote, F. W. & Wallace, T. Detection of occult carcinoma in the apparently benign breast biopsy through specimen radiography. *Cancer* **26**, 944–952 (1970).
75. Sgroi, D. C. Preinvasive breast cancer. *Annu. Rev. Pathol.* **5**, 193–221 (2010).
76. Bleiweiss, I. J. Pathology of breast cancer - UpToDate. (2019). Available at: <https://www.uptodate.com/contents/pathology-of-breast-cancer>. (Accessed: 11th March 2019)
77. Allred, D. C. Ductal carcinoma in situ: terminology, classification, and natural history. *J. Natl. Cancer Inst. Monogr* **2010**, 134–138 (2010).
78. Pinder, S. E. & Ellis, I. O. The diagnosis and management of pre-invasive breast disease: ductal carcinoma in situ (DCIS) and atypical ductal hyperplasia (ADH) - current definitions and classification. *Breast Cancer Res.* **5**, 254–257 (2003).
79. Wärnberg, F., Nordgren, H., Bergh, J. & Holmberg, L. Ductal carcinoma in situ of the breast from a population-defined cohort: An evaluation of new histopathological classification systems. *Eur. J. Cancer* **35**, 714–720 (1999).
80. Holland, R. *et al.* Ductal carcinoma in situ: a proposal for a new classification. *Semin. Diagn. Pathol.* **11**, 167–180 (1994).
81. Silverstein, M. J. *et al.* Prognostic classification of breast ductal carcinoma-in-situ. *Lancet* **345**, 1154–1157 (1995).
82. Lagios, M. D., Margolin, F. R., Westdahl, P. R. & Rose, M. R. Mammographically detected duct carcinoma in situ. Frequency of local recurrence following tylectomy and prognostic effect of nuclear grade on local recurrence. *Cancer* **63**, 618–624 (1989).
83. Lakhani, S, Ellis, I, Schnitt, S, Tan, P, van de Vijver, M. WHO Classification of Tumours of the Breast, Fourth Edition. in *IARC WHO Classification of Tumours, No 4* 240 (2012).
84. Silverstein, M. J. & Lagios, M. D. Continued observation of the natural history of low-grade ductal carcinoma in situ reaffirms proclivity for local recurrence even after more than 30 years of follow-up. *Breast Dis. A Year B. Q.* **26**, 319–320 (2015).
85. Ramnani, D. Webpathology.com: A Collection of Surgical Pathology Images. Available at: <https://www.webpathology.com/>. (Accessed: 12th April 2019)
86. Azzopardi, J. G., Ahmed, A. & Millis, R. R. Problems in breast pathology. *Major Probl. Pathol.* **11**, i–xvi, 1–466 (1979).
87. Sandhu, J., Dubey, V., Makkar, M. & Suri, V. Pure primary signet ring cell carcinoma breast: A rare cytological diagnosis. *J. Cytol.* **30**, 204–206 (2013).
88. D'Alfonso, T. M., Ginter, P. S., Liu, Y.-F. & Shin, S. J. Cystic Hypersecretory (In Situ) Carcinoma of the Breast. *Am. J. Surg. Pathol.* **38**, 45–53 (2014).
89. Bombonati, A. & Sgroi, D. C. The molecular pathology of breast cancer progression. *J. Pathol.* **223**, 307–317 (2011).
90. Wellings, S. R. & Jensen, H. M. On the origin and progression of ductal carcinoma in the human breast. *JNCI J. Natl. Cancer Inst.* **50**, 1111–8 (1973).
91. Lopez-Garcia, M. a, Geyer, F. C., Lacroix-Triki, M., Marchió, C. & Reis-Filho, J. S. Breast cancer precursors revisited: molecular features and progression pathways. *Histopathology* **57**, 171–192 (2010).

## REFERENCES

92. Barnes, N. L. P., Boland, G. P., Davenport, A., Knox, W. F. & Bundred, N. J. Relationship between hormone receptor status and tumour size, grade and comedo necrosis in ductal carcinoma in situ. *Br. J. Surg.* **92**, 429–434 (2005).
93. Sarode, V. R. *et al.* A Comparative Analysis of Biomarker Expression and Molecular Subtypes of Pure Ductal Carcinoma In Situ and Invasive Breast Carcinoma by Image Analysis: Relationship of the Subtypes with Histologic Grade, Ki67, p53 Overexpression, and DNA Ploidy. *Int. J. Breast Cancer* **2011**, 1–7 (2011).
94. Han, K. *et al.* Expression of HER2neu in Ductal Carcinoma in situ is Associated with Local Recurrence. *Clin. Oncol.* **24**, 183–189 (2012).
95. Kerlikowske, K. *et al.* Biomarker expression and risk of subsequent tumors after initial ductal carcinoma in situ diagnosis. *JNCI J. Natl. Cancer Inst.* **102**, 627–637 (2010).
96. Gao, Y., Niu, Y., Wang, X., Wei, L. & Lu, S. Genetic changes at specific stages of breast cancer progression detected by comparative genomic hybridization. *J. Mol. Med.* **87**, 145–152 (2009).
97. Kietzman, W., Riegel, A. T. & Ory, V. Early-Stage Progression of Breast Cancer. in *Breast Cancer - From Biology to Medicine* (InTech, 2017). doi:10.5772/65633
98. Abba, M. C. *et al.* A Molecular Portrait of High-Grade Ductal Carcinoma In Situ. *Cancer Res.* **75**, 3980–3990 (2015).
99. Lukas, J., Niu, N. & Press, M. F. p53 mutations and expression in breast carcinoma in situ. *Am. J. Pathol.* **156**, 183–91 (2000).
100. Pang, J. M. B. *et al.* Breast ductal carcinoma in situ carry mutational driver events representative of invasive breast cancer. *Mod. Pathol.* **30**, 952–963 (2017).
101. Hannemann, J. F. J. *et al.* Classification of ductal carcinoma in situ by gene expression profiling. *Breast Cancer Res.* **8**, R61 (2006).
102. Ma, X.-J. *et al.* Gene expression profiles of human breast cancer progression. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5974–5979 (2003).
103. Seth, A. *et al.* Gene expression profiling of ductal carcinomas in situ and invasive breast tumors. *Anticancer Res.* **23**, 2043–2051 (2003).
104. Schuetz, C. S. *et al.* Progression-specific genes identified by expression profiling of matched ductal carcinomas in situ and invasive breast tumors, combining laser capture microdissection and oligonucleotide microarray analysis. *Cancer Res.* **66**, 5278–5286 (2006).
105. Lesurf, R. *et al.* Molecular Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast Cancer. *Cell Rep.* **16**, 1166–1179 (2016).
106. Youngswirth, L., Boughey, J. C. & Hwang, S. Surgery versus monitoring and endocrine therapy for low-risk DCIS: The COMET Trial. *Bulletin of The American College of Surgeons* (2017). Available at: <http://bulletin.facs.org/2017/01/surgery-versus-monitoring-and-endocrine-therapy-for-low-risk-dcis-the-comet-trial/>. (Accessed: 16th March 2019)
107. Francis, A. *et al.* Addressing overtreatment of screen detected DCIS; the LORIS trial. *Eur. J. Cancer* **51**, 2296–2303 (2015).
108. Elshof, L. E. *et al.* Feasibility of a prospective, randomised, open-label, international multicentre, phase III, non-inferiority trial to assess the safety of active surveillance for low risk ductal carcinoma in situ – The LORD study. *Eur. J. Cancer* **51**, 1497–1510 (2015).
109. Ryser, M. D. *et al.* Outcomes of Active Surveillance for Ductal Carcinoma in Situ: A Computational Risk Analysis. *JNCI J. Natl. Cancer Inst.* **108**, djv372 (2016).
110. Doke, K., Butler, S. & Mitchell, M. P. Current Therapeutic Approaches to DCIS. *J. Mammary Gland Biol. Neoplasia* **23**, 279–291 (2018).
111. Solin, L. J. *et al.* A Multigene Expression Assay to Predict Local Recurrence Risk for Ductal Carcinoma In Situ of the Breast. *JNCI J. Natl. Cancer Inst.* **105**, 701–710 (2013).
112. Rakovitch, E. *et al.* A population-based validation study of the DCIS Score predicting recurrence risk in individuals treated by breast-conserving surgery alone. *Breast Cancer Res. Treat.* **152**, 389–398 (2015).
113. Nofech-Mozes, S., Hanna, W. & Rakovitch, E. Molecular Evaluation of Breast Ductal Carcinoma in Situ with Oncotype DX DCIS. *Am. J. Pathol.* **189**, 975–980 (2019).
114. Raldow, A. C., Sher, D., Chen, A. B., Recht, A. & Punglia, R. S. Cost Effectiveness of the Oncotype DX DCIS Score for Guiding Treatment of Patients With Ductal Carcinoma In Situ. *J. Clin. Oncol.* **34**, 3963–3968 (2016).
115. Breast screening: the Sloane Project - GOV.UK. Available at: <https://www.gov.uk/guidance/breast-screening-the-sloane-project>. (Accessed: 18th March 2019)
116. Definition of invasive cancer - NCI Dictionary of Cancer Terms - National Cancer Institute. Available at: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/invasive-cancer>. (Accessed: 25th

- March 2019)
117. Micalizzi, D. S., Farabaugh, S. M. & Ford, H. L. Epithelial-Mesenchymal Transition in Cancer: Parallels Between Normal Development and Tumor Progression. *J. Mammary Gland Biol. Neoplasia* **15**, 117–134 (2010).
  118. Krakhmal, N. V., Zavyalova, M. V., Denisov, E. V., Vtorushin, S. V & Perelmuter, V. M. Cancer Invasion: Patterns and Mechanisms. *Acta Naturae* **7**, 17–28 (2015).
  119. Mcsherry, E. A., Donatello, S., Hopkins, A. M. & McDonnell, S. Molecular basis of invasion in breast cancer. *Cell. Mol. Life Sci.* **64**, 3201–3218 (2007).
  120. Badruddoja, M. Ductal Carcinoma In Situ of the Breast: A Surgical Perspective. *Int. J. Surg. Oncol.* **2012**, (2012).
  121. Stomper, P. C., Geradts, J., Edge, S. B. & Levine, E. G. Mammographic predictors of the presence and size of invasive carcinomas associated with malignant microcalcification lesions without a mass. *AJR. Am. J. Roentgenol.* **181**, 1679–84 (2003).
  122. Rakha, E. A. *et al.* Invasion in breast lesions: the role of the epithelial-stroma barrier. *Histopathology* **72**, 1075–1083 (2018).
  123. Sue, G. R., Lannin, D. R., Killelea, B. & Chagpar, A. B. Predictors of microinvasion and its prognostic role in ductal carcinoma in situ. *Am. J. Surg.* **206**, 478–481 (2013).
  124. Champion, C. D. *et al.* DCIS with Microinvasion: Is It In Situ or Invasive Disease? *Ann. Surg. Oncol.* (2019). doi:10.1245/s10434-019-07556-9
  125. Giampieri, S., Pinner, S. & Sahai, E. Intravital Imaging Illuminates Transforming Growth Factor  $\beta$  Signaling Switches during Metastasis. *Cancer Res.* **70**, 3435–3439 (2010).
  126. Clark, A. G. & Vignjevic, D. M. Modes of cancer cell invasion and the role of the microenvironment. *Curr. Opin. Cell Biol.* **36**, 13–22 (2015).
  127. Friedl, P. & Alexander, S. Cancer Invasion and the Microenvironment: Plasticity and Reciprocity. *Cell* **147**, 992–1009 (2011).
  128. Christiansen, J. J. & Rajasekaran, A. K. Reassessing epithelial to mesenchymal transition as a prerequisite for carcinoma invasion and metastasis. *Cancer Res.* **66**, 8319–8326 (2006).
  129. Definition of metastasis - NCI Dictionary of Cancer Terms - National Cancer Institute. Available at: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/metastasis>. (Accessed: 25th March 2019)
  130. Weigelt, B., Peterse, J. L. & van't Veer, L. J. Breast cancer metastasis: markers and models. *Nat. Rev. Cancer* **5**, 591–602 (2005).
  131. Fidler, I. J. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat. Rev. Cancer* **3**, 453–458 (2003).
  132. Paget, S. The distribution of secondary growths in cancer of the breast. *Lancet* **133**, 571–573 (1889).
  133. Narod, S. A. & Sopik, V. Is invasion a necessary step for metastases in breast cancer? *Breast Cancer Research and Treatment* **169**, 9–23 (2018).
  134. Narod, S. A., Ahmed, H. & Sopik, V. Wherein the authors attempt to minimize the confusion generated by their study 'Breast cancer mortality after a diagnosis of ductal carcinoma in situ' by several commentators who disagree with them and a few who don't: a qualitative study. *Curr. Oncol.* **24**, e255–e260 (2017).
  135. Osako, T., Iwase, T., Kimura, K., Horii, R. & Akiyama, F. Detection of occult invasion in ductal carcinoma *in situ* of the breast with sentinel node metastasis. *Cancer Sci.* **104**, 453–457 (2013).
  136. Bissell, M. J. & Hines, W. C. Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. *Nat. Med.* **17**, 320–9 (2011).
  137. Definition of tumor microenvironment - NCI Dictionary of Cancer Terms - National Cancer Institute. Available at: <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/tumor-microenvironment>. (Accessed: 25th March 2019)
  138. Ursini-Siegel, J. & Park, M. It takes two to tango in the microenvironment! *Breast Cancer Res.* **15**, 102 (2013).
  139. Ma, X.-J., Dahiya, S., Richardson, E., Erlander, M. & Sgroi, D. C. Gene expression profiling of the tumor microenvironment during breast cancer progression. *Breast Cancer Res.* **11**, R7 (2009).
  140. Nelson, A. C., Machado, H. L. & Schwertfeger, K. L. Breaking through to the Other Side: Microenvironment Contributions to DCIS Initiation and Progression. *J. Mammary Gland Biol. Neoplasia* **23**, 207–221 (2018).
  141. McCormack, V. A. & dos Santos Silva, I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiol. Biomarkers Prev.* **15**, 1159–69 (2006).
  142. Boyd, N. *et al.* The origins of breast cancer associated with mammographic density: a testable biological hypothesis. *Breast Cancer Res.* **20**, 17 (2018).
  143. Frantz, C., Stewart, K. M. & Weaver, V. M. The extracellular matrix at a glance. *J. Cell Sci.* **123**, 4195–200 (2010).
  144. Shiga, K. *et al.* Cancer-Associated Fibroblasts: Their

## REFERENCES

- Characteristics and Their Roles in Tumor Growth. *Cancers (Basel)*. **7**, 2443–58 (2015).
145. Kuperwasser, C. *et al.* Reconstruction of functionally normal and malignant human breast tissues in mice. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4966–4971 (2004).
  146. Dvorak, H. F. Tumors: Wounds that do not heal. *N. Engl. J. Med.* **315**, 1650–1659 (1986).
  147. Sugimoto, H., Mundel, T. M., Kieran, M. W. & Kalluri, R. Identification of fibroblast heterogeneity in the tumor microenvironment. *Cancer Biol. Ther.* **5**, 1640–6 (2006).
  148. Provenzano, P. P. *et al.* Collagen reorganization at the tumor-stromal interface facilitates local invasion. *BMC Med.* **4**, 38 (2006).
  149. Vong, S. & Kalluri, R. The Role of Stromal Myofibroblast and Extracellular Matrix in Tumor Angiogenesis. *Genes Cancer* **2**, 1139–1145 (2011).
  150. Morgan, M. M. *et al.* Mammary fibroblasts reduce apoptosis and speed estrogen-induced hyperplasia in an organotypic MCF7-derived duct model. *Sci. Rep.* **8**, 7139 (2018).
  151. Gil Del Alcazar, C. R. *et al.* Immune Escape in Breast Cancer During *In Situ* to Invasive Carcinoma Transition. *Cancer Discov.* **7**, 1098–1115 (2017).
  152. Agahozo, M. C., Hammerl, D., Debets, R., Kok, M. & van Deurzen, C. H. M. Tumor-infiltrating lymphocytes and ductal carcinoma in situ of the breast: friends or foes? *Mod. Pathol.* **31**, 1012–1025 (2018).
  153. Hendry, S. *et al.* Relationship of the Breast Ductal Carcinoma *In Situ* Immune Microenvironment with Clinicopathological and Genetic Features. *Clin. Cancer Res.* **23**, 5210–5217 (2017).
  154. Disis, M. L. & Park, K. H. Immunomodulation of Breast Cancer via Tumor Antigen Specific Th1. *Cancer Res. Treat.* **41**, 117–121 (2009).
  155. DeNardo, D. G. *et al.* CD4(+) T cells regulate pulmonary metastasis of mammary carcinomas by enhancing protumor properties of macrophages. *Cancer Cell* **16**, 91–102 (2009).
  156. Campbell, M. J. *et al.* Characterizing the immune microenvironment in high-risk ductal carcinoma in situ of the breast. *Breast Cancer Res. Treat.* **161**, 17–28 (2017).
  157. Thompson, E. *et al.* The immune microenvironment of breast ductal carcinoma in situ. *Mod. Pathol.* **29**, 249–258 (2016).
  158. Lowenfeld, L. *et al.* Dendritic Cell Vaccination Enhances Immune Responses and Induces Regression of HER2<sup>pos</sup> DCIS Independent of Route: Results of Randomized Selection Design Trial. *Clin. Cancer Res.* **23**, 2961–2971 (2017).
  159. Pandey, P. R., Saidou, J. & Watabe, K. Role of myoepithelial cells in breast tumor progression. *Front. Biosci.* **15**, 226–36 (2010).
  160. Sternlicht, M. D., Kedeshian, P., Shao, Z. M., Safarians, S. & Barsky, S. H. The human myoepithelial cell is a natural tumor suppressor. *Clin. Cancer Res.* **3**, 1949–58 (1997).
  161. Jones, J., Shaw, J., Pringle, J. & Walker, R. Primary breast myoepithelial cells exert an invasion-suppressor effect on breast cancer cells via paracrine down-regulation of MMP expression in fibroblasts and tumour cells. *J. Pathol.* **201**, 562–572 (2003).
  162. Nguyen, M. *et al.* The human myoepithelial cell displays a multifaceted anti-angiogenic phenotype. *Oncogene* **19**, 3449–3459 (2000).
  163. Allen, M. D. *et al.* Altered Microenvironment Promotes Progression of Preinvasive Breast Cancer: Myoepithelial Expression of  $\alpha v \beta 6$  Integrin in DCIS Identifies High-risk Patients and Predicts Recurrence. *Clin. Cancer Res.* **20**, 344–357 (2014).
  164. Man, Y.-G. & Sang, Q.-X. A. The significance of focal myoepithelial cell layer disruptions in human breast tumor invasion: a paradigm shift from the “protease-centered” hypothesis. *Exp. Cell Res.* **301**, 103–118 (2004).
  165. Macmillan, C. D., Chambers, A. F. & Tuck, A. B. Breast Cancer Metastasis and Drug Resistance. in *Breast Cancer Metastasis and Drug Resistance* (ed. Ahmad, A.) 143–159 (Springer New York, 2013). doi:10.1007/978-1-4614-5647-6
  166. Zubeldia-Plazaola, A. *et al.* Glucocorticoids promote transition of ductal carcinoma in situ to invasive ductal carcinoma by inducing myoepithelial cell apoptosis. *Breast Cancer Res.* **20**, 65 (2018).
  167. Russnes, H. G., Lingjærde, O. C., Børresen-Dale, A.-L. & Caldas, C. Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. *Am. J. Pathol.* **187**, 2152–2162 (2017).
  168. Goldhirsch, A. *et al.* Strategies for subtypes - dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Ann. Oncol.* **22**, 1736–1747 (2011).
  169. Coates, A. S. *et al.* Tailoring therapies-improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann. Oncol.* **26**, 1533–1546 (2015).
  170. Sørli, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10869–10874 (2001).

171. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
172. Sørlie, T. Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. *Eur. J. Cancer* **40**, 2667–2675 (2004).
173. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat. Med.* **15**, 907–913 (2009).
174. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **12**, R68 (2010).
175. Prat, A. & Perou, C. M. Mammary development meets cancer genomics. *Nat. Med.* **15**, 842 (2009).
176. Parker, J. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160–1167 (2009).
177. Wallden, B. *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genomics* **8**, 54 (2015).
178. Nielsen, T. *et al.* Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. *BMC Cancer* **14**, 177 (2014).
179. Ohnstad, H. O. *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 Res.* (2017). doi:10.1186/s13058-017-0911-9
180. Prat, A. *et al.* Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast* **24**, S26–S35 (2015).
181. Fleischer, T. *et al.* DNA methylation signature (SAM40) identifies subgroups of the Luminal A breast cancer samples with distinct survival. *Oncotarget* **8**, 1074–1082 (2017).
182. Cheang, M. C. U. *et al.* Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin. Cancer Res.* **14**, 1368–1376 (2008).
183. Nielsen, T. O. *et al.* Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin. Cancer Res.* **10**, 5367–5374 (2004).
184. Badowska-Kozakiewicz, A. M. & Budzik, M. P. Immunohistochemical characteristics of basal-like breast cancer. *Contemp. Oncol. (Poznan, Poland)* **20**, 436–443 (2016).
185. Herschkowitz, J. I. *et al.* Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol.* **8**, R76 (2007).
186. Sabatier, R. *et al.* Claudin-low breast cancers: Clinical, pathological, molecular and prognostic characterization. *Mol. Cancer* **13**, 228 (2014).
187. Dias, K. *et al.* Claudin-low breast cancer; clinical & pathological characteristics. *PLoS One* **12**, e0168669 (2017).
188. Yu, K.-D. *et al.* Different Distribution of Breast Cancer Subtypes in Breast Ductal Carcinoma in situ (DCIS), DCIS with Microinvasion, and DCIS with Invasion Component. *Ann. Surg. Oncol.* **18**, 1342–1348 (2011).
189. Tamimi, R. M. *et al.* Comparison of molecular phenotypes of ductal carcinoma in situ and invasive breast cancer. *Breast Cancer Res.* **10**, R67 (2008).
190. Meijnen, P., Peterse, J. L., Antonini, N., Rutgers, E. J. T. & van de Vijver, M. J. Immunohistochemical categorisation of ductal carcinoma in situ of the breast. *Br. J. Cancer* **98**, 137–142 (2008).
191. Allred, D. C. *et al.* Ductal carcinoma in situ and the emergence of diversity during breast cancer evolution. *Clin. Cancer Res.* **14**, 370–378 (2008).
192. Muggerud, A. A. *et al.* Molecular diversity in ductal carcinoma in situ (DCIS) and early invasive breast cancer. *Mol. Oncol.* **4**, 357–368 (2010).
193. Manning, H. C., Buck, J. R. & Cook, R. S. Mouse models of breast cancer: Platforms for discovering precision imaging diagnostics and future cancer medicine. *J. Nucl. Med.* **57**, 60S–68S (2016).
194. Dontu, G. & Ince, T. A. Of mice and women: a comparative tissue biology perspective of breast stem cells and differentiation. *J. Mammary Gland Biol. Neoplasia* **20**, 51–62 (2015).
195. Visvader, J. E. Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. *Genes and Development* **23**, 2563–2577 (2009).
196. Paine, I. S. & Lewis, M. T. The Terminal End Bud: the Little Engine that Could. *J. Mammary Gland Biol. Neoplasia* **22**, 93–108 (2017).
197. Cardiff, R. D. & Wellings, S. R. The Comparative Pathology of Human and Mouse Mammary Glands. *J. Mammary Gland Biol. Neoplasia* **4**, 105–122 (1999).
198. Chaffin, C. L. & VandeVoort, C. A. Follicle growth, ovulation, and luteal formation in primates and rodents: A comparative perspective. *Exp. Biol. Med.* **238**, 539–548 (2013).
199. Swiatnicki, M. R. & Andrechek, E. R. How to Choose a Mouse Model of Breast Cancer, a Genomic Perspective. *J. Mammary Gland Biol. Neoplasia* **1–13** (2019). doi:10.1007/s10911-019-09433-3
200. Cardiff, R. D. *et al.* The mammary pathology of genetically engineered mice: the consensus report and recommendations from the Annapolis meeting. *Oncogene* **19**, 968–988 (2000).

## REFERENCES

201. Richmond, A. & Su, Y. Mouse xenograft models vs GEM models for human cancer therapeutics. *Dis. Model. Mech.* **1**, 78–82 (2008).
202. Kittrell, F. *et al.* Mouse Mammary Intraductal (MIND) Method for Transplantation of Patient Derived Primary DCIS Cells and Cell Lines. *Bio-protocol* **6**, e1744 (2016).
203. Behbod, F. *et al.* An intraductal human-in-mouse transplantation model mimics the subtypes of ductal carcinoma in situ. *Breast Cancer Res.* **11**, R66 (2009).
204. Sflomos, G. *et al.* A Preclinical Model for ER $\alpha$ -Positive Breast Cancer Points to the Epithelial Microenvironment as Determinant of Luminal Phenotype and Hormone Response. *Cancer Cell* **29**, 407–422 (2016).
205. Behbod, F., Gomes, A. M. & Machado, H. L. Modeling Human Ductal Carcinoma In Situ in the Mouse. *J. Mammary Gland Biol. Neoplasia* **23**, 269–278 (2018).
206. Borowsky, A. D. Choosing a mouse model: experimental biology in context—the utility and limitations of mouse models of breast cancer. *Cold Spring Harb. Perspect. Biol.* **3**, a009670 (2011).
207. Aldaz, C. M., Liao, Q. Y., LaBate, M. & Johnston, D. A. Medroxyprogesterone acetate accelerates the development and increases the incidence of mouse mammary tumors induced by dimethylbenzanthracene. *Carcinogenesis* **17**, 2069–2072 (1996).
208. Liu, Y. *et al.* Mammalian models of chemically induced primary malignancies exploitable for imaging-based preclinical theragnostic research. *Quant. Imaging Med. Surg.* **5**, 708–729 (2015).
209. Abba, M. C. *et al.* DMBA induced mouse mammary tumors display high incidence of activating Pik3caH1047 and loss of function Pten mutations. *Oncotarget* **7**, 64289–64299 (2016).
210. Fougner, C., Bergholtz, H., Kuiper, R., Norum, J. H. & Sørli, T. Claudin-low-like mouse mammary tumors show distinct transcriptomic patterns uncoupled from genomic drivers. *Breast Cancer Res.* **21**, 85 (2019).
211. Pfefferle, A. D. *et al.* Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. *Genome Biol.* **14**, R125 (2013).
212. Bergholtz, H. *et al.* A Longitudinal Study of the Association between Mammographic Density and Gene Expression in Normal Breast Tissue. *J. Mammary Gland Biol. Neoplasia* **24**, 163–175 (2019).
213. Haakensen, V. D. *et al.* Expression levels of uridine 5'-diphospho-glucuronosyltransferase genes in breast tissue from healthy women are associated with mammographic density. *Breast Cancer Res.* **12**, R65 (2010).
214. Haakensen, V. D. *et al.* Gene expression profiles of breast biopsies from healthy women identify a group with claudin-low features. *BMC Med. Genomics* **4**, 77 (2011).
215. Haakensen, V. D. *et al.* Serum estradiol levels associated with specific gene expression patterns in normal breast tissue and in breast carcinomas. *BMC Cancer* **11**, 332 (2011).
216. Aure, M. R. *et al.* Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. *Breast Cancer Res.* **19**, 44 (2017).
217. Elston, E. W. & Ellis, I. O. Method for grading breast cancer. *J. Clin. Pathol.* **46**, 189–90 (1993).
218. O'Connell, P. *et al.* Analysis of loss of heterozygosity in 399 premalignant breast lesions at 15 genetic loci. *JNCI J. Natl. Cancer Inst.* **90**, 697–703 (1998).
219. Kim, S. Y. *et al.* Genomic differences between pure ductal carcinoma in situ and synchronous ductal carcinoma in situ with invasive breast cancer. *Oncotarget* **6**, 7597–7607 (2015).
220. Agilent Technologies, A. *Two-color microarray-based gene expression analysis: Low input quick Amp labeling. Agilent Technologies User Manual* (2010).
221. Bumgarner, R. Overview of dna microarrays: Types, applications, and their future. *Curr. Protoc. Mol. Biol.* **Chapter 22**, Unit 22.1. (2013).
222. Segundo-Val, I. S. & Sanz-Lozano, C. S. Introduction to the Gene Expression Analysis. *Methods Mol. Biol.* **1434**, 29–43 (2016).
223. Tsang, H.-F. *et al.* NanoString, a novel digital color-coded barcode technology: current and future applications in molecular diagnostics. *Expert Rev. Mol. Diagn.* **17**, 95–103 (2017).
224. NanoString Technologies. *Using the nCounter® Analysis System with FFPE Samples for Gene Expression Analyses.* (2012).
225. Reimers, M. An Opinionated Guide to Microarray Data Analysis. Available at: <http://www.people.vcu.edu/~mreimers/OGMDA/normalize.expression.html>. (Accessed: 10th May 2019)
226. Bolstad, B. M. Pre-processing DNA microarray data. in *Fundamentals of Data Mining in Genomics and Proteomics* 51–78 (Springer US, 2007). doi:10.1007/978-0-387-47509-7-3
227. Bolstad, B. Probe Level Quantile Normalization of High Density Oligonucleotide Array Data. *Unpubl.*

- Manuscr.* (2001).
228. Román-Pérez, E. *et al.* Gene expression in extratumoral microenvironment predicts clinical outcome in breast cancer patients. *Breast Cancer Res.* **14**, R51 (2012).
  229. Sun, X. *et al.* Relationship of mammographic density and gene expression: Analysis of normal breast tissue surrounding breast cancer. *Clin. Cancer Res.* **19**, 4972–4982 (2013).
  230. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 6567–6572 (2002).
  231. Liu, Y., Hayes, D. N., Nobel, A. & Marron, J. S. Statistical Significance of Clustering for High-Dimension, Low-Sample Size Data. *J Am Stat Assoc* **103**, 1281–1293 (2008).
  232. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Source J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
  233. Narula, S. C. Predictive mean square error and stochastic regressor variables. *J. Appl. Stat.* **23**, 11–17 (1974).
  234. D’Aurizio, R. *et al.* Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. *Nucleic Acids Res.* **44**, e154–e154 (2016).
  235. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16910–16915 (2010).
  236. Van Loo, P. *et al.* Analyzing cancer samples with SNP arrays. *Methods Mol. Biol.* **802**, 57–72 (2012).
  237. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–74 (2007).
  238. The International HapMap Consortium. The international HapMap project. *Nature* **426**, 789–796 (2003).
  239. Diskin, S. J. *et al.* Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **36**, e126 (2008).
  240. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
  241. Veeck, J. & Esteller, M. Breast Cancer Epigenetics: From DNA Methylation to microRNAs. *J. Mammary Gland Biol. Neoplasia* **15**, 5–17 (2010).
  242. Weisenberger, D. J. *et al.* *Comprehensive DNA Methylation Analysis on the Illumina® Infinium® Assay Platform. APPLICATION NOTE: ILLUMINA® EPIGENETIC ANALYSIS* (2008).
  243. Kurdyukov, S. & Bullock, M. DNA methylation analysis: Choosing the right method. *Biology (Basel).* **5**, 3 (2016).
  244. Touleimat, N. & Tost, J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4**, 325–341 (2012).
  245. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).
  246. Laurent-Puig, P. *Ion AmpliSeq Cancer Hotspot Panel v2.* (2015).
  247. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
  248. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
  249. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
  250. Zhou, W. *et al.* Full sequencing of TP53 identifies identical mutations within in situ and invasive components in breast cancer suggesting clonal evolution. *Mol. Oncol.* **3**, 214–219 (2009).
  251. Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P. & Tyson, G. W. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput. Biol.* **9**, e1003031 (2013).
  252. Simon, R. M. Personalized Cancer Genomics. *Annu. Rev. Stat. Its Appl.* **5**, 169–182 (2018).
  253. Noble, W. S. How does multiple testing correction work? *Nat. Biotechnol.* **27**, 1135–1137 (2009).
  254. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
  255. van Iterson, M., Boer, J. M. & Menezes, R. X. Filtering, FDR and power. *BMC Bioinformatics* **11**, 450 (2010).
  256. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 5116–5121 (2001).
  257. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
  258. Amrhein, V., Greenland, S. & McShane, B. Scientists rise up against statistical significance. *Nature* **567**, 305–307 (2019).
  259. Wasserstein, R. L. & Lazar, N. A. The ASA’s

## REFERENCES

---

- Statement on *p*-Values: Context, Process, and Purpose. *Am. Stat.* **70**, 129–133 (2016).
260. Kolesnikov, N. *et al.* ArrayExpress update—simplifying data submissions. *Nucleic Acids Res.* **43**, D1113–D1116 (2015).
261. Lappalainen, I. *et al.* The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* **47**, 692–695 (2015).
262. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
263. Sechidis, K. *et al.* Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics* **34**, 3365–3376 (2018).
264. Bethune, G., Bethune, D., Ridgway, N. & Xu, Z. Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. *J. Thorac. Dis.* **2**, 48–51 (2010).
265. Nasr, R. *et al.* Eradication of acute promyelocytic leukemia-initiating cells through PML-RARA degradation. *Nat. Med.* **14**, 1333–1342 (2008).
266. Song, Q., Merajver, S. D. & Li, J. Z. Cancer classification in the genomic era: five contemporary problems. *Hum. Genomics* **9**, 27 (2015).
267. Knudsen, E. S. *et al.* Progression of ductal carcinoma in situ to invasive breast cancer is associated with gene expression programs of EMT and myoepithelia. *Breast Cancer Res. Treat.* **133**, 1009–1024 (2012).
268. Williams, K. E. *et al.* Molecular phenotypes of DCIS predict overall and invasive recurrence. *Ann. Oncol.* **26**, 1019–1025 (2015).
269. Groen, E. J. *et al.* Finding the balance between over- and under-treatment of ductal carcinoma in situ (DCIS). *The Breast* **31**, 274–283 (2017).
270. Morita, M. *et al.* CD8+ tumor-infiltrating lymphocytes contribute to spontaneous “healing” in HER2-positive ductal carcinoma in situ. *Cancer Med.* **5**, 1607–1618 (2016).
271. Buerger, H. *et al.* Comparative genomic hybridization of ductal carcinoma in situ of the breast—evidence of multiple genetic pathways. *J. Pathol.* **187**, 396–402 (1999).
272. Esserman, L. J. *et al.* Use of Molecular Tools to Identify Patients With Indolent Breast Cancers With Ultralow Risk Over 2 Decades. *JAMA Oncol.* **27**, 4515–4521 (2017).
273. Doebar, S. C. *et al.* Extent of ductal carcinoma in situ according to breast cancer subtypes: a population-based cohort study. *Breast Cancer Res. Treat.* **158**, 179–187 (2016).
274. Clark, S. E. *et al.* Molecular subtyping of DCIS: heterogeneity of breast cancer reflected in pre-invasive disease. *Br. J. Cancer* **104**, 120–127 (2011).
275. Shin, S. J. *et al.* Molecular evidence for progression of microglandular adenosis (MGA) to invasive carcinoma. *Am. J. Surg. Pathol.* **33**, 496–504 (2009).
276. Geyer, F. C. *et al.* Microglandular adenosis or microglandular adenoma? A molecular genetic analysis of a case associated with atypia and invasive carcinoma. *Histopathology* **55**, 732–743 (2009).
277. Schreiner, D. & Weiner, J. A. Combinatorial homophilic interaction between  $\alpha$ -protocadherin multimers greatly expands the molecular diversity of cell adhesion. *Proc. Natl. Acad. Sci.* **107**, 14893–14898 (2010).
278. Huang, R. Y. J., Guilford, P. & Thiery, J. P. Early events in cell adhesion and polarity during epithelial-mesenchymal transition. *J. Cell Sci.* **125**, 4417–4422 (2012).
279. Gheldof, A. & Berx, G. Cadherins and epithelial-to-mesenchymal transition. in *Progress in Molecular Biology and Translational Science* **116**, 317–336 (Academic Press, 2013).
280. Heine, P., Ehrlicher, A. & Käs, J. Neuronal and metastatic cancer cells: Unlike brothers. *Biochim. Biophys. Acta - Mol. Cell Res.* **1853**, 3126–3131 (2015).
281. Norum, J. H. *et al.* GLI1 induced mammary gland tumours are transplantable and maintain major molecular features. *Int. J. Cancer* (2019).
282. Nickel, B., Moynihan, R., Barratt, A., Brito, J. P. & McCaffery, K. Renaming low risk conditions labelled as cancer. *BMJ* **362**, k3322 (2018).
283. Esserman, L. J. & Varma, M. Should we rename low risk cancers? *BMJ* **364**, k4699 (2019).
284. Esserman, L. J. *et al.* Addressing overdiagnosis and overtreatment in cancer: a prescription for change. *Lancet Oncol.* **15**, e234–e242 (2014).



## Paper I

### **A longitudinal study of the association between mammographic density and gene expression in normal breast tissue**


Helga Bergholtz, Tonje Gulbrandsen Lien, Giske Ursin, Marit Muri Holmen, Åslaug Helland, Therese Sørli and Vilde Drageset Haakensen.

*Journal of Mammary Gland Biology and Neoplasia* 2019, 24, 163–175.





# A Longitudinal Study of the Association between Mammographic Density and Gene Expression in Normal Breast Tissue

Helga Bergholtz<sup>1</sup> · Tonje Gulbrandsen Lien<sup>1</sup> · Giske Ursin<sup>2,3,4</sup> · Marit Muri Holmen<sup>5</sup> · Åslaug Helland<sup>1,6,7</sup> · Therese Sørli<sup>1,8</sup> · Vilde Drageset Haakensen<sup>1,7</sup> 

Received: 16 May 2018 / Accepted: 5 December 2018  
© The Author(s) 2019

## Abstract

High mammographic density (MD) is associated with a 4–6 times increase in breast cancer risk. For post-menopausal women, MD often decreases over time, but little is known about the underlying biological mechanisms. MD reflects breast tissue composition, and may be associated with microenvironment subtypes previously identified in tumor-adjacent normal tissue. Currently, these subtypes have not been explored in normal breast tissue. We obtained biopsies from breasts of healthy women at two different time points several years apart and performed microarray gene expression analysis. At time point 1, 65 samples with both MD and gene expression were available. At time point 2, gene expression and MD data were available from 17 women, of which 11 also had gene expression data available from the first time point. We validated findings from our previous study; negative correlation between *RBL1* and MD in post-menopausal women, indicating involvement of the TGF $\beta$  pathway. We also found that breast tissue samples from women with a large decrease in MD sustained higher expression of genes in the histone family H4. In addition, we explored the previously defined *active* and *inactive* microenvironment subtypes and demonstrated that normal breast samples of the *active* subtype had characteristics similar to the claudin-low breast cancer subtype. Breast biopsies from healthy women are challenging to obtain, but despite a limited sample size, we have identified possible mechanisms relevant for changes in breast biology and MD over time that may be of importance for breast cancer risk and tumor initiation.

**Keywords** Normal breast biology · Mammographic density · Gene expression · *RBL1* · Microenvironment

## Background

Breast cancer cells are extensively influenced by their non-cancerous surroundings, the microenvironment. The microenvironment consists of cells (such as fibroblasts, immune cells, endothelial cells and normal epithelial cells) and extracellular

matrix (ECM) including collagen, which all may influence initiation and progression of cancer [1, 2]. Mammographic density (MD) is a measure of radiologic density of the breast [3]. It varies extensively between individuals and may be seen as a radiologic reflection of breast tissue composition; epithelial and non-epithelial cells as well as collagen increase MD

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s10911-018-09423-x>) contains supplementary material, which is available to authorized users.

✉ Vilde Drageset Haakensen  
vdd@ous-hf.no

<sup>1</sup> Department of Cancer Genetics, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway

<sup>2</sup> Cancer Registry of Norway, Oslo, Norway

<sup>3</sup> Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway

<sup>4</sup> University of Southern California, Los Angeles, CA, USA

<sup>5</sup> Department of Radiology, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway

<sup>6</sup> Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

<sup>7</sup> Department of Oncology, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway

<sup>8</sup> Centre for Cancer Biomarkers CCBio, Dep. of Clinical Medicine, University of Bergen, Bergen, Norway

whereas fatty tissue reduces MD [4]. High MD is a strong independent risk factor for breast cancer, but the underlying mechanisms are still unclear [5–8]. Reduction in MD has been linked to a reduction in breast cancer incidence for women using Tamoxifen as primary prevention [9] and for patients receiving adjuvant hormonal therapy [10].

Normal breast tissue changes throughout life and is influenced by different hormonal events such as menarche, pregnancy, lactation and menopause [11]. The composition of breast tissue is also influenced by heritability [12, 13], use of hormonal therapy [14], nutrition [15, 16] and changes in Body Mass Index (BMI) [17]. MD decreases with age [18] and continues to decrease after menopause [19, 20]. The paradox of decreasing MD in parallel with increasing breast cancer incidence with age, can be explained by the model proposed by Pike et al. [11] which states that biological “breast tissue age” is determined by the cumulative exposure of damaging events to the breast tissue. High MD can reflect such damaging exposure, and thus contribute to increased breast cancer risk. It is important to note that MD is not a single biological state by itself, but recapitulates complex physiological and pathological conditions [2, 21].

Breast tissue from healthy women not undergoing surgery is extremely hard to obtain. The women in this study had previously donated tissue to research when they were examined at breast diagnostic centers. In order to allow a longitudinal study, these women agreed to undergo a second invasive procedure, which allowed us to present the first data on gene expression changes in normal breast tissue over time.

In our previous studies of normal breast tissue [22, 23], we identified a group of normal breast tissue samples exhibiting upregulation of mesenchymal and stem cell genes and downregulation of epithelial markers and adhesion genes, a trait identified in tumors of the claudin-low breast tumor subtype. Furthermore, we identified 24 genes that were negatively correlated to MD, including *RBL1* (Retinoblastoma-like protein 1, p107) and three uridine 5'-diphospho-glucuronosyltransferase (UGT) genes whose protein products are known to inactivate estrogen metabolites. *RBL1* is expressed at high levels in normal breast epithelium [24], and is thought to have similar tumor suppressive effects as its cousin gene *RB1*. In addition to acting as gatekeepers of the G1-S transition, the RB proteins may play roles in preservation of chromosomal stability, induction and maintenance of senescence, and regulation of apoptosis, cellular differentiation and angiogenesis [25].

The microenvironment is known to be crucial to cancer initiation and progression [26, 27]. Román-Pérez et al. proposed a method for extratumoral microenvironment subtyping based on gene expression patterns, classifying tumor adjacent normal tissue as *active* or *inactive* [28]. The *active* subtype is characterized by features such as inflammatory response, fibrosis and cellular movement; features similar to the claudin-low breast cancer subtype, proposed by Herschkowitz et al. [29]. The *inactive* subtype is characterized by maturation,

differentiation of epithelial cells, and high cell adhesion. This subtype was later shown to correlate with high MD [21]. These microenvironment subtypes have not been explored in individuals without cancer, but if present in healthy breast tissue, they could potentially influence breast cancer initiation differently.

The aim of this study was to investigate the changes in gene expression that take place in normal breast tissue over a time period of several years, especially in relation to changes in MD and to validate correlations between gene expression and MD identified in our previous study. We validated a negative correlation between *RBL1* expression and mammographic density in postmenopausal women and found an association between change in MD and change in expression of histone-related genes. We also demonstrated that the previously defined *active* and *inactive* microenvironment subtypes are present in normal breast tissue.

## Methods

### Subjects

Two separate breast biopsies from healthy volunteering women (i.e. without cancer disease) were obtained with 5–8 years between sample times. The present study is based on our previous study, Mammographic Density and Genetics 1 (MDG1) [22], of women attending the National Breast Cancer Screening Program. The included women were referred to one of several breast diagnostic centers for biopsies due to suspicious findings on mammograms or abnormal clinical findings, and biopsies from breasts without any malignant disease were obtained. Only women without signs of malignant disease were included in this study and biopsies were taken from the contralateral breast of the suspected lesions. MD was determined from mammograms. A total of 120 healthy women were included. Of these, gene expression profiles were available for 79 and MD for 113, with overlapping data for 65 women. Five to eight years later, women who revisited the breast diagnostic center were invited to participate in a follow-up study (MDG2) where new biopsies were obtained, new mammograms taken and new MD assessments performed. A total of 25 women revisited the center at the second time point. All women agreed to participate and completed a questionnaire providing information like height, weight and menopausal status. With regard to menopause status in MDG1, the women were estimated to be pre-, post- or peri-menopausal based on serum levels of FSH, LH and estradiol as previously described [30]. All women provided a signed informed consent. The study was approved by the local ethical committee and local authorities (IRB approval no S-02036).

## Biopsies

In both studies, biopsies were obtained as previously described [22]. Briefly, ultrasound guided core biopsies using a 14 gauge needle was performed in an area of some MD to avoid biopsies consisting purely of adipose tissue. Most biopsies were sampled in the upper, lateral quadrant at both time points. The biopsies were snap frozen and stored in  $-80^{\circ}\text{C}$  until RNA isolation. Since healthy breast tissue express less mRNA than tumor tissue, the entire biopsy was required for mRNA extraction. Therefore, no tissue was left for histological or immunohistochemical evaluation.

## RNA Isolation and Expression Arrays

Gene expression data for the samples from the previous study (MDG1) are deposited in NCBI's Gene Expression Omnibus [31] and are accessible through GEO Series accession number GSE18672 [32]. Two additional gene expression datasets were retrieved from GEO: GSE72644 comprises data from breast cancer patients, where multiple biopsies from unaffected normal ducts in the same breast were retrieved for several patients [33]. GSE4823 [34] contains data from normal breast tissue microdissected into epithelium and stroma cellular compartments. The platform used for all three datasets was Agilent Human Gene Expression 4x44K microarrays (G4110A, two colors) (Agilent, Technologies, Santa Clare, USA).

From the new set of biopsies (MDG2), total RNA was isolated using Qiagen miRNeasy Mini kit (Qiagen, Hilden, Germany). The tissue was homogenized by manually mincing on ice with a scalpel followed by Mixer Mill for 40 seconds until complete homogenization. RNA extraction including DNase treatment was performed according to the protocol provided by the supplier. RNA concentrations were measured by NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA) and RNA quality was analyzed using Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA).

To obtain whole genome expression data, Agilent Sureprint G3 Human Gene Expression 8x60K microarrays (G4851A) (Agilent, Technologies, Santa Clare, USA) with Low Input Quick Amp Labeling protocol were used. RNA input was 40 ng and Cy3 was used as fluorophore. Quality Control was performed in Agilent's Feature Extraction software. The microarray expression data for MDG2 have been deposited in the ArrayExpress database at EMBL-EBI under accession number E-MTAB-5885 [35, 36].

In the current study, a total of 25 biopsies were obtained. Of these, three samples were excluded due to too low RNA concentration for expression analysis, and three samples failed Cy3-labeling. Nineteen samples were successfully run on arrays and passed all quality control criteria. For controls, one sample of commercially available normal breast RNA

(Ambion Human Breast Total RNA, Thermo Fisher Scientific, Wilmington, DE, USA) and one tumor sample were included throughout the whole pipeline. Two of the samples had no associated MD data. In total, data from 17 samples were complete with both gene expression and MD data. From the previous study (MDG1), 65 samples were complete with gene expression and MD data. For six women, gene expression data were obtained at time point two only. In total, paired data were available from 11 women.

## Mammographic Density

Digital craniocaudal mammograms were obtained at routine mammographic centers using a standard protocol. Mammographic density was estimated using the University of Southern California Madena assessment method as described by Ursin et al. [37]: Using the Madena computer software, the reader (GU) outlined the total area of the breast, and the number of pixels was counted by the software. This represents the total breast area. MD was assessed as follows: First, a region of interest that includes all dense areas except those representing the pectoralis muscle or scanning artifacts was identified. Then, a yellow tint was applied to all pixels within the region of interest shaded at or above a threshold intensity of gray. The software then counted the tinted pixels, which represent the area of absolute density. Percent density was determined by dividing the absolute dense area by the total breast area, and multiplying by 100 [5]. Test-retest reliability was 0.99 for absolute density. For cases with mammograms for both breasts available (14 out of 17), the correlation of MD was very high between the right and left breast (Pearson correlation  $r = 0.97$ ,  $p < 0.001$ ,  $n = 14$ ), thus, for these women the average MD was used. For the remaining three women, MD was calculated for the breast with available scans. As a measure of MD change, both absolute change ( $MD2 - MD1$ ) and relative change ( $\frac{MD2 - MD1}{MD1}$ ) was calculated. Since women with low MD in the first study may potentially have a lower absolute decrease than women with high MD, relative change was used for comparison to gene expression and clinical parameters.

## Statistical Analysis

Analysis of the relationship between MD at time point one (MD1) and two (MD2) was performed using Pearson correlation. Out of the 24 genes whose expression were identified as significantly associated with MD1 in our first study (MDG1), 16 genes, represented by 23 probes, were present on the array used in the second study (MDG2). To investigate the association between gene expression of these genes and MD in the second study, Pearson correlation was used. Different versions of whole genome expression arrays were used for MDG1 and

MDG2; notably Agilent Human Gene Expression 4x44K and 8x60K. To avoid introducing bias, the two expression datasets were analyzed separately and then compared using a rank based approach: For the 11 samples with expression data at both time points, after collapsing to gene level using the median expression of the probes, the genes overlapping in both datasets ( $n = 15,107$ ) were extracted. For each time point separately, the genes in each sample were ranked based on their expression value. We then calculated rank change as a proxy for change in expression. Spearman correlation analysis was performed to investigate the association between relative change in MD ( $\frac{MD2-MD1}{MD1}$ ) and relative change in rank of gene expression ( $\frac{Rank2-Rank1}{Rank1}$ ). The top and bottom 200 genes from this analysis were used for gene ontology analyses. As a sensitivity analysis, we checked gene ontology terms associated with the top 500 genes as well.

The breast microenvironment subtypes (*active/inactive*) were calculated using the Chreighton correlation method as described in Sun et al. [21, 28, 38], separately for the two datasets: The signature consisting of 3194 genes was retrieved from Sun et al. with +1 assigned to up-regulated and -1 to down-regulated genes. Expression values for genes overlapping with the signature were extracted (for MDG1 2786 genes, for MDG2 2444 genes) and the Pearson correlation coefficients to the signature were calculated. The samples were classified as *active* if the correlation coefficient was positive, and *inactive* if it was negative. A Welch two sample t-test was used to find differentially expressed genes between *active* and *inactive* subtype in MDG1 followed by gene ontology analyses. The association between microenvironment subtypes and MD was tested using the non-parametric Wilcoxon-Mann-Whitney test. This test was also used when exploring associations between microenvironment subtypes and relevant genes in both MDG1 and MDG2. All statistical tests were two-sided with significance level  $\alpha = 0.05$ . Spearman correlation was used where associations between ranks were explored, otherwise Pearson correlation was used accompanied by visualization of the data. All statistical analyses were performed in Rstudio version 1.0.136 [39]. PAM50-subtypes were estimated using the R Package “genefu” [40] and for power analyses the R Package “pwr” was used [41]. To be able to discover similarly strong correlations between MD and age/BMI as previously reported ( $-0.56/-0.21$ ) [42] with a power above 0.8 and significance level  $\alpha = 0.05$ , at least 174/21 samples would be needed. Thus, the size of our cohort is too small to draw any firm conclusions of an association between MD, BMI and age (Online resource 1: Fig.S1 A and B). These parameters were therefore not adjusted for in the analyses to prevent introducing unnecessary noise. A power of 0.66 was obtained in the analyses of the association between microenvironment subtype and MD ( $n_1 = 28$ ,  $n_2 = 37$ ,  $d = 0.61$  (effect size as reported in Sun et al. [21]),  $\alpha =$

0.05). Gene Ontology (GO) analyses were performed in the web-based functional annotation tool DAVID 6.8 [43, 44] which performs enrichment analyses on gene sets enabling exploration of biological systems and pathways.

Scores for epithelial-to-mesenchymal transition (EMT) [28], proliferation [45] and fibrosis (gene signature associated with desmoid type fibromatosis) [46] were calculated using a standard (Z) score approach: For every gene in each signature, a standardized expression value was calculated by subtracting the mean across all samples, then dividing by the standard deviation. The sample's score was calculated by taking the mean of the standardized expression values of all genes in the signature (Online resource 2).

## Gene Set Enrichment Analyses

Gene set enrichment analyses were carried out using the Hallmark gene sets from the Molecular Signatures Database (MSigDB [47, 48]) on the MDG1 dataset: For each sample, genes were ranked by their expression values. Wilcoxon-Mann-Whitney test was used to test difference in rank between the genes in each gene set compared to those not in the gene set. The resulting  $p$  value was transformed using this formula:  $-10 \times \log_{10}(p \text{ value})$  and the sign was changed according to the direction of enrichment of genes (i.e. whether the genes were highly or lowly expressed) resulting in an enrichment score for each sample and each gene set (Online resource 3). This enrichment score was used for subsequent statistical testing.

## Results

### Cohort Description

A total of 24 women included in the first MDG study accepted participation in the second study. For 17 of these, both MD and gene expression data was available and used for further analyses. None of the women experienced breast cancer after they were included in the first study. Relevant clinical information is presented in Table 1. Age at second biopsy ranged from 55 to 66 and all the women were at this time point postmenopausal. Mammograms were obtained and MD was estimated as described in Methods. As expected, MD1 and MD2 were highly dependent (Pearson correlation  $r = 0.80$ ,  $p < 0.001$ ,  $n = 17$ ) (Online resource 1: Fig.S1 C). MD decreased from the first to the second measurement for all but one woman. There was no difference in relative MD change between women who had passed menopause between sampling times ( $n = 5$ , mean relative MD change =  $-41.7\%$ ) compared to those who already were postmenopausal at the first time point ( $n = 8$ , mean relative MD change =  $-42.6\%$ ).

**Table 1** Clinical information at time point two including mammographic density at both time points

Sample	BMI	Menopause change	Expression data in both studies	Months between biopsies	MD1 (%)	MD2 (%)	MD Absolute difference	MD Relative difference (%)
NORM-11	28.65	Yes	Yes	74	23.16	15.29	-7.87	-33.98
NORM-17	30.46	No	Yes	79	13.54	8.6	-4.94	-36.49
NORM-24	28.84	No	Yes	76	17.61	9.93	-7.69	-43.64
NORM-26	23.39	Yes	Yes	77	34.5	14.58	-19.92	-57.73
NORM-31	21.78	NA	No	94	12.72	1.94	-10.78	-84.76
NORM-32	31.25	NA	No	89	7.28	3.01	-4.27	-58.65
NORM-33	30.82	NA	No	90	20.42	4.12	-16.3	-79.8
NORM-34	26.57	NA	Yes	99	20.02	8.53	-11.49	-57.4
NORM-38	23.23	No	No	92	32.86	40.07	7.21	21.93
NORM-39	17.99	No	Yes	96	25.69	10.8	-14.88	-57.94
NORM-44	27.01	No	Yes	96	41.61	22.85	-18.77	-45.1
NORM-49	19.47	Yes	Yes	75	15.59	9.44	-6.15	-39.43
NORM-50	34.29	Yes	No	72	28.06	10.61	-17.45	-62.17
NORM-56	22.41	No	No	78	60.82	34.97	-25.85	-42.5
NORM-61	24.46	No	Yes	73	17.08	4.3	-12.79	-74.85
NORM-64	19.37	Yes	Yes	69	18.02	14.45	-3.57	-19.82
NORM-66	34.6	No	Yes	76	9.94	4.58	-5.37	-53.99
<b>Mean</b>	<b>26.15</b>			<b>82.65</b>	<b>23.47</b>	<b>12.83</b>	<b>-10.64</b>	<b>-48.61</b>
<b>Min</b>	<b>17.99</b>			<b>69</b>	<b>7.28</b>	<b>1.94</b>	<b>-25.85</b>	<b>-84.76</b>
<b>Max</b>	<b>34.60</b>			<b>99</b>	<b>60.82</b>	<b>40.07</b>	<b>7.21</b>	<b>21.93</b>

BMI: Body mass index Menopause change: No = post-menopausal at both time points; Yes = pre-/peri-menopausal at time point one, post-menopausal at time point two. NA = not available in MDG1. MD1 and MD2: Percent mammographic density at time point 1 and 2, respectively. Age is omitted from the table as it is considered a sensitive parameter

Summary statistics are written in bold italics

### Associations between RBL1 Expression and Mammographic Density Were Validated in the Second Biopsies

Probes for two of the genes identified in our previous study, were significantly associated to MD also in our second study: Retinoblastoma-like protein 1 (*RBL1*) and Leucine-rich repeat-containing 2 (*LRRC2*) (Table 2). *RBL1* was represented by two probes and both confirmed the previously identified negative association to MD, however, only one of these reached statistical significance (Fig. 1). For the UGT genes that were found to be negatively correlated to MD in MDG1, a negative association was found at the second time point as well, although statistical significance was not reached. (Online resource 1: Fig.S2).

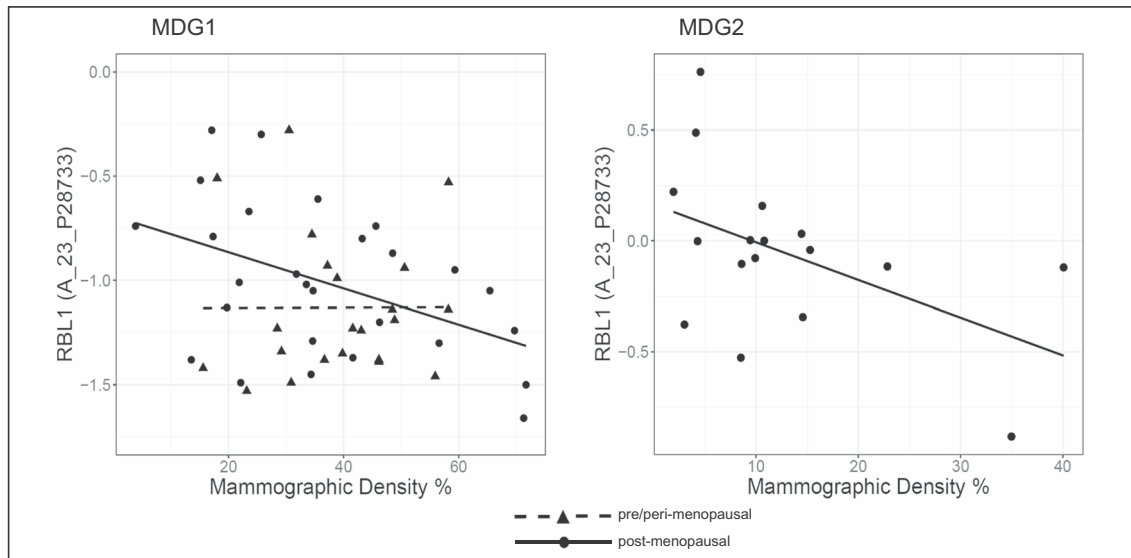
When stratifying MDG1 based on menopause status, the correlation between *RBL1* expression and MD was evident in post-menopausal women only (Fig. 1), indicating that processes relevant for breast tissue composition may change with menopause. To assess the effect of menopausal status on overall gene expression, we identified differentially expressed genes between biopsies from post- and pre-/peri-menopausal women in the largest cohort (MDG1), and found only five differentially expressed genes. Next, we correlated overall gene expression with MD separately in the two menopausal groups, and found substantially more genes associated to MD in the postmenopausal group than the pre-/peri-menopausal (1169 vs. 436 genes) with only 14 genes overlapping between the two groups.

**Table 2** Correlation between gene expression and mammographic density at time point 2 (MDG2, n = 17). Probes included are those whose expression was correlated with mammographic density at time point 1 and present on the arrays used at time point 2

Gene Name	Probe name	r	p-value
<i>ATG7</i>	A_24_P944827	-0.1669	0.5220
<i>ATG7</i>	A_23_P143987	0.0589	0.8224
<i>CABP7</i>	A_33_P3348061	0.2097	0.4193
<i>CD86</i>	A_24_P131589	0.1967	0.4491
<i>ESR1</i>	A_24_P383478	0.2055	0.4288
<i>ESR1</i>	A_33_P3379356	0.0806	0.7584
<i>ESR1</i>	A_23_P309739	0.0585	0.8234
<i>H2AFJ</i>	A_33_P3379391	0.4062	0.1057
<i>H2AFJ</i>	A_23_P204277	0.3783	0.1343
<i>HMBOX1</i>	A_24_P932736	0.254	0.3252
<i>LMOD1</i>	A_33_P3368879	0.291	0.2572
<i>LMOD1</i>	A_33_P3295261	0.1755	0.5005
<i>LRRC2</i>	A_23_P334798	-0.6889	0.0022 *
<i>LRRC2</i>	A_23_P155463	0.1637	0.5302
<i>NPY1R</i>	A_23_P69699	0.2385	0.3566
<i>PIK3R5</i>	A_23_P66543	0.2439	0.3454
<i>PPP6R1</i>	A_23_P119448	0.2639	0.3061
<i>RBL1</i>	A_23_P28733	-0.4909	0.0454 *
<i>RBL1</i>	A_24_P276102	-0.3373	0.1855
<i>RPA4</i>	A_23_P254212	0.3781	0.1346
<i>UGT2B10</i>	A_23_P7342	-0.1826	0.4829
<i>UGT2B11</i>	A_23_P212968	-0.1405	0.5906
<i>UGT2B7</i>	A_23_P136671	-0.2846	0.2682

r: Pearson correlation coefficients

\* :p-value <0.05)



**Fig. 1** Expression of *RBL1* as a function of mammographic density in MDG1 (stratified by menopause status) and MDG2 (all post-menopausal). Pearson correlation: MDG1: Post-menopausal ( $n = 28$ ),

$r = -0.51$ ,  $p = 0.0061$ ; pre/peri-menopausal ( $n = 22$ ),  $r = 0.0039$ ,  $p = 0.99$ . MDG2 ( $n = 17$ ):  $r = -0.49$ ,  $p = 0.045$

Seeing that *RBL1*-expression showed a consistent negative correlation to MD over time, prompted us to examine the association between *RBL1* expression and enrichment of Hallmark gene sets from the Molecular Signature Database (Online resource 3). We correlated expression values for *RBL1* in the MDG1 dataset with Gene Set Enrichment analysis (GSEA) enrichment scores and found that the enrichment scores of *WNT/β-catenin signaling* and *MYC-targets* were significantly negatively correlated to *RBL1* expression (Spearman correlation,  $\rho = -0.397$ ,  $p = 0.0011/\rho = -0.259$ ,  $p = 0.037$ ). Further, we wanted to investigate whether MD could be associated with processes relevant for cancer development. To this end, we correlated enrichment scores from GSEA to MD for the samples in the MDG1 dataset and found that gene sets related to *Apoptosis* and *Estrogen response* were significantly negatively correlated to MD (Spearman correlation,  $p = 0.0268/0.0343$ ,  $\rho = -0.277/-0.265$ ), while *TGFβ-signaling* was marginally not significant (Spearman correlation,  $p = 0.0638$ ,  $\rho = -0.233$ ).

Intra-individual variation of gene expression may be a complicating factor in all studies where only one biopsy is analyzed. To assess the intra-individual variability of *RBL1* expression, we made use of a separate dataset (GSE72644) with gene expression data from two biopsies of normal ductal tissue obtained from different parts of the breast from several patients. We found low correlation between *RBL1* expression in different ducts of the same patient (Spearman correlation,  $p = 0.67$ ,  $\rho = 0.167$ ), indicating some degree of *intra*-individual variability of *RBL1* expression; however the *inter*-individual variability was small, as demonstrated by a low standard deviation of *RBL1* (SD for *RBL1*: 0.21 vs. mean SD for all genes: 0.68).

### Gene Expression in Normal Breasts Changes over Time

From 11 of the women, tissue biopsies were obtained at both time points. A rank-based approach (see Methods section) was taken to overcome the challenge of analyzing gene expression data from two different platforms. To identify biological processes changing in breast tissue over time in parallel with changes in MD, normalized gene expression values were ranked from lowest to highest within each sample, separately for time point one and two. This was followed by Spearman correlation to identify genes with positive or negative correlation between relative change of MD and relative change in gene expression ranks (Online resource 4). Gene ontology analysis of the top 200 genes with a negative correlation between change in gene expression and relative change in MD, revealed involvement of several genes in the histone family H4. Sensitivity analysis using the top 500 genes confirmed these results. In other words, breast tissue samples with a large decrease in MD from the first to the second time point sustained a high expression of these genes (Online resource 1: Fig.S3).

### Identifying Microenvironment Subtypes in Normal Breast Tissue

To investigate whether the microenvironment subtypes proposed by Román-Pérez [28] could be identified in normal tissue from healthy breasts, we assigned all tissue samples to a microenvironment subtype (*active/inactive*) (Online resource 5). In the MDG1 study, 27 samples (41.5%) were of the *active* subtype, while 38 samples



(58.5%) were assigned to the *inactive* subtype, whereas for the MDG2 study, 8 samples (47.1%) were *active*, and 9 (52.9%) were *inactive*. Of the 11 samples with data at both time points, five kept their subtype (45.5%) (one *active* and four *inactive*), while six samples (44.5%) changed subtype (three from *active* to *inactive*, and three from *inactive* to *active*). There was no difference in distribution of menopause status between the two subtypes (Fisher exact test,  $p = 0.314$ ).

As previously noted by Sun et al., the two microenvironment subtypes may differ in characteristics such as adhesion, stem cell features and TGF $\beta$ -signalling. We confirmed these results in our data from normal breasts. In our largest study, the MDG1 study, 3104 genes were significantly differentially expressed between the two subtypes (1390 up and 1714 down in *active* vs. *inactive*, Welch two sample t-test, FDR <1%). Gene ontology (GO) analysis showed enrichment of GO-terms related to cell-cell adhesion and tight junctions among the genes that were lower expressed in the *active* subtype compared to the *inactive*, while for genes higher expressed in the *active* subtype, we found GO-terms related to stem cell-like features such as Aldehyde dehydrogenase and Wnt-signaling (Online resource 6). There was a clear distinction between the subtypes in both cohorts with regard to the expression of genes relevant for the claudin-low tumor subtype [28, 49]; the adhesion genes (e.g. *CLDN3*, *CLDN4*, *CLDN7*, *CDH1* and *OCLN*) were lower expressed in the *active* subtype compared to the *inactive*, while the EMT-related genes (e.g. *TWIST*, *ZEB1* and *ZEB2*) were higher expressed (Fig. 2). To consolidate these findings, we tested whether gene signatures from the GSEA analyses were differently enriched between *active* and *inactive* microenvironment subtypes in the MDG1 dataset. As many as 28 (out of 50) Hallmark gene sets were differently enriched, confirming the extensive differences between the subtypes (Online resource 3). Most notably were genes involved in *Adipogenesis*, *TGF $\beta$ -signaling* and *Epithelial to Mesenchymal Transition* higher expressed in the *active* subtype compared to the *inactive* (Mann Whitney U tests,  $p < 0.001$ ).

In contrast to the findings in Sun et al. [21], we did not find a significant association between microenvironment subtype and MD in any of the cohorts (Online resource 1: Fig. S4). Neither was *RBL1* differently expressed between the subtypes (Fig. 3). However, since we found a negative correlation between *RBL1* and MD, we wanted to investigate whether genes that may be influenced by *RBL1* expression (through its role as a co-repressor together with the transcriptional repressor E2F4) were differentially expressed between the two subtypes. In this context, *MYC* is particularly interesting, as it is highly relevant in cancer and involved in proliferation [50]. We found that *MYC* was significantly differently expressed between the subtypes in both cohorts (Fig. 3). There was, however, no significant correlation between *MYC* and *RBL1* (Spearman correlation, MDG1:  $p = 0.114$ ,  $\rho = -0.198$ ,

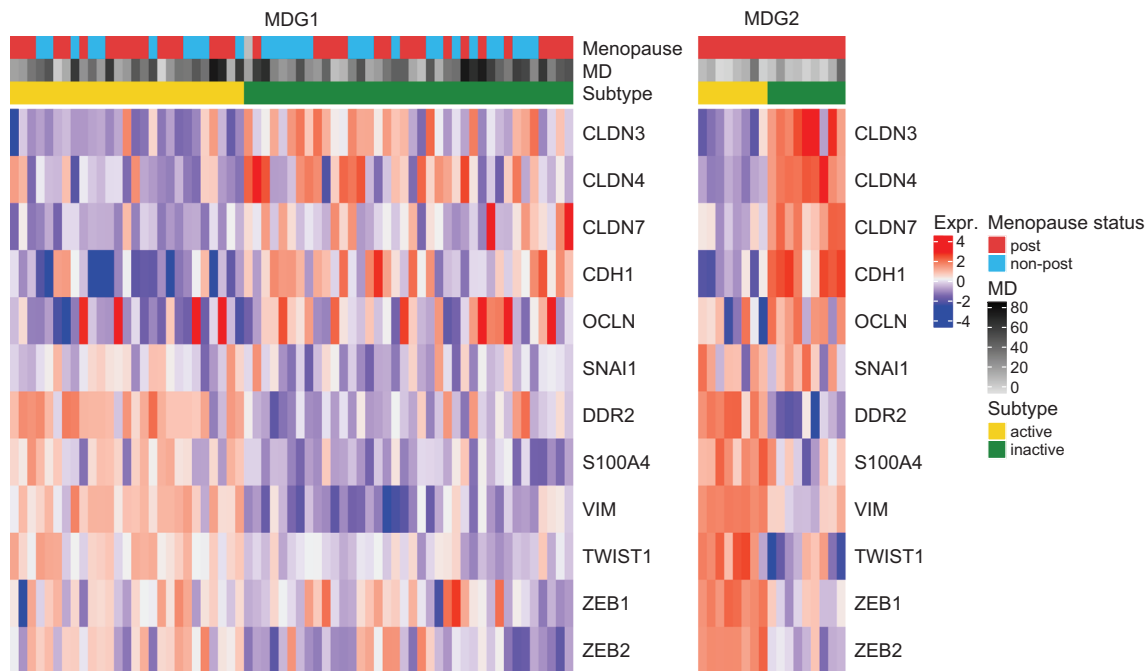
MDG2: 0.503,  $\rho = 0.174$ ). We therefore wanted to identify E2F4 target genes that were both differentially expressed between the microenvironment subtypes and negatively correlated to *RBL1* expression. Platelet derived growth factor subunit A (*PDGFA*) fulfilled both these criteria in MDG1 (Fig. 3, Spearman correlation *PDGFA* vs. *RBL1*,  $p$  value = 0.007,  $\rho = -0.33$ ). *PDGFA* was also differentially expressed between the subtypes in MDG2. In both cohorts, *PDGFA* was higher expressed in the *inactive* compared to the *active* subtype.

To further explore the differences in properties between the microenvironment subtypes that could be relevant for mammographic density, we calculated standardized z-scores for EMT, fibrosis and proliferation (Online resource 1, Fig. S5 and Online resource 2). Both EMT- and fibrosis signatures were significantly higher in the *active* compared to the *inactive* subtype. In MDG2, there were significantly higher proliferation scores in the *inactive* subtype and the same tendency was also seen in MDG1, however not significant. In addition, scores for EMT and fibrosis in MDG1 were significantly positively correlated in *active*, but not for the *inactive* samples (Fig. 4).

Since there were extensive differences between the microenvironment subtypes in the Gene Set Enrichment Analyses, we wanted to investigate whether the GSEA enrichment scores were differentially associated with MD between the two subtypes in MDG1 (Online resource 3). We found that, in the *active* subtype, the enrichment scores for the pathway *MYC-targets* were significantly positively correlated with MD (Spearman correlation,  $\rho = 0.385$ ,  $p = 0.0476$ ). For the *inactive* subtype, several gene sets involved in hormonal processes (i.e. *Estrogen response* and *Androgen response*) were negatively correlated with MD. In addition, *TGF $\beta$ -signaling* was negatively correlated to MD in the *inactive* subtype, although significance was not reached ( $p = 0.063$ ). These results suggest that target genes of the TGF $\beta$  pathway may be involved in processes relevant for MD in both microenvironment subtypes.

### Spatial Distribution of *RBL1* in Normal Breast Tissue

The microenvironment subtypes most likely reflect interplay between stromal and epithelial cells. In this context, it was of interest to investigate whether there was a spatial difference in gene expression of relevant genes between epithelial and stromal cellular compartments. As additional tissue from our cohort was not available for analyses, the spatial distribution of *RBL1*-expression was studied in a separate dataset (GSE4823) comprising data from normal breast tissue microdissected into epithelial and stromal cellular compartments. These data showed higher expression of *RBL1* and a trend toward high expression of *PDGFA* in the epithelial cells compared to the stromal cells (Online resource 1, Fig. S6). For *MYC*, there was



**Fig. 2** Expression of selected claudin-low and EMT relevant genes in MDG1 and MDG2. Menopausal status, mammographic density and microenvironment subtype are indicated in the colored boxes above the heatmaps

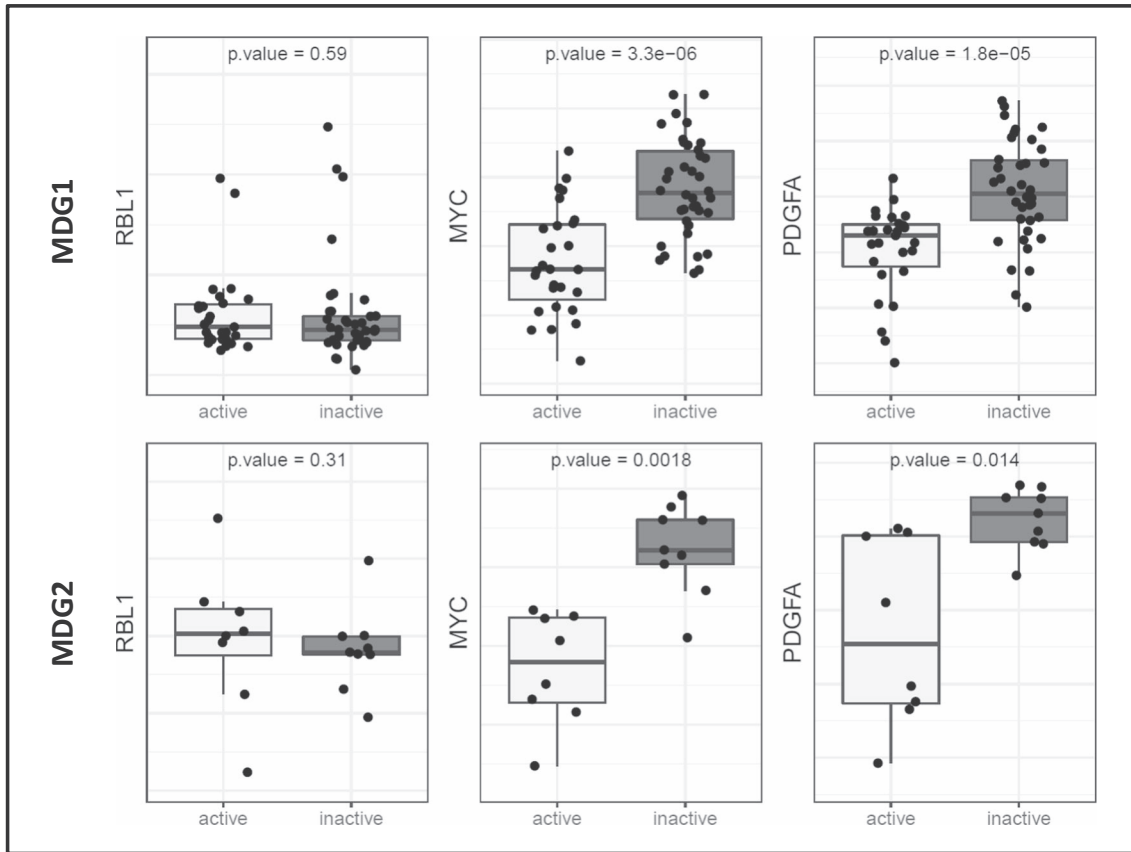
substantial variation of expression in the epithelial cell compartment, with higher expression in stromal cells in general. This difference was, however, not significant. The spatial distribution of the corresponding proteins were validated using the Human Protein Atlas where protein expression in epithelial cells were confirmed for all three proteins [51–53].

## Discussion

Our study confirmed that low expression of *RBL1* in normal breast tissue in repeated measurements years apart was associated with high MD in postmenopausal women. *RBL1* closely resembles *RBI* and functions as a tumor suppressor gene involved in cell cycle regulation [54, 55]. The inverse relationship between *RBL1* expression and MD harmonizes with its presumptive role as a tumor suppressor through regulation of epithelial cell proliferation and modification of the ECM. *RBL1* acts as a co-repressor of transcription as part of the SMAD complex downstream of TGF $\beta$  in the TGF $\beta$ -signaling pathway [56, 57]. TGF $\beta$  has a pleiotropic role in cancer development, contributing to regulating cell proliferation, epithelial-to-mesenchymal transition (EMT) and ECM formation in a highly context dependent manner [58]. Increased TGF $\beta$ -signaling in the normal breast is known to inhibit proliferation of epithelial cells [59] and TGF $\beta$ -signaling has previously been shown to be reduced in dense mammary tissue [60]. Paradoxically, TGF $\beta$  enhances the synthesis of collagen crosslinking enzymes, which increases the rigidity of the collagen network in the ECM [61] and

contribute to MD [62, 63]. Adding to the complexity, is the fact that high activity of the TGF $\beta$  pathway may have a tumor suppressive role in the initiation and early progression of cancer, and later switch to have a pro-tumorigenic and prometastatic role [64]. Reduced TGF $\beta$ -signaling may lead to decreased repression of several target genes involved in cell proliferation (possibly affecting MD) and neoplastic transformation [50, 54]. We did not find a significant correlation between *RBL1* and *MYC* expression. However, Gene Set Enrichment Analyses (GSEA) indicated a relationship between *RBL1* and *MYC*-related pathways as both *WNT*/ $\beta$ -catenin and *MYC*-target gene sets were negatively correlated to *RBL1* expression. These pathways are involved in epithelial proliferation [50, 65].

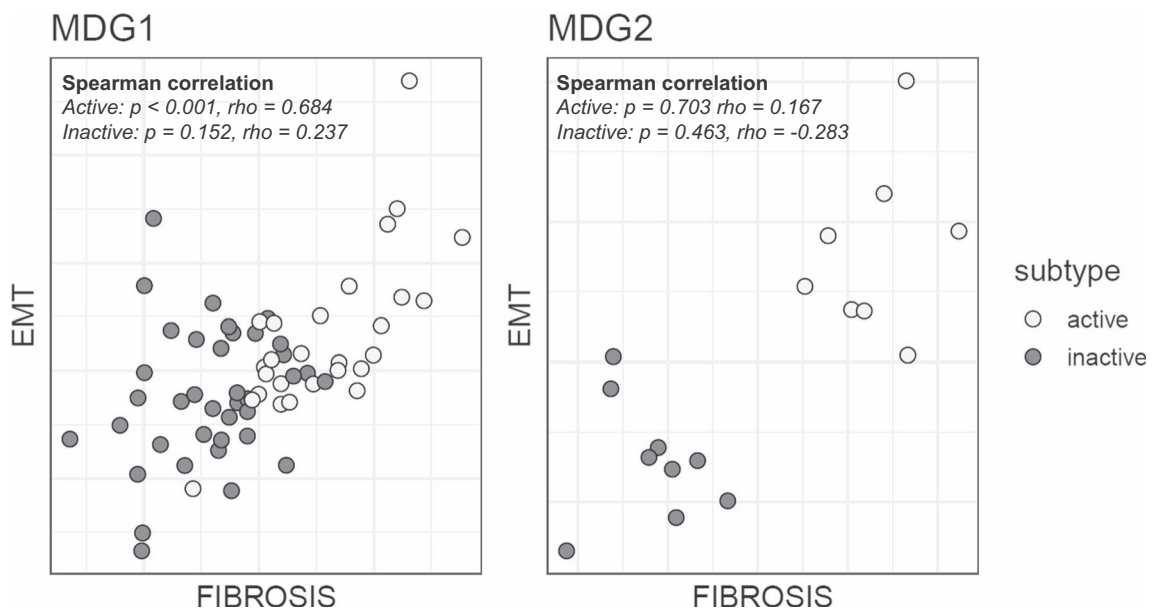
In postmenopausal women, the estrogen-mediated cell proliferation is lower than in pre-menopausal women [66]. In this study, we saw that low expression of *RBL1* was associated with high MD only in postmenopausal women. The explanation for this may lie in the cross-talk between the ER $\alpha$  and TGF- $\beta$  signaling pathways as ER $\alpha$  represses SMAD3-function in an estradiol-dependent manner [67, 68]. There was a distinct difference in the number of genes whose expression correlated to MD in postmenopausal breast tissue compared to the pre/peri-menopausal breast tissue in the MDG1 cohort. This is in contrast to the low number of differentially expressed genes between breast biopsies from post- and pre/peri-menopausal women at a genome-wide level which has also been reported from other studies [60, 69]. The low number of genes whose expression correlated to MD in pre/peri-menopausal



**Fig. 3** RBL1, MYC and PDGFA expression in the microenvironment subtypes active and inactive in MDG1 and MDG2. P-values from Wilcoxon-Mann-Whitney tests

women may be a reflection of more heterogeneity in genes relevant for MD as these may fluctuate substantially due to hormonal changes through the menstrual cycle, potentially masking such associations [70].

Low expression of histones may lead to a more open chromatin structure which is thought to cause higher genomic instability and inappropriate gene expression possibly contributing to carcinogenesis [71]. In accordance with this, we



**Fig. 4** Fibrosis score vs. EMT score in MDG1 and MDG2. P-values and rho from Spearman correlation tests

found that breast tissue with a large decrease in MD over time showed sustained or higher expression of histone proteins of the H4 family compared to those with a smaller decrease in MD. Interestingly, high expression of histone genes has been shown to slow down the aging process in cells as high availability of histone proteins contributes to a tighter chromatin structure [71].

The microenvironment subtypes proposed by Sun et al. and Román-Pérez et al. [21, 28] were observed in the normal breast samples in our study. The samples classified as *active* subtype showed features such as high expression of EMT-related genes, low expression of genes involved in cell-cell adhesion and upregulation of GO-terms related to stem cell-like characteristics similar to what is found in the claudin-low breast tumor subtype [49, 72–74]. Although the claudin-low subtype was initially discovered in breast tumors, we have previously found evidence of claudin-low characteristics in normal breast tissue from MDG1 [23]. All of these were in the present study determined as *active* subtype. There was no difference in *RBL1* expression between the subtypes, but both *MYC* and *PDGFA* were significantly higher expressed in the *inactive* samples compared to the *active*, indicating higher activation of the TGF $\beta$  pathway in the *active* compared to the *inactive* subtype. The samples of the *active* subtype also showed enrichment of fibrosis-related genes shown by Beck et al. to be enriched in a subset of breast carcinomas associated with longer survival [46]. The presence of increased EMT features, TGF $\beta$  activation and fibrosis in the *active* subtype may indicate the presence of a “wound healing” phenotype even without any tumor initiation [75, 76].

In Sun et al. the *inactive* subtype was associated with slightly higher MD. We could not detect the same association between MD and microenvironment subtypes in our data. However, this may be a question of insufficient power. The high degree of fibrosis seen in the *active* subtype does not harmonize with higher MD in the *inactive* samples as was observed by Sun et al., since one would suspect that a high degree of fibrosis would lead to higher density. However, normal fibroblasts may inhibit proliferation of epithelial cells [77], and as mammographic density is a product of both different cell types and ECM constituents, a higher content of epithelial cells in the *inactive* subtype could explain this discrepancy. Additionally, the samples analyzed by Sun et al. were tumor adjacent tissue, while in our study, the biopsies were normal breast tissue from healthy individuals. This is an important distinction, as dynamic interactions between tumor cells, tumor adjacent normal epithelium, and stroma may influence gene expression patterns.

Mammographic density is a comprehensive measurement representing the whole breast and may have limited ability of capturing local differences, which may further explain the lack of association between MD and microenvironment subtype in our study. Also, intra-breast heterogeneity, such as presence of

stem cell niches [78], may explain differences between two biopsies from the same breast. We found, however, a low degree of intra-individual variability of expression of relevant genes in normal breasts using an external dataset, which strengthens our finding of a negative association between MD and *RBL1*-expression.

## Conclusions

This is the first study of gene expression in two normal breast biopsies from the same healthy individuals taken several years apart. We have validated a negative correlation between *RBL1* expression and mammographic density in postmenopausal women, and found that breast tissue samples from women with a large decrease in mammographic density over time sustained higher expression of histone family genes. We also identified the previously defined *active* and *inactive* microenvironment subtypes and characterized their biological differences in normal breast tissue. Our data indicated an association between MD and target genes in the TGF $\beta$ -signaling pathway regardless of microenvironment subtype. This study has identified mechanisms relevant for normal breast tissue biology and MD over time that may be of importance for breast cancer risk and tumor initiation.

**Acknowledgements** We would like to thank the women volunteering for an extra biopsy for their time and cooperation participating in this study, Phuong Vu and Tone Olsen for assistance and support in the lab, and Laxmi Silwal-Pandit for valuable input to the bioinformatic analyses.

**Availability of Data and Material** Gene expression data for the samples from the previous study (MDG1) are deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE18672. The microarray expression data for MDG2 have been deposited in the ArrayExpress database at EMBL-EBI (<http://www.ebi.ac.uk/arrayexpress/>) under accession number E-MTAB-5885.

**Author's contribution** The study was designed by ÅH, GU, MMH and VD. ÅH, VD and TS ensured funding. MMH and VD assisted in data collection. HB performed the laboratory work. GU estimated the amount of mammographic density. HB and TGL performed statistical analyses of the data. HB, TGL, TS and VDH interpreted the results and wrote the paper. All authors read and approved the final manuscript.

**Funding** This work was supported by grants from the South-Eastern Norway Regional Health Authority (2012056 to TS and 2010046 to ÅH) and The Norwegian Cancer Society (794926 to ÅH).

## Compliance with Ethical Standards

**Ethics Approval and Consent to Participate** All women provided a signed informed consent. The study was approved by the Norwegian Regional Committee for Medical and Health Research Ethics, Region South-East (IRB approval no S-02036) and the Protocol Committee of Oslo University Hospital, Oncological department (ref: 2002–15) and finally by the Norwegian Centre for Research Data (NSD) (ref: 2004/00008 CBR/-).

**Competing Interests** All authors declare that they have no potential conflicts of interest.

**Abbreviations** BMI, body mass index; ECM, extracellular matrix; EMT, epithelial-to-mesenchymal transition; FDR, false discovery rate; FSH, follicle-stimulating hormone; GO, gene ontology; LH, luteinizing hormone; MD, mammographic density; MDG, mammographic density and genetics (name of study); RBL1, Retinoblastoma-like protein 1; SD, standard deviation; TGF $\beta$ , Transforming Growth Factor Beta

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

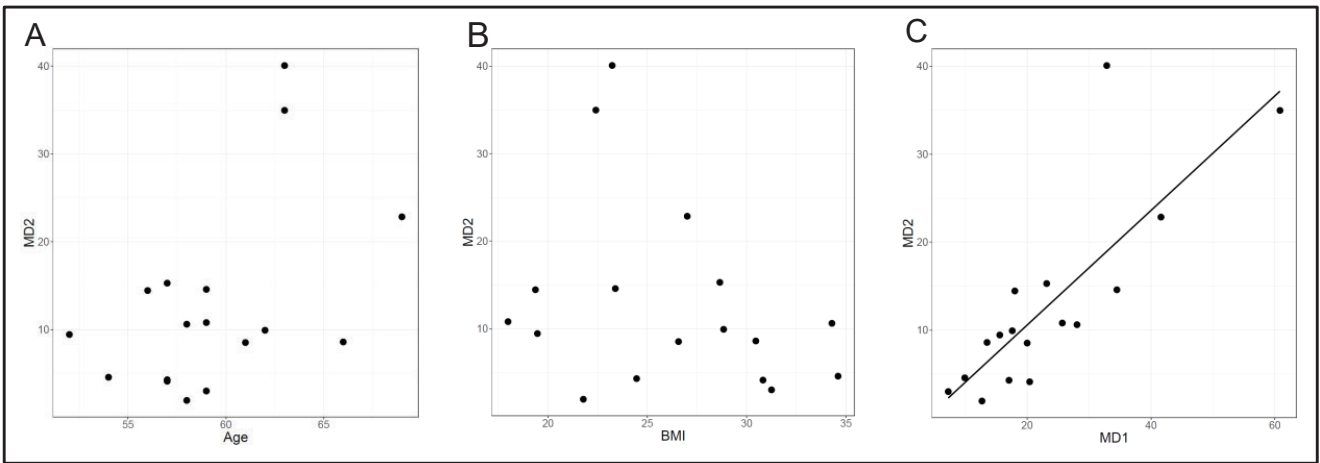
- Bissell MJ, Radisky D. Putting tumours in context. *Nat Rev Cancer*. 2001;1:46–54. <https://doi.org/10.1038/35094059>.
- Boyd N, Berman H, Zhu J, Martin LJ, Yaffe MJ, Chavez S, et al. The origins of breast cancer associated with mammographic density: a testable biological hypothesis. *Breast Cancer Res*. 2018;20(17):17. <https://doi.org/10.1186/s13058-018-0941-y>.
- Yaffe MJ. Mammographic density. Measurement of mammographic density. *Breast Cancer Res*. 2008;10:209. <https://doi.org/10.1186/bcr2102>.
- Martin LJ, Boyd NF. Mammographic density. Potential mechanisms of breast cancer risk associated with mammographic density: hypotheses based on epidemiological evidence. *Breast Cancer Res*. 2008;10:201. <https://doi.org/10.1186/bcr1831>.
- McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiol Biomark Prev*. 2006;15:1159–69. <https://doi.org/10.1158/1055-9965.EPI-06-0034>.
- Pettersson A, Graff RE, Ursin G, Santos Silva ID, McCormack V, Baglietto L, et al. Mammographic density phenotypes and risk of breast Cancer: a meta-analysis. *JNCI J Natl Cancer Inst*. 2014;106:dju078–dju078. <https://doi.org/10.1093/jnci/dju078>.
- Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic density and the risk and detection of breast Cancer. *N Engl J Med*. 2007;356:227–36. <https://doi.org/10.1056/NEJMoa062790>.
- Li T, Sun L, Miller N, Nicklee T, Woo J, Hulse-Smith L, et al. The association of measured breast tissue characteristics with mammographic density and other risk factors for breast cancer. *Cancer Epidemiol Biomark Prev*. 2005;14:343–9. <https://doi.org/10.1158/1055-9965.EPI-04-0490>.
- Cuzick J, Warwick J, Pinney E, Duffy SW, Cawthorn S, Howell A, et al. Tamoxifen-induced reduction in mammographic density and breast Cancer risk reduction: a nested case-control study. *JNCI J Natl Cancer Inst*. 2011;103:744–52. <https://doi.org/10.1093/jnci/djr079>.
- Li J, Humphreys K, Eriksson L, Edgren G, Czene K, Hall P. Mammographic density reduction is a prognostic marker of response to adjuvant Tamoxifen therapy in postmenopausal patients with breast Cancer. *J Clin Oncol*. 2013;31:2249–56. <https://doi.org/10.1200/JCO.2012.44.5015>.
- Pike MC, Krailo MD, Henderson BE, Casagrande JT, Hoel DG. “Hormonal” risk factors, “breast tissue age” and the age-incidence of breast cancer. *Nature*. 1983;303:767–70.
- Ziv E, Shepherd J, Smith-Bindman R, Kerlikowske K. Mammographic breast density and family history of breast cancer. *J Natl Cancer Inst*. 2003;95:556–8. <https://doi.org/10.1093/JNCI/95.7.556>.
- Boyd NF, Dite GS, Stone J, Gunasekara A, English DR, McCredie MRE, et al. Heritability of mammographic density, a risk factor for breast Cancer. *N Engl J Med*. 2002;347:886–94. <https://doi.org/10.1056/NEJMoa013390>.
- Couto E, Qureshi SA, Hofvind S, Hilsen M, Aase H, Skaane P, et al. Hormone therapy use and mammographic density in postmenopausal Norwegian women. *Breast Cancer Res Treat*. 2012;132:297–305. <https://doi.org/10.1007/s10549-011-1810-x>.
- Knight JA, Martin LJ, Greenberg CV, Lockwood GA, Byng JW, Yaffe MJ, et al. Macronutrient intake and change in mammographic density at menopause: results from a randomized trial. *Cancer Epidemiol Biomark Prev*. 1999;8:123–8.
- Ursin G, Sun C-L, Koh W-P, Khoo K-S, Gao F, Wu AH, et al. Associations between soy, diet, reproductive factors, and mammographic density in Singapore Chinese women. *Nutr Cancer*. 2006;56:128–35. [https://doi.org/10.1207/s15327914nc5602\\_2](https://doi.org/10.1207/s15327914nc5602_2).
- Boyd NF, Martin LJ, Sun L, Guo H, Chiarelli A, Hislop G, et al. Body size, mammographic density, and breast Cancer risk. *Cancer Epidemiol Biomark Prev*. 2006;15:2086–92. <https://doi.org/10.1158/1055-9965.EPI-06-0345>.
- Checka CM, Chun JE, Schnabel FR, Lee J, Toth H. The relationship of mammographic density and age: implications for breast Cancer screening. *Am J Roentgenol*. 2012;198:W292–5. <https://doi.org/10.2214/AJR.10.6049>.
- Sterns EE, Zee B. Mammographic density changes in perimenopausal and postmenopausal women: is effect of hormone replacement therapy predictable? *Breast Cancer Res Treat*. 2000;59:125–32. <https://doi.org/10.1023/A:1006326432340>.
- Boyd N, Martin L, Stone J, Little L, Minkin S, Yaffe M. A longitudinal study of the effects of menopause on mammographic features. *Cancer Epidemiol Biomark Prev*. 2002;11:1048–53.
- Sun X, Gierach GL, Sandhu R, Williams T, Midkiff BR, Lissowska J, et al. Relationship of mammographic density and gene expression: analysis of normal breast tissue surrounding breast cancer. *Clin Cancer Res*. 2013;19:4972–82. <https://doi.org/10.1158/1078-0432.CCR-13-0029>.
- Haakensen VD, Biong M, Lingjærde OC, Holmen MM, Frantzen JO, Chen Y, et al. Expression levels of uridine 5'-diphosphoglucuronosyltransferase genes in breast tissue from healthy women are associated with mammographic density. *Breast Cancer Res*. 2010;12:R65. <https://doi.org/10.1186/bcr2632>.
- Haakensen VD, Lingjærde OC, Lüders T, Riis M, Prat A, Troester MA, et al. Gene expression profiles of breast biopsies from healthy women identify a group with claudin-low features. *BMC Med Genet*. 2011;4(77). <https://doi.org/10.1186/1755-8794-4-77>.
- Lapenna S, Giordano A. Cell cycle kinases as therapeutic targets for cancer. *Nat Rev Drug Discov*. 2009;8:547–66. <https://doi.org/10.1038/nrd2907>.
- Indovina P, Marcelli E, Casini N, Rizzo V, Giordano A. Emerging roles of RB family: new defense mechanisms against tumor progression. *J Cell Physiol*. 2013;228:525–35. <https://doi.org/10.1002/jcp.24170>.
- Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144:646–74. <https://doi.org/10.1016/j.cell.2011.02.013>.
- Place AE, Jin Huh S, Polyak K. The microenvironment in breast cancer progression: biology and implications for treatment. *Breast Cancer Res*. 2011;13:227. <https://doi.org/10.1186/bcr2912>.
- Román-Pérez E, Casbas-Hernández P, Pirone JR, Rein J, Carey LA, Lubet RA, et al. Gene expression in extratumoral

- microenvironment predicts clinical outcome in breast cancer patients. *Breast Cancer Res.* 2012;14:R51. <https://doi.org/10.1186/bcr3152>.
29. Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol.* 2007;8:R76. <https://doi.org/10.1186/gb-2007-8-5-r76>.
  30. Haakensen VD, Bjørø T, Lüders T, Riis M, Bukholm IK, Kristensen VN, et al. Serum estradiol levels associated with specific gene expression patterns in normal breast tissue and in breast carcinomas. *BMC Cancer.* 2011;11:332. <https://doi.org/10.1186/1471-2407-11-332>.
  31. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;30:207–10. <https://doi.org/10.1093/NAR/30.1.207>.
  32. Gene Expression Omnibus Series GSE18672. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18672>
  33. Gene Expression Omnibus Series GSE72644. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72644>. Accessed 29 Aug 2018.
  34. Gene Expression Omnibus Series GSE4823. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4823>. Accessed 3 Sep 2018.
  35. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. *Nucleic Acids Res.* 2015;43:D1113–6. <https://doi.org/10.1093/nar/gku1057>.
  36. ArrayExpress Database. Accession number E-MTAB-5885. <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5885>
  37. Ursin G, Astrahan MA, Salane M, Parisky YR, Pearce JG, Daniels JR, et al. The detection of changes in mammographic densities. *Cancer Epidemiol Biomark Prev.* 1998;7:43–7.
  38. Creighton CJ, Casa A, Lazard Z, Huang S, Tsimelzon A, Hilsenbeck SG, et al. Insulin-like growth factor-I activates gene transcription programs strongly associated with poor breast cancer prognosis. *J Clin Oncol.* 2008;26:4078–85. <https://doi.org/10.1200/JCO.2007.13.4429>.
  39. Team RS. RStudio: integrated development for R. Boston: RStudio, Inc; 2016.
  40. Gendoo DMA, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. Genefu: an R/bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics.* 2016;32:1097–9. <https://doi.org/10.1093/bioinformatics/btv693>.
  41. Champely S Basic Functions for Power Analysis [R package pwr version 1.2–1].
  42. Skippage P, Wilkinson L, Allen S, Roche N, Dowsett M, a'Hern R. Correlation of age and HRT use with breast density as assessed by Quantra™. *Breast J.* 2013;19:79–86. <https://doi.org/10.1111/tbj.12046>.
  43. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2008;4:44–57. <https://doi.org/10.1038/nprot.2008.211>.
  44. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009;37:1–13. <https://doi.org/10.1093/nar/gkn923>.
  45. 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. *Clin Cancer Res.* 2010;16:5222–32. <https://doi.org/10.1158/1078-0432.CCR-10-1282>.
  46. Beck AH, Espinosa I, Gilks CB, van de Rijn M, West RB. The fibromatosis signature defines a robust stromal response in breast carcinoma. *Lab Invest.* 2008;88:591–601. <https://doi.org/10.1038/labinvest.2008.31>.
  47. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci.* 2005;102:15545–50. <https://doi.org/10.1073/pnas.0506580102>.
  48. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database Hallmark gene set collection. *Cell Syst.* 2015;1:417–25. <https://doi.org/10.1016/j.cels.2015.12.004>.
  49. 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 Res.* 2010;12:R68. <https://doi.org/10.1186/bcr2635>.
  50. Dang CV. MYC on the path to cancer. *Cell.* 2012;149:22–35. <https://doi.org/10.1016/j.cell.2012.03.003>.
  51. Tissue expression of RBL1 - Staining in breast - The Human Protein Atlas. <https://www.proteinatlas.org/ENSG00000080839-RBL1/tissue/breast>. Accessed 12 Nov 2018.
  52. Tissue expression of MYC - Staining in breast - The Human Protein Atlas. <https://www.proteinatlas.org/ENSG00000136997-MYC/tissue/breast>. Accessed 12 Nov 2018.
  53. Tissue expression of PDGFA - Staining in breast - The Human Protein Atlas. <https://www.proteinatlas.org/ENSG00000197461-PDGFA/tissue/breast>. Accessed 12 Nov 2018.
  54. O'Connor RJ, Schaley JE, Feeney G, Hearing P. The p107 tumor suppressor induces stable E2F DNA binding to repress target promoters. *Oncogene.* 2001;20:1882–91. <https://doi.org/10.1038/sj.onc.1204278>.
  55. Di Fiore R, D'Anneo A, Tesoriere G, Vento R. RB1 in cancer: different mechanisms of RB1 inactivation and alterations of pRb pathway in tumorigenesis. *J Cell Physiol.* 2013;228:1676–87. <https://doi.org/10.1002/jcp.24329>.
  56. Chen C-R, Kang Y, Siegel PM, Massagué J. E2F4/5 and p107 as Smad cofactors linking the TGFβ receptor to c-myc repression. *Cell.* 2002;110:19–32. [https://doi.org/10.1016/S0092-8674\(02\)00801-2](https://doi.org/10.1016/S0092-8674(02)00801-2).
  57. Ikushima H, Miyazono K. TGFβ signalling: a complex web in cancer progression. *Nat Rev Cancer.* 2010;10:415–24. <https://doi.org/10.1038/nrc2853>.
  58. Papageorgis P, Stylianopoulos T. Role of TGFβ in regulation of the tumor microenvironment and drug delivery (review). *Int J Oncol.* 2015;46:933–43. <https://doi.org/10.3892/ijo.2015.2816>.
  59. Moses H, Barcellos-Hoff MH. TGF-beta biology in mammary development and breast cancer. *Cold Spring Harb Perspect Biol.* 2011;3:a003277. <https://doi.org/10.1101/cshperspect.a003277>.
  60. Yang WT, Lewis MT, Hess K, Wong H, Tsimelzon A, Karadag N, et al. Decreased TGFβ signaling and increased COX2 expression in high risk women with increased mammographic breast density. *Breast Cancer Res Treat.* 2010;119:305–14. <https://doi.org/10.1007/s10549-009-0350-0>.
  61. Egeblad M, Rasch MG, Weaver VM. Dynamic interplay between the collagen scaffold and tumor evolution. *Curr Opin Cell Biol.* 2010;22:697–706. <https://doi.org/10.1016/j.ccb.2010.08.015>.
  62. Ironside AJ, Jones JL. Stromal characteristics may hold the key to mammographic density: the evidence to date. *Oncotarget.* 2016;7:31550–62. <https://doi.org/10.18632/oncotarget.6912>.
  63. Huo CW, Chew G, Hill P, Huang D, Ingman W, Hodson L, et al. High mammographic density is associated with an increase in stromal collagen and immune cells within the mammary epithelium. *Breast Cancer Res.* 2015;17(79):79. <https://doi.org/10.1186/s13058-015-0592-1>.
  64. Tang B, Vu M, Booker T, Santner SJ, Miller FR, Anver MR, et al. TGF-beta switches from tumor suppressor to prometastatic factor in

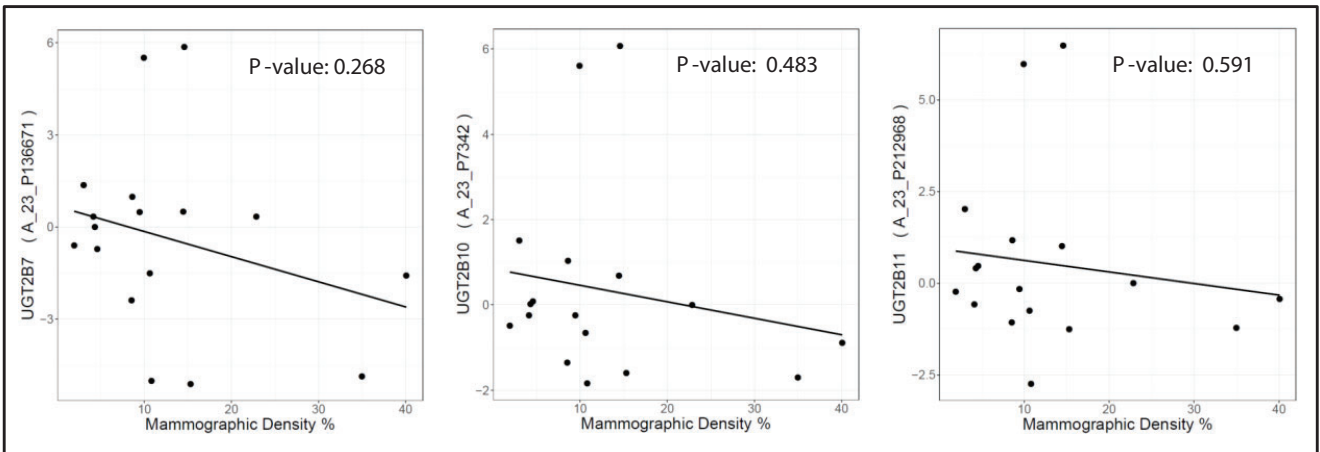
- a model of breast cancer progression. *J Clin Invest*. 2003;112:1116–24. <https://doi.org/10.1172/JCI18899>.
65. Kaldis P, Pagano M. Wnt Signaling in Mitosis. *Dev Cell*. 2009;17:749–50. <https://doi.org/10.1016/j.devcel.2009.12.001>.
  66. Christov K, Chew KL, Ljung BM, Waldman FM, Duarte LA, Goodson WH, et al. Proliferation of normal breast epithelial cells as shown by in vivo labeling with bromodeoxyuridine. *Am J Pathol*. 1991;138:1371–7.
  67. Cherlet T, Murphy LC. Estrogen receptors inhibit Smad3 transcriptional activity through Ap-1 transcription factors. *Mol Cell Biochem*. 2007;306:33–42. <https://doi.org/10.1007/s11010-007-9551-1>.
  68. Matsuda T, Yamamoto T, Muraguchi A, Saatcioglu F. Cross-talk between transforming growth factor-beta and estrogen receptor signaling through Smad3. *J Biol Chem*. 2001;276:42908–14. <https://doi.org/10.1074/jbc.M105316200>.
  69. Pirone JR, D'Arcy M, Stewart DA, Hines WC, Johnson M, Gould MN, et al. Age-associated gene expression in normal breast tissue mirrors qualitative age-at-incidence patterns for breast cancer. *Cancer Epidemiol Biomark Prev*. 2012;21:1735–44. <https://doi.org/10.1158/1055-9965.EPI-12-0451>.
  70. Atashgaran V, Wrin J, Barry SC, Dasari P, Ingman WV. Dissecting the biology of menstrual cycle-associated breast Cancer risk. *Front Oncol*. 2016;6:267. <https://doi.org/10.3389/fonc.2016.00267>.
  71. Feser J, Truong D, Das C, Carson JJ, Kieft J, Harkness T, et al. Elevated histone expression promotes life span extension. *Mol Cell*. 2010;39:724–35. <https://doi.org/10.1016/j.molcel.2010.08.015>.
  72. Dias K, Dvorkin-Gheva A, Hallett RM, Wu Y, Hassell J, Pond GR, et al. Claudin-low breast cancer; clinical & pathological characteristics. *Clin Pathol Charact PLoS ONE*. 2017;12. <https://doi.org/10.1371/journal.pone.0168669>.
  73. Pohl S-G, Brook N, Agostino M, Arfuso F, Kumar A, Dharmarajan A. Wnt signaling in triple-negative breast cancer. *Oncogenesis*. 2017;6:e310. <https://doi.org/10.1038/oncsis.2017.14>.
  74. Ma I, Allan AL. The role of human aldehyde dehydrogenase in Normal and Cancer stem cells. *Stem Cell Rev Reports*. 2011;7:292–306. <https://doi.org/10.1007/s12015-010-9208-4>.
  75. Cheng F, Shen Y, Mohanasundaram P, Lindström M, Ivaska J, Ny T, et al. Vimentin coordinates fibroblast proliferation and keratinocyte differentiation in wound healing via TGF- $\beta$ -slug signaling. *Proc Natl Acad Sci U S A*. 2016;113:E4320–7. <https://doi.org/10.1073/pnas.1519197113>.
  76. Klass BR, Grobelaar AO, Rolfe KJ. Transforming growth factor beta 1 signalling, wound healing and repair: a multifunctional cytokine with clinical implications for wound repair, a delicate balance. *Postgrad Med J*. 2009;85:9–14. <https://doi.org/10.1136/pgmj.2008.069831>.
  77. Sadlonova A, Bowe DB, Novak Z, Mukherjee S, Duncan VE, Page GP, et al. Identification of molecular distinctions between Normal breast-associated fibroblasts and breast Cancer-associated fibroblasts. *Cancer Microenviron*. 2009;2:9–21. <https://doi.org/10.1007/s12307-008-0017-0>.
  78. Roberts KJ, Kershner AM, Beachy PA. The stromal niche for epithelial stem cells: a template for regeneration and a brake on malignancy. *Cancer Cell*. 2017;32:404–10. <https://doi.org/10.1016/j.ccell.2017.08.007>.

# Online resource 1: Figure S1-S6

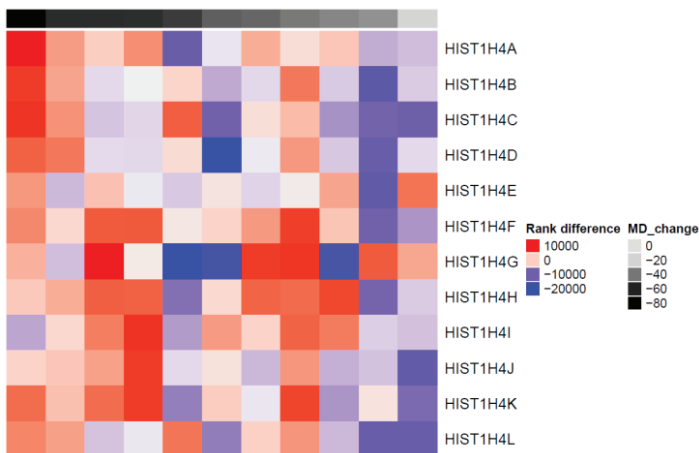
## Figure S1



## Figure S2



## Figure S3



## Figure S4

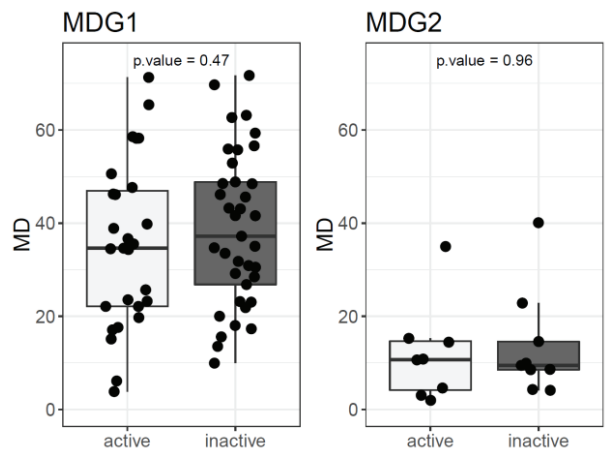




Figure S5

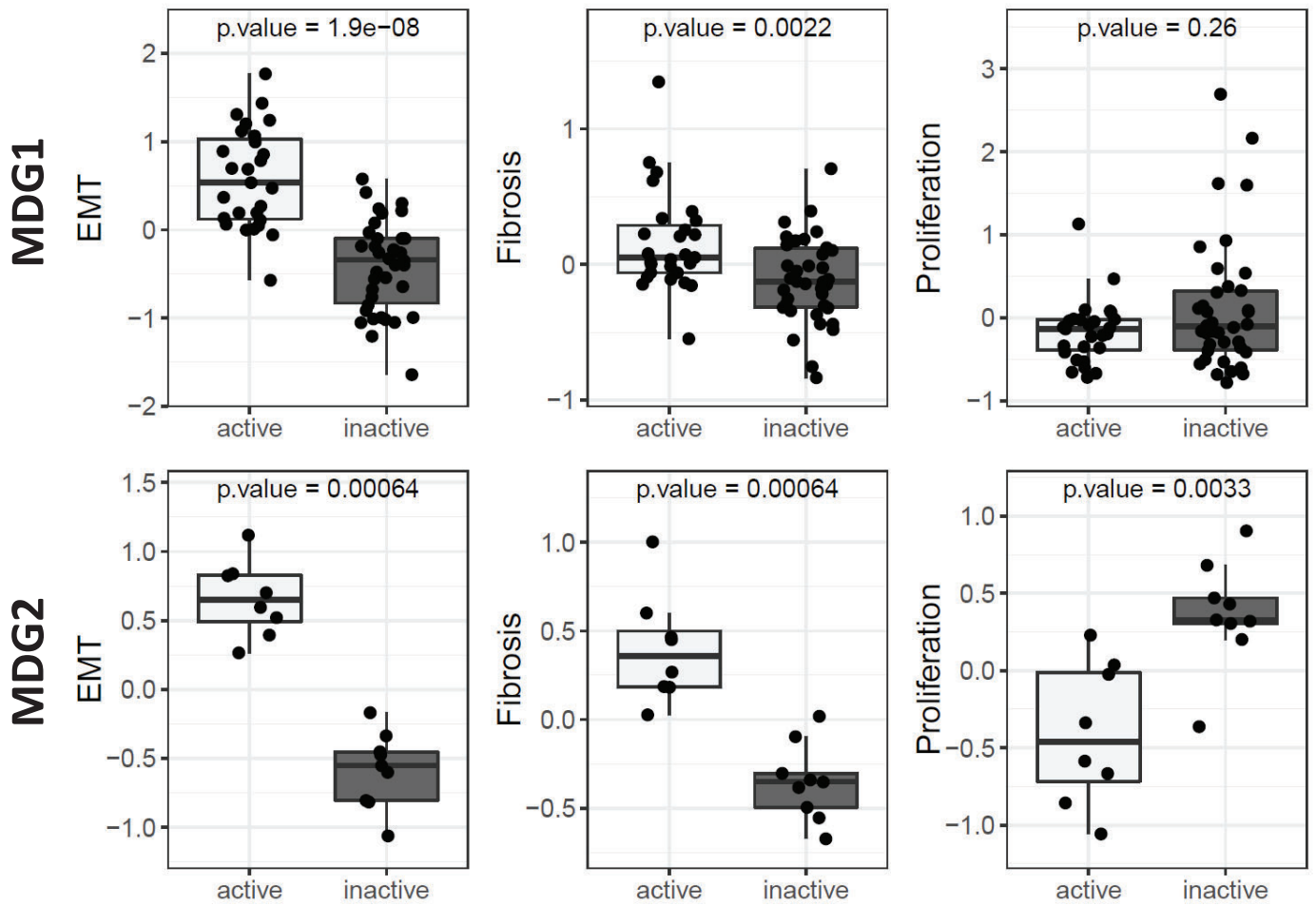
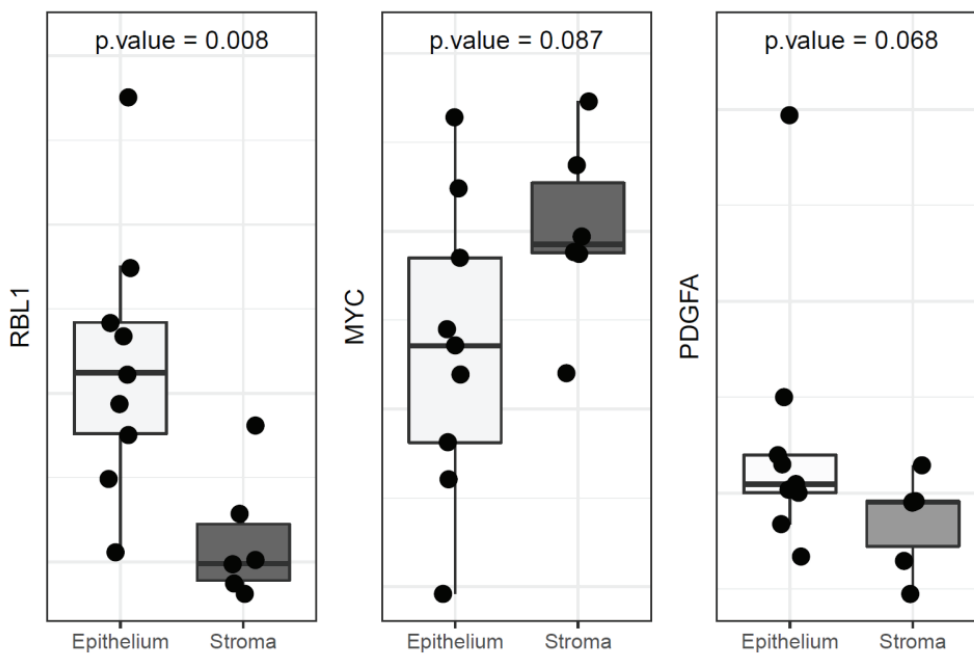


Figure S6



Online resource 2-6 are large tables that are made available electronically



## Paper II

### **Claudin-low-like mouse mammary tumors show distinct transcriptomic patterns uncoupled from genomic drivers**

Christian Fougner, Helga Bergholtz, Raoul Kuiper, Jens Henrik Norum and Therese Sørliie.

*Breast Cancer Research 2019. 21, 85.*



RESEARCH ARTICLE

Open Access



# Claudin-low-like mouse mammary tumors show distinct transcriptomic patterns uncoupled from genomic drivers

Christian Fougner<sup>1</sup>, Helga Bergholtz<sup>1</sup>, Raoul Kuiper<sup>2</sup>, Jens Henrik Norum<sup>1</sup> and Therese Sørlie<sup>1,3,4\*</sup> 

## Abstract

**Background:** Claudin-low breast cancer is a molecular subtype associated with poor prognosis and without targeted treatment options. The claudin-low subtype is defined by certain biological characteristics, some of which may be clinically actionable, such as high immunogenicity. In mice, the medroxyprogesterone acetate (MPA) and 7, 12-dimethylbenzanthracene (DMBA)-induced mammary tumor model yields a heterogeneous set of tumors, a subset of which display claudin-low features. Neither the genomic characteristics of MPA/DMBA-induced claudin-low tumors nor those of human claudin-low breast tumors have been thoroughly explored.

**Methods:** The transcriptomic characteristics and subtypes of MPA/DMBA-induced mouse mammary tumors were determined using gene expression microarrays. Somatic mutations and copy number aberrations in MPA/DMBA-induced tumors were identified from whole exome sequencing data. A publicly available dataset was queried to explore the genomic characteristics of human claudin-low breast cancer and to validate findings in the murine tumors.

**Results:** Half of MPA/DMBA-induced tumors showed a claudin-low-like subtype. All tumors carried mutations in known driver genes. While the specific genes carrying mutations varied between tumors, there was a consistent mutational signature with an overweight of T>A transversions in TG dinucleotides. Most tumors carried copy number aberrations with a potential oncogenic driver effect. Overall, several genomic events were observed recurrently; however, none accurately delineated claudin-low-like tumors. Human claudin-low breast cancers carried a distinct set of genomic characteristics, in particular a relatively low burden of mutations and copy number aberrations. The gene expression characteristics of claudin-low-like MPA/DMBA-induced tumors accurately reflected those of human claudin-low tumors, including epithelial-mesenchymal transition phenotype, high level of immune activation, and low degree of differentiation. There was an elevated expression of the immunosuppressive genes *PTGS2* (encoding COX-2) and *CD274* (encoding PD-L1) in human and murine claudin-low tumors.

**Conclusions:** Our findings show that the claudin-low breast cancer subtype is not demarcated by specific genomic aberrations, but carries potentially targetable characteristics warranting further research.

**Keywords:** Breast cancer, Claudin-low, Subtypes, Genomics, Transcriptomics, Mouse models, DMBA, MPA

\* Correspondence: therese.sorlie@rr-research.no

<sup>1</sup>Department of Cancer Genetics, Oslo University Hospital, Oslo, Norway

<sup>3</sup>Centre for Cancer Biomarkers CCBIO, University of Bergen, Bergen, Norway

Full list of author information is available at the end of the article



© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

## Background

The claudin-low subtype of breast cancer (BC) is a distinct disease entity associated with a relatively poor prognosis, and with an inadequately understood clinical significance [1–3]. It is characterized by low expression of tight junction and cell-cell adhesion genes, low degree of differentiation, epithelial-mesenchymal transition (EMT) phenotype, and high level of immune cell infiltration [2]. The claudin-low subtype represents 7–14% of all breast cancers, and despite its unique biological features, there are no therapies specifically targeting the subtype [2–5]. While claudin-low tumors are found in several large-scale studies, there is a paucity of information regarding their specific genomic characteristics [6–9]. Thus, significant gaps remain in the understanding of the biology of claudin-low tumors, and there is a need for further research to explore how their unique features may be therapeutically targeted.

Accurate preclinical models are vital for research into novel treatment options. Mouse mammary tumors may be induced through exposure to medroxyprogesterone acetate (MPA) and 7,12-dimethylbenzanthracene (DMBA) [10]. The tumors generated by this protocol are diverse, and a subset of these show similarities to the human claudin-low subtype [11, 12]. A homogeneous primary in vivo model of claudin-low breast cancer does not currently exist [11]. While the mechanisms of MPA [10, 13] and DMBA [14–17] have been described, there is still contention regarding the suitability of a chemically induced model of cancer for a disease that is not primarily caused by carcinogens in humans [18]. Evaluating the claudin-low subset of MPA/DMBA-induced tumors as a model for human disease is therefore an important step toward advancing preclinical research of claudin-low breast cancer.

In this study, we identified and comprehensively characterized claudin-low-like mouse mammary tumors generated by MPA/DMBA-induced carcinogenesis. Through genomic and transcriptomic analyses, we evaluated these tumors as a model for human claudin-low breast cancer and showed these tumors to be phenotypically accurate representations of their human counterparts. In parallel, we analyzed the previously unexplored genomic features of human claudin-low breast cancer. Our findings highlighted several features of claudin-low breast cancer with potential therapeutic implications, including a low tumor mutational burden, high expression of the immune checkpoint gene *CD274* (encoding PD-L1), and high expression of *PTGS2* (encoding cyclooxygenase-2).

## Methods

### Mouse strains and tumor induction

Double transgenic mice, *Lgr5-EGFP-Ires-CreERT2;R26R-Confetti* [19], were generated by crossing heterozygous *Lgr5-EGFP-Ires-CreERT2* mice with heterozygous *R26R-*

*Confetti* mice. These transgenes are considered biologically inert and all female offspring, including wild type, single, or double transgenic mice, were used for MPA/DMBA-treatment experiments. All mice were locally bred and maintained within a specific pathogen-free barrier facility according to local and national regulations, with food and water ad libitum. Female mice were treated with medroxyprogesterone acetate (MPA) and 7, 12-dimethylbenzanthracene (DMBA) in accordance with the established protocol [10]. In brief, 90-day release MPA pellets (50 mg/pellet, Innovative Research of America cat.# NP-161) were implanted subcutaneously at 6 and 19 weeks after birth. One microgram of DMBA (Sigma Aldrich cat.# D3254) dissolved in corn oil (Sigma Aldrich cat.# C8267) was administered by oral gavage at 9, 10, 12, and 13 weeks after birth. Tumor growth was regularly monitored by manual palpation and measured by a caliper. Tumor volume was estimated using the following formula: volume = (width<sup>2</sup> × length)/2. When the tumors reached the maximum allowed size of 1000 mm<sup>3</sup>, or at the age of 32 weeks, tissue was collected at necropsy and fixed in 4% paraformaldehyde (PFA) or snap frozen and stored at –80 °C. Eighteen tumors from 14 mice, of which four mice carried two mammary tumors, were subject to genomic and transcriptomic analyses. Six normal mammary glands collected from mice not undergoing MPA/DMBA treatment were included as controls. Mouse features and histopathological tumor features can be found in Additional file 1.

### Histopathology and immunohistochemistry

Mouse tissue was fixed overnight in 4% PFA, routinely processed and paraffin embedded. Formalin-fixed paraffin-embedded tissue was sectioned and stained with hematoxylin and eosin (HE). HE-stained tissue was classified by a certified veterinary pathologist. Immunohistochemical staining was performed as previously described [20] with primary antibodies against K5 (Covance cat.# PRB-160P), K18 (Progen cat.# 61028), Ki67 (Novocastra cat.# NCL-Ki67p), ERα (Millipore cat.# 06-935), PR (Abcam cat.# ab131486), and Her2/Erbb2 (Millipore cat.# 06-562).

### DNA and RNA isolation

DNA isolation for exome sequencing was carried out at Theragen Etx Bio Institute (Seoul, South Korea). DNA was isolated using QIAamp DNA Mini Kit (Qiagen cat.# 51306) per the manufacturer's protocol. DNA from two samples (*S159\_14\_11* and *S176\_14\_11*) was isolated using CTAB Extraction Solution (Biosesang cat.# C2007) per the manufacturer's protocol. DNA integrity was assessed by electrophoresis, and concentration was determined using the Nanodrop ND-1000 spectrophotometer (Thermo Scientific cat.# ND-1000) and Qubit fluorometer (Thermo Scientific cat.# Q33226). Total RNA and DNA

isolation for gene expression microarrays was carried out using the QIAcube system (Qiagen cat.# 9001292) with the AllPrep DNA/RNA Universal Kit (Qiagen cat.# 80224) according to the protocol provided by the supplier, with 30-mg tissue as input. The tissue was manually minced with a scalpel on ice followed by lysis and homogenization using TissueLyzer LT (Qiagen cat.# 85600) and Qiashredder (Qiagen cat.# 79654), respectively. Nucleic acid concentrations were measured by NanoDrop ND-1000 spectrophotometer, and RNA integrity was analyzed using Agilent 2100 Bioanalyzer (Agilent Technologies cat.# G2939BA).

#### Gene expression microarrays

Gene expression profiling was performed using RNA isolated from 18 snap-frozen MPA/DMBA-induced tumors and six normal/untreated mouse mammary gland samples. Whole genome expression data was obtained using Agilent SurePrint G3 Mouse Gene Expression 8x60K microarrays (Agilent Technologies cat.# G4852B) with Low Input Quick Amp Labeling protocol (Agilent Technologies cat.# 5190-2331) and the Cy3 fluorophore. Forty nanogram RNA was used for input. Microarrays were scanned using an Agilent SureScan Microarray Scanner (Agilent Technologies cat.# G4900DA), and data was extracted using Agilent Feature Extraction software. One tumor sample (*S422\_15\_2*) failed quality control and was excluded from further gene expression analyses.

#### Gene expression analyses

Gene expression data was analyzed using Qlucore Omics Explorer 3.2 (Qlucore AB) and R version 3.3.2 [21]. Gene expression values were quantile normalized, and probes with a standard deviation of less than 2.8% of the largest observed standard deviation were filtered out. For genes represented by more than one probe, mean expression values were calculated to obtain one gene expression value per gene. Principal component analysis was performed to assess data quality, and one normal mammary gland sample (*S178\_14\_2*) was identified as an outlier and removed from further analysis. Murine subtypes were determined by first calculating centroids for each subtype using the original data from Pfeifferle et al. [11], followed by calculating Spearman correlation for every sample to each of the subtype centroids. The subtype with the highest correlation coefficient was assigned as the sample's subtype. Two tumor clusters were identified by hierarchical clustering using the murine intrinsic gene list [11], and SigClust [22] was used to test the significance of the difference between the clusters.

Unsupervised hierarchical clustering was performed using average linkage and Spearman correlation as the distance metric. Immune cell infiltration was inferred using ESTIMATE [23]. Scores for gene signatures relevant to

the claudin-low subtype (adhesion, EMT, luminalness, proliferation, vascular content, immunosuppression, and interferons [2, 24–27]) were calculated using a standard (*Z*) score approach: for every gene in each signature, a standardized expression value was calculated by subtracting the mean across all samples, then dividing by the standard deviation. Calculation of the mean of the standardized expression values across all genes in the signature yielded the score. Gene lists included in the different signatures are found in Additional file 2. The degree of differentiation was calculated using a differentiation predictor [2]. Two-tailed Wilcoxon rank-sum tests were used for statistical testing of differences in scores between two groups.

#### Whole exome sequencing

Whole exome sequencing was carried out at Theragen Etx Bio Institute. Library preparation and target enrichment was carried out using the SureSelect XT Mouse All Exon Kit (Agilent cat.# 5190-4641) per the manufacturer's instructions. Sequencing was performed on an Illumina HiSeq 2500 (Illumina cat.# SY-401-2501). DNA was sequenced to an average depth of 58. Quality control was performed with FastQC [28].

#### Sequence alignment and processing

Adapter sequences were removed using CutAdapt, version 1.10 [29]. Low-quality reads were trimmed using Sickle version 1.33 [30], in paired end mode with quality threshold set to 20 and length threshold set to 50 base pairs. Reads were aligned to the mm10 reference genome using the Burrows-Wheeler MEM aligner (BWA-MEM), version 0.7.12 [31]. Following alignment, duplicate reads were marked using Picard (<https://broadinstitute.github.io/picard/>) version 2.0.1. Base quality scores were then recalibrated using GATK version 3.6.0 [32–34]. Lists of known single nucleotide polymorphisms and indels for the FVB/N mouse strain were downloaded from the Mouse Genomes Project, dbSNP release 142, and used for base quality score recalibration and mutation filtering [35].

#### Mutation calling and analysis

Somatic mutations were called using the MuTect2 algorithm in GATK [32–34] with a minimum allowed base quality score of 20. Mutations were filtered against variants found in matched normal liver tissue and known single nucleotide polymorphisms for the FVB/N mouse strain. Candidate somatic mutations which did not pass the standard MuTect2 filters were removed from further analysis. Mutations not meeting the following requirements were also removed from further analysis: minimum allele depth of 10, minimum allele frequency of 0.05, and presence of the mutation in both forward and reverse strands. Mutations were annotated using SnpEff

[36] and filtered for downstream analysis using SnpSift [37]. Candidate driver mutations were defined as moderate or high impact mutations, as defined by SnpEff, in driver genes as identified by the COSMIC cancer gene census [38]. To identify hotspot mutations, mouse amino acid positions were aligned to the orthologous human amino acid position using Clustal Omega [39] through UniProtKB [40] and used to query mutations found in the COSMIC database [38]. Mutational spectrum and signature analysis was performed using the deconstructSigs framework [41] modified to allow the use of the mm10 mouse reference genome. The COSMIC mutational signatures were used for reference [42].

### Copy number aberration analyses

Copy number aberrations were identified from exome sequence data using EXCAVATOR2 [43] using the mm10 reference genome. CNA calling was performed using standard settings and a window size of 20000 bp. Potential driver CNAs were identified by filtering for CNAs associated with cancer in the COSMIC cancer gene census [38].

### Analyses of human breast cancer data

Processed data from the METABRIC [6, 7] and TCGA [44] cohorts were downloaded from or analyzed directly on the cBioportal platform [45, 46].

### Plot generation

Plots were created using R version 3.3.2 [21]. Heatmaps were created using ComplexHeatmap [47]. Mutational spectrum histograms were created using the deconstructSigs package [41]. All other plots were generated using the ggplot2 package [48].

## Results

### Gene expression subtyping reveals two distinct tumor clusters

We determined the murine transcriptomic subtypes of 17 MPA/DMBA-induced mammary tumors from 13 mice (Additional file 1) by calculating each tumor's Spearman correlation to the murine subtype centroids [11]. This revealed nine murine subtypes in the cohort (Table 1, Additional file 3), which separated into two distinct clusters upon hierarchical clustering (Fig. 1,  $p = 0.044$ , SigClust [22]). One cluster consisted of claudin-low<sup>Ex</sup> and squamous-like<sup>Ex</sup> tumors, both of which have been shown to resemble the human claudin-low subtype [11]; this is therefore referred to as the claudin-low-like cluster. The other cluster contained tumors from seven different subtypes and is referred to as the mixed cluster. In four instances, two tumors from different mammary glands were harvested from the same mouse. These were classified as

**Table 1** Subtype distribution of MPA/DMBA-induced tumors and normal mouse mammary gland tissue

No. of samples	Murine subtype	Cluster
6	Claudin-low <sup>Ex</sup>	Claudin-low-like
2	Squamous-like <sup>Ex</sup>	Claudin-low-like
3	PyMT <sup>Ex</sup>	Mixed
1	Class3 <sup>Ex</sup>	Mixed
1	Class8 <sup>Ex</sup>	Mixed
1	Class14 <sup>Ex</sup>	Mixed
1	ErbB2-like <sup>Ex</sup>	Mixed
1	Wnt1-Early <sup>Ex</sup>	Mixed
1	Wnt1-Late <sup>Ex</sup>	Mixed
5 (normal mammary)	Normal <sup>Ex</sup>	Normal

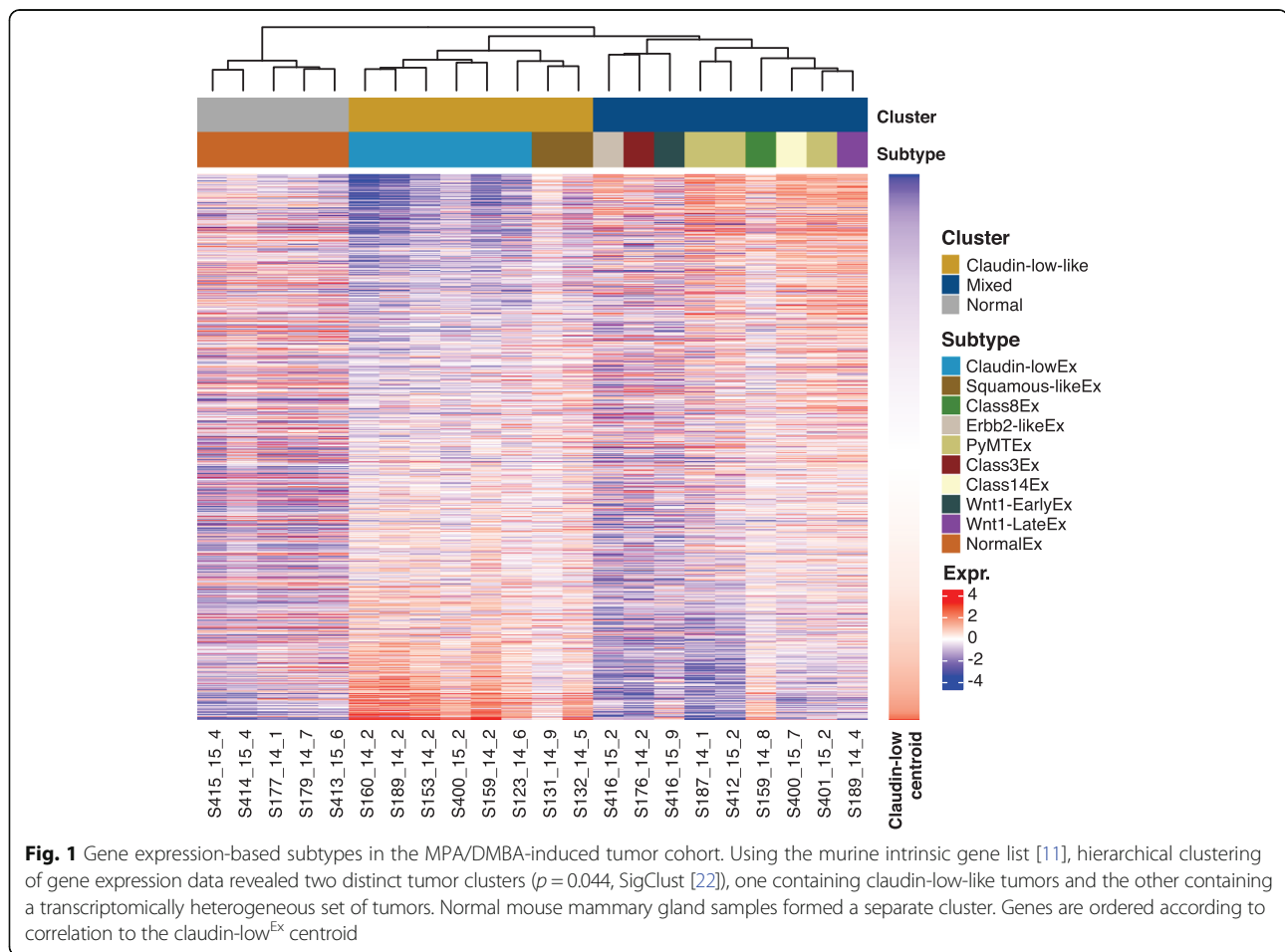
different subtypes in all cases and are presumed to be distinct primary tumors. All normal mammary gland samples were classified as normal-like<sup>Ex</sup> and clustered separately from the tumors.

Histopathological analysis corroborated the intertumor heterogeneity that was demonstrated by subtyping (Additional file 1). Five of the eight claudin-low-like tumors, including both squamous-like<sup>Ex</sup> tumors, showed a squamous appearance, while no tumors in the mixed cluster displayed this histological phenotype ( $p = 0.009$ , Fisher's exact test). There was also a higher frequency of claudin-low-like tumors showing marked neutrophil infiltration ( $p = 0.002$ , Fisher's exact test) and displaying a marked or partial spindle appearance ( $p = 0.050$ , Fisher's exact test) compared to tumors in the mixed cluster.

### Mutations in MPA/DMBA-induced mammary tumors are independent of gene expression subtype

To determine the genetic characteristics of the tumors, we performed exome sequencing to a mean depth of 58, with 84% of bases being sequenced to a coverage of 20× or higher. We identified a mean of 589 mutations per tumor (range 288 to 1795), corresponding to a mean mutation rate of 11.9 mutations per megabase (range 5.8 to 36.2) (Fig. 2a). This was substantially higher than the average 1.3 mutations per megabase found in human breast cancer [49]. The mutational rate in MPA/DMBA-induced mammary tumors was also relatively high when compared to other chemically induced murine tumors (range 1.4 to 13.0 mutations per megabase) [50–52] and when compared to tumors arising in genetically engineered mouse models (range 0.1 to 0.7 mutations per megabase) [52–57]. There was no significant difference in mutational burden between the tumors in the claudin-low-like and the mixed cluster, and the only subtype-specific trend was a particularly high mutational burden in the two squamous-like<sup>Ex</sup> tumors (Fig. 2a).





**Fig. 1** Gene expression-based subtypes in the MPA/DMBA-induced tumor cohort. Using the murine intrinsic gene list [11], hierarchical clustering of gene expression data revealed two distinct tumor clusters ( $p = 0.044$ , SigClust [22]), one containing claudin-low-like tumors and the other containing a transcriptomically heterogeneous set of tumors. Normal mouse mammary gland samples formed a separate cluster. Genes are ordered according to correlation to the claudin-low<sup>Ex</sup> centroid

All tumors carried mutations in driver genes defined by the COSMIC cancer gene census [38], with a mean of 13.8 driver genes carrying mutations per tumor (range 4 to 29) (Fig. 2b). Several driver genes were recurrently mutated, including *Trp53*, *Kras*, and *Kmt2c* (Additional file 4), but no driver genes carried mutations at a significantly different rate between the two clusters. We did, however, identify two notable trends which did not reach statistical significance: an elevated rate of *Trp53* mutations in the claudin-low-like cluster (50% vs. 11%,  $p = 0.13$ , two-tailed Fisher's exact test) and an elevated rate of *Zfmx3* mutations also in the claudin-low-like cluster (37.5% vs. 0%,  $p = 0.08$ , two-tailed Fisher's exact test). No mutations were significantly associated with histological features.

#### MPA/DMBA-induced tumors and human breast cancers display disparate gene mutational profiles

To narrow down potential driver mutations in the MPA/DMBA-induced tumors, we compared amino acid changes caused by mutations in driver genes to known amino acid changes in human cancers [38] (Table 2, Additional file 5). There were hotspot amino acid

changes in all *Ras* genes, including *Kras* G12C, G13R, Q61H, *Hras* Q61L, and *Nras* Q61L. In total, 8 of 18 tumors carried hotspot amino acid changes in *Ras* genes. There was one *Pik3ca* mutation in the cohort causing an H1047R amino acid change. This mutation is frequently found in human breast cancer and has previously been reported in DMBA-induced mouse mammary tumors [58].

There were marked disparities between the gene mutational profiles of human breast cancer [44] and MPA/DMBA-induced tumors (Fig. 2c, Additional file 6). The two most frequently mutated genes in breast cancer are *PIK3CA* and *TP53*. While *TP53* showed comparable mutation rates between human breast cancer and MPA/DMBA-induced tumors (34% and 28%, respectively), *PIK3CA* mutation does not appear to be a common event in MPA/DMBA-induced tumors (35% in BC, 6% in MPA/DMBA). Several frequently mutated genes in breast cancer, such as *CDH1*, *GATA3*, and *MAP3K1*, were not mutated in any MPA/DMBA-induced tumors. Conversely, many genes frequently mutated in MPA/DMBA-induced tumors, such as *ATR*, *FAT1*, and *KRAS*, are rarely mutated in breast cancer.

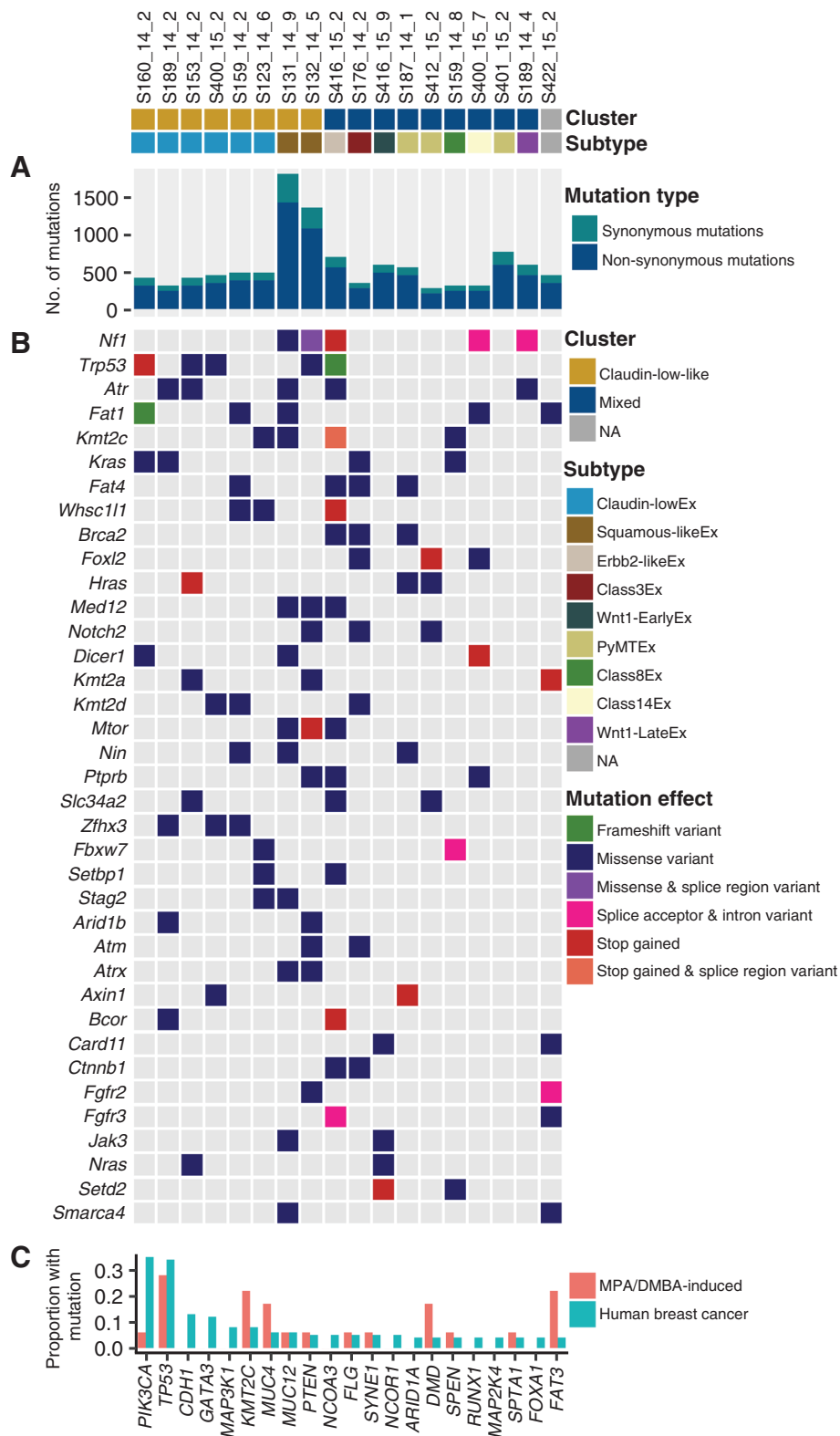


Fig. 2 (See legend on next page.)

(See figure on previous page.)

**Fig. 2** Somatic mutations in MPA/DMBA-induced mouse mammary tumors. **a** The MPA/DMBA-induced tumors carried between 288 and 1795 exonic mutations. No significant differences in mutational burden were found between the clusters; however, a high mutational rate was observed in the two squamous-like<sup>Ex</sup> tumors. **b** *Nf1*, *Trp53*, *Atr*, and *Fat1* were the most frequently mutated driver genes in the MPA/DMBA-induced tumor cohort. No specific mutations accurately delineated the tumor clusters. **c** MPA/DMBA-induced tumors generally showed divergent mutational rates compared to human breast cancer in the genes most frequently mutated in human breast cancer. *TP53* mutations occurred at a similar rate in MPA/DMBA-induced tumors and human breast cancer

### DMBA induces a characteristic mutational spectrum with a high frequency of T>A transversions in TG dinucleotides

To characterize the mutagenic profile of DMBA, we analyzed the mutational spectra of the MPA/DMBA-induced tumors. Mutations showed a majority of T>A transversions, which accounted for 63% of all mutations (Additional file 7A). In their trinucleotide context, thymine mutations (T>N) were overrepresented in positions with a 3' guanine nucleotide (Additional file 7B and C, Additional file 8). This was statistically significant when compared to the proportion of thymine nucleotides in an NTG context in the mouse reference genome ( $p < 0.001$  in all cases, two-tailed Wilcoxon rank-sum test). There was a similar overrepresentation of cytosine mutations in positions with a 3' adenine. This was statistically significant for C>A and C>G mutations ( $p < 0.001$ ), but not for C>T mutations ( $p = 0.089$ ), when compared to the proportion of cytosine nucleotides in an NCA context in the mouse reference genome.

Mutation signature analysis revealed evidence of signatures 4, 6, 22, 24, and 25 [42] in the MPA/DMBA-induced tumors (Additional file 7D). All tumors were associated with signature 22, while signatures 4 and 25 were found in 17 and 11 of the 18 tumors, respectively.

**Table 2** Selected hotspot mutations in MPA/DMBA-induced tumors

Sample	Gene	Amino acid change
S176_14_2	<i>Ctnnb1</i>	Asp32Asn
S416_15_2	<i>Ctnnb1</i>	Thr41Ile
S187_14_1	<i>Hras</i>	Gln61Leu
S412_15_2	<i>Hras</i>	Gln61Leu
S159_14_8	<i>Kras</i>	Gly12Cys
S160_14_2	<i>Kras</i>	Gly12Cys
S176_14_2	<i>Kras</i>	Gly13Arg
S189_14_2	<i>Kras</i>	Gln61His
S153_14_2	<i>Nras</i>	Gln61Leu
S416_15_9	<i>Nras</i>	Gln61Leu
S187_14_1	<i>Pik3ca</i>	His1047Arg
S132_14_5	<i>Trp53</i>	His211Pro
S153_14_2	<i>Trp53</i>	Lys129Met
S400_15_2	<i>Trp53</i>	Gln141Pro
S400_15_2	<i>Trp53</i>	His211Pro

Signatures 24 and 6 were only found in four and one tumor(s), respectively. Notably, none of the signatures found in MPA/DMBA-induced tumors have been associated with human breast cancer [42].

### MPA/DMBA-induced tumors have diverse copy number profiles

Breast cancer is largely driven by copy number aberrations (CNAs) [59], yet the copy number profiles of MPA/DMBA-induced mammary tumors have not previously been described. We found a mean of 1299 genes with CNA per tumor (range 90–3057), of which a mean of 65% were amplifications. There was a tendency for claudin-low-like tumors to have a lower burden of CNAs, with a mean of 919 genes carrying CNA, compared to the mixed group of tumors, with a mean of 1637 genes carrying CNA (Fig. 3a). This trend did however not reach statistical significance ( $p = 0.139$ , two-tailed Wilcoxon rank-sum test).

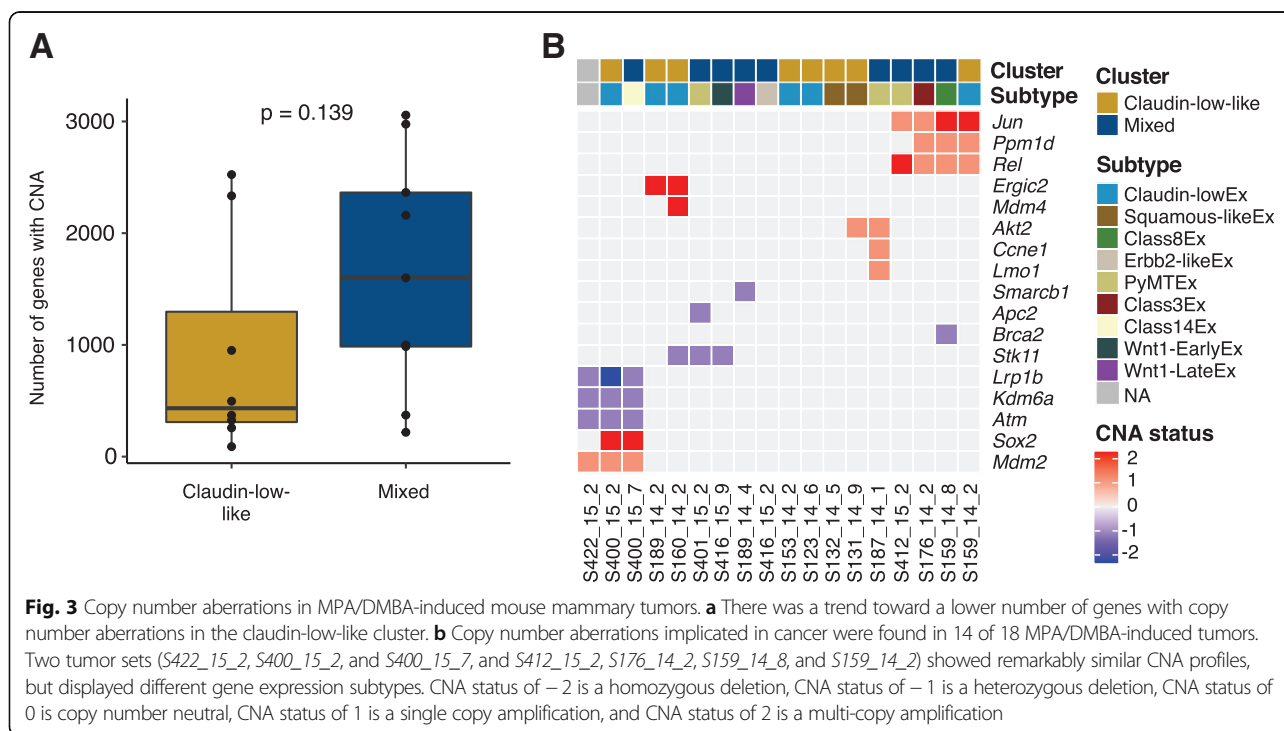
To determine CNAs in the MPA/DMBA-induced tumors with a potential oncogenic driver effect, we identified amplifications and deletions known to be associated with cancer [38] (Fig. 3b). We found that 14 of the 18 tumors carried potential driver CNAs (range 0 to 4, mean 2.6). Three of the four tumors not carrying potential driver CNAs were claudin-low-like. There was however no statistically significant difference in the number of potential driver CNAs between the clusters. Several genes had recurrent CNAs, but none occurred at a statistically significant different rate in one cluster versus the other.

Only two of the CNA events identified in MPA/DMBA-induced tumors occur at a notable rate in human breast cancer; *MDM4* is amplified in 25%, and *PPM1D* is amplified in 10% of human BC [6, 7].

We observed two sets of tumors carrying remarkably similar CNA profiles (Fig. 3b). None of the tumors in these two sets displayed the same murine subtype as any other tumor within the same set.

### The human claudin-low breast cancer genome is characterized by a low mutational burden, frequent *TP53* mutations, and a low rate of CNA

Little has been published specifically describing the genomic characteristics of human claudin-low breast cancer. We therefore analyzed the 218 claudin-low tumors found in the METABRIC dataset, for which DNA



sequence data from 173 genes and whole genome copy number data is available [6, 7].

Across the 173 sequenced genes, claudin-low tumors carried a mean of 4.7 mutations per tumor, significantly lower than the mean of 7.3 mutations per tumor for all other tumors ( $p < 0.001$ , two-tailed Wilcoxon rank-sum test) (Fig. 4a). Claudin-low tumors share several characteristics with basal-like tumors and are often classified as such by the PAM50 assay [2, 6, 7]; however, basal-like tumors showed a significantly higher mutational burden than claudin-low tumors (mean 8.1 mutations per tumor,  $p < 0.001$ , two-tailed Wilcoxon rank-sum test).

There was a high degree of overlap between the genes most frequently mutated in claudin-low breast cancers and the genes most frequently mutated in all other breast cancers (Fig. 4b). Most of these genes carried mutations at similar rates between claudin-low and non-claudin-low tumors, albeit with a tendency toward a slightly lower rate in claudin-low tumors. There were however two notable differences in mutational frequency: a significantly higher rate of *TP53* mutations and a significantly lower rate of *PIK3CA* mutations in claudin-low tumors compared to other tumors. Similarly, basal-like tumors also carried a high frequency of *TP53* mutations and a low frequency of *PIK3CA* mutations [7, 44].

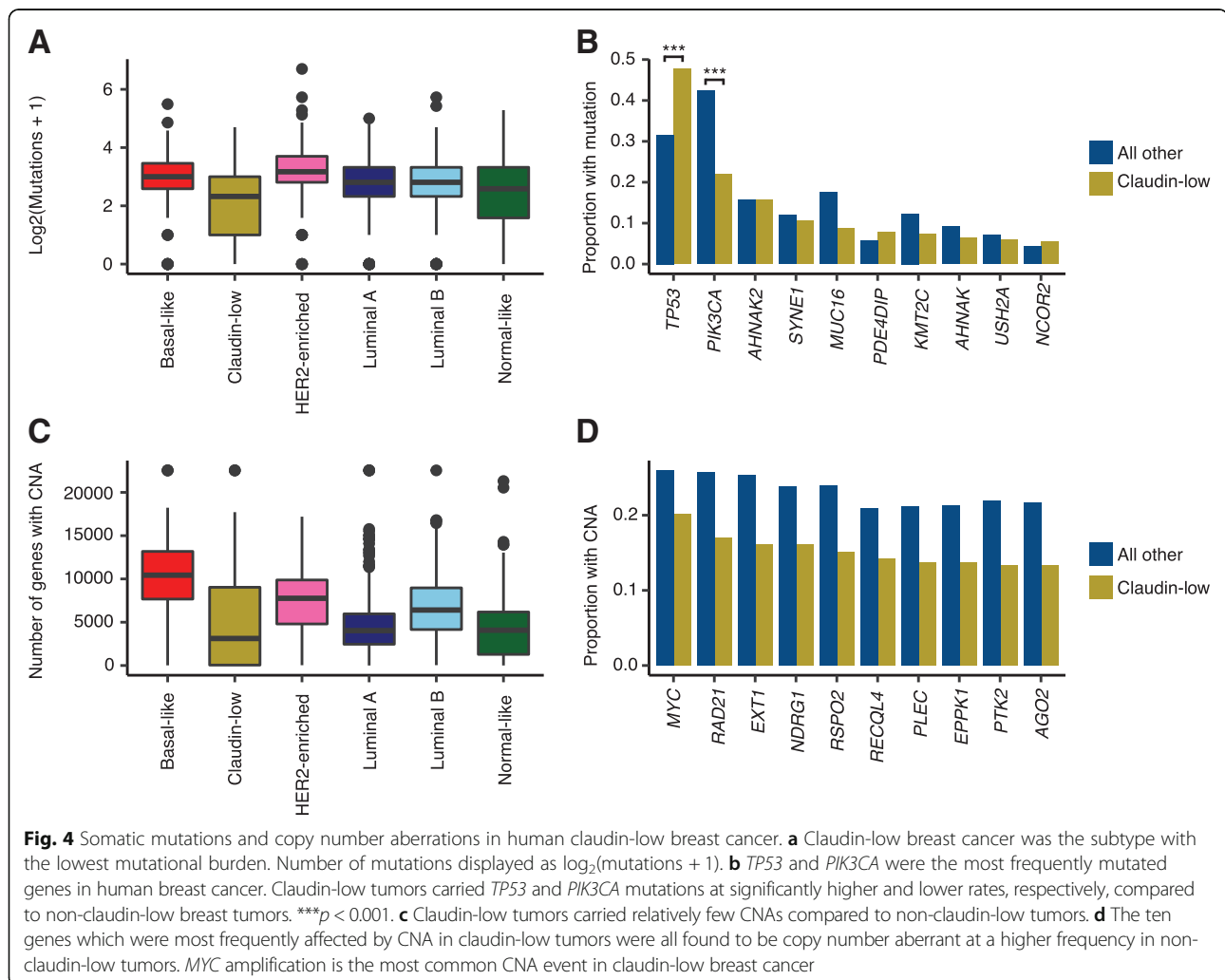
Human claudin-low breast tumors carried significantly fewer genes with copy number aberration (mean 4879) compared to all other tumors (mean 6247;  $p < 0.001$ , two-tailed Wilcoxon rank-sum test) (Fig. 4c). This

difference was also marked when comparing claudin-low tumors with basal-like tumors (mean 10,175 genes per tumor;  $p < 0.001$ , two-tailed Wilcoxon rank-sum test).

By gene, the most frequent copy number event in claudin-low breast cancer was *MYC* amplification, found in 20% of cases (Fig. 4d). In comparison, this event was found in 26% of all other breast tumors. The ten most frequently amplified genes in claudin-low breast cancer were all located at chromosomal position 8q24, a region also frequently amplified in basal-like breast cancers [6, 7].

#### Claudin-low-like MPA/DMBA-induced mammary tumors accurately reflect the gene expression characteristics of their human counterpart

We explored several established gene expression features of the claudin-low subtype and found that MPA/DMBA-induced claudin-low-like tumors accurately mirrored their human counterpart. Specifically, claudin-low-like tumors had low expression of genes involved in cell-cell adhesion, low expression of luminal genes, and high expression of genes related to EMT (Fig. 5a, Additional file 9). Claudin-low-like tumors also showed a markedly lower degree of differentiation compared to tumors in the mixed cluster. In particular, the claudin-low-like cluster expressed significantly higher and lower levels of *Cd44* and *Cd24a*, respectively, indicating a stem cell-like phenotype in these tumors [2, 60] (Additional file 10). There was no significant difference in the expression of proliferation-related genes between the two clusters. Vascular content-related genes were expressed at a significantly higher level in



claudin-low-like tumors compared to the tumors in the mixed cluster (Additional file 9), indicating a higher degree of neoangiogenesis in these tumors.

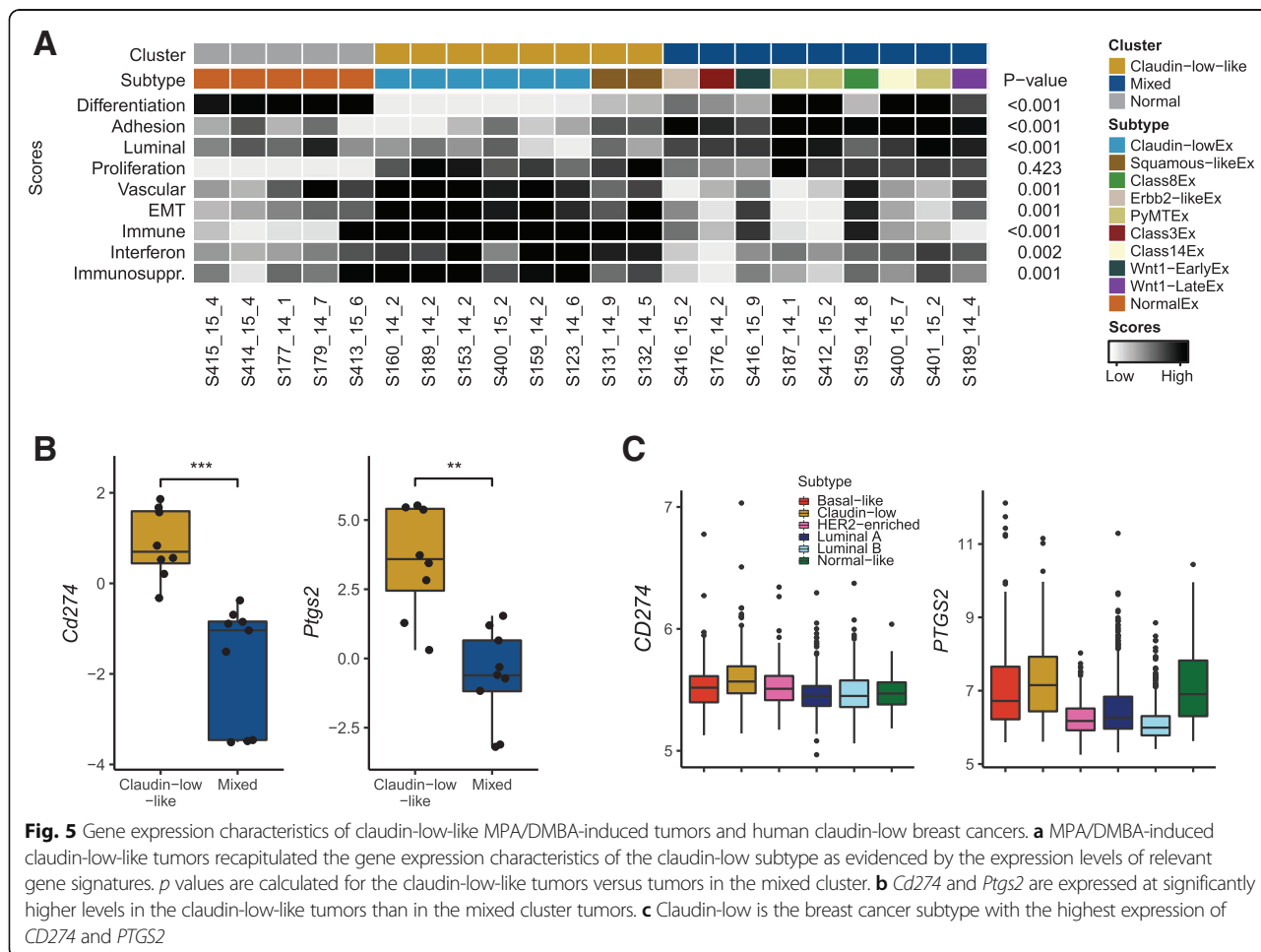
Immune cell admixture was significantly higher in the claudin-low-like tumors compared to tumors in the mixed cluster ( $p < 0.001$ , two-tailed Wilcoxon rank-sum test) and compared to normal mammary gland samples ( $p = 0.006$ ). We also found higher expression of genes related to immunosuppression and interferons in the claudin-low-like cluster compared to both the mixed cluster and normal mammary gland samples. In combination, high immune cell infiltration and high expression of type 1 interferon-related and immunosuppressive genes are characteristics of tumors that may respond to immunotherapeutics [61, 62].

We identified a significantly elevated expression of two potentially actionable genes related to immunosuppression in the claudin-low-like tumors: the immune checkpoint encoding gene *Cd274* and the cyclooxygenase encoding gene *Ptgs2* (Fig. 5b). These features were also

characteristic of human claudin-low tumors in the METABRIC cohort [6, 7], which showed significantly higher expression levels of both *PTGS2* and *CD274* compared to non-claudin-low breast tumors ( $p < 0.001$  for both, two-tailed Wilcoxon rank-sum test) and compared specifically to basal-like tumors ( $p = 0.004$  and  $p < 0.001$ , respectively) (Fig. 5c). These characteristics may indicate a susceptibility to immune checkpoint inhibitors and cyclooxygenase inhibitors in human claudin-low breast cancer [63, 64].

## Discussion

In this study, we have performed a comprehensive analysis of mutations, copy number aberrations, and gene expression characteristics of MPA/DMBA-induced mouse mammary tumors. We found marked intertumor heterogeneity and showed that half of the tumors displayed a claudin-low-like phenotype, in line with a previous report [11]. Our findings demonstrate that these tumors provide a transcriptomically accurate representation of human



claudin-low breast tumors, reflecting key features such as an EMT phenotype, high level of immune infiltration, and a low degree of differentiation.

MPA/DMBA-induced tumors carried a mutational burden multiple times that of human breast cancer, a high frequency of activating *Ras*-mutations, and a characteristic mutational spectrum. The specific genes carrying mutations varied widely between tumors; however, all tumors had a consistent mutational signature. This indicates that the dominant mutational process in these tumors is DMBA-induced mutagenesis, and not aberrations occurring after tumor initiation, as a result of, e.g., disrupted DNA repair. Copy number aberrations in MPA/DMBA-induced tumors have not previously been explored, and we show here that most tumors carry potential driver CNAs. However, while we noted several genomic trends, such as a higher rate of *Trp53* mutation and a lower burden of CNA in MPA/DMBA-induced claudin-low-like tumors, no individual genomic features accurately delineated the two gene expression-based tumor clusters. Further, several tumors carried similar sets of mutations and/or CNAs but displayed different

subtypes. This suggests that no specific genomic event determines tumor subtype and that other etiological models may be more appropriate, such as different cells-of-origin [65] or microenvironmental factors [66]. This finding concurs with recent reports showing that transgenic mouse mammary tumors display histological and transcriptomic phenotypes largely uncoupled from their underlying driver mutations [67–69]. One possible model for MPA/DMBA-induced tumorigenesis is therefore as follows: first, MPA induces a RANK-I-mediated mammary gland proliferation [10, 13]. DMBA then induces mutations in mammary cells in a pattern as elucidated by our mutation signature analysis, predominantly in TG and CA dinucleotides, stochastically distributed throughout the genome. The tumor is initiated when one or more driver mutations occur, for example, *Trp53* or *Ras*-mutation, with the tumor phenotype, however, determined by non-genomic factors. The biochemical mechanism of DMBA-induced mutagenesis has been described [14, 15], whereas no causal mechanism for DMBA-induced copy number aberration is known; it is therefore likely that CNAs arise after tumor initiation.

Previous genomic analyses which included human claudin-low breast tumors have either not included specific analyses of the subtype [6, 7], included few samples [3], or have been restricted to the triple-negative [70, 71] or metaplastic [72] subsets of claudin-low tumors. We show here that human claudin-low tumors are characterized by a low number of mutations and a low burden of CNAs. This finding is surprising, given the apparent inverse correlation between CNA and mutational burden in cancer [59], and indicates that the claudin-low subtype is relatively genomically stable compared to other breast cancers. We also find similarities in genomic characteristics between claudin-low tumors and basal-like tumors, in particular a high frequency of *TP53* mutations, a low frequency of *PIK3CA* mutations, and 8q24 amplifications as a common event. While the transcriptional similarity between these two subtypes is established [2], these findings illustrate that there are also marked genomic similarities between claudin-low and basal breast cancer, albeit with a lower burden of genomic aberrations in claudin-low tumors.

Claudin-low tumors show high expression of immune-related genes and a high level of immune cell infiltration [2, 3, 73]. However, claudin-low tumors also express high levels of immunosuppressive genes. In MPA/DMBA-induced claudin-low-like tumors, we observed an elevated expression of two particularly notable genes involved in immunosuppression: *Ptgs2* (encoding COX-2) and *Cd274* (encoding PD-L1). This observation was consistent in human claudin-low breast cancer. COX-2 may be implicated in cancer development through several mechanisms: reducing apoptosis, increasing epithelial cell proliferation, promoting angiogenesis, and increasing invasiveness of tumor cells and immunosuppression [74–76]. COX-2 may also be involved in vasculogenic mimicry, a process in which epithelial tumor cells form vascular channel-like structures without participation of endothelial cells, allowing nutrients to reach tumor cells without the need for neoangiogenesis [77]. Vasculogenic mimicry has previously been shown to occur in claudin-low tumors [24]. COX-2 and PD-L1 are clinically actionable through the use of COX inhibitors [63] and checkpoint inhibitors [78], respectively. Further research into the potential use of checkpoint inhibitors and COX inhibitors in claudin-low breast cancer is warranted, with promising future avenues including combinatorial Treg depletion [73].

## Conclusions

In summary, we have found that claudin-low-like MPA/DMBA-induced mouse mammary tumors are a transcriptionally accurate model for human claudin-low breast cancer. We did not find strong evidence that claudin-low-like MPA/DMBA-induced tumors are delineated by any specific genomic features; however, the relatively small

number of samples included in this study may have obscured possible associations. By analyzing publicly available data, we showed that human claudin-low breast cancer is a relatively genomically stable subtype. There is a high expression of genes related to immunosuppression in claudin-low breast cancers, a feature which is evident in claudin-low-like MPA/DMBA-induced tumors. Our observations suggest immunosuppression as a potential therapeutic target in claudin-low breast cancer and indicate MPA/DMBA-induced claudin-low-like tumors as an appropriate model for continued research.

## Additional files

**Additional file 1:** Mouse characteristics and histopathological data. (XLSX 14 kb)

**Additional file 2:** Gene lists used for gene expression scores. (XLSX 11 kb)

**Additional file 3:** Subtype correlations for MPA/DMBA-induced tumors. (XLSX 17 kb)

**Additional file 4:** Mutations observed in MPA/DMBA-induced tumors. (XLSX 405 kb)

**Additional file 5:** Driver gene mutations in MPA/DMBA-induced tumors observed in the COSMIC database. (XLSX 37 kb)

**Additional file 6:** Comparative mutation rates in MPA/DMBA-induced tumors and human breast tumors in the TCGA cohort. (XLSX 27 kb)

**Additional file 7:** The mutational spectra and mutational signatures of MPA/DMBA-induced mammary tumors. **a** T>A transversions were the most frequent mutation type in MPA/DMBA-induced tumors, followed by C>A transversions. **b** Heatmap of mutational frequencies by trinucleotide context. There was an overrepresentation of T>N mutations in positions with a 3' guanine and C>N mutations in positions with a 3' adenine. **c** Histogram of C>A and T>A transversions by trinucleotide context in a representative tumor (*S159\_14\_8*). **d** Mutation signature 22 was the predominant mutational signature in the MPA/DMBA-induced tumors and was evident in all tumors in the cohort. (PDF 214 kb)

**Additional file 8:** Mutational signatures for all MPA/DMBA-induced tumors. (ZIP 142 kb)

**Additional file 9:** Gene expression scores by cluster for genes related to differentiation, adhesion, luminal features, proliferation, vascular content, EMT, immune features, interferon signaling and immunosuppression. Two-tailed Wilcoxon rank-sum test. ns = not significant,  $p > 0.05$ . \* $p < 0.05$ . \*\* $p < 0.01$ . \*\*\* $p < 0.001$ . (PDF 9 kb)

**Additional file 10:** Expression of *Cd24a* and *Cd44* by cluster in MPA/DMBA-induced tumors. Claudin-low-like tumors had a lower expression of *Cd24a* and a higher expression of *Cd44* compared to the mixed cluster of tumors ( $p = 0.003$  and  $p = 0.005$ , respectively, two-tailed, Wilcoxon rank-sum test), indicating a stem cell-like phenotype in the claudin-low-like tumors. (PDF 5 kb)

## Abbreviations

BC: Breast cancer; CNA: Copy number aberration; DMBA: 7,12-Dimethylbenzanthracene; EMT: Epithelial-mesenchymal transition; HE: Hematoxylin and eosin; MPA: Medroxyprogesterone acetate; PFA: Paraformaldehyde

## Acknowledgements

We thank Phuong Vu, Eldri Undlien Due, and Tina Brinks for helping with the laboratory work; Prof. Rune Toftgård for providing the transgenic mouse lines; and the support staff at the Department of Comparative Medicine, Oslo University Hospital Norwegian Radium Hospital, for the help with the animal work. We are grateful to the members of the Department of Cancer

Genetics, Institute for Cancer Research, Oslo University Hospital, for insightful discussions, and in particular thank Tonje G. Lien for the statistical input.

#### Authors' contributions

CF, HB, JHN, and TS contributed to the conceptualization. CF, HB, RK, JHN, and TS contributed to the methodology. CF and HB contributed to the formal analysis. CF, HB, RK, JHN, and TS contributed to the investigation. JHN and TS contributed to the resources. CF and HB wrote the original draft of the manuscript. CF, HB, RK, JHN, and TS wrote, reviewed, and edited the manuscript. CF and HB contributed to the visualization. JHN and TS contributed to the supervision. TS contributed to the funding acquisition. All authors read and approved the final manuscript.

#### Funding

This work was supported by grants from the Norwegian Research Council ([www.forskingsradet.no/](http://www.forskingsradet.no/)) (250459 to TS), South-Eastern Norway Regional Health Authority ([www.helse-sorost.no/](http://www.helse-sorost.no/)) (2012056 to TS), and the Medical Student Research Program at the University of Oslo ([www.med.uio.no](http://www.med.uio.no)) (to CF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

#### Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the European Nucleotide Archive, accession number PRJEB29718, and ArrayExpress, accession number E-MTAB-7507.

#### Ethics approval and consent to participate

The Norwegian Food Safety Authority approved all experiments in advance of their implementation (approval number 4385).

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Department of Cancer Genetics, Oslo University Hospital, Oslo, Norway. <sup>2</sup>Department of Laboratory Medicine, Karolinska Institutet, Stockholm, Sweden. <sup>3</sup>Centre for Cancer Biomarkers CCBIO, University of Bergen, Bergen, Norway. <sup>4</sup>Institute for Clinical Medicine, University of Oslo, Oslo, Norway.

Received: 5 March 2019 Accepted: 17 July 2019

Published online: 31 July 2019

#### References

- Herschkwitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol.* 2007;8(5):R76.
- 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 Res.* 2010;12(5):R68.
- Sabatier R, Finetti P, Guille A, Adelaide J, Chaffanet M, Viens P, et al. Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization. *Mol Cancer.* 2014;13(1):228.
- Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. *Mol Oncol.* 2011;5(1):5–23.
- Dias K, Dvorkin-Gheva A, Hallett RM, Wu Y, Hassell J, Pond GR, et al. Claudin-low breast cancer; clinical & pathological characteristics. *PLoS One.* 2017;12(1):e0168669.
- Curtis C, Shah SP, Chin S-F, 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.
- Pereira B, Chin S-F, Rueda OM, Vollan H-KM, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat Commun.* 2016;7:11479.
- Hennessy BT, Gonzalez-Angulo A-M, Stemke-Hale K, Gilcrease MZ, Krishnamurthy S, Lee J-S, et al. Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. *Cancer Res.* 2009;69(10):4116–24.
- Prat A, Adamo B, Cheang MCU, Anders CK, Carey LA, Perou CM. Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. *Oncologist.* 2013;18(2):123–33.
- Aldaz CM, Liao QY, LaBate M, Johnston DA. Medroxyprogesterone acetate accelerates the development and increases the incidence of mouse mammary tumors induced by dimethylbenzanthracene. *Carcinogenesis.* 1996;17(9):2069–72.
- Pfefferle AD, Herschkowitz JI, Usary J, Harrell J, Spike BT, Adams JR, et al. Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. *Genome Biol.* 2013;14(11):R125.
- Yin Y, Bai R, Russell RG, Beildeck ME, Xie Z, Kopelovich L, et al. Characterization of medroxyprogesterone and DMBA-induced multilineage mammary tumors by gene expression profiling. *Mol Carcinog.* 2005;44(1):42–50.
- Gonzalez-Suarez E, Jacob AP, Jones J, Miller R, Roudier-Meyer MP, Erwert R, et al. RANK ligand mediates progesterin-induced mammary epithelial proliferation and carcinogenesis. *Nature.* 2010;468(7320):103.
- Baird WM, Hooven LA, Mahadevan B. Carcinogenic polycyclic aromatic hydrocarbon-DNA adducts and mechanism of action. *Environ Mol Mutagen.* 2005;45(2–3):106–14.
- Frenkel K. 7,12-dimethylbenz[a]anthracene induces oxidative DNA modification in vivo. *Free Radic Biol Med.* 1995;19(3):373–80.
- Dean JH, Ward EC, Murray MJ, Lauer LD, House RV. Mechanisms of dimethylbenzanthracene-induced immunotoxicity. *Clin Physiol Biochem.* 1985;3(2–3):98–110.
- Miyata M, Furukawa M, Takahashi K, Gonzalez FJ, Yamazoe Y. Mechanism of 7, 12-dimethylbenz[a]anthracene-induced immunotoxicity: role of metabolic activation at the target organ. *Jpn J Pharmacol.* 2001;86(3):302–9.
- Trichopoulos D, Adami H, Ekbohm A, Hsieh C, Lagiou P. Early life events and conditions and breast cancer risk: from epidemiology to etiology. *Int J Cancer.* 2008;122(3):481–5.
- Snippert HJ, Van Der Flier LG, Sato T, Van Es JH, Van Den Born M, Kroon-Veenboer C, et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell.* 2010;143(1):134–44.
- Norum JH, Bergström Å, Andersson AB, Kuiper RV, Hoelzl MA, Sørle T, et al. A conditional transgenic mouse line for targeted expression of the stem cell marker LGR5. *Dev Biol.* 2015;404(2):35–48.
- Team RC, Computing RF, for S. R. A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017.
- Liu Y, Hayes DN, Nobel A, Marron JS. Statistical significance of clustering for high-dimension, low-sample size data. *J Am Stat Assoc.* 2008;103(483):1281–93.
- Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun.* 2013;4:2612.
- Chuck Harrell J, Pfefferle AD, Zalles N, Prat A, Fan C, Khramtsov A, et al. Endothelial-like properties of claudin-low breast cancer cells promote tumor vascular permeability and metastasis. *Clin Exp Metastasis.* 2014;31:33–45.
- Kardos J, Chai S, Mose LE, Selitsky SR, Krishnan B, Saito R, et al. Claudin-low bladder tumors are immune infiltrated and actively immune suppressed. *JCI insight.* 2016;1(3):e85902.
- TO N, 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. *Clin Cancer Res.* 2010;16(21):5222–32.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci.* 2005;102(43):15545–50.
- Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 2011;17(1):10–2.
- Joshi NA, Sickle: a sliding-window, adaptive, quality-based tool for FastQ files; 2011.
- Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:13033997*; 2013.
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high-confidence variant calls: the



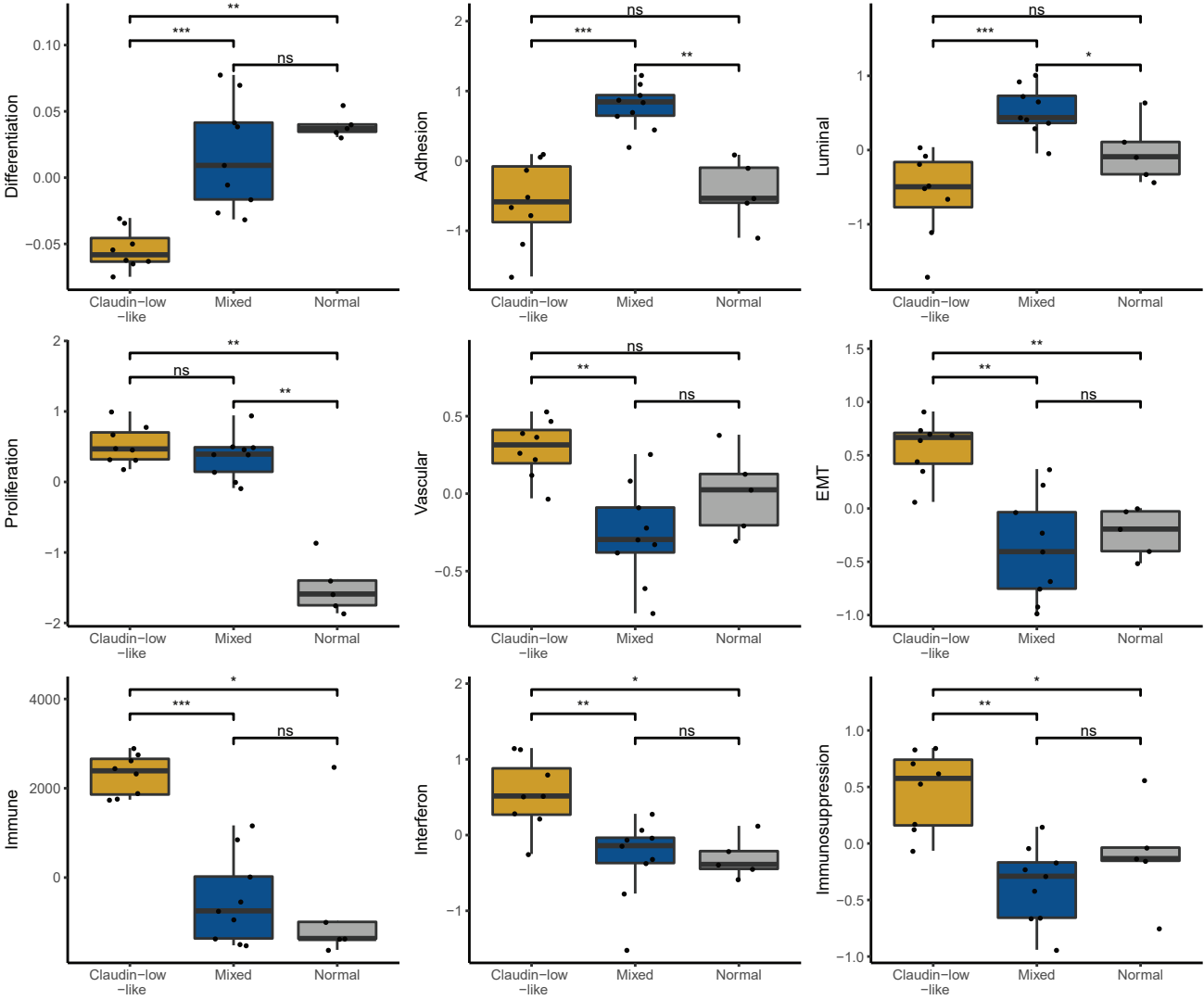
- genome analysis toolkit best practices pipeline. In: *Current protocols in bioinformatics*: Wiley; 2013. <https://doi.org/10.1002/0471250953.bi1110s43>.
33. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–303.
  34. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43(5):491–8.
  35. Wong K, Bumpstead S, Van Der Weyden L, Reinholdt LG, Wilming LG, Adams DJ, et al. Sequencing and characterization of the FVB/NJ mouse genome. *Genome Biol.* 2012;13(8):1–12.
  36. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* 2012;6(2):80–92.
  37. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program. *SnpSift Front Genet.* 2012;3:35.
  38. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 2016;45(D1):D777–83.
  39. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 2014;7(1):539.
  40. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2017;45(D1):D158–69.
  41. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* 2016;17(1):1.
  42. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. *Nature.* 2013; 500(7463):415–21.
  43. D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. *Nucleic Acids Res.* 2016;44(20):e154.
  44. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012;490(7418):61.
  45. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal.* 2013;6(269):pl1.
  46. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *ACR; 2012.* <https://doi.org/10.1158/2159-8290.CD-12-0095>.
  47. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics.* 2016;32(18): 2847–9.
  48. Wickham H. *ggplot2*. New York: Springer New York; 2009.
  49. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.* 2013;499(7457):214–8.
  50. McCreery MQ, Halliwill KD, Chin D, Delrosario R, Hirst G, Vuong P, et al. Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. *Nat Med.* 2015;21(12):1514.
  51. Westcott PMK, Halliwill KD, To MD, Rashid M, Rust AG, Keane TM, et al. The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. *Nature.* 2015;517(7535):489–92.
  52. Nassar D, Latil M, Boeckx B, Lambrechts D, Blanpain C. Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. *Nat Med.* 2015;21(8):946.
  53. Francis JC, Melchor L, Campbell J, Kendrick H, Wei W, Armisen-Garrido J, et al. Whole-exome DNA sequence analysis of Brca2-and Trp53-deficient mouse mammary gland tumours. *J Pathol.* 2015;236(2):186–200.
  54. Pfefferle AD, Agrawal YN, Koboldt DC, Kanchi KL, Herschkowitz JI, Mardis ER, et al. Genomic profiling of murine mammary tumors identifies potential personalized drug targets for p53-deficient mammary cancers. *Dis Model Mech.* 2016;9(7):749–57.
  55. Liu H, Murphy CJ, Karreth FA, Emdal KB, White FM, Elemento O, et al. Identifying and targeting sporadic oncogenic genetic aberrations in mouse models of triple-negative breast cancer. *Cancer Discov.* 2018;8(3):354–69.
  56. McFadden DG, Politi K, Bhutkar A, Chen FK, Song X, Pirun M, et al. Mutational landscape of EGFR-, MYC-, and Kras-driven genetically engineered mouse models of lung adenocarcinoma. *Proc Natl Acad Sci.* 2016;113(42):E6409–17.
  57. McFadden DG, Papagiannakopoulos T, Taylor-Weiner A, Stewart C, Carter SL, Cibulskis K, et al. Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell.* 2014;156(6):1298–311.
  58. Abba MC, Zhong Y, Lee J, Kil H, Lu Y, Takata Y, Simper MS, Gaddis S, Shen J, Aldaz CM. DMBA induced mouse mammary tumors display high incidence of activating Pik3caH1047 and loss of function Pten mutations. *Oncotarget.* 2016;7(39):64289.
  59. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet.* 2013;45(10):1127–33.
  60. Visvader JE, Stingl J. Mammary stem cells and the differentiation hierarchy: current status and perspectives. *Genes Dev.* 2014;28(11):1143–58.
  61. Hegde PS, Karanikas V, Evers S. The when, the where, and the how of immune monitoring for cancer immunotherapies in the era of checkpoint inhibition. *Clin Cancer Res.* 2016;22(8):1865–74.
  62. Jamieson NB, Maker AV. Gene-expression profiling to predict responsiveness to immunotherapy. *Nat Publ Gr.* 2016;24(3):134–40.
  63. Zelenay S, Van Der Veen AG, Böttcher JP, Snelgrove KJ, Rogers N, Acton SE, et al. Cyclooxygenase-dependent tumor growth through evasion of immunity. *Cell.* 2015;162(6):1257–70.
  64. Chokr N, Chokr S. Immune checkpoint inhibitors in triple negative breast cancer: what is the evidence? *J Neoplasm.* 2018;3(2):6.
  65. Prat A, Perou CM. Mammary development meets cancer genomics. *Nat Med.* 2009;15(8):842.
  66. Hanahan D, Coussens LM. Accessories to the crime: functions of cells recruited to the tumor microenvironment. *Cancer Cell.* 2012;21(3):309–22.
  67. Hollern DP, Swiatnicki MR, Andrechek ER. Histological subtypes of mouse mammary tumors reveal conserved relationships to human cancers. *PLoS Genet.* 2018;14(1):e1007135.
  68. Rennhack J, Swiatnicki M, Zhang Y, Li C, Bylett E, Ross C, Szczepanek K, Hanrahan W, Jayatissa M, Hunter K, Andrechek E. Integrated sequence and gene expression analysis of mouse models of breast cancer reveals critical events with human parallels. *bioRxiv.* 2018:375154. <https://www.biorxiv.org/content/10.1101/375154v1.full>.
  69. Hollern DP, Andrechek ER. A genomic analysis of mouse models of breast cancer reveals molecular features of mouse models and relationships to human breast cancer. *Breast Cancer Res.* 2014;16(3):R59.
  70. Morel A-P, Ginestier C, Pommier RM, Cabaud O, Ruiz E, Wicinski J, et al. A stemness-related ZEB1–MSRB3 axis governs cellular pliancy and breast cancer genome stability. *Nat Med.* 2017;23(5):568.
  71. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua SAW, et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clin Cancer Res.* 2015;21(7):1688–98.
  72. Weigelt B, Ng CKY, Shen R, Popova T, Schizas M, Natrajan R, et al. Metastatic breast carcinomas display genomic and transcriptomic heterogeneity. *Mod Pathol.* 2015;28(3):340.
  73. Taylor NA, Vick SC, Iglesia MD, Brickey WJ, Midkiff BR, McKinnon KP, et al. Treg depletion potentiates checkpoint inhibition in claudin-low breast cancer. *J Clin Invest.* 2017;127(9):3472–83.
  74. Zarghi A, Arfaei S. Selective COX-2 inhibitors: a review of their structure-activity relationships. *Iran J Pharm Res IJPR.* 2011;10(4):655–83.
  75. Dannenberg AJ, DuBois RN. COX-2: a new target for cancer prevention and treatment. *Karger; 2003.* p. 291. <https://scholar.google.com/scholar?cluster=4132316902324774708>.
  76. Tsujii M, DuBois RN. Alterations in cellular adhesion and apoptosis in epithelial cells overexpressing prostaglandin endoperoxide synthase 2. *Cell.* 1995;83(3):493–501.
  77. Basu GD, Liang WS, Stephan DA, Wegener LT, Conley CR, Pockaj BA, et al. A novel role for cyclooxygenase-2 in regulating vascular channel formation by human breast cancer cells. *Breast Cancer Res.* 2006;8(6):R69.
  78. Yan X, Zhang S, Deng Y, Wang P, Hou Q, Xu H. Prognostic factors for checkpoint inhibitor based immunotherapy: an update with new evidences. *Front Pharmacol.* 2018;9:1050.

## Publisher's Note

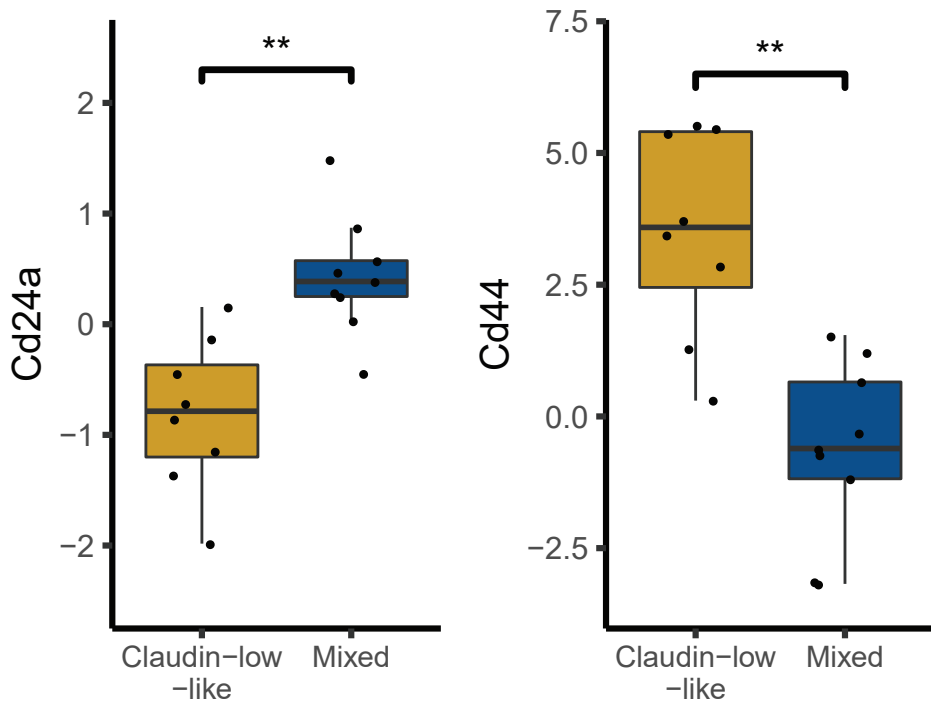
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Additional file 9



# Additional file 10





## Paper III

### **Contrasting DCIS and invasive breast cancer by subtype suggests basal-like DCIS as distinct lesions**

Helga Bergholtz, Tonje Gulbrandsen Lien, David M. Swanson, Arnaldo Frigessi, Oslo Breast Cancer Research Consortium (OSBREAC), Jörg Tost, Maria Grazia Daidone, Fredrik Wärnberg, and Therese Sørliie.

*Manuscript*



# Contrasting DCIS and invasive breast cancer by subtype suggests basal-like DCIS as distinct lesions

Helga Bergholtz<sup>1,2</sup>, Tonje Gulbrandsen Lien<sup>1</sup>, David Swanson<sup>3</sup>, Arnoldo Frigessi<sup>3,4</sup>, Oslo Breast Cancer Research Consortium (OSBREAC)<sup>1</sup>, Jörg Tost<sup>5</sup>, Maria Grazia Daidone<sup>6</sup>, Fredrik Wärnberg<sup>7,8</sup>, and Therese Sørli<sup>1,2</sup> ✉

<sup>1</sup>Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Oslo, Norway

<sup>2</sup>Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

<sup>3</sup>Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway

<sup>4</sup>Department of Biostatistics, University of Oslo, Oslo, Norway

<sup>5</sup>Laboratory for Epigenetics and Environment, Centre National de Recherche en Génomique Humaine, CEA-Institut de Biologie Francois Jacob, Evry, France

<sup>6</sup>Department of applied Research and Technical development, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

<sup>7</sup>Department of Surgical Sciences, Uppsala University, Uppsala, Sweden

<sup>8</sup>Department of Surgery, Uppsala Academic Hospital, Uppsala, Sweden

Ductal carcinoma in situ (DCIS) is a non-invasive type of breast cancer with highly variable potential of becoming invasive and affecting mortality. Currently, many DCIS are overtreated due to lack of specific biomarkers that distinguish low risk lesions from those that are of higher risk of progression. In this study, we present data from 57 DCIS and 313 invasive breast cancers (IBC) on three genomic levels; gene expression, DNA methylation and DNA copy number. We performed subtype stratified analyses and identified differences between DCIS and IBC that suggest subtype specific progression. The most prominent differences were found in tumors of the basal-like subtype: Basal-like DCIS were less proliferative and had a higher degree of differentiation than basal-like IBC. Also, core basal tumors (characterized by high correlation to the basal-like centroid) were not identified amongst DCIS as opposed to IBC. At the copy number level, the basal-like DCIS exhibited fewer copy number aberrations compared to basal-like IBC. An intriguing finding through analysis of DNA methylation was hyper-methylation of multiple protocadherin genes in basal-like IBC compared to basal-like DCIS, possibly caused by long range epigenetic silencing. This points to silencing of cell adhesion-related genes specifically in IBC of the basal-like subtype. Our work affirms that subtype stratification is important when studying progression from DCIS to IBC, and we provide the first evidence that basal-like DCIS show less aggressive characteristics and may be a different entity than basal-like IBC.

DCIS | breast cancer | molecular subtypes | breast tumor progression | gene expression | copy number | methylation

Correspondence: [therese.sorlie@rr-research.no](mailto:therese.sorlie@rr-research.no)

## Introduction

Ductal carcinoma in situ (DCIS) is a non-invasive, non-obligate precursor to invasive breast cancer (IBC) with low risk of progression (1). As breast cancer screening has become widespread, more DCIS lesions are being detected (2–4). Autopsy studies and studies of DCIS from non-treated patients show that many lesions, if left alone, would never progress to invasive disease (5–9). However, there is currently no robust method to distinguish DCIS with invasive potential from those that may be left untreated. DCIS is a heterogeneous disease and may at time of diagnosis vary from indolent lesions to tumors on the verge of becoming invasive. Clinical, histopathological and molecular characteristics may

also vary extensively (10, 11). As a consequence of this uncertainty, treatment for DCIS is often extensive, resulting in substantial overtreatment (12–15).

Knowledge on the underlying mechanisms of progression from DCIS to IBC is still limited. In order to balance risk and benefit for each patient, it is important to determine the tumor's invasive potential. Several studies have observed few genomic differences between DCIS and IBC (16–18). However, most breast cancer progression studies have not taken into account the significance of molecular subtype in DCIS. In IBC, molecular subtypes have distinct characteristics and also provide valuable prognostic and predictive information (19). In a previous study, we found evidence of subtype specific progression from DCIS to IBC suggesting that each molecular subtype undergoes a distinct evolutionary disease course (20). In DCIS, grade and growth pattern provide some information on risk of recurrence, yet, there is still a need for more precise risk prediction (21–23). For this purpose, Oncotype DX Breast DCIS score has been developed to predict individual risk of recurrence after breast conserving surgery (BCS) (24). This assay, however, does not take into account the vast heterogeneity of DCIS and the low risk group still experienced a relatively high risk of recurrence of 10% after 10 years (25). Nevertheless, this illustrates the potential of molecular-based assays for risk prediction in DCIS.

In this study, we explore the differences between DCIS and IBC in a subtype specific manner using data from three genomic levels: gene expression, DNA copy number and DNA methylation. We observed disparate associations between DCIS and IBC across the subtypes and found that basal-like DCIS might represent a different molecular entity than their invasive counterpart. We hypothesize that tumors of different molecular subtypes may have different modes of progression, and by comparing DCIS and IBC for each subtype separately, we aspire to obtain insight that may be used to elucidate further the mechanisms of breast cancer invasion and progression.

## Results

**Diverging subtype characteristics between ductal carcinoma in situ and invasive breast cancer.** Gene expression data were available from 57 DCIS and 313 IBC. Number of samples for each type of data and clinical information is presented in table 1 and suppl. file 1.

	DCIS	IBC
Number of tumors	57	313
Number of expression arrays	57	313
Number of SNP arrays	48	290
Number of Methylation arrays	41	273
Age in years, median (range)	54 (26-82)	54 (26-83)
Size in mm, median (range)	28 (7-90)	18 (2-130)
ELSTON grade (1/2/3/NA)	-	44/115/122/32
EORTC grade (1/2/3/NA)	0/8/21/28	-

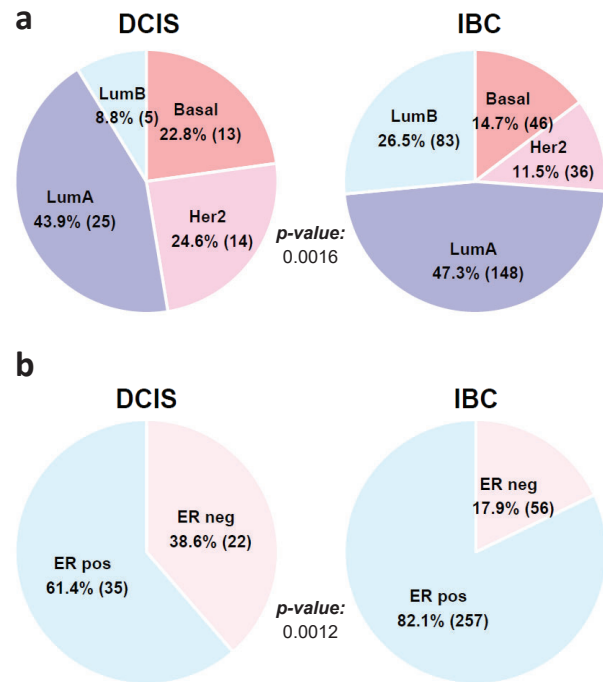
**Table 1.** Summary of available data for analysis including age, size and grade. ELSTON grading applies to invasive tumors (IBC), EORTC grading applies to DCIS tumors.

We determined the PAM50 intrinsic subtypes using the nearest centroid classification method (26) and found significantly different distribution of the subtypes between DCIS and IBC ( $P=0.0016$ , Fisher exact test, Fig. 1a). Most notably was there a higher frequency of the HER2-enriched subtype and a lower frequency of Luminal B tumors in DCIS compared to IBC. This was also reflected by a significantly different distribution of *ESR1* gene expression between the two stages ( $P=0.0012$  Fisher exact test, Fig. 1b). Centroid based subtyping tools such as the PAM50 method, provide each tumor's correlation to all centroids and the tumor is assigned to the subtype with the highest correlation coefficient. In general, we observed that DCIS tumors showed lower correlation coefficients to the subtype centroids compared to IBC; this was particularly evident for DCIS of the basal-like subtype (Table 2). To investigate whether differences in tumor cell content between DCIS and IBC influenced the subtype distribution, we used the ASCAT algorithm (27) to calculate tumor purity based on copy number data. We found no significant difference in tumor cell content between DCIS and IBC (Basal-like:  $P=0.86$ , HER2:  $P=0.2$ , LumA:  $P=0.88$ , LumB:  $P=0.19$ , Mann Whitney U tests, Suppl. Fig. 1a).

	DCIS		IBC	
	Median	Range	Median	Range
<b>Basal</b>	0.26	(0.12-0.46)	0.76	(0.03-0.88)
<b>HER2</b>	0.35	(0.14-0.64)	0.55	(0.15-0.72)
<b>LumA</b>	0.50	(0.20-0.71)	0.56	(0.13-0.82)
<b>LumB</b>	0.33	(0.15-0.39)	0.45	(0.13-0.69)

**Table 2.** Median subtype correlation coefficients to corresponding PAM50 subtypes and range for DCIS and IBC.

The overall lower correlation to the PAM50 centroids in DCIS compared to IBC prompted us to explore the expression of the PAM50 genes in each subtype and tumor stage to identify the contribution of each gene to the subtyping output (Suppl. Fig. 1b). Only one gene (Matrix metalloproteinase 11, *MMP11*, also named stromelysin) clearly delineated DCIS and IBC. *MMP11* is expressed in stromal cells

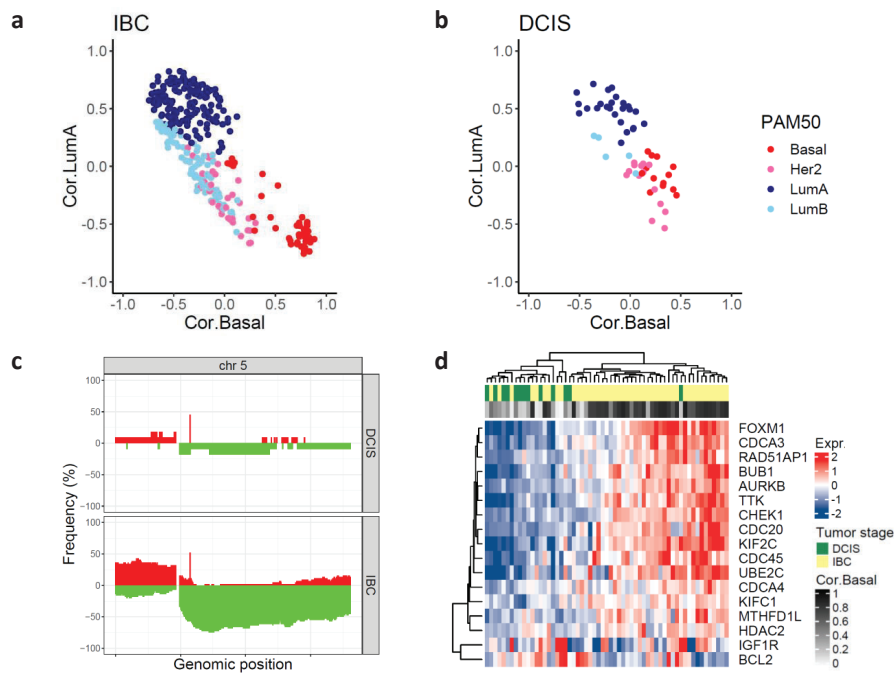


**Fig. 1.** Distribution of PAM50 subtypes (a) and *ESR1* gene expression (b) in DCIS and IBC. The difference between DCIS and IBC is significant for both PAM50 subtypes ( $p=0.0016$ ) and *ESR1* gene expression ( $p=0.0012$ , Fisher exact tests)

and favors cancer cell survival and tumor progression through cleavage of collagen VI (28). *MMP11* was markedly lower expressed in DCIS of all subtypes compared to IBC, in accordance with its non-invasive state. All other PAM50 genes showed expression patterns characteristic of the subtypes, independent of tumor stage. Luminal genes (e.g. *ESR1*, *PGR*, *NAT1*, *BCL2*, *SLC39A6*) were higher expressed in luminal tumors of both DCIS and IBC compared to tumors of basal-like and HER2-enriched subtypes. Basal-like IBC showed markedly higher expression of genes associated with proliferation compared to all other subtypes (including basal-like DCIS). Both DCIS and IBC of the HER2-enriched subtype showed elevated expression of genes typically highly expressed in this subtype (*ERBB2*, *GRB7* and *TMEM45B*). Of note were keratins associated with basal epithelium (*KRT5*, *KRT14* and *KRT17*) markedly higher expressed in DCIS of non-basal subtypes compared to their invasive counterpart while for the basal-like subtype, these keratins were highly expressed both in DCIS and IBC. This observation may be explained by gene expression contribution from a retained myoepithelial cell layer at the DCIS stage.

Interestingly, we identified a distinct group of basal-like IBCs with high correlation to the basal-like centroid and correspondingly low correlation to the luminal A centroid (Fig. 2a), which was not found among basal-like DCIS (Fig. 2b). These invasive tumors may correspond to so-called “core basal” tumors, characterized by deletions on chromosome 5q and high expression of specific genes associated “in trans” with such deletions (29, 30). In accordance with this, we





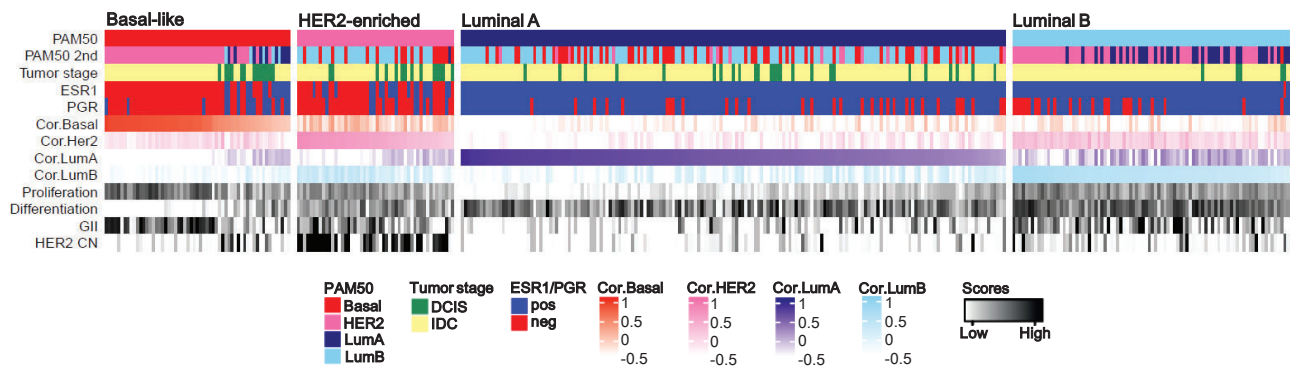
**Fig. 2. Core basal characteristics.** (a) and (b): Association between correlation coefficient to basal-like centroid on the x-axis vs. correlation coefficient to luminal A centroid on the y-axis for IBC and DCIS. (c): Frequencyplot of copy number data on chromosome 5. Genomic position is shown on x-axis. Y-axis show frequency of losses (downward in green) or amplifications (upward in red) in DCIS and IBC separately. (d): Gene expression of core basal genes in DCIS and IBC tumors of basal-like subtype. The heatmap shows genes whose expression previously has been shown to be correlated “in trans” with deletion of 5q in core basal samples.

found 5q deletions at high frequency in basal-like IBC, while only in a minority of basal-like DCIS (Fig. 2c). Clustering gene expression values of the core basal-defining genes revealed two distinct clusters: one consisting of mostly IBC tumors with high correlation to basal-like subtype (i.e. the core basal tumors), and a second cluster including most of the DCIS tumors and IBC tumors with low correlation to basal-like subtype (Fig. 2d). The absence of core basal tumors in DCIS suggests that basal-like DCIS may be a different biological entity than most basal-like IBC.

**Extensive genomic differences between basal-like DCIS and basal-like IBC.** We found few gene expression differences between DCIS and IBC when performing genome wide principal component analysis (PCA) across all subtypes (Suppl. Fig. 2a). This is in accordance with previous studies (17, 18). However, PCA after subtype stratification clearly separated IBC from DCIS in the basal-like and HER2-enriched subtypes, while not in the luminal subtypes (Suppl. Fig. 2b). Also with respect to copy number aberrations, differences between DCIS and IBC varied between subtypes. DCIS exhibited overall fewer copy number changes compared to IBC as demonstrated by overall lower genomic instability index (GII) in all subtypes, and the difference was significant for all subtypes except luminal B (Suppl. Fig. 3a and Suppl. File 1). Nevertheless, the specific copy number changes in DCIS are reminiscent of invasive tumors, including 17q12 amplification in the HER2-enriched subtype and deletions of 16q in luminal A DCIS (Suppl. Fig. 4). Again, the largest difference between DCIS and IBC was found for basal-like tumors with DCIS showing substantially

fewer copy number aberrations compared to basal-like IBC

To further explore subtype specific differences between DCIS and IBC, we included information on the strength of the correlation to all other subtype centroids (Fig. 3, Suppl. File 1). We found that basal-like IBC correlated highly to the basal-like centroid, and next, to the HER2-enriched centroid, while basal-like DCIS showed overall lower correlation to the basal-like centroid and more often had luminal subtypes as their second subtype (Fig. 3). On the contrary, luminal A tumors, both DCIS and IBC, showed relatively high correlation to the luminal A centroid and a similar distribution of the second best subtype (mostly basal-like and luminal B). Next, we calculated gene expression-based proliferation-, differentiation-, immune-, stromal-, and epithelial-to-mesenchymal transition (EMT)-scores, as well as HER2-copy number status (Fig. 3, Suppl. Fig. 3 and Suppl. File 1). All tumors at both disease stages showed subtype specific characteristics such as higher proliferation and lower differentiation in basal-like and HER2-enriched subtypes when compared to luminal A. In general, DCIS received lower stromal and EMT scores compared to IBC. The differences between DCIS and IBC were most pronounced in basal-like tumors: Basal-like DCIS displayed significantly lower median proliferation score compared to basal-like IBC (Suppl. Fig. 3b), while the median differentiation score was significantly higher in basal-like DCIS compared to IBC (Suppl. Fig. 3c), although still lower than in DCIS of any other subtype. There was no statistically significant difference in median immune score, median stromal score or median EMT score between basal-like DCIS and IBC (Suppl.



**Fig. 3. Genomic characteristics of DCIS and IBC tumors.** Each column represents one tumor and are sorted according to PAM50 subtype and ordered by correlation coefficient to the tumor's subtype. Relevant characteristics that commonly differ between molecular subtypes are selected and revealed great variation between subtypes with regards to the difference between DCIS and IBC, with most pronounced differences in the basal-like subtype. PAM50: the sample's subtype, PAM50 2nd: the subtype with second highest correlation, Tumor stage: DCIS (green), IBC (yellow), *ESR1*: Estrogen receptor 1 gene expression, *PGR*: Progesterone receptor gene expression, Cor.Basal/Cor.HER2/Cor.LumA/Cor.LumB: correlation coefficients to the four PAM50 subtypes, Proliferation: gene expression based proliferation score, Differentiation: gene expression based differentiation score, GII: Genomic Instability Index based on copy number data, HER2 CN: Erb-B2 Receptor Tyrosine Kinase 2 copy number.

Fig. 3d, e and f). Overall, these findings show that subtype profiles of DCIS are comparable to those found in IBC, except in the basal-like subtype where DCIS appears to be associated with less aggressive gene expression characteristics.

### Long Range Epigenetic Silencing of cPCDH genes occurs in basal-like IBC.

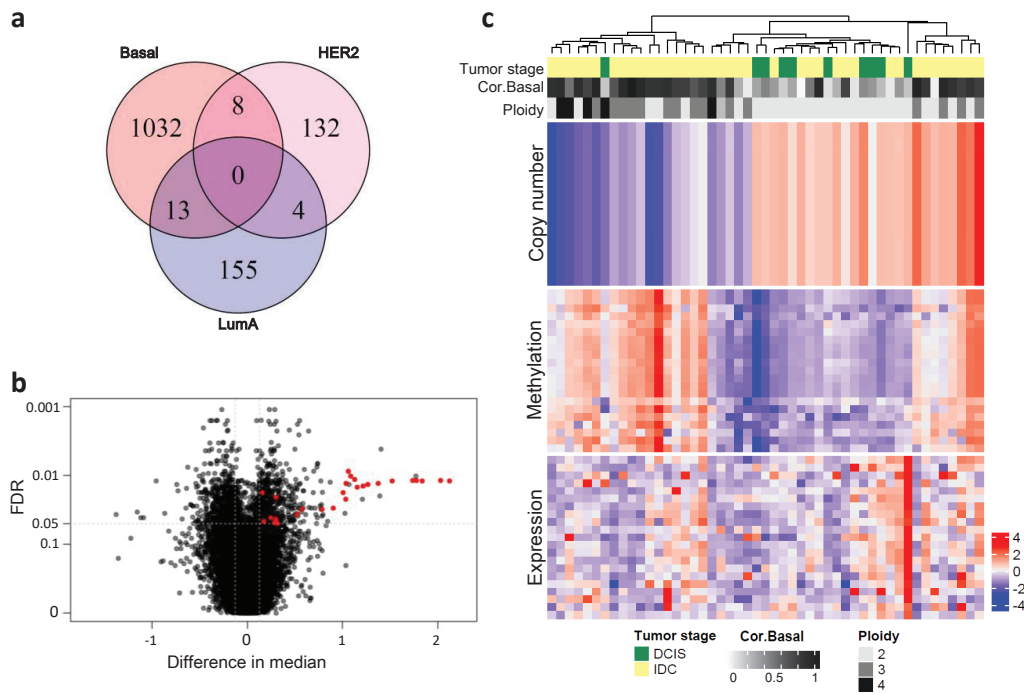
We identified numerous genes with significantly different methylation profile between DCIS and IBC (Suppl. file 2). For the basal-like subtype, 1053 genes showed statistically significant different methylation profile between DCIS and IBC, while for the HER2-enriched and luminal A subtypes, only 144 and 172 genes, respectively, showed significantly different methylation profiles (Fig. 4a). Due to low sample size, no genes with statistically significant different methylation signatures were identified for the luminal B subtype. No differentially methylated genes were common between the other three subtypes. Functional enrichment analysis of genes with significant different methylation profile between DCIS and IBC revealed that cell junction processes were prominent for the basal-like subtype (Suppl. file 3); most notable were nineteen clustered protocadherin genes (cPCDHs, 18 PCDH $\gamma$  and 1 PCDH $\beta$  genes) hypermethylated in basal-like IBC compared to DCIS (figure 4b). These genes are located close together on chromosome 5q31 and are involved in cell-cell adhesion (31, 32). Long Range Epigenetic Silencing (LRES) has previously been shown to occur in cancer across the cPCDH gene clusters (33–35). To explore whether LRES could explain the observed hypermethylation in basal-like IBC, we clustered all basal-like tumors based on the actual methylation status of each CpG in the 800kb window spanning the clustered protocadherin genes (Suppl. Fig. 5). This revealed two distinct clusters of tumors: One group characterized by hypermethylation of CpGs located in the cPCDH genes, including most of the IBC tumors with high correlation to the basal-like centroid, and a second group characterized by lower methylation, consisting of most of the DCIS tumors and the IBC tumors with lower correlation to the basal-like centroid. These findings indicate

that LRES of cPCDHs may be a trait of basal-like IBC, but not of basal-like DCIS.

When compiling methylation, copy number and gene expression data of the cPCDHs for the basal-like tumors, it appeared that invasive tumors with hypermethylation of the cPCDH genes often exhibited deletions of the same genes, and that these changes corresponded well with correlation to the basal-like centroid (Fig. 4c). Importantly, the cluster of tumors with concurrent hypermethylation and deletion of the cPCDH genes consisted mainly of aneuploid tumors, while the sub-cluster containing most DCIS consisted of diploid tumors only. In summary, the notable differences in cPCDH methylation between basal-like DCIS and IBC support our previous results that basal-like DCIS may be a different entity than basal-like IBC.

## Discussion

In this study, we have explored differences between DCIS and IBC in a subtype specific manner using gene expression, copy number and DNA-methylation data derived from fresh frozen tumor material. The study was instigated by findings from our previous study where we hypothesized that progression of DCIS to invasive cancer differ between molecular subtypes (20). The indolent nature of many in situ tumors and the fact that many of these tumors never progress to invasive or metastatic disease harmonize poorly with the results from several studies showing remarkably few genomic differences between DCIS and IBC (16–18). This lack of genomic dissimilarity may be explained by inherent differences between the molecular subtypes: In most breast cancer cohorts, the majority of tumors are of luminal subtypes; hence, characteristics that differentiate disease stages in unstratified analyses are confounded by subtypes. The different distribution of molecular subtypes observed between IBC and DCIS may in part be explained by underrepresentation of small DCIS lesions included in the cohort. However, the frequency of tumors of the least aggressive subtype (luminal A) is similar in DCIS and IBC, indicating that the observed difference in



**Fig. 4. Methylation differences between DCIS and IBC.** (a): Genes with significantly different methylation profiles between DCIS and IBC in basal-like, HER2-enriched and Luminal A subtypes. Luminal B is not shown since no genes showed significantly differentially methylated profiles between DCIS and IBC in this subtype. (Mann Whitney U test,  $FDR < 0.05$ , effect size  $> 0.127$ ), (b): Volcano plot showing the results from Mann Whitney U test comparing methylation profiles in basal-like DCIS vs. basal-like IBC. Difference in median (IBC-DCIS) is shown on the x-axis and False Discovery Rate is shown on the y-axis. Genes colored in red are clustered protocadherins (cPCDHs) (hypermethylated in basal-like IBC compared to basal-like DCIS), (c): Copy number, methylation and gene expression of the 19 cPCDHs significantly differentially methylated between basal-like DCIS and basal-like IBC. cPCDHs are plotted in the same order in all three panels.

subtype distribution between the two tumor stages represents a true distinction.

Interestingly, the most pronounced stage differences were found for the basal-like subtype. Basal-like DCIS showed lower correlation to the basal-like centroid (i.e. low “basalness”) compared to basal-like IBC, and there were no “core basal” DCIS in our data. This is in accordance with a previous integrative clustering analysis that showed genomic isolation of basal-like IBC, and not basal-like DCIS (36). In the present study we showed that the basal-like DCIS tumors exhibited higher correlation to Luminal A subtype, higher degree of differentiation, lower proliferation and lower genomic instability than basal-like IBC. Also with respect to alterations of DNA methylation did basal-like tumors show prominently more differences between DCIS and IBC compared to all other subtypes. Most notable was the marked hypermethylation of CpGs mapping to the clustered protocadherin genes (cPCDHs) in basal-like IBC compared to DCIS and a positive association between hypermethylation of cPCDHs and degree of “basalness”. Hypermethylation of DNA in the genomic location spanning the cPCDH genes through long range epigenetic silencing (LRES) (37) has been shown to increase with progression of cervical cancer (35) and has also been seen in breast cancer (34), colorectal cancer (33) and Wilm’s tumor (38). Interestingly, the chromosomal region of the cPCDH genes (5q31) is frequently

deleted in basal-like IBCs and is a defining feature of core basal IBC tumors (39, 40). Clustered protocadherins are molecules involved in cell-cell adhesion and have also been shown to inhibit cell growth and suppress oncogenic pathways, features consistent with a role as tumor suppressors (41). Loss of intraepithelial cell-cell adhesion is a key feature during tumor cell invasion (42, 43) and it is tempting to speculate that loss of cPCDH tumor suppressor function through LRES may contribute to driving the invasion process specifically in basal-like cancer.

We have shown that the difference between DCIS and IBC is greater for the basal-like subtype compared to all other subtypes. Despite that the intrinsic subtypes were defined in invasive breast cancer, we believe that basal-like DCIS are truly basal-like since firstly, the PAM50 subtyping showed that they correlate the most to the basal-like centroid, albeit to a lower degree than IBC. Secondly, several genomic features of basal-like tumors are also present in basal-like DCIS, including low degree of differentiation, high expression of basal keratins, low expression of luminal genes and immune cell infiltration. Despite these similarities, basal-like DCIS may not be true precursors to basal-like IBC. Basal-like breast cancer is an aggressive disease that develops rapidly. Especially the core basal tumors have an aggressive phenotype with poorer prognosis than non-core basal tumors (30, 44). Although all core basal invasive tumors at some point must

have had an intraductal stage, the transition from DCIS to an invasive stage may occur so rapidly that the probability of “capturing” such a tumor at the DCIS stage is very small as also proposed by Kurbel (45). This hypothesis is supported by the fact that basal-like invasive tumors have fewer concurrent DCIS lesions compared to other subtypes (46). By this follows a hypothesis that those DCIS that are identified as basal-like may be indolent or precursors to the less aggressive non-core basal tumors.

A limitation of this study is the lack of follow-up information on recurrence or survival. Hence, our results need to be validated in a DCIS cohort with more extensive clinical follow-up information. Nevertheless, our study has reaffirmed the necessity of taking a subtype specific approach when studying progression of DCIS and we have demonstrated that there are substantial differences between basal-like DCIS and IBC that may question basal-like DCIS as precursor lesions to invasive breast carcinoma.

## Material and Methods

**Tumor samples.** This study includes data from 57 DCIS and 313 IBC obtained from three different patient cohorts, of which two (“Uppsala” and “Oslo2”) are previously published (47, 48). Data from the third cohort, (“Milano”) is not previously published and was generated from fresh frozen tissue from a total of 34 breast tumors. Histopathological evaluation of H&E stained tissue sections was performed by a trained pathologist.

**DNA and RNA isolation.** Total RNA and DNA was isolated using the QIAcube system with the AllPrep DNA/RNA Universal Kit (cat.no. 80224, Qiagen, Hilden, Germany) with 30mg tissue as input. The tissue was manually minced with a scalpel on ice followed by homogenization using Tissue-Lyzer LT and Qiasredder (Qiagen). RNA and DNA extraction was performed according to the protocol provided by the supplier. Nucleic acid concentrations were measured on a NanoDrop ND-1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and RNA integrity was analyzed using Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA).

**Gene expression analysis.** To obtain whole genome expression data, Agilent Sureprint G3 Human Gene Expression 8x60K microarrays (G4851A) (Agilent, Technologies, Santa Clara, USA) with the Low Input Quick Amp Labeling protocol were used. RNA input was 40ng and Cy3 was used as fluorophore. Quality Control (QC) was performed in Agilent’s Feature Extraction software. From the Milano cohort, five invasive breast carcinomas and 28 DCIS were successfully analyzed and passed all quality control criteria while one DCIS failed QC. As a control, one sample of commercially available normal breast RNA (Ambion Human Breast Total RNA, Thermo Fisher Scientific, Wilmington, DE, USA) was included throughout the whole experimental pipeline. The same microarray platform had been used for the two other patient cohorts. Data from all three cohorts were normalized

together using quantile normalization. For genes represented with more than one probe, mean expression was calculated to obtain one gene expression value per gene.

**Whole genome methylation.** DNA-methylation data was obtained using the Illumina Infinium HumanMethylation450 microarray (Illumina, Inc. CA, USA) following the manufacturer’s instructions. Data was preprocessed using subset quantile normalization (49). The resulting  $\beta$  value represents the fraction of methylated DNA molecules at a specific CpG. Quality control of  $\beta$  values was performed as presented in Wilhelm Benartzi et al. (50):  $\beta$ -values with detection p-values higher than 0.05 (0.225% of the  $\beta$ -values) were replaced by NA. CpG sites where more than 25% of the  $\beta$  values failed quality control, were removed from the analysis resulting in 436 162 reliable CpGs in the final dataset. NA values were imputed using the R-function `impute.knn` with default parameters. To obtain one value per gene, principal component analysis was performed on all CpGs within, or 50kB upstream or downstream from the gene for each sample. The value of the first principal component represents the gene’s methylation profile.

**Copy number aberrations analysis.** Copy number data was obtained using Affymetrix SNP 6.0 arrays (Affymetrix, Santa Clara, CA, USA) at Aros Applied Biotechnology (Aarhus, Denmark) following the manufacturer’s instructions. CEL-files were processed using the PennCNV-Affy library (51) with the HapMap samples as reference set (52) and corrected for GC content (53). The data was segmented using the PCF algorithm with arguments `kmin=5`, `gamma=100` in the R `copynumber` package (54). The copy number of the segment overlapping the gene the most was set as a gene’s copy number. Ploidy and tumor percentage were calculated using the ASCAT algorithm (27). Genome instability index (GII) was derived by calculating the fraction of the genome affected by copy number change.

**PAM50 subtyping.** The tumors were assigned a PAM50 gene expression subtype using the centroid based method from Parker et al. (26) with four subtypes: Basal-like, HER2-Enriched, Luminal A and Luminal B. DCIS and IBC tumors were subtyped together after data normalization. To account for different fractions of estrogen receptor (ER) positive tumors between the training set (from which the centroids were calculated) and test set, the mean values were weighted by the proportion of ER+ tumors. ER-status by IHC was unavailable for some tumors, thus ER-status was determined using the *ESR1* gene expression value. *ESR1* expression showed a distinct bimodal distribution enabling a reliable cut-off to be set. Progesterone receptor (PR) status was derived by *PGR*-expression the same way as for ER. Consistency between ER status derived by IHC and expression was high, with 98% of the tumors (320/327) concurring. After gene centering, we calculated spearman correlation between expression of the PAM50 genes and each of the four subtype centroids and assigned each tumor to a subtype by its highest correlation (Suppl. file 1).

**Gene expression based tumor scores.** Proliferation scores were calculated using an 11-gene proliferation signature (55) and EMT scores were calculated using an EMT signature based on four adhesion genes (weighted negatively) and seven EMT-genes (weighted positively) (Suppl. file 1): For each gene and sample, a standard (Z) score was calculated, then the proliferation/EMT-scores were obtained for every tumor by calculating the mean of all Z-scores across all genes in the signature. Differentiation scores were derived using the differentiation predictor described in Prat et al. (56) and immune and stromal infiltration scores were calculated using ESTIMATE (57).

**Differential methylation.** Genes differentially methylated between DCIS and IBC were identified using Mann-Whitney U tests separately for each subtype. False discovery rate was used to correct for multiple testing. To identify gene lists for functional enrichment analyses, cut-offs were set at both FDR and effect size (defined as the absolute difference in median between DCIS and IBC) to increase the likelihood of finding the biological relevant differences between the two groups. We included genes with FDR<0.05 and effect size within the top 20% (corresponds to a cut-off > 0.127). Functional enrichment analyses of differentially methylated genes were performed using WebGestalt 2019 (WEB-based Gene Set Analysis Toolkit) (58).

**Statistical and bioinformatic analyses.** All statistical analyses were conducted in R (59) unless otherwise specified. Heatmaps were created using the R package ComplexHeatmaps (60) and other plots were created using the package ggplot2 (61). Fisher exact tests were used to compare distribution of subtype and ER-status between the two tumor stages. Mann Whitney U-tests were used to compare tumor content, GII, proliferation scores, differentiation scores, immune scores, stromal scores and EMT scores between DCIS and IBC separately for each subtype.

#### ABBREVIATIONS

BCS: breast conserving surgery; cPCDH: clustered protocadherin; DCIS: ductal carcinoma in situ; EMT: epithelial-to-mesenchymal transition; GII: genomic instability index; IBC: invasive breast cancer; LRES: long range epigenetic silencing; MMP11: matrix metalloproteinase 11; PCA: principal component analysis; PR: progesterone receptor.

#### ETHICAL APPROVAL AND CONSENT TO PARTICIPATE

All women provided a signed informed consent for future biomarker research study. The study complies with the Declaration of Helsinki, and was approved by the each institution's internal review and ethics board (approval numbers: 2016/433 (Oslo, Norway), PG/U-25/01/2012-00001497 (Milan, Italy), 2005/118 (Uppsala, Sweden).

#### DATA AVAILABILITY

Data will be deposited in a suitable repository prior to publication.

#### ACKNOWLEDGEMENTS

This research was supported by funds from Helse Sør-Øst (2012056) and the Norwegian Cancer Society (420056) to TS. We want to thank Eldri U. Due and Phuong Vu for assisting in DNA/RNA isolation and gene expression microarray experiments.

#### AUTHOR CONTRIBUTIONS

Conceptualization: HB, TL, TS, AF, DS; Laboratory work: HB; Data analysis: HB, TS; Funding acquisition: TS; Visualization: HB; Writing manuscript: HB, TL, TS; All authors have read and approved final manuscript

#### COMPETING INTERESTS

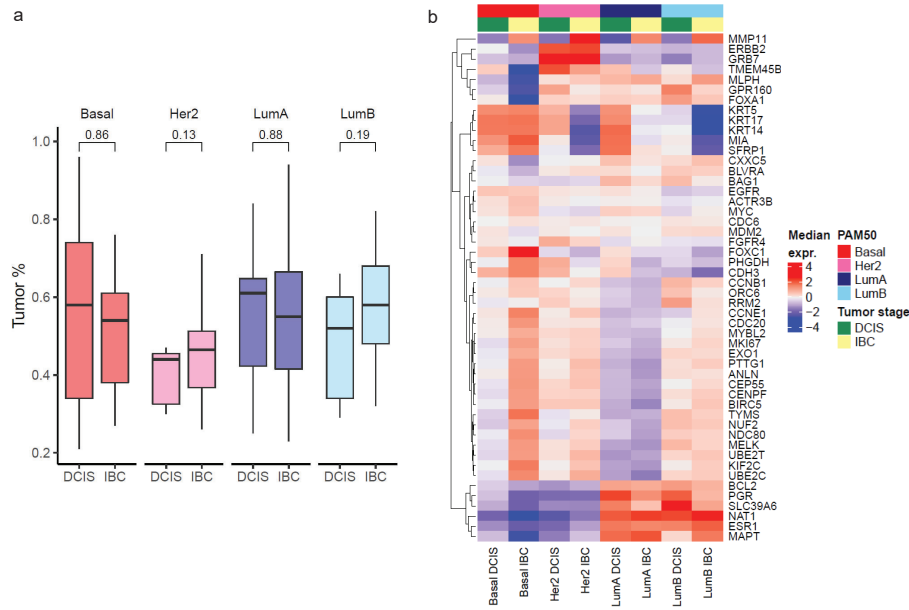
The authors declare no competing interests.

## References

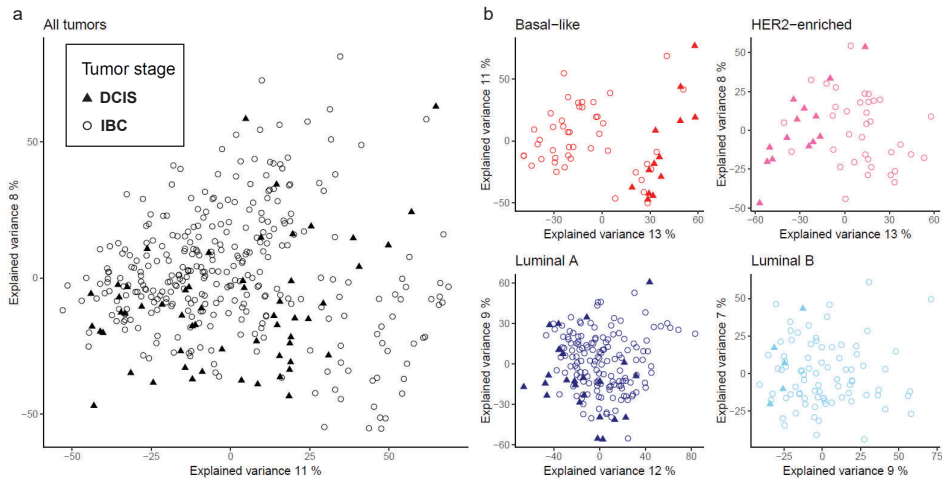
- Cowell, C. F. *et al.* Progression from ductal carcinoma in situ to invasive breast cancer: Revisited. *Molecular oncology* **7**, 1–11 (2013).
- Ernster, V. L. *et al.* Detection of Ductal Carcinoma In Situ in Women Undergoing Screening Mammography. *JNCI: Journal of the National Cancer Institute* **94**, 1546–1554 (2002).
- Seely, J. M. & Alhassan, T. Screening for breast cancer in 2018—what should we be doing today? *Current Oncology* **25**, S115–S124 (2018).
- Virnig, B. A., Tuttle, T. M., Shamlivan, T. & Kane, R. L. Ductal Carcinoma In Situ of the Breast: A Systematic Review of Incidence, Treatment, and Outcomes. *Journal of the National Cancer Institute* **102**, 170–178 (2010).
- Collins, L. C. *et al.* Outcome of patients with ductal carcinoma in situ untreated after diagnostic biopsy: results from the Nurses' Health Study. *Cancer* **103**, 1778–1784 (2005).
- Nielsen, M., Jensen, J. & Andersen, J. Precancerous and cancerous breast lesions during lifetime and at autopsy. A study of 83 women. *Cancer* **54**, 612–615 (1984).
- Page, D. L., Dupont, W. D., Rogers, L. W. & Landenberger, M. Intraductal carcinoma of the breast: follow-up after biopsy only. *Cancer* **49**, 751–758 (1982).
- Page, D. L., Dupont, W. D., Rogers, L. W., Jensen, R. A. & Schuyler, P. A. Continued local recurrence of carcinoma 15–25 years after a diagnosis of low grade ductal carcinoma in situ of the breast treated only by biopsy. *Cancer* **76**, 1197–1200 (1995).
- Sanders, M. E., Schuyler, P. A., Dupont, W. D. & Page, D. L. The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. *Cancer* **103**, 2481–2484 (2005).
- Burstein, H. J., Polyak, K., Wong, J. S., Lester, S. C. & Kaelin, C. M. Ductal Carcinoma in Situ of the Breast. *New England Journal of Medicine* **350**, 1430–1441 (2004).
- Gorringe, K. L. & Fox, S. B. Ductal Carcinoma In Situ Biology, Biomarkers, and Diagnosis. *Frontiers in Oncology* **7**, 248 (2017).
- Esserman, L. J. *et al.* Addressing overdiagnosis and overtreatment in cancer: a prescription for change. *Lancet Oncol.* **15**, e234–e242 (2014).
- Groen, E. J. *et al.* Finding the balance between over- and under-treatment of ductal carcinoma in situ (DCIS). *The Breast* **31**, 274–283 (2017).
- Narod, S. A., Iqbal, J., Giannakeas, V., Sopik, V. & Sun, P. Breast Cancer Mortality After a Diagnosis of Ductal Carcinoma In Situ. *JAMA Oncology* **1**, 888–896 (2015).
- Sagara, Y., Julia, W., Golshan, M. & Toi, M. Paradigm shift toward reducing overtreatment of ductal carcinoma in situ of breast. *Frontiers in Oncology* **7**, 192 (2017).
- Hwang, E. S. *et al.* Patterns of chromosomal alterations in breast ductal carcinoma in situ. *Clinical Cancer Research* **10**, 5160–5167 (2004).
- Ma, X.-J. *et al.* Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 5974–5979 (2003).
- Vincent-Salomon, A. *et al.* Integrated genomic and transcriptomic analysis of ductal carcinoma in situ of the breast. *Clinical Cancer Research* **14**, 1956–1965 (2008).
- Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 10869–10874 (2001).
- Lesurf, R. *et al.* Molecular Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast Cancer. *Cell Reports* **16**, 1166–1179 (2016).
- Onega, T. *et al.* The diagnostic challenge of low-grade ductal carcinoma in situ. *European Journal of Cancer* **80**, 39–47 (2017).
- Wallis, M. G. *et al.* The effect of DCIS grade on rate, type and time to recurrence after 15 years of follow-up of screen-detected DCIS. *British Journal of Cancer* **106**, 1611–1617 (2012).
- Wang, S. Y., Shamlivan, T., Virnig, B. A. & Kane, R. Tumor characteristics as predictors of local recurrence after treatment of ductal carcinoma in situ: A meta-analysis. *Breast Cancer Research and Treatment* **127**, 1–14 (2011).
- Rakovitch, E. *et al.* A population-based validation study of the DCIS Score predicting recurrence risk in individuals treated by breast-conserving surgery alone. *Breast Cancer Research and Treatment* **152**, 389–398 (2015).
- Hanna, W. M. *et al.* Ductal carcinoma in situ of the breast: an update for the pathologist in the era of individualized risk assessment and tailored therapies. *Modern Pathology* **32**, 896–915 (2019).
- Parker, J. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160–1167 (2009).
- Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 16910–16915 (2010).
- Motrescu, E. R. *et al.* Matrix metalloproteinase-11/stromelysin-3 exhibits collagenolytic function against collagen VI under normal and malignant conditions. *Oncogene* **27**, 6347–6355 (2008).
- Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
- Tischkowitz, M. *et al.* Use of immunohistochemical markers can refine prognosis in triple negative breast cancer. *BMC Cancer* **7**, 134 (2007).
- Chen, W. V. & Maniatis, T. Clustered protocadherins. *Development (Cambridge)* **140**, 3297–3302 (2013).
- Gul, I. S., Hulpiau, P., Saeys, Y. & van Roy, F. Evolution and diversity of cadherins and catenins. *Experimental Cell Research* **358**, 3–9 (2017).
- Dallosso, A. R. *et al.* Long-range epigenetic silencing of chromosome 5q31 protocadherins is involved in early and late stages of colorectal tumorigenesis through modulation of oncogenic pathways. *Oncogene* **31**, 4409–4419 (2012).
- Novak, P. *et al.* Agglomerative epigenetic aberrations are a common event in human breast cancer. *Cancer Research* **68**, 8616–8625 (2008).
- Wang, K. H. *et al.* Global methylation silencing of clustered proto-cadherin genes in cervical cancer: Serving as diagnostic markers comparable to HPV. *Cancer Medicine* **4**, 43–55 (2015).
- Swanson, D. M., Lien, T., Bergholtz, H., Sorlie, T. & Frigessi, A. A Bayesian two-way latent structure model for genomic data integration reveals few pan-genomic cluster subtypes in a

- breast cancer cohort. *Bioinformatics* (2019).
37. Forn, M. *et al.* Long range epigenetic silencing is a trans-species mechanism that results in cancer specific deregulation by overriding the chromatin domains of normal cells. *Molecular Oncology* **7**, 1129–1141 (2013).
  38. Dallosso, A. R. *et al.* Frequent long-range epigenetic silencing of protocadherin gene clusters on chromosome 5q31 in Wilms' tumor. *PLoS Genetics* **5**, e1000745 (2009).
  39. Bergamaschi, A. *et al.* Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes Chromosomes and Cancer* **45**, 1033–1040 (2006).
  40. Yu, W., Kanaan, Y., Baed, Y. K. & Gabrielson, E. Chromosomal changes in aggressive breast cancers with basal-like features. *Cancer Genetics and Cytogenetics* **193**, 29–37 (2009).
  41. Van Roy, F. Beyond E-cadherin: Roles of other cadherin superfamily members in cancer. *Nature Reviews Cancer* **14**, 121–134 (2014).
  42. Gheldof, A. & Bex, G. Cadherins and epithelial-to-mesenchymal transition. In *Progress in Molecular Biology and Translational Science*, vol. 116, 317–336 (Academic Press, 2013).
  43. Huang, R. Y. J., Guilford, P. & Thiery, J. P. Early events in cell adhesion and polarity during epithelial-mesenchymal transition. *Journal of Cell Science* **125**, 4417–4422 (2012).
  44. Cheang, M. C. U. *et al.* Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clinical Cancer Research* **14**, 1368–1376 (2008).
  45. Kurbel, S. In search of triple-negative DCIS: Tumor-type dependent model of breast cancer progression from DCIS to the invasive cancer. *Tumor Biology* **34**, 1–7 (2013).
  46. Doebar, S. C. *et al.* Extent of ductal carcinoma in situ according to breast cancer subtypes: a population-based cohort study. *Breast Cancer Research and Treatment* **158**, 179–187 (2016).
  47. Aure, M. R. *et al.* Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. *Breast Cancer Research* **19**, 44 (2017).
  48. Mugggerud, A. A. *et al.* Molecular diversity in ductal carcinoma in situ (DCIS) and early invasive breast cancer. *Molecular oncology* **4**, 357–368 (2010).
  49. Touleimat, N. & Tost, J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4**, 325–341 (2012).
  50. Wilhelm-Benartzi, C. S. *et al.* Review of processing and analysis methods for DNA methylation array data. *British Journal of Cancer* **109**, 1394–1402 (2013).
  51. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research* **17**, 1665–74 (2007).
  52. The International HapMap Consortium. The international HapMap project. *Nature* **426**, 789–796 (2003).
  53. Diskin, S. J. *et al.* Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research* **36**, e126 (2008).
  54. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
  55. Nielsen, T. O. *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* **16**, 5222–5232 (2010).
  56. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research* **12**, R68 (2010).
  57. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications* **4**, 2612 (2013).
  58. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: An integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research* **33**, W741–W748 (2005).
  59. R Core Team & R Foundation for Statistical Computing. R: A language and environment for statistical computing. (2017). /[www.R-project.org](http://www.R-project.org).
  60. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
  61. Wickham, H. *Ggplot2 : elegant graphics for data analysis* (2016).

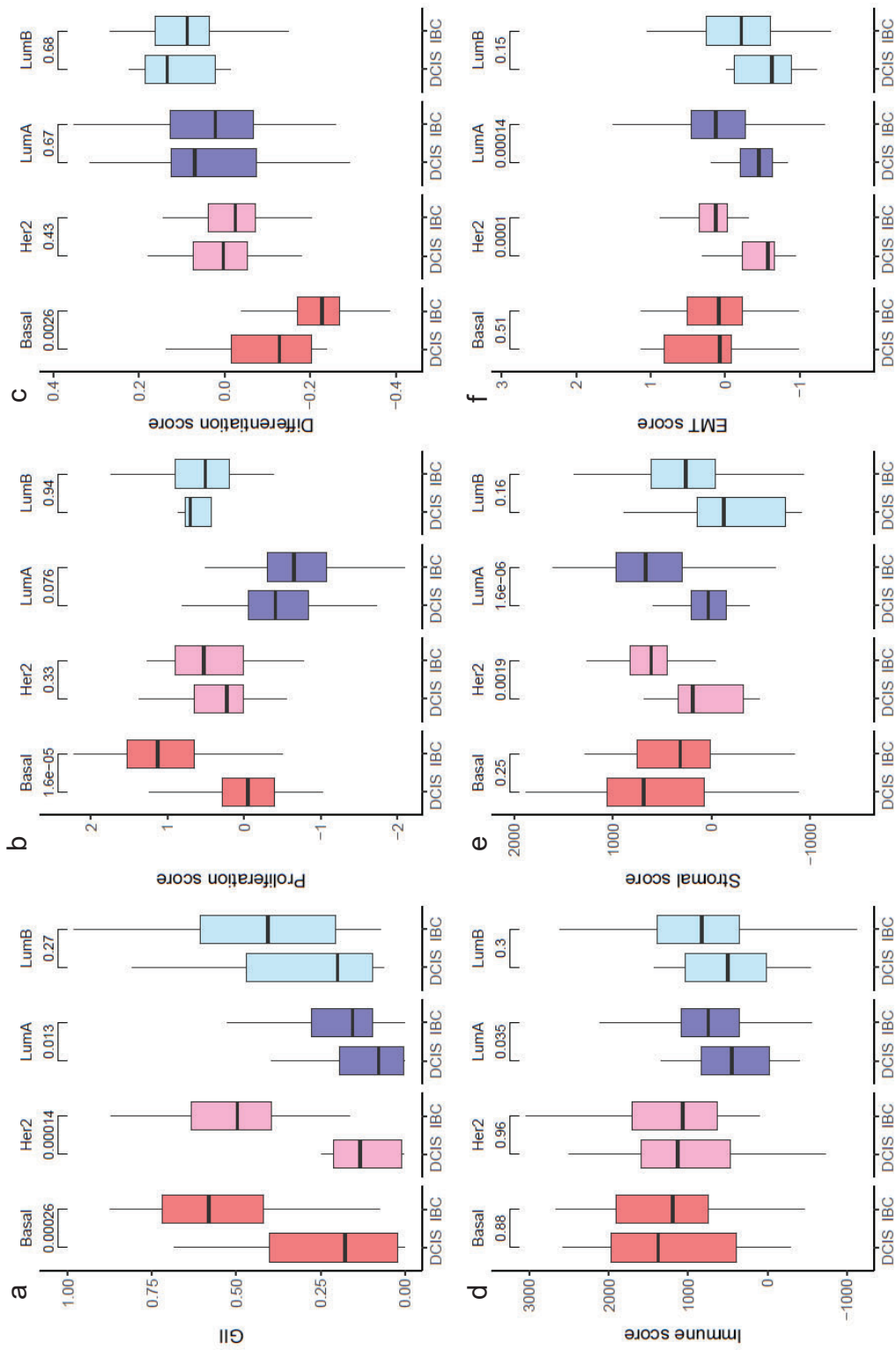
## Supplementary Figures



**Fig. S1.** (a): Tumor content boxplot separated by tumor stage and PAM50 subtype. P-values obtained by Mann Whitney U tests, DCIS vs. IBC in each subtype separately. (b): Heatmap showing median gene expression value for genes included in the PAM50 centroid for each subtype and tumor stage separately.

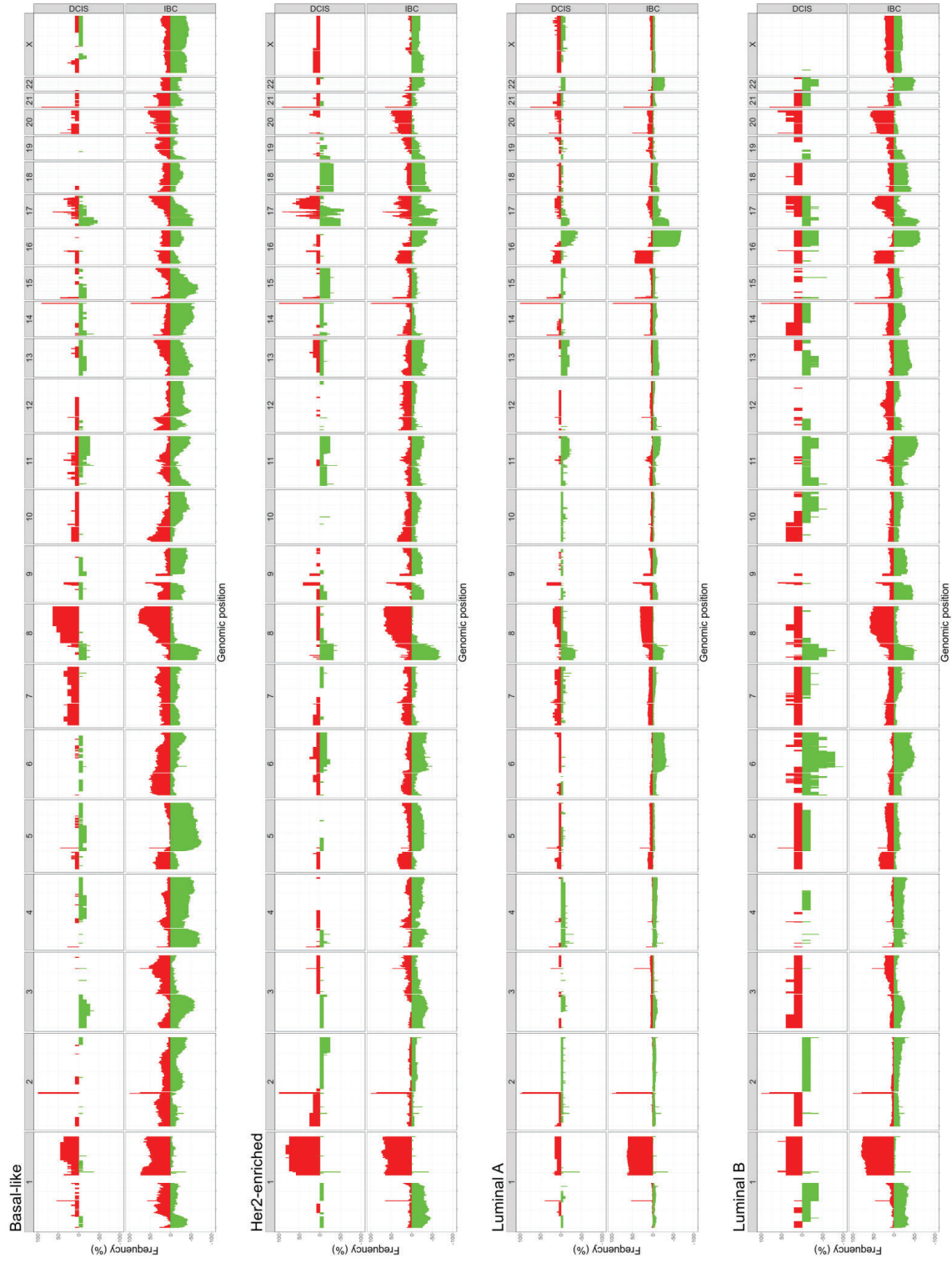


**Fig. S2.** Principal Component Analyses plots based on genome wide gene expression data of all samples together (a) and separately for the PAM50 subtypes (b). Principal component 1 is shown on the x-axis and principal component 2 on the y-axis. The explained variance in each PCA analyses is indicated.

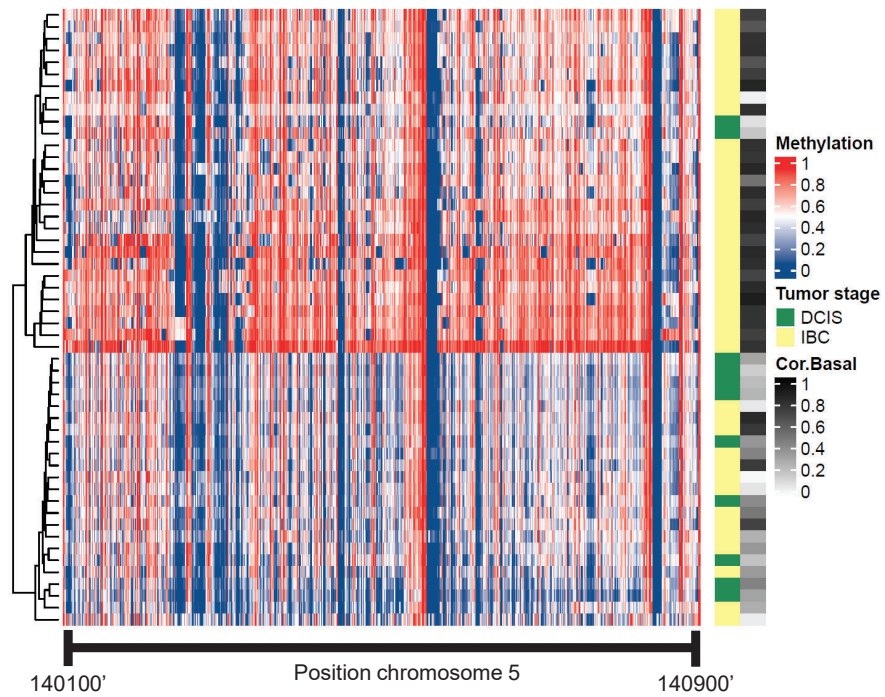


**Fig. S3.** Boxplots showing genomic instability score (GII) based on copy number (a), gene expression based proliferation score (b), differentiation score (c), immune score (d), stromal score (e) and EMT score (f). P-values obtained by Mann Whitney U tests, DCIS vs. IBC in each subtype separately.





**Fig. S4.** Frequencyplot of DCIS and IBC separately for the PAM50 subtypes. The x-axis show genomic position and the y-axis show frequency of losses (downward in green) or amplifications (upward in red).



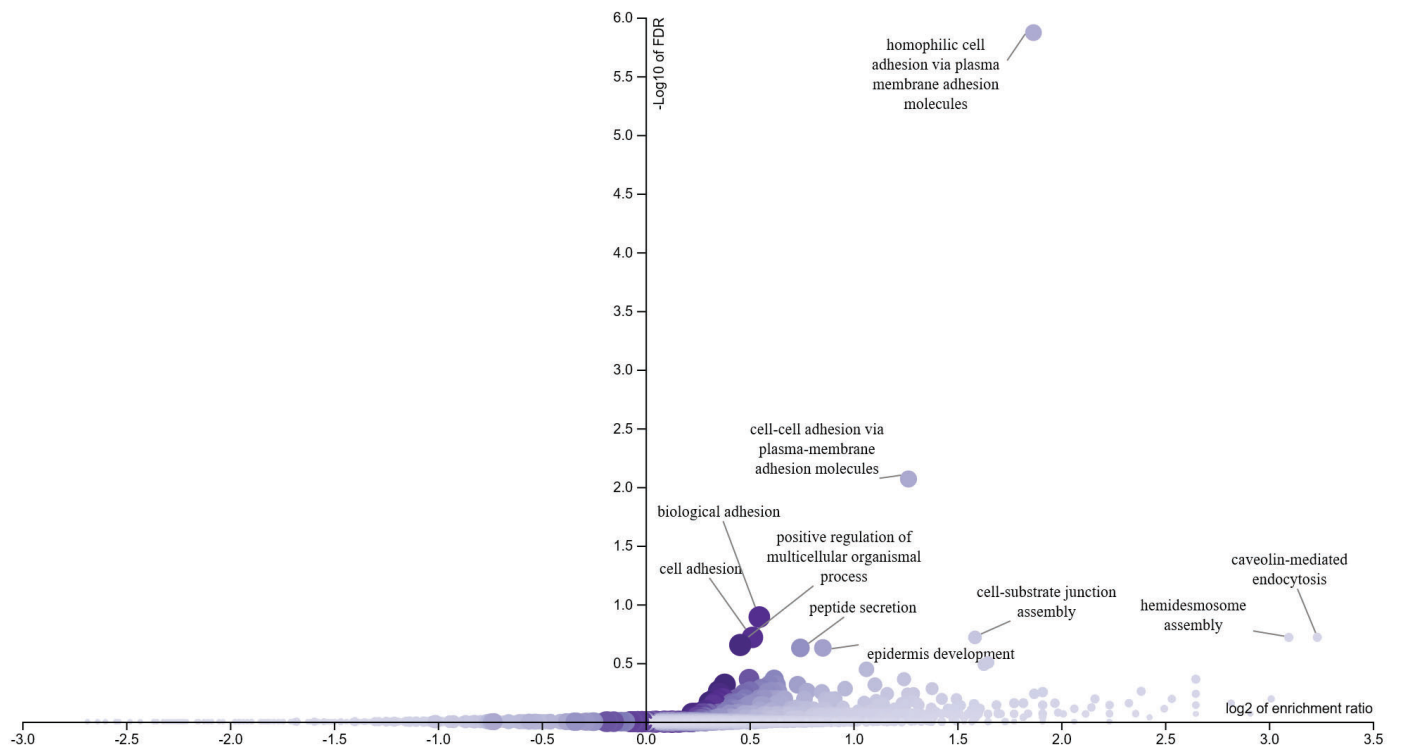
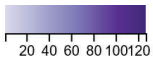
**Fig. S5.** Heatmap showing methylation status ( $\beta$ -values) of all CpGs in the 800kb window spanning the cPCDH genes on chromosome 5q. CpGs (columns) are ordered according to genomic position (indicated below). Rows (tumors) are clustered.

Supplementary files 1 and 2 are large tables that are made available electronically.

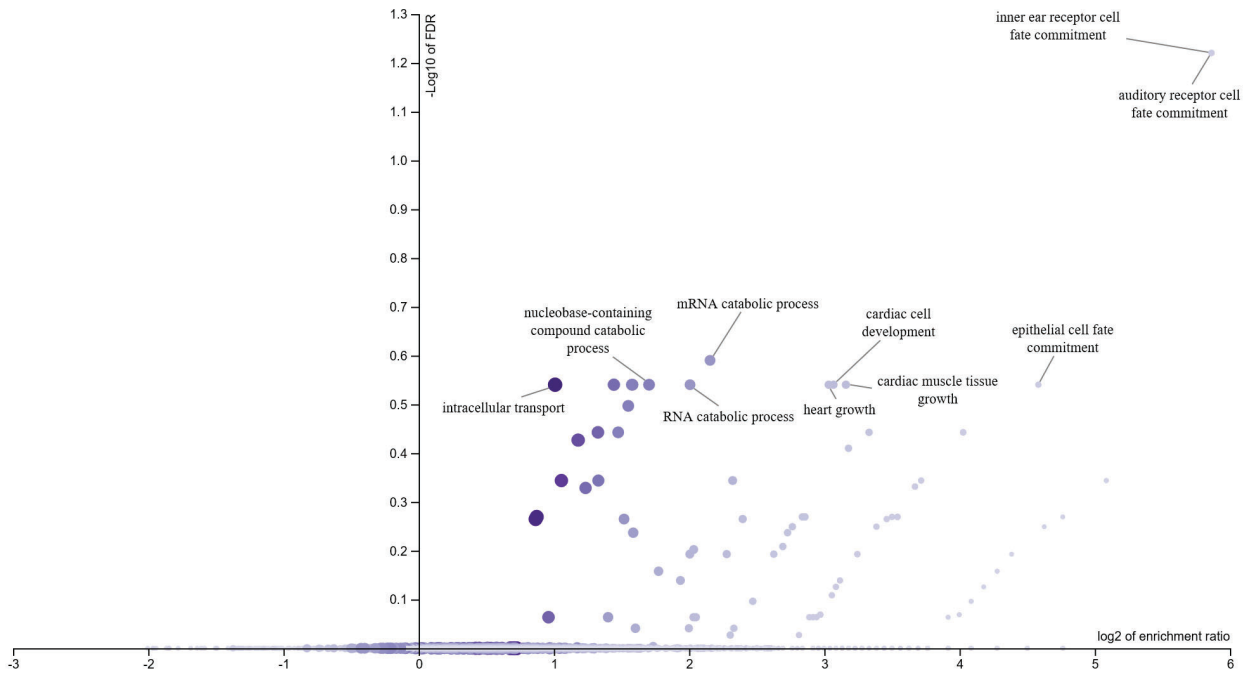
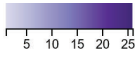
# Supplementary file 3

Gene set analyses using WEB-based GENE Set Analysis Toolkit (<http://www.webgestalt.org/>) Input are genes with significant different DNA methylation profile between DCIS and IBC (FDR<0.05 and effect size within top 20%). Analyses are performed separately for each PAM50 subtype (LumB not included since no genes had significantly different methylation profiles in this analyses).

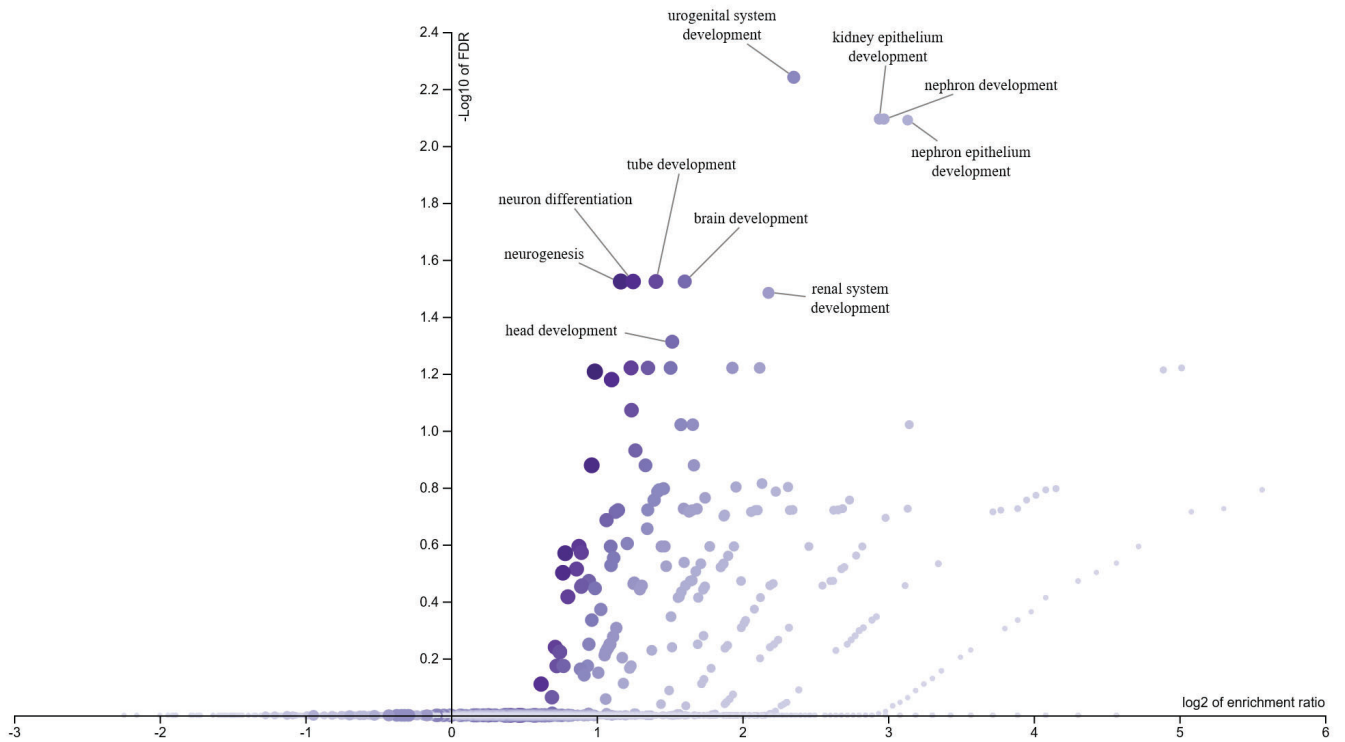
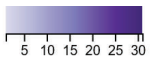
## Basal-like



# HER2-enriched



# Luminal A



## **Paper IV**

### **Comparable cancer–relevant mutation profiles in synchronous ductal carcinoma in situ and invasive breast cancer**

Helga Bergholtz, Surendra Kumar, Fredrik Wärnberg, Torben Lüders, Vessela Kristensen and Therese Sørliie.

*Manuscript*



# Comparable cancer–relevant mutation profiles in synchronous ductal carcinoma in situ and invasive breast cancer

Helga Bergholtz<sup>1</sup>, Surendra Kumar<sup>1</sup>, Fredrik Wärnberg<sup>3,4</sup>, Torben Lüders<sup>5</sup>,  
Vessela Kristensen<sup>1,2,5\*</sup>, and Therese Sørlie<sup>1,2\*✉</sup>

<sup>1</sup>Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Oslo, Norway

<sup>2</sup>Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

<sup>3</sup>Department of Surgical Sciences, Uppsala University, Uppsala, Sweden

<sup>4</sup>Department of Surgery, Uppsala Academic Hospital, Uppsala, Sweden

<sup>5</sup>Department of Clinical Molecular Biology (EpiGen), Division of Medicine, Akershus University Hospital, Lørenskog, Norway

\*VK and TS should be considered joint senior author

**BACKGROUND** Ductal carcinoma in situ (DCIS) comprises a diverse group of pre-invasive lesions in the breast and poses a considerable clinical challenge due to lack of markers of progression. Genomic alterations are to a large extent similar in DCIS and invasive carcinomas, although differences in copy number aberrations, gene expression patterns and mutations have been found. In mixed tumors with synchronous invasive breast cancer (IBC) and DCIS, it is still unclear whether invasive tumor cells are directly derived from the DCIS cells.

**AIMS** Our aim was to compare cancer-relevant mutation profiles of different cellular compartments in mixed DCIS/IBC and pure DCIS tumors.

**METHODS AND RESULTS** We performed targeted sequencing of 50 oncogenes in microdissected tissue from three different epithelial cell compartments (in situ, invasive and normal adjacent epithelium) from 26 mixed breast carcinomas. In total, 44 tissue samples (19 invasive, 16 in situ, 9 normal) were subjected to sequencing using the Ion Torrent platform and the AmpliSeq™ Cancer Hotspot Panel v2. For comparison, 10 additional, pure DCIS lesions were sequenced. Across all mixed samples, we detected 22 variants that previously have been described in cancer and are present in the COSMIC database. The most commonly affected genes were *TP53*, *PIK3CA* and *ERBB2*. The *PIK3CA*:p.H1047R variant was found in 9 samples from six patients. Most variants detected in the invasive compartment of a tumor were also found in the corresponding in situ cell compartment indicating a clonal relationship between the tumor stages. Across 10 pure DCIS lesions, only three variants were identified.

**CONCLUSION** Similar mutation profiles between in situ and invasive cell compartments indicate a similar origin of the two tumor stages in mixed breast tumors. The low number of potential driver variants found in pure DCIS compared to the in situ cell compartments of mixed tumors implies that these two in situ lesion types may be different entities.

DCIS | invasive breast cancer | mutations | targeted sequencing | breast tumor progression

Correspondence: [therese.sorlie@rr-research.no](mailto:therese.sorlie@rr-research.no)

## Introduction

Ductal carcinoma in situ (DCIS) is a non-invasive breast cancer. In DCIS, abnormal cells are contained within the milk

ducts while the basement membrane is intact, and there is no invasion of surrounding stroma (1). Today, DCIS comprises about 20% of all breast carcinoma diagnoses, usually detected in the context of mammography screening (2). In situ lesions are generally accepted as non-obligate precursors to invasive breast cancer (IBC), but importantly, not all in situ lesions progress to become invasive. There is however, an increased risk of developing invasive breast cancer subsequent to an in situ carcinoma if left untreated (3, 4). The clinical challenge is therefore to distinguish high risk from low risk lesions in order to offer optimal treatment to these patients (5). Much remains to be learned about the pathogenesis of DCIS to be able to predict disease progression of these non-obligate breast cancer precursors.

Many cases of breast cancer present as mixed lesions, i.e. synchronous invasive ductal carcinoma and ductal carcinoma in situ. In such cases, the in situ lesion is thought to be the precursor of the invasive tumor and studies have reported an overall high degree of similarity of genetic aberrations between DCIS and IBC (6–10). Nevertheless, differences in type and frequency of mutations have also been reported (11). It has been hypothesized that DCIS and IBC originate from the same ancestor cell, but have deviated prior to the in situ stage following separate tumor progression paths (12). In tumors without an in situ compartment the invasive carcinoma may have arisen *de novo* (13), or the pre-invasive stage has been a brief, transient phase along the progression to invasive breast carcinoma (14). More, in-depth sequencing studies are required to investigate the intra-lesion heterogeneity in DCIS and whether progression to IBC is a result of clonal selection (6, 15).

In this study we have sequenced microdissected cell compartments from 26 mixed breast tumors using the Ion AmpliSeq™ Cancer Hotspot Panel v2. The mutation spectrum across forty-four samples of carcinoma in situ, invasive carcinoma and adjacent normal tissue showed a high degree of similarity between synchronous DCIS and IBC and a higher mutation frequency in the in situ cell compartment in mixed tumors compared to pure DCIS.

## Material and Methods

**Tumor tissue samples.** Fresh frozen tissue from patients with mixed tumors (i.e. invasive ductal carcinoma with synchronous in situ lesion) or pure DCIS was collected at the Fresh Tissue Biobank, Department of Pathology, Uppsala University Hospital, Sweden. Histopathological evaluation of all cases was performed by a pathologist.

**Laser capture microdissection.** Invasive, in situ and normal cell areas were microdissected using laser capture microdissection (LCM) on a Zeiss inverted microscope PALM Laser Micro-Beam System (Carl Zeiss, Germany) as previously described (8). Frozen 14  $\mu\text{m}$ -thick sections were mounted on polyethylene membrane (PALM) covered slides and stained with hematoxylin (60  $\mu\text{l}$ ) mixed with RNasin for 1 min, incubated in Zincfix (60  $\mu\text{l}$ ) for 30 sec, followed by a series of 30-sec incubation steps in 75%, 95% and 100% ethanol, respectively. Adjacent 4  $\mu\text{m}$ -thick sections were cut and stained by a routine hematoxylin and eosin protocol to locate the areas to be microdissected. Cells were captured into collecting caps and preserved in 50  $\mu\text{l}$  Trizol at  $-80^{\circ}\text{C}$  for DNA extraction. The number of cells obtained was estimated by the operator during microdissection and between 100 and 4000 cells were obtained for each sample. Pure DCIS samples were not microdissected; for these samples, whole FFPE tumor sections were used for DNA isolation.

**DNA purification.** DNA was isolated using Qiagen (Hilden, Germany) DNeasy Blood and Tissue Mini Kit. Samples were thawed and centrifuged at 16,000  $\times$  g for 15 min to precipitate DNA. After complete removal of Trizol, 180  $\mu\text{l}$  buffer ATL and 20  $\mu\text{l}$  protease was added and the tubes incubated at  $56^{\circ}\text{C}$  overnight before addition of 200  $\mu\text{l}$  buffer AL. Samples were mixed well by vortexing before 200  $\mu\text{l}$  ethanol was added and the samples were again mixed well by vortexing. The samples were then transferred to DNeasy Mini spin columns and further processed as per the manufacturer's instructions before DNA was eluted in 100  $\mu\text{l}$  buffer AE. To improve recovery of the DNA, the elution buffer was left on the columns for 5 minutes before a final centrifugation step. For quantification and quality assessment of the DNA, quantitative polymerase chain reaction (qPCR) was performed with the KAPA hgDNA Quantification and QC Kit (KAPA Biosystems, Wilmington, MA) as per the manufacturer's instructions. Isolation of DNA from pure DCIS tumors were performed using the QIAcube system with the AllPrep DNA/RNA Universal Kit (cat.no. 80224, Qiagen, Hilden, Germany) according to protocol provided by the supplier.

**Library preparation.** Sequencing libraries for Ion Torrent sequencing were prepared using the Ion Torrent AmpliSeq™ Library Kit 2.0 (Thermo Fisher Scientific, Waltham, MA), and the Ion AmpliSeq™ Sample ID Panel as per the manufacturer's instructions. Briefly, approx. 100 pg DNA was mixed with Ion AmpliSeq™ HiFi Master Mix and the two primer pools and amplified for 27 cycles followed by partial

digestion of the primer sequences and ligation of barcoded adapters. The libraries were purified using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA) and amplified by polymerase chain reaction (PCR) for 5 cycles followed by a two-round purification process with AMPure XP beads. The final libraries were quantified on Agilent Bioanalyzer instrument (Agilent Technologies, Santa Clara, CA) with the Agilent High Sensitivity DNA Kit and stored at  $-20^{\circ}\text{C}$ . The AmpliSeq™ Cancer Hotspot Panel yields 207 amplicons covering hotspot regions of 50 relevant cancer genes (Suppl. file 1).

**Template preparation and sequencing.** Libraries were normalized to 100 pM in Low TE and equal amounts of each library were pooled. Each pool was diluted 10 times and 20  $\mu\text{l}$  were clonally amplified on the Ion OneTouch system using the Ion OneTouch 200 Template Kit v2 DL and enriched with the Ion OneTouch ES as per the manufacturer's instructions. Sequencing was carried out on the Ion Torrent Personal Genome Machine (PGM) using the Ion PGM 200 Sequencing Kit and Ion 314 or Ion 316 Chips for 400 cycles according to the manufacturer's instructions. For the microdissected samples, mean number of mapped reads was 395584 (range 126272-933608), mean read length 108bp (range 73-115bp) while mean depth was 1655 (range 411-3922). For the pure DCIS samples mean number of mapped reads was 273769 (range 227290-345207), mean read length 113 (range 112-116bp) and mean depth 1262 (range 1046-1585).

**Variant calling.** Data was analyzed using the AmpliSeq™ Variant Caller plug-in within the Ion Torrent Suite software (version 5.0.4, Thermo Fisher Scientific). Forty-seven samples were sequenced in total. Three samples were excluded from further analysis after quality assessment and in all, 44 microdissected samples from 26 mixed tumors and 10 pure DCIS were successfully sequenced. Due to low input and varying sample quality for the microdissected samples, a strict cut-off was applied; only variants with maximum allele frequency  $>10\%$  and quality  $>100$  across all microdissected samples were included. Variants were manually assessed in Integrative Genomics Viewer (16) to evaluate strand bias and potential technically induced artifacts. Finally, to include only variants that affect function and which may be of clinical significance, variants were filtered by including only those present in Catalogue of Somatic Mutations in Cancer (COSMIC, version 77 accessed May 2016) (17), and excluding SNPs present in the variant database in the 1000 Genomes Project (accessed May 2016) (18).

**Validation by digital droplet PCR.** Digital droplet PCR was performed using the RainDrop system (RainDance technologies) to validate the *PIK3CA*:H1047R variant found in nine samples on the IonTorrent platform. DNA was isolated from separate FFPE tumor sections using DNeasy Mini spin columns as described above. An assay with two color fluorescent TaqMan probes was used to discriminate between droplets containing mutant and wild type alleles. A 50  $\mu\text{l}$



reaction mix containing 2x KAPA Probe Force qPCR Master Mix (Sigma Aldrich, St. Louis, MO), 25x Droplet Stabilizer (RainDance Technologies), 13,3 µl nuclease free H<sub>2</sub>O, 9 µl DNA sample, and 0,7 µl primer/probe mix with 500 nM fwd/rev primer and 200 nM WT/mutant probe was made for each sample. The total reaction mix was loaded onto the RainDance Source chip for partitioning of the mix into millions of single droplets. Each droplet contains a PCR mix – oil emulsion and a single DNA fragment (positive) or no target molecule (negative). After partitioning, a PCR amplification was performed, where each droplet acts as an individual PCR reaction. The PCR conditions were as follows: 98°C (3 min), 55 cycles of 95°C (10 sec) and 60°C (1 min) with ramp speed of 0,5°C /second, 72°C (10 min), 98°C (10 min), 12°C (10 min), and keep at 12°C . The samples were transferred to the RainDrop Sense instrument for automatic counting of positive and negative droplets depending on the presence or absence of a fluorescent signal enabling calculation of the absolute number of targets present in the original sample.

**Statistical analysis.** Fisher’s exact tests were performed to test whether there was any statistically significant association between variants in genes and estrogen receptor (ER) or progesterone receptor (PR) status and to test the difference in frequency of samples carrying variants between synchronous and pure DCIS.

## Results and discussion

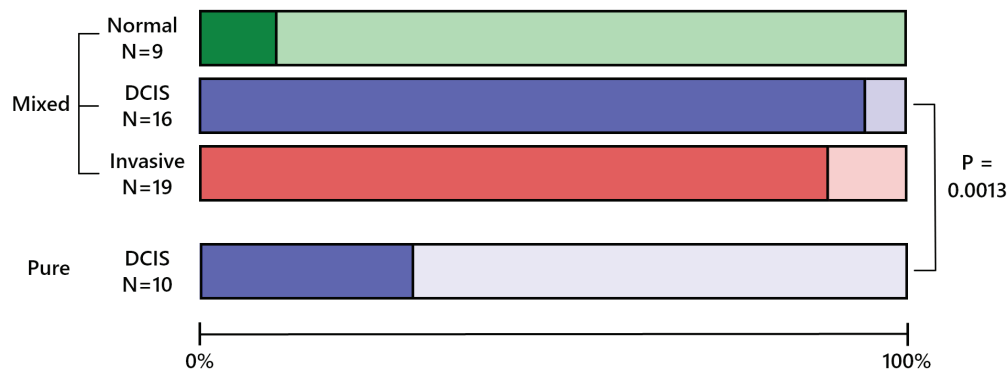
In total, 19 invasive carcinoma, 16 DCIS and nine normal, microdissected tissue samples from 26 patients with mixed tumors, were subjected to targeted sequencing of 50 oncogenes and tumor suppressor genes. Amongst the samples were three triplets (i.e. in situ, invasive and normal epithelial cells from the same tumor), and thirteen in situ/invasive pairs. In addition, we sequenced 10 pure DCIS samples. An overview of relevant clinical information is shown in Suppl. file 2. Mean age at time of diagnosis was 52 years (min/max: 30/81) and mean tumor size 23.7 mm (min/max: 1/80). Of all tumors were 28/36 (78%) ER positive and 23/36 (64%) PR positive. All PR-positive tumors were ER-positive.

Across all mixed tumor samples we identified twenty-two different, potentially pathogenic variants in eight genes (*AKT1*, *CDH1*, *CDKN2A*, *ERBB2*, *MET*, *PIK3CA*, *STK11* and *TP53*) (Suppl. file 3). *PIK3CA* and *TP53* were the most commonly mutated genes in our cohort. Most variants were present in only one patient; but for two of the genes (*AKT1* and *PIK3CA*) identical variants were identified in more than one patient. The number of variants in each in situ or invasive cell compartment ranged from zero to four, and most tumors carried only one variant (11/16 in situ, 12/19 invasive). The most common variant (*PIK3CA*:p.H1047R) was found in nine samples from six patients. Across the mixed tumors, four different *PIK3CA* variants were detected in 12 patients. In contrast, no *TP53* variant was found in more than one patient. Among the thirteen cases of pairs, for which both

		CDH1	CDKN2A	ERBB2	MET	PIK3CA	STK11	TP53	Total
No of variants:		1	1	4	1	4	1	7	19
UPP027	D					1			1
	I					1			1
UPP038	D			2					2
	I			1				1	2
UPP047	D	1					1		2
	I	1					1		2
UPP050	D					1			1
	I					1		1	2
UPP058	D	1		2				1	4
	I			2	1			1	4
UPP065	D					1			1
	I					1			1
UPP087	D								0
	I					1			1
UPP110	D					1			1
	I					1			1
UPP126	D					1			1
	I								0
UPP147	D							2	2
	I							1	1
UPP158	D					1			1
	I					1			1
UPP208	D							1	1
	I							1	1
UPP244	D							1	1
	I					1		1	2
Total	D	1	1	4		6	1	5	18
	I		1	3	1	7	1	6	19

Fig. 1. Overview of genes with pathogenic variants identified in the 13 available DCIS/IBC sample pairs. DCIS (blue), IBC (red).

in situ and invasive samples were available from the same patient, we found nineteen variants in seven different genes (Figure 1). In six of the cases, the variant(s) were identical in both compartments. Interestingly, we found four different *ERBB2* variants; two variants (p.D769H and p.V777L) resided in both the in situ and invasive tumor compartments of the same tumor. The other two *ERBB2* variants were found in a second tumor. One of these (p.D769Y) was found in both the in situ and invasive tumor compartments and the other (p.L755S) was found only in the in situ compartment. Altogether, these findings demonstrate large inter-tumor heterogeneity in mutation pattern in synchronous DCIS and IBC and indicate that *ERBB2* variants also are present early in tumorigenesis.



**Fig. 2.** Relative frequency of samples with pathogenic variants in the different cell compartments in mixed tumors compared to pure DCIS. Dark color represents samples carrying at least one potentially pathogenic variant, light color represent samples without any identified pathogenic variant. There was a statistically significant difference between synchronous DCIS and pure DCIS ( $P=0.0013$ , Fisher's exact test).

In three tumors (UPP027, UPP208 and UPP244), normal epithelium was successfully sequenced in addition to invasive and in situ tumor tissue. Two of these tumors (UPP027 and UPP208) carried only one variant each (*PIK3CA*:p.H1047R and *TP53*:p.A84fs, respectively) and none of the corresponding normal compartments carried these variants. The last tumor with three samples (UPP244) carried two different variants; one of these (*TP53*:p.R175H) was found in both the in situ and invasive compartments, while the other (*PIK3CA*:H1047R) was found in the normal compartment at a frequency of 30%, absent in the in situ compartment and present at a very low frequency (3%) in the invasive compartment. Across the cohort, nine normal breast epithelium samples were sequenced, and amongst these, only the one case described above (UPP244), carried any of the variants.

We found a significant association between *PIK3CA* variants and positive PR status ( $P=0.039$ , Fisher's exact test), which has been previously noted (19–21). A similar association was not seen for ER ( $P=0.44$ ), however; the low number of samples in this study may have prevented the identification of any such association.

In addition to microdissected tissue from mixed tumors, we sequenced ten pure (non-microdissected) DCIS. Only three variants in three different tumors were detected; *PIK3CA*:p.C420R, *PIK3CA*:p.E542K, and *TP53*:p.R213X (Suppl. file 3). Two of these were not detected in any of the microdissected samples, while the third variant (*PIK3CA*:p.C420R) was found in one of the invasive tumor cell compartments, but was filtered out due to low sequencing depth. Interestingly, there was a notable difference in the number of variants across the 50 genes between in situ cell compartments from mixed tumors compared with pure DCIS. Almost all in situ cell compartments from the mixed tumors, 15/16 (94%), carried at least one variant while only 3 out of 10 (30%) of the pure DCIS tumors carried any of the variants. This difference is significant ( $P = 0.0013$ , Fisher's exact test, Figure 2). Noticeably, targeted sequencing as performed here includes only a limited number of genes and therefore we cannot exclude the possibility that mutation spectra across other putative driver genes might be similar between the two

different types of in situ cancers. Nevertheless, our findings indicate that in situ cells from a tumor with synchronous IBC have a more invasive-like mutational phenotype compared to pure DCIS and consequently that synchronous DCIS and pure DCIS could be different entities. These findings confirm those of other studies (22, 23) and highlight the importance of being conscious about distinguishing synchronous DCIS from pure DCIS lesions when studying tumor progression.

When DCIS presents synchronous with invasive disease, it is unclear whether these multiple stage-specific cell populations have a common ancestor or develop from multiple clones. Previous sequencing studies have reported similar mutation profiles in DCIS and IBC, with *PIK3CA*, *TP53* and *GATA3* as the most commonly affected genes (8–10, 22, 24–28). However, different prevalence of *PIK3CA* variants has been observed between DCIS and IBC. One study reported *PIK3CA* variants restricted to the in situ compartment in two cases of synchronous DCIS and IBC, while in a third case, a reduced frequency of a specific *PIK3CA* variant was found in invasive cells relative to the cells from the in situ compartment (9). In one tumor in our study, we found a *PIK3CA* variant in the in situ cells, and not in the invasive cell compartment, while in two tumors, we found a *PIK3CA* variant in the invasive cells while not in the corresponding in situ cell compartment. In our study, the sequencing panel did not include *GATA3*, so the high frequency of *GATA3* variants previously found in DCIS could not be confirmed (26).

Ion semiconductor sequencing is a “sequencing by synthesis” method based upon detection of hydrogen ions that are released during polymerization of DNA. The technology is well suited for targeted sequencing of samples with minute amounts of DNA which is often the challenge with microdissected tissue. This has allowed us to sequence a panel of the most frequently mutated genes in cancer, in relatively few cells from stored Trizol cell fractions after microdissection. To validate our findings, we used digital droplet PCR to quantify the most frequently detected variant in this study, *PIK3CA*:p.H1047R and found similar frequencies as by sequencing (Suppl. file 4). Three of the *TP53* mutated samples in this study were included in a previous study of *TP53* mu-

tations in synchronous DCIS and IBC (8). We detected two of these variants in this study, while the third, a 10bp deletion of codons 106-109, was not called by the Ion Torrent analysis pipeline. However, we identified the deletion in our data by manual inspection. This discrepancy could be due to inaccurate flow-calls, a known artifact of PGM, which may cause homopolymers to be under-called (29).

## Conclusions

In this study we performed targeted sequencing of microdissected tissue from in situ and invasive tumor cell compartments from 26 patients with mixed DCIS/IBC tumors, in addition to 10 pure DCIS tumors. Across the 50 cancer-relevant genes included in the panel, we found that the spectrum of variants was similar between synchronous DCIS and IBC indicating clonal relationship between the two tumor stages and selection of subclones during tumor progression. *PIK3CA* and *TP53* were the most frequently mutated genes and alterations occurred at the DCIS stage or earlier. Pure DCIS showed significantly lower number of variants compared to synchronous DCIS.

### ABBREVIATIONS

COSMIC: Catalogue of somatic mutations in cancer; DCIS: Ductal carcinoma in situ; ER: Estrogen receptor; IBC: Invasive breast cancer; LCM: Laser capture microdissection; PALM: Polyethylene membrane; PCR: Polymerase chain reaction; PGM: Personal genome machine; PR: Progesterone receptor; qPCR: Quantitative polymerase chain reaction

### ETHICAL CONSIDERATIONS

The study complies with the Declaration of Helsinki, and was approved by the Ethics Committee at Uppsala University Hospital, Sweden (approval number 2005/118).

### AUTHOR CONTRIBUTIONS

All authors take responsibility for the integrity of the data and all authors approved the final manuscript. Conceptualization, V.K., T.S.; Methodology, H.B., S.K.; Investigation, H.B., S.K.; Formal Analysis, H.B. and S.K.; Resources, F.W., V.K., T.S.; Writing - Original Draft, H.B., T.S.; Writing - Review Editing, H.B., F.W., V.K., T.S.; Visualization, H.B.; Supervision, V.K., T.S.; Funding Acquisition, V.K., T.S.

### COMPETING INTERESTS

The authors have stated explicitly that there are no conflicts of interest in connection with this article.

### ACKNOWLEDGEMENTS

We thank Inger Bergheim and Helen Vålerhaugen for assisting in performing sequencing and ddPCR experiments. This work was supported by funds from the Regional Health Authorities in the South East of Norway.

## Bibliography

- Lakhani, S, Ellis, I, Schnitt, S, Tan, P, van de Vijver, M. *WHO Classification of Tumours of the Breast, Fourth Edition* (2012).
- Ward, E. M. *et al.* Cancer statistics: Breast cancer in situ. *CA: A Cancer Journal for Clinicians* **65**, 481–495 (2015).
- Erbas, B., Provenzano, E., Armes, J. & Gertig, D. The natural history of ductal carcinoma in situ of the breast: a review. *Breast cancer research and treatment* **97**, 135–44 (2006).
- Sanders, M. E., Schuyler, P. a., Simpson, J. F., Page, D. L. & Dupont, W. D. Continued observation of the natural history of low-grade ductal carcinoma in situ reaffirms proclivity for local recurrence even after more than 30 years of follow-up. *Modern Pathology* **28**, 662–669 (2015).
- Gorringe, K. L. & Fox, S. B. Ductal Carcinoma In Situ Biology, Biomarkers, and Diagnosis. *Frontiers in Oncology* **7**, 248 (2017).
- Cowell, C. F. *et al.* Progression from ductal carcinoma in situ to invasive breast cancer: Revisited. *Molecular oncology* **7**, 1–11 (2013).
- Rane, S. U., Mirza, H., Grigoriadis, A. & Pinder, S. E. Selection and evolution in the genomic landscape of copy number alterations in ductal carcinoma in situ (DCIS) and its progression to invasive carcinoma of ductal/no special type: a meta-analysis. *Breast Cancer Research and Treatment* **153**, 101–121 (2015).
- Zhou, W. *et al.* Full sequencing of TP53 identifies identical mutations within in situ and invasive components in breast cancer suggesting clonal evolution. *Molecular oncology* **3**, 214–219 (2009).
- Hernandez, L. *et al.* Genomic and mutational profiling of ductal carcinomas in situ and matched adjacent invasive breast cancers reveals intra-tumour genetic heterogeneity and clonal selection. *The Journal of pathology* **227**, 42–52 (2012).
- Miron, A. *et al.* PIK3CA mutations in in situ and invasive breast carcinomas. *Cancer Research* **70**, 5674–5678 (2010).
- Sakr, R. A. *et al.* PI3K pathway activation in high-grade ductal carcinoma in situ-implications for progression to invasive breast carcinoma. *Clinical Cancer Research* **20**, 2326–2337 (2014).
- Kuerer, H. M. *et al.* Ductal carcinoma in situ: State of the science and roadmap to advance the field (2009).
- Wong, H., Lau, S., Yau, T., Cheung, P. & Epstein, R. J. Presence of an in situ component is associated with reduced biological aggressiveness of size-matched invasive breast cancer. *British journal of cancer* **102**, 1391–6 (2010).
- Kurbel, S. In search of triple-negative DCIS: Tumor-type dependent model of breast cancer progression from DCIS to the invasive cancer (2013).
- Kader, T. *et al.* Atypical ductal hyperplasia is a multipotent precursor of breast carcinoma. *Journal of Pathology* **248**, 326–338 (2019).
- Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24–26 (2011).
- Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* **43**, D805–D811 (2015).
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- Pang, B. *et al.* Prognostic role of PIK3CA mutations and their association with hormone receptor expression in breast cancer: a meta-analysis. *Scientific reports* **4**, 6255 (2014).
- Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
- Saal, L. H. *et al.* PIK3CA mutations correlate with hormone receptors, node metastasis, and ERBB2, and are mutually exclusive with PTEN loss in human breast carcinoma. *Cancer research* **65**, 2554–9 (2005).
- Kim, S. Y. *et al.* Genomic differences between pure ductal carcinoma in situ and synchronous ductal carcinoma in situ with invasive breast cancer. *Oncotarget* **6**, 7597–7607 (2015).
- Schorr, M. C., Pedrini, J. L., Savaris, R. F. & Zettler, C. G. Are the pure in situ breast ductal carcinomas and those associated with invasive carcinoma the same? *Applied Immunohistochemistry and Molecular Morphology* **18**, 51–54 (2010).
- Dunlap, J. *et al.* Phosphatidylinositol-3-kinase and AKT1 mutations occur early in breast carcinoma. *Breast Cancer Research and Treatment* **120**, 409–418 (2010).
- Agahozo, M. C. *et al.* PIK3CA mutations in ductal carcinoma in situ and adjacent invasive breast cancer. *Endocrine-Related Cancer* **26**, 471–482 (2019).
- Pang, J. M. B. *et al.* Breast ductal carcinoma in situ carry mutational driver events representative of invasive breast cancer. *Modern Pathology* **30**, 952–963 (2017).
- Doebbar, S. C. *et al.* Progression of ductal carcinoma in situ to invasive breast cancer: comparative genomic sequencing. *Virchows Archiv* **474**, 247–251 (2019).
- Li, H. *et al.* PIK3CA mutations mostly begin to develop in ductal carcinoma of the breast. *Experimental and Molecular Pathology* **88**, 150–155 (2010).
- Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P. & Tyson, G. W. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS computational biology* **9**, e1003031 (2013).

**Supplementary file 1. Overview of genes included in Ion AmpliSeq™ Hotspot Cancer Panel v2.**  
 Shaded grey are genes where potentially pathogenic variants are identified in this cohort

<i>ABL1</i>	<i>EGFR</i>	<i>GNAS</i>	<i>KRAS</i>	<i>PTPN11</i>
<i>AKT1</i>	<i>ERBB2</i>	<i>GNAQ</i>	<i>MET</i>	<i>RB1</i>
<i>ALK</i>	<i>ERBB4</i>	<i>HNF1A</i>	<i>MLH1</i>	<i>RET</i>
<i>APC</i>	<i>EZH2</i>	<i>HRAS</i>	<i>MPL</i>	<i>SMAD4</i>
<i>ATM</i>	<i>FBXW7</i>	<i>IDH1</i>	<i>NOTCH1</i>	<i>SMARCB1</i>
<i>BRAF</i>	<i>FGFR1</i>	<i>JAK2</i>	<i>NPM1</i>	<i>SMO</i>
<i>CDH1</i>	<i>FGFR2</i>	<i>JAK3</i>	<i>NRAS</i>	<i>SRC</i>
<i>CDKN2A</i>	<i>FGFR3</i>	<i>IDH2</i>	<i>PDGFRA</i>	<i>STK11</i>
<i>CSF1R</i>	<i>FLT3</i>	<i>KDR</i>	<i>PIK3CA</i>	<i>TP53</i>
<i>CTNNB1</i>	<i>GNA11</i>	<i>KIT</i>	<i>PTEN</i>	<i>VHL</i>

Supplementary file 2 - Clinical information all patients

**Mixed DCIS and IBC tumors**

Patient	Age	SIZE	ER	PR	EORTC grade	ELSTON grade	HER2 (IHC)	P53 (IHC)	Norm	DCIS	Inv
UPP059	44	20	pos	pos	3	3	pos	neg			x
UPP077	65	18	pos	pos	2	1	pos	neg			x
UPP078	57	13	pos	neg	2	1	pos	neg	x		
UPP081	60	15	pos	pos	2	1	pos	neg	x		
UPP117	57	13	pos	neg	2	1	pos	neg	x		
UPP124	44	60	neg	neg	3	3	neg	NA		x	
UPP136	54	16	pos	pos	2	2	pos	neg		x	
UPP150	51	11	pos	neg	1	1	NA	NA	x		
UPP152	39	1	neg	neg	2	1	NA	NA			x
UPP202	53	13	pos	pos	3	2	NA	NA			x
UPP233	52	30	pos	pos	3	3	NA	NA			x
UPP038	48	9	neg	neg	2	2	pos	pos		x	x
UPP047	81	13	pos	pos	2	2	neg	pos		x	x
UPP050	49	10	pos	pos	2	1	neg	neg		x	x
UPP058	52	NA	neg	neg	3	2	neg	pos		x	x
UPP065	46	10	pos	pos	2	1	pos	neg		x	x
UPP087	81	NA	pos	pos	1	1	neg	neg		x	x
UPP110	31	NA	pos	pos	3	3	NA	NA		x	x
UPP126	80	5	pos	pos	NA	1	NA	NA		x	x
UPP147	54	16	pos	neg	3	3	NA	NA		x	x
UPP158	41	30	pos	pos	LCIS	2	NA	NA		x	x
UPP216	42	NA	pos	pos	LCIS	2	NA	NA	x		x
UPP224	44	15	pos	pos	2	1	NA	NA	x	x	
UPP027	51	26	pos	pos	2	2	pos	neg	x	x	x
UPP208	56	14	pos	neg	3	3	NA	neg	x	x	x
UPP244	55	80	neg	neg	3	3	NA	NA	x	x	x
									9	16	19

**Pure DCIS tumors**

Patient	Age	SIZE	ER	PR	EORTC grade	ELSTON grade	HER2 (IHC)	P53 (IHC)	Norm	DCIS	Inv
UPP001	47	NA	pos	pos	2	not appl.	neg	neg		x	
UPP008	55	20	neg	neg	3	not appl.	pos	pos		x	
UPP116	30	20	pos	pos	3	not appl.	pos	pos		x	
UPP123	48	50	pos	pos	2	not appl.	neg	pos		x	
UPP142	44	17	neg	neg	3	not appl.	pos	neg		x	
UPP143	49	35	pos	pos	2	not appl.	neg	pos		x	
UPP177	30	25	pos	pos	2	not appl.	neg	neg		x	
UPP210	60	30	neg	neg	3	not appl.	neg	pos		x	
UPP220	45	60	pos	pos	3	not appl.	neg	pos		x	
UPP250	81	40	pos	pos	2	not appl.	neg	NA		x	
									10		



Supplementary file 3 (cont.) - Allele frequency of variants in pure tumors

Variant information	Gene Name	<i>TP53</i>	<i>PIK3CA</i>	<i>PIK3CA</i>
	Accession number	NM 000546	NM 006218	NM 006218
	Chromosome	chr17	chr3	chr3
	Position	7578212	178936082	178927980
	DNA change	c.C637T	c. G1624A	C.T1258C
	Protein change	P.R213X	P.E542K	P.C420R
	Mutation type	stopgain	non-synonymous SNV	non-synonymous SNV
Allele frequency	UPP001			
	UPP008			0.21
	UPP116			
	UPP123			
	UPP142	0.12		
	UPP143			
	UPP177		0.29	
	UPP210			
	UPP220			
	UPP250			
No of samples with variant		1	1	1

Supplementary file 4 - Validation results ddPCR. *PIK3CA* :p.H1047R

SampleID	Input (ng DNA)	WT droplets	mut droplets	Mut.frequency ddPCR	Mut. frequency IonTorrent	Comments
UPP027 Normal	0.054	66	0	0%	0%	
UPP027 DCIS	0.297	108	50	32%	33%	
UPP027 Invasive	0.01	58	20	26%	30%	
UPP050 Normal	NA	10187	2	0.02%	NA	Ion Torrent sequencing failed
UPP050 DCIS	0.021	2639	1279	33%	43%	
UPP050 Invasive	0.058	7330	20	0.27%	0%	
UPP087 Normal	NA	10957	567	5%	NA	Ion Torrent sequencing failed
UPP087 DCIS	0.016	0	0	NA	0%	ddPCR Failed
UPP087 Invasive	0.062	1102	272	20%	40%	
UPP158 DCIS	0.085	169	88	34%	41%	
UPP158 Invasive	0.156	288	184	39%	18%	
UPP233 Invasive	0.144	14	6	30%	27%	
UPP244 Normal	0.063	1	0	NA	29%	ddPCR Failed
UPP244 DCIS	0.144	2115	0	0%	0%	
UPP244 Invasive	0.081	43281	1	0.002%	3%	
WT-PIK3CA_ctr	10	7596	0	0%	not applicable	Negative control
PIK3CA_5%_ctr	50	40733	3091	7%	not applicable	Positive control
PIK3CA_5%_ctr	50	43064	3139	7%	not applicable	Positive control

## **ERRATA LIST**

**Doctoral candidate:** Helga Bergholtz

**Title of thesis:** Deciphering molecular heterogeneity and relevance of subtypes in breast cancer progression

### **Abbreviations for different types of corrections:**

**Cor** – correction of language

**Cpltf** – change of page layout or text format

<b>Page</b>	<b>Line</b>	<b>Original text</b>	<b>Type of correction</b>	<b>Corrected text</b>
16	8	originating from <b>tissue</b> several cell types	Cor	originating from several cell types
32	25	direct comparison of (...) <b>were</b> more challenging	Cor	direct comparison of (...) <b>was</b> more challenging
39	26	Using a more targeted approach, (...) may, in many cases be more appropriate.	Cor	Using a more targeted approach, (...) may, in many cases, be more appropriate.