On the genetic determinants of cancer phenotypes

Christian Fougner

Faculty of Medicine
University of Oslo

Department of Cancer Genetics
Institute for Cancer Research
Oslo University Hospital

Oslo University Hospital

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

ISBN 978-82-8377-731-4

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

Acknowledgements

The work presented in this thesis was carried out at the Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital. The research was made possible by funding from the University of Oslo Medical Student Research Program, the Norwegian Research Council and the South-Eastern Norway Regional Health Authority. I am grateful to the Faculty of Medicine at the University of Oslo for admitting me to the PhD program.

My research journey started when I got in touch with Therese in December 2012. I wrote an unsolicited application to work as a lab assistant and sent it by post. It was the first time I had applied for anything resembling a job, and I was certain that all important things in the adult world happened on paper. In my motivation letter, I wrote that I wanted to discover oncogenes, and mentioned three of Therese's papers on the topic that I thought were interesting. Two of the papers I didn't understand, and I later realized had nothing to do with oncogenes. The final paper, it turned out, wasn't even a research article, but a preface to a special edition of a journal. Therese, thank you so much for seeing past my cluelessness and welcoming me into your group. I have always appreciated your open-door policy and your unremitting efforts to be the best possible leader to all of us in the group (and the department). Your late-night e-mails catching up on things others would have waited a week to get done, and your inexplicable ability to conjure *just one more* period of funding, has not gone unnoticed. I've always been glad to have a supervisor who's invariably the last one to go home from a party, and I'm sure we'll keep in touch long after any professional connection has ended.

Jens Henrik, thank you for your patience in ultimately having to deal with my cluelessness those first few years. I've heard talk of wet-lab researchers with good hands and

wet-lab researchers with bad hands, and you definitely had the fortune of training one of the latter when Therese passed me on to you. Nonetheless, our weekly meetings always helped me process whatever hadn't quite worked since last time we spoke, and your unrelenting optimism and curiosity, in the end, led us down interesting paths. Thank you for getting me through the steepest part of the learning curve.

Ole Christian, trying to keep up with you and Tonje discussing the intricacies of statistical methods for several hours every Tuesday always left my brain feeling like soup for the rest of the afternoon. But, your unparalleled ability to explain complex ideas made the year of working together an absolute pleasure. Thanks for ensuring that none of my statistical cluelessness ever made its way outside the confines of *Forskningsbygget*.

Thank you to everyone at the Department of Cancer Genetics for creating a fun and social work environment filled with discussions of varying degrees of insightfulness. In particular, I would like to express my gratitude to all current and former colleagues in the *Breast tumor initiation and progression* group: Anna, Anne-Marthe, Eldri, Elen, Hedda, Helene, Helga, Kristine, Margit, Nirma, Phoung, Silje, Simen, Store-Tonje, Lille-Tonje, Torbjørn, and Veronica. Group meetings with you definitely kept me sane, and always gave me much-needed weekly assurances that what I was doing was actually *research*, and not just concocting conspiracy theories in *R*. Helga, I especially appreciated our collaborations and the early discussions about *claudin-low* (in italics) that gave me the confidence to fully flesh out the idea. Tonje, thanks for playing the role of the sparring partner that Ole Christian really needed for the final study to take shape. An immense gratitude goes out to my officemates throughout the years – Anne-Marthe, Astrid, Elen, Hedda and Torbjørn – all of whom had impeccable personal hygiene and were always up for a closed-door ranting session. Daniel, thanks for always being able to sort out any computer issue in five minutes, that otherwise would have taken me five days to figure out. Gry, thanks for keeping the department running all these years, and for all the times you've said *"we'll figure something out"* – invariably with a mischievous grin – every time a question about funding came up. Anne-Lise, thank you for your leadership in the earlier years, and for creating the open and inquisitive culture that continues to permeate the department.

Mamma og Pappa, where do I even begin to thank you? There have been countless times where I questioned whether this whole research thing was really worth it, especially when doing it in parallel with medical school. Our phone calls every Sunday – the majority of which have ended with *"stå på!"* this past year – have undoubtedly played a giant role in keeping me going. I hope that I'll be able to come home a lot more often once the world is a little more normal and that we'll be able to have those conversations over a meal and a glass

of wine instead of over FaceTime. You've always instilled the value of hard work on me, and that's by far the biggest reason why I've managed to complete this PhD. Christopher, I wouldn't be where I am today if it weren't for the math and physics tutoring back at ICS, and if you hadn't gone ahead and proved that coding was something mere mortals could learn. Thank you for always setting a high bar to live up to. Kikki, thank you for letting me know that *Forskerlinja* was mandatory, three years before I even got into medical school. There's no way I'd have ended up in research if it weren't for your unequivocal assurances that it was a good idea. James, thanks for proving that it's possible to spend most of your free time acting like an absolute idiot, whilst still being pretty clever and making something of yourself in life. If it weren't for our periodical weekends together blowing off steam, I'd probably have sold all my belongings and escaped to an island off the coast of South America halfway through the PhD. I can't wait until we make it back to Weisses. Erlend, I can't believe it's been almost a decade since the two of us first entered the realm of *MedFak* – it's finally truly over! Thanks for taking care of *han Anton* in Oslo. Lydia, you of all people have had to suffer the most from my workaholic tendencies. If it weren't for you, I would have ended up malnourished and have forgotten how to have a normal human interaction. Thanks for holding out this last year. And to everyone else who in any way, shape or form has been there for me these past seven years – you know who you are – thank you!

Christian Fougner
Oslo, May 2020

Table of contents

Acknowledgements 1

List of studies I

Abbreviations III

Sammendrag på norsk V

Introduction 1

The central dogma of biology and transcriptional regulation 1

The biology of cancer 3

Breast cancer epidemiology and risk factors 7

Clinical considerations in breast cancer 9

Molecular classification of breast carcinomas 11

Genetic and epigenetic characteristics of breast cancers 14

Mouse models of cancer 16

Aims 23

Results in brief 25

Claudin-low-like mouse mammary tumors show distinct transcriptomic patterns uncoupled from genomic drivers 25

Re-definition of *claudin-low* as a breast cancer phenotype 27

A pan-cancer atlas of transcriptional dependence on DNA methylation and copy number aberrations 29

Methodological considerations 31
Cohorts 31
Transcriptomic analyses 34
Copy number analyses 37
Somatic mutation analyses 40
Methylation analyses 43
Statistical analyses 46
Ethical considerations 49
Open data and privacy 49
Animal welfare 50
Reproducibility 50
Open access and rapid dissemination 51
Discussion 53
Claudin-low and the molecular classification of breast carcinomas 53
Determinants of cancer phenotypes 58
Future perspectives 65
Concluding remarks 67
References 69

List of studies

Study I

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 Sørlie.
Breast Cancer Research 21, 85 (2019).

Study II

Re-definition of *claudin-low* as a breast cancer phenotype

Christian Fougner, Helga Bergholtz, Jens Henrik Norum and Therese Sørlie.
Nature Communications 11, 1787 (2020).

Study III

A pan-cancer atlas of transcriptional dependence on DNA methylation and copy number aberrations

Christian Fougner, Elen K. Höglander, Tonje G. Lien, Therese Sørlie, Silje Nord, Ole Christian Lingjærde.
BioRxiv 2020.05.04.076901 (2020).

Abbreviations

ASCAT	Allele-Specific Copy Number Analysis of Tumors
cDNA	Complimentary DNA
CNA	Copy Number Aberration
CpG	Cytosine-Phosphate-Guanine
DCIS	Ductal Carcinoma In Situ
DMBA	7,12-Dimethylbenzanthracene
DNA	Deoxyribonucleic Acid
EMT	Epithelial-Mesenchymal Transition
ER	Estrogen Receptor
GEMM	Genetically Engineered Mouse Model
HER2	Human Epidermal Growth Factor Receptor 2
MPA	Medroxyprogesterone Acetate
mRNA	Messenger RNA
PAH	Polycyclic Aromatic Hydrocarbon
PARP	Poly(ADP-Ribose) Polymerase
PCA	Principal Component Analysis
RNA	Ribonucleic Acid
SDI	Socio-Demographic Index
TNM	Tumor, Lymph Node, Metastasis
VEGF	Vascular Endothelial Growth Factor

Sammendrag på norsk

Kreft er en heterogen sykdomsgruppe, forårsaket av forandringer i cellenes arvemateriale. Disse forandringene kan være genetiske (altså i DNA-sekvensen, for eksempel punktmutasjoner og kopitallsaberrasjoner) eller epigenetiske (arvelige forandringer som ikke påvirker nukleotidsekvensen, for eksempel DNA-metylering). Denne avhandlingen omfatter tre studier som utforsker sammenhengen mellom slike genetiske og epigenetiske faktorer, og fenotypen til en tumor.

Brystkreft kan, ut fra genekspresjonsprofiler, deles inn i subtyper. Disse subtypene har forskjellig etiologi, molekylære drivere og prognose, og krever forskjellig behandling. Egnede dyremodeller er nødvendig for å utvikle nye behandlinger for brystkreft. I *Studie I* utforsket vi de molekylære egenskapene til mammatumorer fra en karsinogen-indusert musemodell for brystkreft. Genekspresjonsanalyser avdekket at halvparten av tumorene viste likheter til en kjent subtype av brystkreft som kalles *claudin-low*. Vi sekvenserte de kodende regionene av genomet til tumorene, som viste at tumorene hadde en mutasjonsbyrde opp til ti ganger høyere enn det man observerer i human brystkreft. Vi sammenlignet mutasjoner og kopitallsaberrasjoner i claudin-low og ikke-claudin-low musetumorer, men fant få forskjeller mellom de to tumorgruppene. Vi analyserte videre humane brysttumorer, og fant at claudin-low-tumorer hadde relativt lavt antall mutasjoner og kopitallsaberrasjoner sammenlignet med tumorer fra andre subtyper. Det var betydelig overlapp mellom mutasjonene og kopitallsaberrasjonene i humane claudin-low- og basal-like-brysttumorer. Vi analyserte videre genekspresjonsprofilene til claudin-low tumorer (fra mennesker og mus) og fant uttalte tegn på immunsuppresjon. I sum antyder funnene våre at faktorer utenom forandringer i DNA-sekvensen lå til grunn for de observerte claudin-

low-karakteristika, og at immunsuppresjon kan være et behandlingsmål verdt å undersøke videre.

Når human brystkreft klassifiseres etter *intrinsic*-systemet brukes genekspresjonsdata til å først dele tumorer inn i fem subtyper (basal-like, normal-like, luminal A, luminal B, og normal-like). Deretter deles tumorer inn etter hvorvidt de klassifiseres som claudin-low eller ikke, i et separat trinn som også bruker genekspresjonsdata. Hvis en tumor klassifiseres som claudin-low blir den opprinnelige subtypen slettet, og claudin-low-tumorer blir dermed ansett som én gruppe uavhengig av hvilke subtyper de først klassifiseres som. Samtidig viser claudin-low-tumorer betydelig heterogenitet. I *Studie II* analyserte vi claudin-low tumorer i tre humane brystkreftkohorter med hypotesen at *claudin-low* ikke egentlig er en subtype, men heller en fenotype som kan ses i tillegg til en underliggende subtype. Vi identifiserte claudin-low tumorer (n = 87) i en kohort bestående av nesten to tusen brystkreftpasienter, og delte disse inn etter deres underliggende subtype. De fleste claudin-low tumorer ble opprinnelig klassifisert som basal-like, normal-like, eller luminal A. Vi sammenlignet claudin-low tumorer av forskjellige subtyper, og fant at mange tumorkarakteristika var reflektert i tumorens subtype (fremfor claudin-low-status). Blant annet, har claudin-low tidligere blitt beskrevet som en sykdomsgruppe med dårlig prognose, men da vi delte claudin-low tumorer inn etter subtype fant vi ingen forskjeller i overlevelse mellom claudin-low- og ikke-claudin-low-tumorer. De egenskaper som var karakteristiske for claudin-low-tumorer var en lav byrde av mutasjoner og kopitallsaberrasjoner, og høy infiltrasjon av immun- og stromaceller. Videre utforsket vi en ny metode for å identifisere claudin-low-tumorer, hvilket antydet at den etablerte metoden muligens feil-klassifiserer en gruppe basal-like tumorer med høy grad av immun- og stromainfiltrasjon som claudin-low. Til sist analyserte vi claudin-low-brysttumorer i to ytterligere kohorter, som til stor grad validerte funnene våre fra den første kohorten. Vi bemerket imidlertid at det var vesentlig forskjell mellom kohortene i prevalens av claudin-low-tumorer, samtidig som det var forskjellige inklusjonskriterier for tumorcelleprosent. Varierende prevalens av claudin-low-tumorer mellom kohortene kunne trolig forklares av at claudin-low-tumorer i varierende grad ble ekskludert grunnet lav tumorcelleprosent. I sum viser funnene fra *Studie II* at *claudin-low* er en fenotype, og ikke en genuin subtype (som tidligere antatt). Funnene antydet, i likhet med *Studie I*, at claudin-low egenskapene ikke kan tilskrives enkelte mutasjoner eller kopitallsaberrasjoner.

I *Studie III* prøvde vi å tallfeste effekten av DNA-metylering og kopitall på genekspresjon på tvers av genomet, og på tvers av tumortyper. Dette krevde utvikling av nye analytiske metoder. Kopitallet og ekspresjonen av et gen kan uttrykkes med en enkelt

tallverdi, mens metyleringsstatusen til mange CpGer kan være relevant for ekspresjonen av et gen. Vi viste at den sammensatte metyleringsstatusen til et gen kan representeres i et redusert antall dimensjoner ved bruken av prinsipalkomponentanalyse. Vi modellerte deretter assosiasjonene mellom ekspresjon og metylering (E-M), og fant at de til stor grad viste ulineære sammenhenger (som kan tyde på en metningseffekt). I motsetning var assosiasjonene mellom ekspresjon og kopitall (E-C) hovedsakelig lineære. Videre måtte analysene korrigeres for variabelt antall prøver i de forskjellige krefttypene, som vi løste med en metode basert på repetert nedskalering til konstant utvalgsstørrelse. Vi anvendte våre metoder på et pan-cancer-datasett, som resulterte i et atlas av E-M- og E-C-assosiasjoner. For å gjøre dette datasettet mest mulig anvendelig utviklet vi et multi-funksjonelt nettbasert verktøy for pan-cancer analyser av E-M- og E-C-assosiasjoner. Våre analyser av atlaset identifiserte betydelige forskjeller mellom krefttyper i graden av E-M- og E-C-assosiasjon. Genekspresjonen i plateepitelkarsinom fra lungene viste høyest assosiasjon til kopitall, og genekspresjonen i testikkel-cancer viste høyest assosiasjon til metylering. Det var en sterk sammenheng mellom byrden av kopitallsforandringer i en tumortype og graden av E-C-assosiasjon, men det var relativt lite sammenheng mellom grad av metyleringsvarianse i en tumortype og grad av E-M-assosiasjon. Det var ingen sammenheng mellom graden av E-M-assosiasjon i en tumortype og grad av E-C-assosiasjon i en tumortype. På enkeltgennivå var det derimot en invers sammenheng i hvorvidt gener hadde høy E-C- eller høy E-M-assosiasjon (for de fleste tumortyper). Mer inngående analyse viste at høyere byrde av kopitallsaberrasjoner i enkeltgener førte til høyere E-C-assosiasjon, uavhengig av E-M assosiasjon i det gitte genet. Derimot var det, i mange tumortyper, slik at høy metyleringsvarianse førte til høy E-M-assosiasjon i gener med lav E-C-assosiasjon, men at høy metyleringsvarianse ikke førte til høy E-M-assosiasjon i gener med høy E-C-assosiasjon. Disse resultatene tyder på at kopitallsforandringer kan overkjøre metyleringsforandringer, men at metyleringsforandringer i mindre grad kan overkjøre kopitallsforandringer.

Til sammen inkluderer denne avhandlingen analyser på tvers av to arter og på tvers av over tjue tumortyper. Studiene beskriver nye metoder og nye datasett, som er gjort offentlig tilgjengelig. Resultatene illustrerer den enorme kompleksiteten i sammenhengen mellom genetiske faktorer og fenotyper, og peker ut spennende muligheter for fremtidig forskning.

Introduction

The central dogma of biology and transcriptional regulation

Biological organisms are composed of cells, each of which carry a molecular blueprint for their own components in the form of deoxyribonucleic acids (DNA)[1]. DNA consists of sequences of four fundamental building blocks – adenine, cytosine, guanine and thymine – which collectively encode the genetic material of a cell. In itself, DNA has relatively few biological effector functions. Rather, DNA acts as the template for the molecules which ultimately differentiate the cells in a tree from the cells in a human, namely proteins. Proteins are themselves composed of smaller building blocks – amino acids – and the sequences of these building blocks in a protein are encoded in an underlying DNA template. The cellular machinery in known organisms does, however, not allow proteins to be generated directly from DNA. Ribonucleic acids (RNA) consist of four building blocks, chemically similar to those found in DNA, which act as an intermediate state for genetic information between DNA and protein. The process of generating RNA from DNA is referred to as *transcription*, or *gene expression*. The process of generating a protein from RNA is referred to as *translation*. The above describes, in abridged form, the central dogma of molecular biology[2]:

In a normal mammalian cell, there are two copies of each chromosome, one from each parent. The number of DNA copies encoding a protein – ordinarily two – should remain constant throughout the lifetime of a cell and throughout all cells in the organism. In contrast, the abundance of proteins must be continuously regulated in order to maintain homeostasis. This regulation also allows two cells in the human body to share identical DNA, but to act as disparately as a myocyte or a neuron.

Regulation of protein levels can occur through transcriptional, translational and post-translational mechanisms. In the studies carried out in this thesis, the *phenotype* (i.e. characteristics) of a cell or tissue is primarily measured by RNA abundance, rather than protein. Discussion here is therefore focused on transcriptional regulation. Studies have, however, shown that post-transcriptional regulation plays a considerable role in determining protein abundance[3,4], and this must be kept in mind throughout.

Transcription is initiated by the binding of an RNA polymerase (the enzyme which synthesizes RNA from DNA), coupled with a transcription factor, to a DNA sequence called a *promoter*[1]. Promoters are located upstream of the gene being transcribed. This catalyzes a complex biochemical process, ultimately generating an RNA strand complimentary to the DNA strand. Countless factors determine the quantity of RNA transcribed in a cell, most of which are operative prior to, and at the point of, transcription initiation.

In eukaryotes, DNA is stored in a condensed form, coupled with histones and non-histone chromosomal proteins, collectively referred to as *chromatin*[1]. DNA is poorly accessible in the form of condensed chromatin. Regulating the transcriptional machinery's access to DNA, in particular the promotor region, is therefore one of the most important methods by which a cell modulates gene expression. DNA access is principally regulated by mechanisms affecting chromatin structure, such as nucleosome remodeling and covalent histone modifications. Transcriptional access to DNA can also be controlled by direct modifications to the DNA itself, for example by addition of a methyl group to cytosine nucleotides with an adjacent downstream guanine (*CpGs)*[5]. CpG methylation is generally thought to inhibit transcription, but has in recent studies been associated with both increased and decreased gene expression[6,7]. Functionally relevant genetic features, such as those described here, which are heritable but do not affect the underlying DNA sequence, are collectively referred to as *epigenetics.*

Transcription factors are essential regulators of gene expression[1,8]. They are a class of protein that bind to specific DNA sequences, and thereby increase or decrease transcription of genes adjacent to their binding sites. Transcription factors modulate expression by several mechanisms, for example by promoting or blocking the recruitment of RNA polymerase to

a gene. Transcription factors are also operative in chromatin remodeling, and thereby modulate access to DNA. At least 1600 genes in the human genome may act as transcription factors, and there is great variability in how these are regulated[8]. Influences affecting the function of transcription factors include their cellular localization, molecular alterations (e.g. phosphorylation), ligand binding, and regulation of synthesis. There is also major variation in the affinity with which DNA-binding domains in transcription factors bind to various transcription factor binding sites. Alterations in these, for example due to somatic mutations or germline variations, may alter the transcriptional effect of transcription factors.

In diseases affecting genomic integrity, such as in certain cancers, the number of copies of a gene in a cell may be altered. In the event that a gene is deleted, the DNA blueprint for that gene no longer exists, and it follows that the gene no longer can be transcribed. Conversely, if there is a gain in the number of copies of a gene, there may be a corresponding increase in gene expression[7,9]. The transcriptional effect of such a gain may however depend on several factors, including the genomic location of the additional gene copy and the epigenetic state of the gene and its surroundings.

In sum, the path from a genetic state to a phenotype is enormously complex. The above describes only a fraction of the elements that affect transcription, and does not begin to discuss post-transcriptional factors, such as microRNAs and protein degradation. Additionally, biological systems display extensive random variability and are affected by environmental exposures. The extent to which observable genetic and epigenetic factors deterministically govern the behavior of cells in a complex organism remains poorly understood.

The biology of cancer

Cancer is a disease of the genome. In cancer, a subpopulation of cells, in a multicellular organism, carries some set of features providing a selective advantage. This selective advantage leads to an expansion of that subpopulation at the expense of the organism as a whole. Cancer must therefore be viewed through the lens of evolution and natural selection[10]. Essentially, the cells in a healthy individual are in balance with one another, sharing resources to ensure the wellbeing of the entire organism. Every time DNA is replicated, or subjected to a mutagenic exposure, a small number of aberrations are introduced to the genome. These aberrations may include point mutations (a single nucleotide is substituted for another), copy number aberrations (the number of copies of a DNA segment is reduced

or increased) and structural re-arrangements (a DNA segment is moved from one place in the genome to another). Epigenetic alterations may also be introduced. Most of these events are functionally insignificant, and do not provide any selective advantage. Some, however, lead to an increased fitness for that individual cell, for example mutations causing a constitutive activation of cell-cycle genes. Aberrations which lead to increased fitness are termed *drivers*, whereas functionally insignificant aberrations which do not confer a selective advantage are termed *passengers*[11]. Once a cell gains a driver aberration, it is likely to expand into a larger cell population in which all daughter cells carry that driver. The extent to which the cell population expands is dependent upon on several factors, including: The exact nature of the driver event (some aberrations have a more powerful *oncogenic* – cancer inducing – effect than others), the inherent propensity of the cell to become cancerous (less differentiated cells generally have a lower barrier for oncogenic transformation[12–15]), and evolutionary competition with other cells (e.g. competition for nutrients, avoiding destruction by immune cells). Oncogenic transformation often requires multiple drivers, and may be a process occurring over several decades[16].

Competition with healthy cells may be the main selective pressure at the early stages of tumorigenesis, but as a cancer progresses, competition between distinct tumor cell populations may increasingly govern its evolution[10]. Cancers frequently carry loss-of-function aberrations in genes which ordinarily maintain the integrity of the genome. When these functions are lost, genetic and epigenetic aberrations accumulate at an increasing rate, and novel subpopulations within a tumor arise. These subpopulations may expand due to increased fitness in competition with other tumor cells (e.g. higher proliferation rate, greater ability to access nutrients), or due to greater fitness when faced with new selective pressures (e.g. one subpopulation is resistant to a treatment whereas other subpopulations are not). Importantly, competition between tumor subpopulations does not necessarily show winner-take-all dynamics, and several subpopulations (or *subclones*) are often present in a tumor at any given time[10]. The presence of multiple subclones in a tumor is referred to as *intratumor heterogeneity*.

In sum, tumors are dynamic and heterogeneous organisms shaped by their environment. Cancers cannot be understood if only considered in light of their characteristics at the time of measurement, but must also be viewed in terms of how they arrived there, and how they might progress if subjected to new selective pressures.

The primary features that characterize cancers were enumerated by Hanahan and Weinberg, and termed the *hallmarks of cancer* (*Figure 1*)[17]. The most fundamental hallmark of cancer

cells is the ability to sustain proliferative signaling. In healthy tissues, cellular proliferation is a carefully regulated process, largely governed by growth factor signaling. In cancer, this process is dysregulated, and cancer cells can gain an increased ability to drive their own proliferation. Mechanisms for this include production of their own growth factors (i.e. autocrine signaling), stimulation of tumor-adjacent cells to produce growth factors (i.e. paracrine signaling), and increased production of growth factor receptors. Signaling proteins involved in regulating proliferation may also gain molecular aberrations (e.g. mutations or structural re-arrangements), which lead to an independence from growth factor signaling. Independence from growth factors may be enabled by constitutive activation of signaling proteins, or disruption of negative feedback loops. Aberrant genes that drive proliferation in cancer are referred to as *oncogenes*. In their non-aberrant state, these are called *proto-oncogenes*.

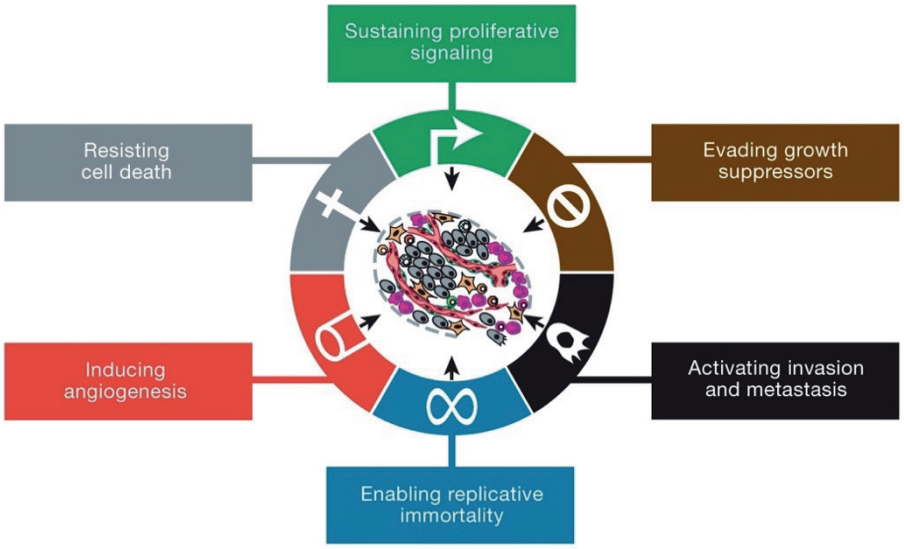


Figure 1: The hallmarks of cancer. Reproduced from Hanahan and Weinberg[17] with permission from Elsevier.

While cellular proliferation is upregulated by aforementioned mechanisms, it can also be downregulated by growth suppressors. Genes whose functions inhibit tumorigenesis (e.g. growth suppression, maintaining genomic stability) are called *tumor suppressor genes*. In essence, proliferative signaling pathways are integrated into several nodes which ultimately determine whether or not a cell proceeds through cell-cycle. These nodes, such as RB or TP53, act as gatekeepers to proliferation. In order to develop, cancers must therefore evade

these growth suppressors. Tumor suppressor genes are often inactivated by mutations, deletions or structural rearrangements, although cancer can still develop with these genes intact[18].

Cells have the ability to induce their own death through apoptosis. Apoptosis is a cancer-protective function, which can be induced by physiological stresses, such as elevated oncogene signaling (for example by RAS or MYC), or as a result of DNA damage. When apoptosis is initiated, the cell is disassembled and consumed by neighboring and phagocytic cells. The possibility for the stressed cell to develop into cancer is thereby obviated. Cancer cells must therefore resist cell death which would ordinarily be provoked by apoptosis-inducing stressors. Several tumor suppressor genes operative in regulating apoptosis may be aberrant in cancer, most notably *TP53* which commonly carries loss-of-function mutations or deletions[17,18].

Most healthy cells are only able to undergo a limited number of cell-cycles before reaching senescence (a viable, but non-proliferative state), or undergoing crisis (involving cell death). This limit is conferred by telomeres, which are repetitive nucleotide sequences on the ends of chromosomes that are progressively shortened upon DNA replication. If telomeres no longer protect chromosome ends, unstable end-to-end chromosome fusions may occur, which threaten cell viability. The limited number of cell-cycles which a healthy cell can undergo is not sufficient for a clinically relevant tumor to emerge. Cancer cells can gain replicative immortality by extending telomere length, using the enzyme *telomerase*, and thereby divide indefinitely. Alternatively, if a cancer cell has gained the ability to evade cell death, the chromosomal instability arising due to eroded telomere length may no longer trigger cell death. This would lead to an increased acquisition of new genomic aberrations, potentially accelerating tumor progression. Importantly though, an excessive rate of genomic instability might lead to so many aberrations being acquired that the tumor cells no longer become viable. It is therefore possible that chromosomal instability due to eroded telomeres may be an early driver of genomic aberration (leading to initial oncogenic transformation), and that telomerase expression is a characteristic acquired at a relatively late stage (once the cancer genome has become sufficiently aberrant)[17].

Cancer cells require nutrients and must dispose of waste, both of which can be transported by blood. If cancer cells could only use existing vasculature for transport, non-perfused tumor areas would become necrotic, and solid tumors would have relatively limited growth potential. Many cancers have the ability to induce angiogenesis by producing signaling proteins such as vascular endothelial growth factor (VEGF). Oncogenic signaling and hypoxia are common factors which may stimulate production of VEGF from tumor cells.

Healthy cells are arranged according to specific anatomical structures, which enable physiological function of organs. Cancer cells, in contrast, do not proliferate to generate ordered structures, but rather expand wherever possible. Initially, this proliferation may respect existing anatomical structures, but eventually cancer cells are likely to become invasive (i.e. grow into surrounding tissue) and metastatic (i.e. spread to other locations in the body). The mechanisms behind invasion and metastasis are varied, and remain inadequately understood.

These hallmarks provide a basic framework for understanding the main features that enable carcinogenesis. However, the evolutionary dynamics that govern tumor progression lead to an astounding diversity in cancers, and any overarching description of the disease group will necessarily be a simplification. Commonalities across tumors can be identified, but it must also be noted that cancer is ultimately disease in which each case is unique.

Breast cancer epidemiology and risk factors

In women, breast cancer is the most common form of cancer world-wide (excluding non-melanoma skin cancer), and accounts for the highest number of deaths and disability-adjusted life years lost[19]. In total, over 600 000 deaths are caused by breast cancer every year, leading to an annual loss of over 17 million disability-adjusted life years. On a national level, breast cancer shows higher incidence in more developed countries (defined by Socio-Demographic Index – SDI). In the countries with highest quintile SDI, 1 in 11 women develop breast cancer over a lifetime, while in countries with lowest quintile SDI, 1 in 38 women develop breast cancer over a lifetime[19].

The trends in breast cancer mortality and incidence in the United States are depicted in *Figure 2*. The trends observed here are broadly similar to those seen in other western countries with widespread mammographic screening (e.g. Norway). From the 1980s, there was a sharp increase in the incidence of breast cancer diagnoses[20]. This increase coincided with the introduction of mammography screening of asymptomatic women. If the increase reflected a genuine trend in cancer development in a population, one would also expect a proportionate increase in the incidence of metastatic cancer. As this was not the case, it is likely that the increased incidence was primarily a result of more vigilant diagnostic practice. In the early 2000s, the incidence of breast cancer stabilized at a level approximately 50% greater than the level prior to the introduction of mammography screening. Breast cancer mortality started declining in the 1990s, which is likely a result of both earlier diagnosis and

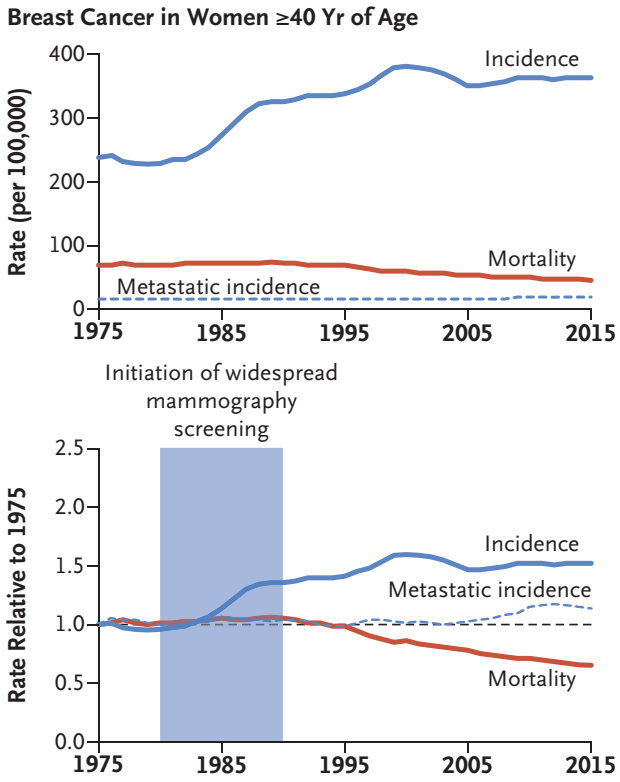


Figure 2: Incidence and mortality in breast cancer over time (U.S.A.). Metastatic incidence refers to patients with metastasis at the time of initial breast cancer diagnosis. Reproduced with permission from Welch et al.[20], copyright Massachusetts Medical Society.

improved treatment[21]. There is however considerable evidence indicating that a major proportion of breast tumors identified by screening would not have progressed to clinical disease if left untreated[22,23]. The identification and treatment of such clinically insignificant tumors are referred to as *overdiagnosis* and *overtreatment*, respectively. Understanding which early-stage breast lesions require treatment, and which do not, remains a significant clinical challenge[22–24].

Numerous factors affect an individual's risk for developing breast cancer, the most important being age (higher incidence in older individuals) and gender (higher incidence in females)[25]. Up to approximately 30% of breast cancer risk may be hereditary[26,27]. Germline variants can be categorized by penetrance, that is, by how great of an increased disease risk the variant confers. High penetrance variants in genes such as *BRCA1* and *BRCA2* can lead to an estimated lifetime breast cancer risk of 65% and 45%, respectively[28]. Other high penetrance variants have been identified in *TP53* (Li-Fraumeni syndrome), *PTEN* (Cowden syndrome), and *STK11* (Peutz-Jegher syndrome). However, only 20-25% of hereditary breast cancer risk can be explained by these well-characterized variants[28]. The majority of hereditary breast cancer risk appears to be mediated in a polygenic manner by a great number of low penetrance variants. Less than half of familial breast cancer risk can, however, be explained by currently known risk variants[29]. Breast cancer risk is also strongly related to hormonal and reproductive factors, including age at first childbirth (younger age at first birth associated with lower risk), parity (greater number of children associated with lower risk) and menopause (greater risk in pre-menopausal women compared to post-menopausal women of the same age)[30]. Important lifestyle factors include body mass index, physical activity and alcohol consumption. Smoking may lead to a slightly

increased risk, however breast cancer is in general not strongly associated with exposure to mutagenic carcinogens[30,31].

Clinical considerations in breast cancer

The human mammary gland is a tree-like branching structure[32]. The internal end of the mammary gland consists of numerous milk-producing alveoli which collectively form lobules. These drain out through lactiferous ducts to the nipple. The mammary gland is enveloped by adipose tissue and a specialized supporting stroma. The mammary gland has two cell layers: a basal myoepithelial layer, and a luminal layer of columnar epithelium.

Breast cancer generally progresses through multiple stages prior to becoming invasive. The path to invasion may differ depending on the histological subtype of the tumor; ductal carcinomas, which account for up to 75% of diagnosed cases[33], are focused upon here. The generally accepted progression model posits that normal mammary epithelial cells may go through the stages of flat epithelial atypia, atypical ductal hyperplasia, and ductal carcinoma *in situ* (DCIS), before becoming invasive[33]. The pre-invasive stages are non-obligate precursors to invasive disease, meaning that a lesion does not necessarily need to go through every stage in order to become invasive. Also, progression may spontaneously halt at any stage. After becoming invasive, breast tumors may metastasize, commonly to lymph nodes, the brain or to bone. Breast tumors are heterogeneous, and this heterogeneity is also evident in pre-invasive lesions[34]. The molecular pathways leading to progression are poorly understood and likely differ extensively between tumors.

The treatment of a breast tumor is primarily dependent upon the stage at which it is identified, and the targetable characteristics displayed by the tumor[35]. Breast cancers are staged using three main criteria: tumor (whether or not the lesion has become invasive, and how large it is), lymph nodes (the extent to which cancer cells have spread to lymph nodes), and metastasis (whether or not the tumor has metastasized to other organs). Collectively, these factors, referred to as the TNM system, capture the main considerations which determine surgical options. DCIS and breast tumors which have not yet metastasized can be treated surgically with curative intent, but this is not the case for metastatic breast cancer. For larger tumors, neo-adjuvant (i.e. before surgery) treatment using chemotherapy or targeted therapy can shrink the tumor, leading to improved surgical outcomes[35].

Medical treatment of breast tumors consists of chemotherapy and targeted therapy. Chemotherapy is a collective term for medical treatments which in some way target cellular

proliferation. Cancer cells proliferate faster than most healthy cells, and such treatment will therefore disproportionately affect cancer cells. The majority of healthy cells divide at some non-zero rate, and certain cell types, such as those in hair follicles or the intestine, proliferate relatively rapidly. Healthy cells will therefore also to some extent be affected by chemotherapy, leading to considerable side effects. Targeted therapies use a more focused approach, selectively affecting specific pathways which are as unique as possible to tumor cells. 70-80% of breast tumors express estrogen receptor (ER)[36,37], which acts as a transcription factor and is a major regulator of mammary gland proliferation. ER signaling can be inhibited, either using selective estrogen receptor modulators (e.g. tamoxifen), or by reducing estrogen production using gonadal suppressors or aromatase inhibitors[38]. Approximately 20%-30% of breast tumors show overexpression of human epidermal growth factor receptor 2 (HER2). HER2 acts as an oncogene when overexpressed, and can be targeted using monoclonal antibodies such as trastuzumab[39]. ER and HER2, in conjunction with progesterone receptor, are routinely analyzed in clinical practice. Tumors which express none of these receptors – referred to as *triple-negative* – are generally more aggressive and lack targeted therapies.

If a deleterious *BRCA* mutation is identified, tumors can be treated using poly(ADP-ribose) polymerase (PARP) inhibitors[40]. In brief, BRCA1 and BRCA2 are tumor suppressors with functions related to repair of double-stranded DNA breaks. If certain mutations arise in *BRCA1/2*, their protein products cease to function. Independently of BRCA function, single-stranded DNA breaks may also occur in cells and become double stranded breaks upon DNA replication. Ordinarily, single stranded breaks are repaired by the protein PARP1, but this mechanism is blocked by PARP-inhibitors. Thus, inhibiting PARP essentially exacerbates the underlying problem to such an extent that cell death occurs. This concept is called *synthetic lethality*.

Immunotherapies have shown limited effect in some early breast cancer trials[41]. This may be ascribed to the immune phenotype and relatively low mutational burden in breast cancer[42]. A PD-L1 inhibitor was, however, recently granted approval for treatment of triple-negative breast cancer with PD-L1 expression[43], and further approaches to stratification and combinatorial therapies continue to be explored[41].

Finally, radiation therapy may be used to treat breast tumors[35]. Radiotherapy can be used after surgery in order to eliminate residual tumor cells, and thereby reduce the risk of relapse. It is also an effective tool for killing tumor cells where surgery is not an option (for example brain or bone metastases).

Molecular classification of breast carcinomas

The goal of personalized medicine is to tailor treatment of disease to the individual patient. When a patient's disease characteristics are profiled, certain features are relatively simple to interpret, such as whether or not a breast tumor expresses estrogen receptor, or if a lesion has become invasive. In modern cancer genetic research, it is possible to identify tumor mutations, copy number aberrations, gene expression levels, protein levels, methylation status and germline polymorphisms, in addition to traditional clinical variables. If all these features are profiled, across the entire genome, it is possible to generate on the order of millions of data points for a single tumor. These features cannot each be assessed individually. However, they often correlate with one another, and it is therefore possible to distil this mass of data into meaningful and interpretable disease groups. In breast cancer, the three most important molecular classification systems are the intrinsic subtypes[44,45], claudin-low status[46], and the IntClust subtypes[37].

Gene expression microarrays were pioneered in the 1990s, allowing, for the first time, the entire transcriptome of a tumor to be characterized[47]. One of the first major applications of this technology was gene expression profiling of breast carcinomas by Perou & Sørlie *et al.*[44,45]. Tumors were sampled before and after chemotherapy. In order to identify inter-tumor variation inherent to the tumors (i.e. not an effect of therapy or random noise), genes were identified which showed high variance across tumors and low variance within repeated samples[44,48]. Hierarchical clustering of expression values from those genes was then performed. This revealed the existence of five robust tumor groups, which could be validated in external cohorts[45,49]: Basal-like, HER2-enriched, luminal A, luminal B, and normal-like. These were named the *intrinsic subtypes*.

Basal-like tumors showed gene expression features similar to the basal epithelial cell layer in mammary ducts, and were named accordingly. These tumors were aggressive, both in terms of survival and proliferation levels, and were mostly triple-negative. HER2-enriched tumors showed certain similarities to basal-like tumors (e.g. they were ER-negative), but were characterized by frequent *ERBB2* (HER2) overexpression. Luminal A and B tumors both showed transcriptomic features reminiscent of the luminal epithelial layer of mammary ducts, and were mostly ER-positive. Luminal A tumors showed relatively good prognosis and low proliferation levels, whereas luminal B tumors showed significantly worse prognosis and higher proliferation levels. Normal-like tumors showed transcriptomic similarities to normal mammary tissue, and were mostly ER-positive (although later studies have shown

somewhat divergent proportions of ER-positivity among normal-like tumors[37,50–52]). It remains unclear whether the normal breast-like features in this tumor group are genuine cancer cell features, or an artefact of non-tumor infiltration.

The intrinsic subtypes have been validated, and shown to be of clinical relevance in numerous cohorts[53,54]. The intrinsic subtypes permeate essentially all tumor characteristics, and it has been proposed that the different subtypes should effectively be viewed as distinct disease entities[48,52,55,56]. Molecular assays for intrinsic subtyping (notably PAM50/Prosigna[53]) are being commercialized and are rapidly entering clinical practice[57].

The claudin-low disease group was identified by Herschkowitz *et al.*[58] when gene expression data from mouse mammary tumors were jointly analyzed with gene expression data from human breast tumors. A previously unidentified cluster emerged in this analysis, which showed low expression of claudins and other genes related to cell-cell adhesion. This tumor group, named *claudin-low,* was later characterized in depth by Prat *et al.*[46] and was proposed as a sixth intrinsic subtype. It is important to note that identification of claudin-low tumors was performed as a distinct second step after intrinsic subtype classification. The original intrinsic subtype was therefore overwritten in tumors classified as claudin-low. Claudin-low tumors were – and continue to be – analyzed as a single group, irrespective of their underlying intrinsic subtype[46,59,60].

Claudin-low tumors showed high expression of genes related to epithelial-mesenchymal transition (EMT), and transcriptomic patterns consistent with a stem cell-like, or less differentiated, state[46]. Claudin-low tumors also showed high levels of immune and stromal infiltration. Breast stroma displays mesenchymal gene expression patterns, which in transcriptomic analyses are difficult to differentiate from an EMT program[61]. Whether the EMT-like transcriptional signature in claudin-low tumors was a result of a genuine EMT process in tumor cells, or a result of stromal admixture, was unclear[46]. This concern was partially addressed by immunofluorescence staining of tumors for cytokeratins and the EMT transcription factor vimentin. Tumor cells (identified by cytokeratins) in many, but not all, claudin-low tumors expressed vimentin. However, numerous non-claudin-low basal-like tumors also expressed vimentin.

Claudin-low tumors were often ER-negative. They were described as a subgroup of basal-like tumors, despite a substantial proportion of claudin-low tumors being classified as normal-like[46]. Claudin-low tumors were relatively heterogeneous, and appeared in many respects to be intermediate to luminal-like and basal-like tumors. For example, survival in patients with claudin-low tumors was greater than in patients with basal-like tumors, but

lesser than in patients with luminal A tumors. Proliferation levels in claudin-low tumors were lesser than in basal-like tumors, but greater than in luminal A and normal-like tumors.

The prevalence of claudin-low tumors in various cohorts analyzed by Prat *et al.*[46] ranged from 7% to 14%. In The Cancer Genome Atlas breast cancer cohort only 1.5% of tumors were classified as claudin-low[56]. In contrast, the distribution of the originally proposed intrinsic subtypes is relatively consistent across cohorts[37,44,53,56].

Claudin-low tumors have been explored by others, and the findings from the initial characterization have been reasonably robust[59,60]. The existence of claudin-low tumors has been validated in numerous cohorts[37,56,59,60,62,63], and an analogous tumor group has been identified in bladder cancer[64,65].

One major advance in the understanding of claudin-low tumors is the concept of *cellular pliancy*. Cellular pliancy is the notion that "each differentiation stage within a defined cellular lineage is associated with a unique susceptibility to malignant transformation when subjected to a specific oncogenic insult"[12]. This concept is illustrated in *Figure 3*, in which embryonic/undifferentiated cells are located at the top of the landscape representing the differentiation hierarchy[12]. As cells differentiate, epigenetic changes lead to a loss of pluripotency, and cells become progressively more committed to differentiated states in specific lineages. This is illustrated by a deepening valley in the landscape. All cells, irrespective of differentiation state, may be subject to oncogenic insults, such as mutations or copy number aberrations. However, the more differentiated a cell is, the greater the oncogenic perturbation needs to be in order to push a cell away from its committed differentiation path over to a malignant state. This concept is partially motivated by the observation that pediatric cancers generally have much lower mutational burden than adult cancers, and is supported by mechanistic studies in mice and cell lines[12]. One important note is that transdifferentiation may occur in

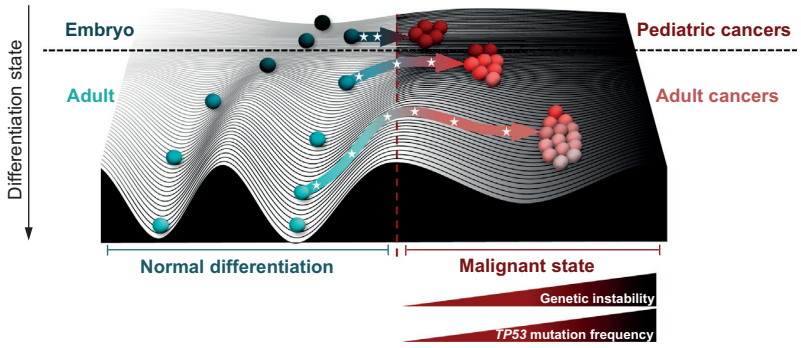


Figure 3: Graphical depiction of cellular pliancy. Movement down the y-axis represents a progression along the differentiation hierarchy, in which deepening valleys represent increasing commitment to specific lineages represented on the x-axis. White stars represent oncogenic insults. Reproduced from Puisiseux et al.[12] with permission from Elsevier.

differentiated cells, for example through EMT (which may be activated by microenvironmental signaling)[12,66,67]. Such processes may therefore reduce the number of genetic perturbations required for malignant transformation. Morel *et al.*[67] explored cellular pliancy in breast cancer, and found that that the EMT transcription factor ZEB1 promoted malignant transformation, while maintaining genomic stability. Triple-negative claudin-low breast tumors showed high expression of *ZEB1* and other EMT-associated genes, and a paucity of copy number aberrations. These findings provided a mechanistic rationale for genomic stability in claudin-low tumors, although analyses did not extend to ER-positive tumors.

The intrinsic subtypes and claudin-low were identified using only phenotypic data (i.e. gene expression). Curtis *et al.*[37] investigated the possibility of generating a breast cancer classification in which both cause (copy number aberration) and effect (gene expression) are considered together. Genes were identified in which there was a *cis* correlation between copy number and expression, and the top thousand most highly correlated genes were used for integrative clustering[68]. This revealed the existence of ten clusters, which were named the *IntClust* subtypes. These displayed distinct copy number profiles, gene expression features, and survival patterns. One of these subtypes – IntClust4 – was notable due to a near absence of copy number aberrations. These tumors had strong immune infiltration, and were heterogeneous in their expression of estrogen receptor. There was a substantial overlap between claudin-low tumors and IntClust4 tumors[37,62,67]. IntClust4 tumors showed relatively homogeneous copy number and gene expression patterns in those genes used for IntClust subtyping. The subtype was, however, later split into IntClust4 ER-positive and ER-negative groups due to substantial differences in several characteristics not fully captured by IntClust classification[37,62,69]. This could be viewed as a luminal/basal split in the copy number derived IntClust4 subtype.

Genetic and epigenetic characteristics of breast cancers

Somatic mutations in breast cancer genomes were first comprehensively profiled in 2012 by Stephens *et al.*[70] and The Cancer Genome Atlas consortium[56]. These findings have later been expanded upon in several other studies[18,62,71–75]. When compared to other cancer types, breast cancers have relatively low mutational burden, on average carrying approximately one mutation per million base pairs[76]. The somatic mutation profiles of breast tumors are

heterogenous, with no single mutation consistently found in more than approximately 40% of cases. Patterns of mutations correlate well with tumor phenotypes (i.e. molecular subtype, hormone receptor status). Basal-like/ER-negative tumors are characterized by frequent *TP53* mutations, and luminal-like/ER-positive tumors are characterized by *PIK3CA* and *GATA3* mutations (*Figure 4*). However, there is moderate overlap between the mutations found in different subtypes, and mutation status cannot currently be used to accurately determine subtype (although this may be possible with approaches involving machine learning[77]). Mutations in breast cancer follow a long-tailed distribution, with *TP53*, *PIK3CA* and *GATA3* among the only genes that are consistently found to carry mutations in over ten percent of breast tumors.

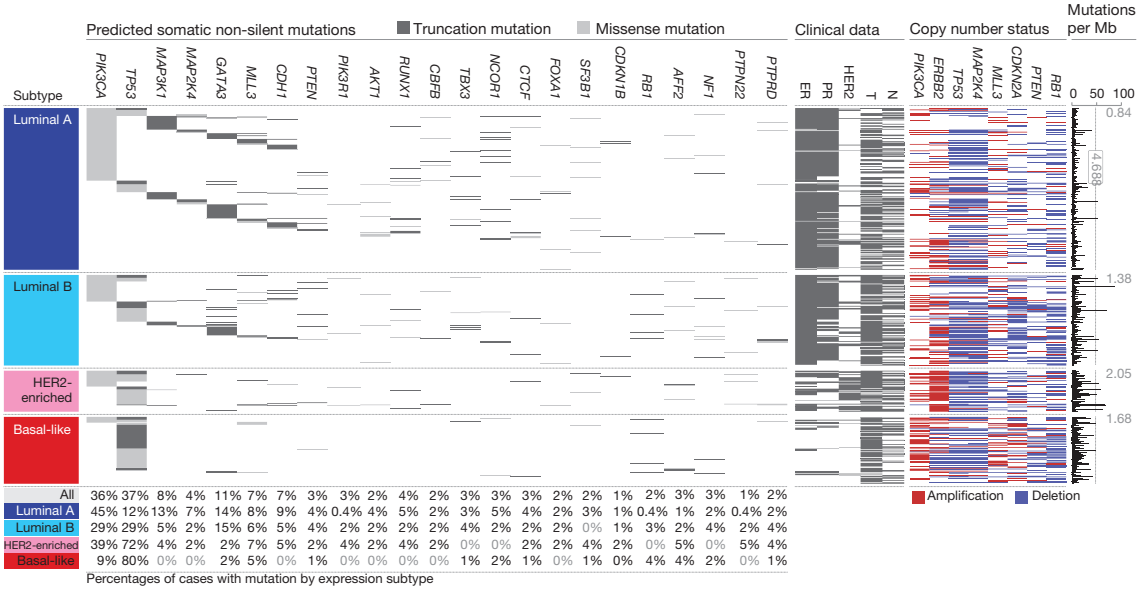


Figure 4: Distribution of mutations and copy number aberrations in The Cancer Genome Atlas breast cancer cohort[56].

Mutations are caused by numerous processes, such as DNA replication errors or carcinogenic exposures (e.g. tobacco smoke, ultraviolet light). Each mutational process has a unique propensity to induce mutations in certain nucleotides, and in certain nucleotide contexts. For example, ultraviolet light has a tendency to induce cytosine-to-thymine (C>T) mutations[31]. An overweight of these C>T mutations are found in positions where there is a thymine nucleotide adjacent (5') to the mutated cytosine. These unique imprints, which can be deconvoluted from sequencing data, are referred to as mutational signatures[31,76,78]. Originally, 21 signatures were identified[76], but more recent analyses have proposed up to 81

signatures[31]. Evidence of several mutational signatures has been found in breast tumors, including signatures associated with deamination of 5-methylcytosine, APOBEC activity and defective DNA mismatch repair[31]. Breast cancer genomes may in some cases show imprint of tobacco smoking, but are mostly unaffected by signatures associated with environmental mutagen exposure.

Breast cancers frequently display gross chromosomal instability[37,56]. It has been proposed that the majority of breast cancers are more heavily driven by copy number aberrations (CNAs) than point mutations[79]. An estimated 12% of transcriptional variation in breast cancer can be attributed to CNA[9], and the IntClust subtypes[37] illustrate the phenotypic importance of these expression-associated CNAs. Several genes, and genomic regions, show recurrent CNA, such as *ERBB2* (17q), *MYC*, (8q) and *TP53* (17p)[37,56]. Certain CNAs show considerable association with intrinsic subtypes, such as *ERBB2* amplification generally found in HER2-enriched tumors, and *MYC* amplification frequently found in basal-like tumors (*Figure 4*). In general, basal-like and HER2-enriched tumors show greater burden, and more complex patterns, of CNA than tumors in the other intrinsic subtypes[37,80–82]. IntClust subtypes are closely associated to copy number patterns, with most of them defined by specific CNAs (such as IntClust1 linked to 17q amplification). Two IntClust subtypes are defined by a near absence of CNAs (IntClust3 and 4)[83]. An early study of CNAs in claudin-low tumors suggested that copy number patterns were essentially the same as in basal-like tumors[59]. A more recent and comprehensive study suggested genomic stability in claudin-low tumors[67].

Methylation patterns of breast carcinomas are often aberrant and linked to molecular subtypes[56,84]. There is a marked contrast between the methylation features of basal-like and of luminal tumors. Normal-like and HER2-enriched tumors are not as clearly demarcated, and may to some extent show methylation patterns similar to either basal-like or luminal tumors[56,84]. When compared to tumor-adjacent breast tissue, luminal/ER-positive tumors are reported to show more aberrant methylation profiles than basal-like/ER-negative tumors[85,86].

Mouse models of cancer

Ethical and practical factors necessitate the use of animal models in biomedical research. Humans and mice share broadly similar physiology, and have a 79% consensus in amino acid sequence in orthologous proteins[87]. Mice have relatively short generation time, with

pregnancy being carried to completion within approximately three weeks after mating. Sexual maturity in females is reached within 6-8 weeks after birth[88]. With litter sizes averaging approximately five to eight (depending on the mouse strain), it is practical to carry out experiments at a relatively high rate in mice. Laboratory mice are usually inbred and are therefore genetically homogenous. Inbreeding thereby reduces random germline variation as a confounding factor in experiments.

Several limitations of mice as models for human disease must be noted. There are numerous differences between the two species in both innate and adaptive immune systems, and there is increasing skepticism against mice as immune system models[88–90]. The genetic homogeneity in laboratory mice may also be considered a weakness, as it does not reflect the genetic diversity found in human populations. Some generally accepted practices for animal experiments may also introduce limitations beyond inherent physiological differences. For example, young mice are commonly used for study of diseases that in humans occur in the elderly. Also, microbial exposure is minimized in animal facilities, which confounds any processes related to the microbiome.

Cancer can be investigated in mice by genetically engineering a predisposition to cancer, exposing mice to carcinogens, or by transplanting cancer cells into mice. Genetically engineered mouse models (GEMMs) are mice with specific modifications to their genome which might predispose for disease development (e.g. cancer). GEMMs were pioneered in the 1970s by injecting viral DNA into explanted mouse blastocysts[91]. Early techniques were crude, and could not control where in the genome the DNA sequence was inserted, nor how many copies of it were inserted. Contemporary methods allow for more precise modification of the genome, including insertions, deletions and single nucleotide substitutions[92]. It is often important that genetic modifications only exert an effect in a specific tissue, or at a specific time. For example, certain genetic modifications, such as knock-out of essential cell-cycle genes, might be incompatible with gestational development. Such a genetic modification could therefore only be studied by knocking out the given gene in post-natal mice. Global effect of a genetic modification might lead to cancer developing in one tissue before developing in the tissue actually intended for study. These issues can be approached using various genetic techniques, such as Cre-Lox recombination (enabling temporal control of a genetic modification), or using tissue-specific promoters, such as mouse mammary tumor virus (enabling tissue-specific expression)[92]. GEMMs are primary cancer models, meaning that the complete tumorigenic process can be studied, including tumor initiation. GEMMs are immunocompetent, enabling the study of tumor-immune interactions. One downside to

GEMMs is that their tumors are driven by relatively few mutations, and therefore do not reflect the diversity of mutations observed in human cancers[93–98]. Despite this genetic homogeneity, GEMM tumors do, however, often display surprisingly diverse phenotypes[58,99–101]. This observation highlights that the phenotypic characteristics of a tumor are not entirely determined by the specific aberrations in the tumor genome. Other factors which may play a role in shaping cancer phenotypes include microenvironmental signaling (including immune response) and tumor cell-of-origin.

Carcinogen exposure is implicated in numerous forms of human cancer, including lung, skin and bladder cancer[31]. Establishing causality between an environmental exposure and cancer development in humans may be difficult. There is often a long latency between carcinogen exposure and the development of clinically observable cancer. There are also countless confounding factors which may affect an individual's lifetime risk of cancer, and the relationship between carcinogen exposure and tumorigenesis is stochastic in nature (e.g. not all smokers develop lung cancer). Carcinogen-induced mouse tumor models are useful for establishing causal relationships between environmental exposures and cancer development. They are also useful as models for carcinogen-induced human cancers. Somewhat controversially though, carcinogen-induced tumor models may be used to study cancers not caused by environmental exposures in humans. For example, 7,12-dimethylbenzanthracene (DMBA) can be used to induce breast cancer in mice[102–104], but mutagenic exposure is not a common cause of breast cancer in humans[31]. One compelling feature in carcinogen-induced models is the heterogeneity of genetic aberrations that can arise, which may more accurately reflect the genetic diversity seen in humans[94,95,105,106]. This heterogeneity might, however, also be a drawback for study designs requiring predictable tumor characteristics. In likeness with GEMMs, carcinogen-induced models are immunocompetent primary tumor models. Carcinogen exposure often leads to high mutational burden[76,94,95,105], which may be relevant for the study of immunotherapeutics[42].

Cancer can also be studied by transplanting previously established tumors, or tumor cells, into mice. This can be done using human cancers (xenografts) or murine cancers (allografts). Human tumor material used for transplantation can either originate directly from patients (i.e. patient derived xenografts[107]) or from immortalized cell lines. Transplantation therefore represents a way in which human cancer can be studied *in vivo* without involving human subjects. One advantage to transplanting is the possibility of comprehensively profiling tumor tissue, or cells, *ex vivo* prior to experiments. Tumor cells can also be genetically modified prior to transplantation, enabling precise manipulation of tumors in animal models. Tumors can be serially passaged in multiple generations of mice,

which may be relevant for studying tumor evolution. To the extent that tumor characteristics are maintained over multiple generations, serially passaging tumors also allows experiments to be performed on, essentially, the same tumor indefinitely. The immune system is a major challenge in the use of transplanted tumor models. Allografts can be transplanted to immunocompetent mice, provided that the primary tumor was established in a mouse with sufficiently similar genetic background. In contrast, human cancer samples are immunologically rejected if transplanted to immunocompetent mice. Human cancer samples must therefore be transplanted to immunodeficient mice, obviating the possibility of studying tumor-immune reactions. Attempts have been made to humanize the immune system of mice, allowing for transplantation of human tumors to partially immunocompetent recipients[107]. Ultimately mice with humanized immune systems succumb to graft-versus-host disease, but there is a window of several weeks in which they can be useful for investigation. Tumor samples can be transplanted subcutaneously, orthotopically (into the tumor's tissue of origin) or into the circulatory system (by intravenous or intracardiac injection). The choice of how tissue and cells are transplanted will affect the tumor microenvironment and ability to disseminate.

There are several notable differences in mammary gland biology between humans and mice[108]. The mammary glands in both are essentially tree-like structures, but human mammary glands terminate in complex branching structures (*lobular end units*), while mouse mammary glands terminate in simpler ductal structures (*terminal end buds*). The mammary glands in both species are surrounded by adipose tissue. In humans, the mammary gland is additionally supported by a specialized stroma, not found in mice, which contains abundant fibroblasts[108]. This implies microenvironmental distinctions between mouse and human mammary glands. Estrogen receptor signaling is a key driver in the majority of human breast cancers, but the response to estrogen receptor signaling differs significantly between humans and mice[109]. Also, primates have peak estradiol concentrations almost ten times that of mice (although differences in estrogen receptor density and binding affinity may modulate this discrepancy)[110]. The significance of estrogen receptor signaling in mouse mammary tumors is therefore unclear.

Molecular stratification of human breast tumors has been studied extensively, and there is an etiological understanding of the intrinsic subtypes[44,45,111]. It is however less clear how the transcriptomic features in mouse mammary tumors translate to subtypes found in humans. This issue was explored by Herschkowitz *et al.*[58], and was later expanded upon by Pfefferle *et al.*[101]. Inspired by the methodology used in discovering the intrinsic subtypes[44],

gene expression data from 27 mouse mammary tumor models were hierarchically clustered[101]. The majority of these tumor models were GEMMs. This analysis revealed 17 murine subtypes in the dataset. 11 of 27 models showed homogeneous gene expression patterns (i.e. over 80% of tumors found in one cluster), while the remaining models showed more heterogeneous gene expression patterns. A subsequent analysis clustered human and murine data together, and the 17 murine subtypes were correlated to the human intrinsic subtypes. This revealed that representative tumor models existed for all intrinsic subtypes, however no homogeneous claudin-low models were identified.

7,12-dimethylbenzanthracene is a potent carcinogen[112]. Humans may be exposed to DMBA, and other similarly structured polycyclic aromatic hydrocarbons (PAHs), through occupational contact in heavy industries, tobacco smoke, or dietary intake[113]. PAHs themselves are chemically inert, but are in mammalian cells metabolized to diol epoxides which covalently bind to DNA[112,113]. This process is understood to cause mutations by interfering with DNA replication mechanisms. DMBA and other PAHs have numerous additional effects which may influence tumorigenesis, including lymphoid toxicity[114] and inhibition of gap-junction intercellular communication[115].

In breast cancer research, DMBA was initially used for chemically induced carcinogenesis in rats. It was observed that a single bolus of orally administered DMBA could induce mammary tumor formation in females[116]. DMBA-induced carcinogenesis was later explored in mice, and it was found that multiple boluses were required for mammary tumor induction[102,103]. There was long latency to tumor formation, and there was high incidence of mortality from causes other than mammary adenocarcinoma. The addition of a progestin – *medroxyprogesterone acetate* (MPA) – to the tumor induction protocol was later found to reduce latency and increase tumor incidence[104]. MPA induces a RANK-ligand mediated proliferation of mammary epithelial cells[104,117]. This leads to greater tumor incidence by increasing mutagenesis in mammary epithelial cells (as DMBA interferes with DNA replication) and by expanding the pool of mammary epithelial cells susceptible to DMBA-induced mutagenesis.

MPA/DMBA-induced mouse mammary tumors show phenotypic heterogeneity, both at a histological and transcriptomic level[118]. When Pfefferle *et al.*[101] jointly analyzed MPA/DMBA-induced tumors with other mammary tumor models, the twelve profiled MPA/DMBA-induced tumors were allocated to five different murine subtypes. When murine subtypes were mapped to human intrinsic subtypes, eight of twelve MPA/DMBA-induced tumors were classified as subtypes resembling claudin-low. DMBA-induced tumors,

with and without MPA-supplementation, have been DNA-sequenced, which revealed mutational heterogeneity[106]. Tumors without MPA-supplementation showed frequent hotspot *Pik3ca*, mutations, which are also commonly observed in human breast tumors[62,70]. Otherwise, mutations mostly occurred in genes not commonly mutated in human breast cancer. Human breast tumors are generally not associated with exposure to chemical carcinogens[31]. Given the poor concordance between genetic characteristics of human breast tumors and MPA/DMBA-induced tumors, the validity and usefulness of the model is uncertain.

Aims

This thesis is an exploration of the genetic and epigenetic factors that determine cancer phenotypes. The etiological factors taken into consideration include somatic mutations, copy number aberrations and DNA methylation. Cancer phenotypes are here primarily defined through gene expression patterns, with some consideration also given to clinical outcomes.

The overarching hypothesis, throughout the studies in this thesis, is that cancer phenotypes are encoded in underlying genetic and epigenetic features.

We explored this hypothesis, directly and indirectly, in studies which aimed to:

- Characterize the interplay between genomic and transcriptomic features in a mouse model of breast cancer, and evaluate its utility as a model for human disease.
- Critically re-evaluate the established understanding of a gene expression subtype in breast cancer based on genomic, transcriptomic and clinical features.
- Comprehensively identify the genetic and epigenetic correlates of gene expression across the entire genomes of 23 cancer types.

Results in brief

Study I

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 Sørlie. Breast Cancer Research 21, 85 (2019).

Claudin-low breast tumors carry distinct and potentially actionable characteristics[46,59], yet no therapies currently exist that target the unique features associated with the disease group. Accurate and well-characterized animal models are vital pre-requisites for developing new targeted therapies.

In this study, we chemically induced mammary tumors by exposing mice to the carcinogen *7,12-dimethylbenzanthracene* (DMBA) and the progestin *medroxyprogesterone acetate* (MPA). We profiled the genomic and transcriptomic characteristics of eighteen mouse mammary tumors using whole exome sequencing and gene expression microarrays. Using an established subtyping method[101], we found that half of MPA/DMBA-induced tumors showed claudin-low-like gene expression features. Histologically, the claudin-low-like tumors displayed spindloid morphology and marked immune infiltration. Similar histological features have been described in human claudin-low breast tumors[46,119], thus corroborating the transcriptomic classification. All tumors carried mutations in known driver genes, and had mutational burdens several times that of human breast cancers. Despite recurrent driver gene mutations, for example in *Kras* and *Trp53*, no mutations

accurately delineated the gene expression subtypes. Similarly, no copy number aberrations were specific to claudin-low-like tumors. There was, however, a slight trend toward lower levels of genomic instability in claudin-low-like tumors.

We compared the genomic characteristics of MPA/DMBA-induced tumors with those of human breast tumors[56,62]. In general, there was poor concordance between the genomic aberrations found in MPA/DMBA-induced tumors and the genomic aberrations found in human breast cancer. Many recurrently aberrant genes in MPA/DMBA-induced tumors, such as *Kras* or *Nf1*, were infrequently aberrant in human breast cancer. Human claudin-low breast tumors showed low mutational burden and low levels of genomic instability. The genes frequently aberrant in human claudin-low tumors, such as *TP53* mutation and *MYC* amplification, were similar to those commonly aberrant in basal-like breast cancer.

Finally, we investigated the gene expression characteristics of MPA/DMBA-induced tumors. The transcriptomic traits in MPA/DMBA-induced claudin-low-like tumors showed strong concordance with those observed in human claudin-low tumors. Notably, we found that both human and murine claudin-low tumors expressed high levels of the immunosuppressive genes *CD274* (encoding the immune checkpoint ligand PD-L1) and *PTGS2* (encoding the cyclooxygenase COX2). These may represent potential targets in claudin-low breast cancer, and can be inhibited using existing drugs.

In sum, our study revealed a surprisingly poor concordance between genomic aberrations and transcriptomic features in MPA/DMBA-induced mouse mammary tumors. Our findings suggest that non-genomic factors, such as cell-of-origin or tumor microenvironment, may be the primary determinants of tumor phenotypes in this model. There were marked contrasts between the genomic features of human and murine mammary tumors, but transcriptomic characteristics indicated that claudin-low-like MPA/DMBA-induced tumors may nonetheless be a useful model for certain applications.

Study II

Re-definition of *claudin-low* as a breast cancer phenotype

Christian Fougner, Helga Bergholtz, Jens Henrik Norum and Therese Sørlie.
Nature Communications 11, 1787 (2020).

When the breast cancer intrinsic subtypes were discovered, five robust subtypes were identified: Basal-like, HER2-enriched, luminal A, luminal B, and normal-like[44,45,49]. Several years later, an additional intrinsic subtype, named *claudin-low*, was proposed[46,58]. Claudin-low tumors showed distinct gene expression characteristics: Low expression of cell-cell adhesion genes, high expression of epithelial-mesenchymal transition genes, and stem cell-like/less differentiated gene expression patterns. In many regards, claudin-low tumors were remarkably heterogeneous.

Tumors can be allocated to one of the original five intrinsic subtypes using the gene expression-based PAM50 assay[53]. Classification according to claudin-low status, however, requires a distinct second analysis (also gene expression-based)[46]. We asked the question: if *claudin-low* represents an intrinsic subtype, analogous to those originally proposed, why is a separate analysis required to identify them?

This study describes a comprehensive analysis of genomic, transcriptomic and clinical data from three breast cancer cohorts, totaling 3349 tumors[37,51,56,62,120]. We analyzed these data with the notion that *claudin-low* is not a subtype analogous to the intrinsic subtypes, but rather a phenotype which tumors may display *in addition* to their intrinsic subtype.

We identified claudin-low tumors in the METABRIC cohort[37,62] using an established classifier[46]. These were then stratified according to the intrinsic subtype to which they were initially assigned by the PAM50 predictor[53]. Stratifying tumors in this manner revealed that claudin-low tumors carried a marked imprint of their intrinsic subtype. For example, basal-like claudin-low tumors were mostly ER-negative, frequently carried *TP53* mutations, and expressed high levels of the proliferation marker *MKI67*. In contrast, luminal A claudin-low tumors were mostly ER-positive, showed infrequent *TP53* mutations, and expressed lower levels of *MKI67*. Claudin-low tumors have previously been associated with poor prognosis[46,59]. When stratified by intrinsic subtype, we found no evidence indicating that claudin-low-status affects disease-specific survival. Rather, disease-specific survival in patients with claudin-low tumors was mostly reflected in the tumor's intrinsic subtype. Some characteristics were consistently found in claudin-low tumors irrespective of intrinsic

subtype: Low levels of genomic instability, low mutational burden, and high levels of immune and stromal infiltration.

We questioned whether the established classifier might be indiscriminately identifying tumors with marked stromal and immune infiltration as claudin-low. We manually selected a list of nineteen genes related to the pathognomonic claudin-low gene expression characteristics. Hierarchical clustering of these genes revealed a tumor cluster with claudin-low gene expression features. Tumors in this cluster, which we refer to as *core claudin-low* (CoreCL), largely overlapped with claudin-low tumors identified by the established classifier. However, a subset of basal-like claudin-low tumors (as identified by the established predictor) were excluded from this CoreCL cluster. We analyzed these excluded tumors, referred to as *other claudin-low* (OtherCL), and found that they had high levels of immune and stromal infiltration. Beyond this, they showed poor concordance with the characteristics of claudin-low tumors. We concluded that the classification of OtherCL tumors as claudin-low may be dubious.

Finally, we analyzed claudin-low tumors in the Oslo2[51] and TCGA-BRCA[56,120] cohorts, and were able to validate most findings from METABRIC. However, no CoreCL cluster emerged in TCGA-BRCA, and the prevalence of claudin-low tumors was lower than in the other cohorts. Cohorts included in our study had varying cut-offs for tumor purity as inclusion criteria, and we observed that cohorts with more stringent cut-offs showed lower prevalence of claudin-low tumors.

In sum, our study revealed that *claudin-low* is not an intrinsic subtype as has previously been portrayed. Rather, *claudin-low* is a complex additional phenotype which may permeate tumors of various intrinsic subtypes. Analyzing claudin-low tumors as a single homogeneous entity, therefore, obscures the features attributable to claudin-low status. It follows that claudin-low tumors should be studied in a subtype-specific manner. Additionally, our study revealed weaknesses in the established claudin-low classifier, and we proposed approaches to mitigating the effects of these limitations. Our findings elucidate the heterogeneity in claudin-low breast tumors, and will enable more accurate and nuanced investigations into this poorly understood form of cancer.

Study III

A pan-cancer atlas of transcriptional dependence on DNA methylation and copy number aberrations

Christian Fougner, Elen K. Höglander, Tonje G. Lien, Therese Sørlie, Silje Nord, Ole Christian Lingjærde.

BioRxiv 2020.05.04.076901 (2020).

Cancer transcriptomes are shaped by genetic and epigenetic factors, such as DNA methylation and copy number aberrations. The extent to which transcription is associated with methylation and copy number, across the genome, remains inadequately understood. How these associations vary between tumor types is also unknown. In this study, we aimed to characterize the individual and combined effects of methylation and copy number on gene expression in cancer.

We first sought to identify the optimal method for modeling expression-methylation (E-M) and expression-copy number (E-C) associations. While the expression and copy number of a gene can be described using a single value, the methylation state of numerous CpGs may be relevant to the transcription of a gene. Methylation data therefore needed to be reduced to a low-dimensional gene-centric form, with a constant number of dimensions used to represent every gene. We found that principal component analysis could be used to reduce the dimensionality of methylation data, with approximately five principal components required to comprehensively describe gene-centric methylation states. We next explored the dynamics of E-M and E-C associations in cancer. Methylation frequently showed non-linear association to gene expression, indicating saturation effects in E-M relationships. In contrast, the associations between copy number and gene expression were mostly linear (provided copy number data were log-transformed). There was therefore no evidence of diminishing transcriptional effect of copy number at extreme amplification levels.

We applied the developed methods to The Cancer Genome Atlas[120] dataset, thereby generating a pan-cancer atlas of transcriptional dependence on DNA methylation and copy number aberrations (PANORAMA). The atlas was made available through a multi-functional web application.

A survey of the atlas revealed considerable differences between tumor types in extent of E-M/E-C association. In the 23 tumor types analyzed, between 2% and 14% of gene

expression was associated with copy number. Tumor types showed mean expression-methylation association ranging from 11% to 39% (although E-M and E-C associations were not directly comparable due to methodological differences). Mean E-C association in a tumor type was closely correlated with mean genomic instability index. In other words, the more frequently copy number aberrations were observed in a tumor type, the more its transcriptome appeared to be shaped by copy number levels. There were complex relationships between the degree of variation in methylation data, the degree of variation in gene expression data, and the strengths of E-M associations. Our analyses indicated that copy number aberrations are instrumental in shaping cancer transcriptomes. Patterns in E-M associations, however, raised questions about the extent to which aberrant methylation truly drives genome-wide transcriptional heterogeneity within tumor types, and if E-M associations might primarily be a reflection of cell-of-origin and normal-cell admixture.

There was no correlation between mean E-M association and mean E-C association in a tumor type. However, for individual genes, there was a pattern of mutual exclusivity in whether they showed high E-M association or high E-C association. This trend varied somewhat between tumor types, and was not observed in testicular germ cell tumors or thymoma, and high E-M associations and high E-C associations coincided (i.e. were mutually inclusive) in lower grade glioma. GO-term enrichment analysis[121–123] revealed that genes with high E-M association were often related to immune infiltration, while genes with high E-C association were often related to transcription and cell-cycle. We further explored how E-M and E-C associations interacted in individual genes, and found indications that methylation, in general, has a limited ability to modulate transcriptional effects of copy number aberrations.

In sum, we developed novel methods for interrogating associations between gene expression, methylation and copy number, and applied these to a dataset of over 7000 tumors. The database of transcriptomic associations was made available through a website, which also enables customizable analyses. Our explorations of the atlas revealed notable trends in E-M/E-C associations, and provide insight into how methylation and copy number shape cancer transcriptomes.

Methodological considerations

Cohorts

MPA/DMBA-induced mouse mammary tumors

In *Study I*, we used mouse mammary tumors induced by medroxyprogesterone acetate and 7,12-dimethylbenzanthracene. This is a commonly used model of breast cancer with an established tumor induction protocol[117]. Transgenic mice (*Lgr5-EGFP-Ires-CreERT2;R26R-Confetti*) were used. These transgenes are related to lineage tracing experiments[124] and were considered biologically inert. The mice were bred on an FVB/N genetic background and were immunocompetent.

Most genetically engineered mouse models are driven by relatively few mutations or copy number aberrations[93,97,98]. In contrast, carcinogen induced models of cancer show greater burden, and diversity, of genetic aberrations[94,95,105,106]. Carcinogen induced models may therefore be more reflective of the heterogeneity found across human breast cancers. This was one major factor in choosing the MPA/DMBA-induced model for investigation in *Study I*. This aspect does, however, also confer a lack of predictability to genotypes and phenotypes that arise in an experiment. Mutations induced by carcinogens, such as DMBA, are distributed stochastically across the genome, and it is therefore not possible to select which genes should become aberrant. This led to mutations occurring in genes not commonly mutated in human breast cancer, raising questions about the applicability of the model. MPA/DMBA-induced tumors showed far greater mutational burden than human breast cancers. This may have implications for findings related to neoantigen-mediated

immune response[42]. From a statistical perspective, the phenotypic diversity in MPA/DMBA-induced tumors increases the number of mice needed to identify trends in tumor groups.

Eighteen tumors, from fourteen mice, were included in the study. Microarray data from one tumor did not pass quality control, and could therefore not be included in analyses. The seventeen tumors with transcriptomic data available were distributed across six murine subtypes. In order to achieve some degree of statistical power, these subtypes were binned into two groups: Claudin-low-like ($n = 8$), and mixed ($n = 9$). Ideally, this cohort would have had a considerably greater sample size. However, we were constrained by tumor latency, animal welfare concerns, and cost of molecular analyses. Sample size could have been increased by including tumors from previously published studies[106,118]. However, this would have introduced significant confounding factors, including variable technologies used, batch effects, and different genetic backgrounds in mice.

METABRIC

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort[37,62] was used for analyses of human breast tumors in *Study I* and *Study II*. This cohort represents one of the largest publicly available molecular datasets for human breast cancer ($n = 2433$)[62]. Available data include: Gene expression, DNA copy number, targeted sequence data, histological classification and disease-specific survival.

Patients in the METABRIC cohort were enrolled from 1977 through 2005, and from five different centers in the United Kingdom and Canada. Due to the location of inclusion centers, there was a strong skew towards patients of European descent. Findings from METABRIC may therefore be of less relevance to individuals of African or Asian descent. Due to the time points at which samples were collected, patients received different treatment than they might have received today. In particular, targeted therapy was not available for treatment of HER2-positive tumors. Survival data from this patient group is therefore not representative of current clinical practice. Histological annotations released with the original METABRIC publication were primary pathology reports and were therefore not consistent over time and between centers[37,125]. These data are of limited value. A central pathology review was conducted in 2018, which attempted to unify histological classifications[125]. These updated classifications were used in *Study II*. Histological data were available for 84% of tumors.

METABRIC was originally subdivided into a discovery and a validation cohort[37]. Inclusion in the discovery cohort required minimum 40% tumor purity, but no such cut-off

was applied in the validation cohort. In *Study II*, we observed that these cut-offs may have affected the prevalence of claudin-low tumors.

In *Study I*, 218 claudin-low tumors were identified, and these were analyzed as a single group. In *Study II*, a different method was used for claudin-low classification (see subheading *Transcriptomic analyses*), and only 87 tumors were identified as claudin-low. These tumors were further stratified by intrinsic subtype in *Study II*, leaving between 2 and 45 samples in each group. HER2-enriched claudin-low (n = 2) and luminal B claudin-low (n = 3) tumors were excluded from analyses due to low sample numbers. There were between 9 and 45 claudin-low tumors in the remaining groups. While analyses would have been improved by greater sample sizes, these groups appeared to be sufficient for most purposes. Survival analyses may, however, not have been sufficiently powered. There were also large imbalances in sample sizes between groups (e.g. luminal A claudin-low n = 9, luminal A non-claudin-low n = 684), which may have affected statistical analyses. These imbalances, however, reflect the prevalence of tumor groups in the general population of primary breast cancers. Avoiding this would require pre-selection for balanced tumor groups upon recruiting patients to the cohort.

Oslo2

The Oslo2 cohort was used as a validation dataset in *Study II*[51]. Patients were enrolled from multiple centers in Oslo, Norway, from 2006 onwards (data published in 2017). There were 381 samples with gene expression data and hormone receptor status available. Copy number data were only available for seven tumors classified as claudin-low, and these data were therefore not used in our analyses. Tumor purity was not used as an inclusion criterion for the cohort.

Patients in Oslo2 were mostly of European descent, limiting applicability of findings for non-European individuals. It should be noted that Oslo2 was used as a validation dataset for the METABRIC cohort, which shared the same limitation. Samples were collected at a relatively recent timepoint, and patients are therefore more likely to have received contemporary treatment (e.g. HER2-targeted therapy).

There were 29 claudin-low tumors in the cohort. When stratified by intrinsic subtype, group sizes ranged from one to eleven. The group sizes were therefore not sufficient to carry out analyses in a subtype-specific manner.

The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) was a pan-cancer project, in which over eleven thousand tumor samples from 33 tumor types were collected and profiled[120]. Publicly available data from TCGA include: Gene expression, DNA methylation, DNA copy number and whole exome/genome sequence.

Tumors were collected from 2005 to 2013 from centers in the United States. Demographics skewed towards individuals of European descent. Samples were required to have a tumor cellularity over 60% for inclusion.

In *Study II*, we used data from breast tumors in TCGA (n = 1082)[56,120] for validation. We identified[46] 32 claudin-low tumors, which was not sufficient for subtype-stratified analyses. The prevalence of claudin-low tumors in TCGA was lower than in Oslo2 and the METABRIC validation cohort, but similar to the prevalence in the METABRIC discovery cohort. Taken together with other findings in *Study II*, we believe that differing cellularity cut-offs may introduce significant selection bias to analyses of claudin-low tumors.

In *Study III*, we used the TCGA pan-cancer dataset[120]. Only tumors with gene expression, methylation and copy number data were included. Due to variable cohort sizes within TCGA, we excluded tumor types with less than one hundred samples from the generated pan-cancer atlas. Ovarian serous cystadenocarcinoma was also excluded due to inadequate methylation data. This left 23 tumor types which could be included in analyses (n = 7619 tumors). All tumor types in TCGA (excluding ovarian serous cystadenocarcinoma) were made available for analysis through the online tool built in *Study III*.

Transcriptomic analyses

RNA-sequencing (RNA-seq) and gene expression microarrays are commonly used methods for high throughput gene expression profiling. Microarrays are solid surfaces, usually glass slides, with DNA oligonucleotides fixed to the surface[1]. These oligonucleotides, referred to as probes, correspond to coding sequences of known genes and are placed at known locations on the slide. For transcriptional profiling, messenger RNA (mRNA) is isolated from a sample, converted to complimentary DNA (cDNA), labelled with a fluorescent molecule, and hybridized to probes on the slide. The abundance of mRNA can be determined by measuring the intensity of light emitted at probes corresponding to known genes. Several methods exist for RNA-sequencing, including sequencing-by-synthesis, as commercialized by Illumina and others[126,127]. In brief, the approach employed by Illumina is

as follows: RNA is first isolated and fragmented, converted to cDNA, and tagged with a DNA-sequence known as an *adapter*. The cDNA is then hybridized to an oligonucleotide complimentary to the adapter sequence, which is attached to a glass slide called a *flow cell*. cDNA then acts as a template for DNA synthesis, using a polymerase. Every time a base is added in the DNA synthesis, one of four nucleotide-specific light frequencies are emitted and registered by camera. The sequence of an RNA-strand is determined by the order in which the various light frequencies are emitted. For transcriptional profiling, the nucleotide sequence is primarily used to determine which gene an RNA-fragment originates from. The mRNA abundance for a given gene will be proportionate to the number of sequenced cDNA-strands (*reads*) belonging to that gene.

Microarrays were introduced in the late 1990s and their primary advantage over RNA-seq, introduced in the mid 2000s, lies in cost-effectiveness[128]. However, microarrays show poor performance in measuring low gene expression levels and may show non-linear saturation effects at high gene expression levels. RNA-seq shows greater dynamic range, and the number of reads generated can be changed depending on the needs of the experiment. Microarrays can only detect gene expression levels in known genes, and results are heavily influenced by probe design choices made by the manufacturer. RNA-seq is essentially sequence-agnostic and can detect previously unknown transcripts. Finally, sequence data from RNA-seq can be used for several additional purposes, such as identifying splice variants, RNA-editing events, and somatic mutations[129]. In sum, RNA-seq is the contemporary method of choice for transcriptomic profiling of bulk samples. Gene expression microarrays, though still informative, may rapidly be approaching obsolescence.

The MPA/DMBA-induced tumors in *Study I* were profiled using Agilent gene expression microarrays. This choice was motivated by improved ease of integrating our data with previously published mammary tumor datasets analyzed using microarrays[58,101]. At the time of profiling, there was a greater price difference between RNA-seq and microarrays than there is today. Cost efficiency was therefore also a factor. If this study were repeated today, RNA-seq would instead have been used. The METABRIC cohort was profiled using Illumina microarrays, the Oslo2 cohort was profiled using Agilent microarrays, and the TCGA cohort was profiled by RNA-seq (Illumina). Gene expression profiles from TCGA may therefore be considered slightly more accurate than gene expression profiles from other cohorts used in this thesis. It is, however, unlikely that conclusions drawn from METABRIC and Oslo2 were compromised by use of microarray technology.

Classification of breast tumors into gene expression subtypes, and phenotypes, was a major focus of *Studies I* and *II*. Several classification methods exist for human breast cancer, most prominently: PAM50 (intrinsic subtypes)[44,45,53], the nine-cell line claudin-low predictor[46], and IntClust[37,130].

The intrinsic subtypes were originally identified by hierarchical clustering of gene expression values for a list of genes with significantly greater variation in expression between different breast tumors than between paired tumor samples pre- and post-chemotherapy[44]. Early subtyping methods[44,45,49] have since been superseded by the centroid-based PAM50 classifier[53]. PAM50 has proven to be robust, however certain factors might influence classification. Most significantly, composition of the cohort will influence normalization of gene expression values. Ideally, the distribution of tumor characteristics (e.g. estrogen receptor status) in the analyzed cohort should be similar to those of the PAM50 training cohort (although this issue can, to an extent, be mitigated computationally[51]). Using PAM50 on a dataset containing only triple-negative breast tumors, for example, would yield unreliable results. It is also crucial that gene expression data from tumor tissue are not centered together with data from non-tumor tissue. Intrinsic subtypes in *Studies I* and *II* were identified using PAM50, with classifications obtained directly from the cohorts' respective publications[37,51,120].

A classifier for identifying claudin-low breast tumors was developed alongside the original characterization of the tumor group[46]. First, gene expression values from 52 breast cancer cell lines were clustered. This revealed a group of nine cell lines with claudin-low features. A centroid was then generated for gene expression values from claudin-low cell lines, and a second centroid was generated based on all other cell lines. Human breast tumors could be classified as claudin-low if their gene expression values were more highly correlated to the claudin-low centroid than the other centroid. *Claudin-low* was presented as an additional intrinsic subtype, and it was therefore recommended that the subtype identified by PAM50 be overwritten in claudin-low tumors. In *Study I*, we also treated *claudin-low* as an intrinsic subtype, and the tumor group was analyzed as a single entity. In *Study II*, we hypothesized that *claudin-low* was not in fact an intrinsic subtype, but rather a phenotype which could permeate tumors from various intrinsic subtypes. Analyzing claudin-low tumors in a subtype-specific manner was the key investigation in the study.

In *Study I*, we used claudin-low classifications published with the METABRIC dataset[37,62]. 218 tumors were classified as claudin-low, but these were not identified using the nine-cell line predictor (ref. personal communication Dr. Oscar M. Rueda). Rather, the authors had derived claudin-low and non-claudin-low centroids from human breast tumors

in Prat *et al.*[46]. Tumors in METABRIC were then classified using these independently derived centroids. In *Study II*, we chose to classify tumors in METABRIC with the more commonly used nine-cell line predictor, which identified 87 claudin-low tumors. The method originally used in METABRIC is reasonable, but based on findings in *Study II*, we believe it to be somewhat promiscuous. The conclusions drawn from human claudin-low tumors in *Study I* do not change materially if the stricter classifications from *Study II* are used.

In *Study II*, we explored an alternative method for identifying claudin-low tumors. The method was based on hierarchical clustering of gene expression values for a condensed claudin-low gene list. This approach was motivated by our hypothesis that the established method[46] was overly influenced by immune and stromal infiltration. Contrasting with the nine-cell line predictor, we chose a biased approach to building the gene list, which was based on certain key studies of claudin-low tumors[12,60,63,66,67]. This experiment yielded insights into the biology of the claudin-low phenotype, and into limitations of the nine-cell line predictor. However, the biased approach may limit the external generalizability of our method. Identification of claudin-low tumors should therefore not be considered a solved matter. Our method should primarily be viewed as an additional tool in the toolkit for studying claudin-low tumors.

The IntClust subtypes were defined by identifying patterns of copy number aberration with *cis* correlation to gene expression in the METABRIC cohort[37]. Ten IntClust subtypes were defined, one of which (IntClust4) was later subdivided into an ER-positive and an ER-negative group[62,69]. Originally, the IntClust classifier used 754 features, of which 39 were copy number features and the remainder were gene expression features. In *Study II*, the IntClust subtypes in METABRIC were queried directly from the original publication[37]. IntClust subtypes in Oslo2 and TCGA were identified using a purely gene expression-based classifier[130]. There is therefore a slight difference between cohorts in the method used for IntClust classification. However, Ali *et al.*[130] report a 98% concordance between the gene expression method and the combined gene expression and copy number method. The discrepancy introduced should therefore be minimal.

Copy number analyses

In healthy tissues, there should be two copies of each chromosome in every cell, and genes should generally only be found at a single location in the genome. If DNA were isolated from such a tissue, one would find an equal amount of DNA at every genomic location. However,

tumor cells are frequently copy number aberrant. DNA isolated from a tumor would therefore show unequal amounts of DNA at genomic regions with differing copy number states. Using this principle, it is possible to generate comprehensive copy number profiles from DNA sequencing data and single nucleotide polymorphism (SNP) microarray data. The molecular principles underlying DNA-sequencing and SNP microarrays are the same as those previously described for RNA-seq and gene expression microarrays. DNA is, however, used as the substrate rather than RNA converted to cDNA. When copy number profiles are generated from sequencing data, whole genome coverage is ideal, but copy number can also be inferred from targeted data (e.g. whole exome).

When DNA-sequencing is used to profile a copy number aberrant sample, some regions of the genome will be represented by a smaller, or greater, number of reads than others. For example, if sequencing yields no reads corresponding to certain genomic region, it is likely that a deletion has occurred. If there is an increased number of reads corresponding to a genomic region (relative to others), it is likely that an amplification has occurred. When SNP microarrays are used, light intensity from oligonucleotides corresponding to known genes are used, instead of sequencing reads.

The advantages and disadvantages of the two platforms are similar to those noted in the corresponding discussion of transcriptomic analyses. SNP microarrays have lower cost than whole genome (or exome) sequencing. Fluorescence from SNP microarrays can show non-linear saturation effects at extreme copy number levels[131] and dynamic range is limited by this. Sequencing data have greater dynamic range, which is only limited by sequencing depth. Both platforms can be used to infer allele-specific copy number levels[132,133]. Whole genome sequencing can be used to infer structural rearrangements, but this is not possible from SNP array data.

In *Study I*, copy number status in MPA/DMBA-induced tumors were inferred using whole exome sequencing data. In METABRIC (*Studies I* and *II*) and TCGA (*Studies II* and *III*) copy number status were inferred from SNP microarray data. In analyses of METABRIC, copy number states were made discrete and only classified as amplification, deletion, or copy number neutral. Consequently, all amplifications were treated identically, and the dynamic range of SNP arrays was not a limiting factor. In *Study III*, copy number was analyzed as a continuous variable. An upper cap was therefore placed on copy number values at the level where non-linear saturation effects have been observed[131].

Accurately inferring copy number from raw data poses several challenges: Data are noisy, tumors are heterogeneous (due to clonality[10] and normal cell admixture), and many tumors

are aneuploid[134]. Provided technically accurate copy number estimates, there are also biological issues that must be taken into consideration. For example, if whole genome doubling has occurred, the quantity of DNA in the cell has doubled but all genes still have the same relative copy number. How does this affect transcription? Should the copy number of genes be considered relative to ploidy (i.e. four) or relative to a normal diploid genome (i.e. two)? Further, there exist algorithms which can correct for normal cell admixture in a sample, and thereby identify the most likely integer copy number state in a tumor[132,135]. However, if copy number and gene expression are used in integrated analyses, the two data types will be affected by the same normal cell infiltration. Should the more technically accurate integer copy number state be used, or should the confounding factor (normal cell admixture) remain uncorrected under the assumption that both data types are equally affected? While excellent algorithms have been developed for copy number analyses[131,132,135–137], arriving at a copy number state for a gene remains a complex task without a single *correct* solution.

The copy number profiles of MPA/DMBA-induced tumors in *Study I* were generated using EXCAVATOR2[138]. This software was specifically designed for exome sequence data. One distinguishing feature is that off-target reads are taken advantage of, which increases coverage to regions outside of the exome. Copy number states were made discrete and binned as: Homozygous deletion, heterozygous deletion, copy number neutral, single copy amplification, or multi-copy amplification. Tumor purity and ploidy could not be identified using EXCAVATOR2, and these factors were therefore not taken into consideration. In general, sampling from mouse mammary tumors allows for consistently high tumor purity. Copy number states were defined relative to a diploid genome.

In the METABRIC cohort[37,62], used in *Studies I* and *II*, copy number segments were generated from SNP microarray data using circular binary segmentation[137]. Purity and ploidy were determined using ASCAT (Allele-Specific Copy Number Analysis of Tumors)[132]. Copy number states were corrected for purity and ploidy, and binned as: amplification, deletion, or copy number neutral.

In the TCGA cohort[120], used in *Study III*, copy number segments were generated from SNP microarray data using circular binary segmentation[137]. Purity and ploidy were determined using ABSOLUTE[135]. Copy number data were primarily used for correlating copy number against gene expression. We reasoned that both data types should be similarly affected by normal cell admixture, and therefore chose not to correct copy number data for purity. When converting copy number segments to gene-level measurements, we used a method named *Ziggurat Deconstruction* implemented in GISTIC2.0[131]. This is a method for

deconstructing a segmented copy number profile into its most likely set of underlying copy number aberrations. Chromosome arm-level copy number aberrations and focal events are thereby separated from one another. Ziggurat Deconstruction can be thought of as a chromosome arm-level ploidy correction. Copy number was analyzed as a continuous variable. Genomic instability index was calculated using ploidy-corrected copy number segments.

Somatic mutation analyses

The principles behind short read DNA sequencing based on sequencing-by-synthesis – as commercialized by Illumina – are analogous to those described earlier for RNA-seq[126,127]. Generated reads range from approximately 35 to 700 base pairs in length[127]. Other sequencing methods exist, such as nanopore sequencing[139], that generate much longer reads. These are mainly used for *de novo* genome assembly (rather than for discovery of somatic variants) and are not discussed here.

Two technical considerations are of particular importance when designing a sequencing experiment: How much of the genome should be sequenced, and to what depth?

Sequencing entire genomes allows mutations to be discovered at any location. Whole genome sequencing also allows for accurate identification of copy number aberrations and structural rearrangements (as previously discussed). Whole genome sequencing may, however, be cost prohibitive. Most mutations of interest are likely to be in coding regions of the genome, and sequencing non-coding regions may introduce unnecessary complexity. An alternative is to perform targeted sequencing of a limited part of the genome. Targeted sequencing may cover the entire exome (i.e. coding regions), or a more narrowly defined panel of genes. Cost and complexity can thereby be reduced, while retaining coverage of the most important genomic regions. The ability to detect copy number aberrations is dependent on the breadth of the targeted gene panel, but is in principle possible. Structural rearrangements can, to some extent, be identified if the targeted panel is specifically designed for that purpose. The primary downside to targeted sequencing is that panels will only cover regions already thought to be of interest. The ability to identify novel regions of interest is therefore compromised.

At least one in every thousand nucleotides are read incorrectly when using contemporary short read sequencing technology[127]. A mutation must therefore be supported by multiple reads in order to be confidently called as a variant, rather than a sequencing error.

Further, tumors often consist of multiple clones, each of which may have unique mutational profiles[10,140]. Tumors are also infiltrated by non-aberrant normal cells. As a result, different cells in a tumor will carry distinct DNA sequences. A sequencing experiment should therefore generate enough reads such that the measured allele frequency of a mutation accurately depicts the cellular diversity within the tumor. The mean number of reads covering sequenced nucleotides is referred to as the *read depth*. For whole genome and whole exome sequencing studies, minimum read depths of 30 and 80, respectively, are recommended[141]. Greater read depth will improve confidence in called variants, and can enable further applications for sequencing data (particularly for analyses of intratumor heterogeneity and tumor evolution[142]). Increasing read depth, however, also increases the cost of the experiment.

In *Study I,* MPA/DMBA-induced tumors were subject to whole exome sequencing, using Illumina HiSeq technology, to a mean depth of 58. Mutations in non-coding regions were not a focus of the study, and the increased cost and complexity of whole genome sequencing was not considered worthwhile. Increased read depth could have improved the technical accuracy of our results, and enabled analyses of intratumor heterogeneity. MPA/DMBA-induced tumors showed extremely high mutational burden. For the purposes of our analyses, improved detection of low-frequency mutations might therefore mainly have increased the difficulty of interpreting results.

Tumors in the METABRIC cohort were sequenced using a custom gene panel targeting the exons of 173 cancer associated genes[62]. Samples were sequenced using Illumina HiSeq technology to a median depth of 152. In *Study I*, we only wished to perform basic, exploratory analyses of somatic mutations in the cohort, and these data were sufficient. In *Study II*, however, we wished to carry out in-depth analyses of the claudin-low tumor genome. In this regard, we were limited by the targeted gene panel. For example, an analysis of mutational signatures[31,76,78] in claudin-low tumors may have been interesting. Given the size of the targeted gene panel and low mutational burden in claudin-low tumors, this analysis would, however, likely not have been fruitful. The analyses of mutational burden would also have been improved by the greater resolution afforded by whole genome, or exome, sequencing. Explorations of tumor evolution and intratumor heterogeneity would also have been interesting, but technically challenging given the genomic stability in claudin-low tumors.

Extensive processing must be carried out in order to identify somatic mutations from raw sequence data. First, each read must be aligned to its correct genomic location. One challenge

is that certain DNA sequences are found at multiple locations in the genome. It is therefore necessary that reads are long enough so that there is low probability of ambiguous mapping (i.e. a read matches multiple locations in the reference genome). Further, reads may not perfectly match the reference genome due to SNPs, mutations, rearrangements or sequencing errors. Alignment must therefore include some tolerance for discrepancy between the reference genome and read sequences. These issues are handled by algorithms such as the Burrows-Wheeler Aligner[143]. Once reads are aligned, mutations are found by identifying differences between the reference genome and sequenced DNA. Here, the primary challenge lies in determining which differences are of interest, and which are not. In this regard, germline variants and sequencing errors complicate analyses. Germline variants can be filtered out by sequencing non-tumor samples, such as blood, from the same individual and removing variants identified in the normal sample. If normal tissue cannot be sequenced, it is also possible to filter out variants from catalogues of known SNPs[144]. Identification and removal of sequencing errors is a complex process, and extensive multi-step pipelines such as the Genome Analysis Toolkit[145] exist for that purpose. After mutations have been called, their biological meaning must be deciphered. The phenotypic effect of a mutation can only truly be determined *in vivo*. However, *in silico* analyses can predict the protein coding effect of a variant, enabling some inferences. Querying public databases of known mutations, such as COSMIC[146] may also aid in understanding functional significance. Finally, allele frequency should be taken into consideration when interpreting a mutation (e.g. a mutation found in 1% of cells should be viewed differently than a mutation found in 100% of cells).

Exome sequence data from MPA/DMBA-induced tumors were aligned using the Burrows-Wheeler Aligner[143]. Further processing prior to variant calling was carried out according to best practices recommended in the Genome Analysis Toolkit[145]. Variant calling was performed using MuTect2[147]. In addition to sequencing tumor tissue, we also sequenced liver tissue as a matched normal so that germline variants could be filtered out. In hindsight, it is interesting to consider whether liver tissue was the correct choice for this purpose. DMBA was administered orally[104,117], and we can therefore assume that it passed through the liver (first pass metabolism, and later through systemic circulation). It is therefore possible that DMBA induced a distinct set of somatic mutations in the liver, which we treated as germline variants. To the best of our knowledge, the mutagenic effect of orally administered DMBA on liver tissue has not been explored, although liver tumors can be induced by intraperitoneal administration[148,149]. Orally administered DMBA is transported systemically, and this same issue could in principle affect all perfused tissues. One solution to this issue

would be securing a matched normal sample prior to systemic DMBA exposure. Fortunately, we also filtered out known SNPs in the FVB/N mouse strain[150], thereby ameliorating this potential issue. Mutations were removed from consideration if they did not meet the following requirements: Minimum read depth of 10, minimum allele frequency of 0.05, and observed on both forward and reverse strand reads. Mutations were annotated using SnpEff[151]. In order to prioritize functionally significant variants, we focused on mutations predicted to affect protein coding sequences in known driver genes[152], and mutations frequently observed in human cancers[153].

Targeted sequence data from tumors in the METABRIC cohort were aligned with Novoalign (https://www.novocraft.com)[62]. Additional processing was carried out using the Genome Analysis Toolkit[145] and variants were called using MuTect[154]. Numerous additional filters were applied to called variants, including filtering out variants from matched normal tissue[62]. Variants were annotated using ANNOVAR[155].

Methylation analyses

DNA methylation is the addition of a methyl group to a cytosine nucleotide in positions with a downstream guanine (i.e. CpGs). Methylation can be measured by taking advantage of the distinct effect of bisulphite exposure on methylated and unmethylated cytosine[5,156]. When treated with bisulphite, unmethylated cytosine is deaminated to uracil. Methylated cytosine, however, remains unaffected. Following bisulphite exposure, uracil can be converted to thymine through a polymerase chain reaction. This process effectively encodes the methylome into the genome, and the same methodological principles as previously described for microarrays and sequencing can be applied. For microarray-based approaches, oligonucleotides must be designed to target known CpGs. Two oligonucleotides are required for each CpG: One which hybridizes to a CG-dinucleotide at the CpG site (identifying methylated cytosine) and one which hybridizes to a TG-dinucleotide at the CpG site (identifying unmethylated cytosine). Methylation is a binary feature; either a cytosine is methylated, or it is not. Methylation analysis of a completely homogeneous sample would yield binary results for each CpG. However, bulk tumor samples display intratumor heterogeneity, and as a result, it is possible to find both methylated and unmethylated copies of a CpG. Methylation is therefore represented as a proportion (methylated signal divided by the sum of all signal). This measurement is called a *beta-value*.

Whole genome bisulphite sequencing allows the entire methylome to be analyzed, and is not restricted to pre-established CpGs. In contrast, methylation arrays can only examine a subset of all CpGs in the genome. Results derived from methylation arrays will therefore be influenced by the manufacturer's probe design choices. Early microarrays from Illumina had probes for approximately 27 000 CpGs[157], whereas contemporary microarrays have probes for up to 850 000 CpGs[158]. Methylation arrays have significantly lower cost than whole genome bisulphite sequencing, and processing of raw data is less complex[5].

Samples in TCGA, used in *Study III*, were profiled using Illumina methylation microarrays[120]. The vast majority of samples were profiled using HumanMethylation450 arrays (450 000 probes), and a subset was profiled using HumanMethylation27 arrays (27 000 probes). Data from separate platforms could have been analyzed together by only using probes which were shared between arrays. However, only 22 601 probes[120] would have remained, which would have been an unacceptable loss of information for the intended analyses. Samples profiled using HumanMethylation27 arrays were therefore excluded. Only ten ovarian serous cystadenocarcinoma samples were profiled using HumanMethylation450 arrays, and the tumor type was therefore removed from further consideration.

In *Study III*, we modeled the correlation between gene expression and methylation across the entire genome, in 23 tumor types. The expression of a gene can be represented by a single numeric value (i.e. a scalar). In contrast, methylation of multiple CpGs can be of relevance to the transcription of a gene[7]. With the probes available from HumanMethylation450 arrays, some genes had in excess of a thousand CpGs within a 50 kilobase window of coding regions. A major challenge in this study was therefore reducing the dimensionality of methylation data, while retaining as much information as possible. For models to be comparable with one another, it was also important that all genes be represented by the same number of covariates.

Principal component analysis (PCA) is a method for dimensionality reduction. In brief, PCA transforms multidimensional data to a new coordinate system such that a maximum amount of variation is captured by a minimal number of dimensions[159]. For example, a breast cancer dataset could consist of measurements from five CpGs associated with proliferation and five CpGs associated with estrogen receptor signaling. We assume that proliferation-associated CpGs are correlated with one another, and estrogen receptor signaling-associated CpGs are correlated with one another. PCA could transform this data such that the majority of variation in proliferation is projected on to a single dimension (i.e. the first principal component) and the majority of variation in estrogen receptor signaling is projected on to

another dimension (i.e. the second principal component). Thus, a ten-dimensional dataset (each dimension representing a single CpG) is effectively transformed into a two-dimensional dataset (each dimension representing a compound methylation signature). Ten dimensions would still be required to capture 100% of variation in the dataset (assuming no CpGs are perfectly correlated), however progressively less variation would be captured by each successive principal component. Factors that affect how efficiently PCA is able to reduce dimensionality, in a high-dimensional dataset, include:

- To what extent do the different dimensions correlate with one another? If all dimensions in a dataset were perfectly correlated, all variation could be captured by a single principal component. If there were absolutely no correlation between dimensions, PCA would not be able to increase the variance captured by any single dimension.
- How many dimensions are in the data? All other things being equal, the first principal component of a two-dimensional dataset would capture a greater proportion of variation than the first principal component of a hundred-dimensional dataset.
- How is variation in the data distributed? There will often be some pre-existing structure to how variation is distributed among the dimensions in a dataset. If the majority of variation, from the start, is concentrated in few dimensions, few principal components will be required to capture the majority of variation. This consideration is of importance for how data are normalized: If data are scaled to unit variance (i.e. Z-score), the existing distribution of variance in the dataset will be erased.

In *Study III*, we defined CpGs as belonging to a gene if they were located in the gene body, or within 50 000 base pairs upstream or downstream of coding regions. Using PCA, we found that across 23 tumor types, the first principal component captured a median of 41% of per-gene variation[160]. The first five principal components captured a median of 78% of per-gene variation. Variance captured showed differences between tumor types. For example, variance in testicular germ cell tumors was effectively captured using relatively few principal components. This contrasts with thyroid tumors, which required a greater number of principal components to capture the same proportion of variance. Ultimately, a trade-off needed to be made when choosing the number of principal components for modeling. If too few principal components were used, the methylation status of the gene would not be comprehensively described by the data. If too many principal components were used,

relative to sample number (n = 100), models would be likely to overfit. We settled on five principle components as an acceptable compromise between these two considerations. An alternative approach to simply setting a fixed number of principal components, could have been to adjust the number of principal components to a fixed proportion of variance captured. For example, number of principal components could be set individually for each gene, and for each tumor type, so that minimum 75% of variance was captured. Using this approach, the number of principal components would differ between genes, and between tumor types. Models would therefore use differing numbers of covariates, which would make them considerably less comparable to one another. Also, when modeling expression-methylation associations, we did not observe a positive correlation between variance captured in methylation data and model fit (as determined by adjusted R^2). Differences in variance captured did therefore not appear to be a systematically confounding factor for modeling.

Statistical analyses

Statistical analyses of genome-wide data are frequently encumbered by low sample numbers in relation to number of features (i.e. genes) measured. Cancer genetic studies often have data from relatively few samples due to: The resource-intensive nature of collecting samples (from patients or animal models), long time periods required for enrolling sufficient patients, high cost of molecular analyses, and potentially due to fine-grained stratification into groups (as in *Study II*). Whole genome analyses may include measurements of over 20 000 genes, and in biological research, statistical significance is generally defined as P-value under 0.05. However, in 20 000 significance tests comparing two identical populations, 5% of tests would be expected to attain a P-value under 0.05 by chance (assuming tests are independent). This translates to 1000 false positive results. When large numbers of significance tests are performed, it is therefore important to correct for multiple hypothesis testing[161]. Correction for multiple hypothesis testing should likely have been applied more actively in *Study I*. However, very few notable differences in genomic characteristics were found between claudin-low-like and other tumors, and correction for multiple hypothesis testing would not have affected the conclusions. In *Study II*, significance tests were generally carried out in a more targeted manner, and Bonferroni-correction was applied (i.e. P-values were multiplied by the number of tests performed) where appropriate. Correlation coefficients (R^2 values) were the primary metric used in *Study III*, which was a deliberate attempt to avoid the

limitations and pitfalls of *P*-values. One such pitfall which should be mentioned is overreliance on arbitrary thresholds, such as $P < 0.05$. While such thresholds are undoubtedly practical, *P*-values should be interpreted with more nuance than simply whether they are above or below a certain numbers[162]. Where possible, statistical significance should preferably be treated a as continuous rather than binary feature, and should be considered in conjunction with effect size (including confidence intervals).

Statistical analyses are influenced by the number of samples used, and separate tests using different groups sizes might therefore not be directly comparable. For example, in *Study II*, groups of basal-like tumors had samples sizes of 45 and 263, while groups of luminal A tumors had sample sizes of 9 and 684. These are inherent characteristics of the cohort and differences in sample size were therefore unavoidable, but it must be noted that test statistics are influenced by this imbalance. In *Study III*, the number of samples from each tumor type ranged from 36 to 750, and it was essential that results from modeling were directly comparable between tumor types. We resolved this issue by downsampling each tumor type (with sufficient samples) to 100 randomly selected tumors, and performing analyses on the reduced dataset. Random selection of tumors and analysis was repeated 100 times, and median values from model statistics were used as the most accurate model estimates. As a consequence, all tumor types with less than 100 samples were removed from consideration, eliminating nine tumor types from the study.

Determining the optimal way to model gene expression as a function of methylation and copy number was a major focus in *Study III*. We explored this issue by modeling expression-methylation and expression-copy number associations using linear regression, and using non-parametric models that allow for non-linearities. Non-linear terms were implemented using splines in generalized additive models[163]. Log-transformation of data was also explored as a method for improving model fit. We used the Akaike Information Criterion (AIC) to evaluate model fit while also penalizing for complexity[164]. Our analyses revealed that the relationship between expression and copy number could be effectively modeled using linear terms, provided that copy number data were log-transformed. The relationship between expression and methylation, however, showed non-linear dynamics. Downsides to using models allowing for non-linearities include difficulty of interpretation, and potential for overfitting. We reduced the risk of overfitting by limiting the number of basis functions to four, so that a sigmoid curve was the most complex model which could be generated.

Our attempts at reducing dimensionality in methylation data revealed that the methylation profiles of most genes could not be adequately captured by a single value (see *Methylation analyses*). We therefore used the first five principal components of methylation data for modeling associations to gene expression. As a consequence, expression-copy number models and expression-methylation models use different numbers of covariates, and are therefore not directly comparable.

Ethical considerations

Open data and privacy

Samples from over 10 000 patients were analyzed throughout the studies detailed in this thesis[37,51,62,120], yet no patients were directly affected by our research activities. Our studies were enabled by the work of several major consortia, and the scientific community's ongoing movement toward open data access. Re-use of publicly available datasets allows any individual research group to produce more novel science, at a faster pace, than if it were necessary to independently generate datasets. Data re-use can save hospital and funding body resources by not replicating work carried out elsewhere. Data from similar studies can be pooled to increase sample size, or can be used to externally validate findings. Data re-use can also spare patients from unnecessary, and potentially invasive, tissue sampling.

Privacy remains a major concern for open data access. Raw DNA sequence and SNP microarray data can contain highly sensitive and immutable information regarding, for example, ancestry and disease risk factors. Protection of this sort of data may not be a major personal concern for some individuals with advanced cancer. However, privacy practices surrounding hereditary data do not only affect the profiled individual, but also all of their relatives. Tumor samples can be de-identified, but an individual's genome sequence is unique (with the exception of identical twins), and can therefore never truly be anonymized[165]. Several promising approaches are being developed for privacy-preserving analyses of sensitive data (e.g. homomorphic encryption), but these are not yet suitable for

complex exploratory analyses of genomic data[166]. Use of personally identifiable data, from human samples, has essentially been avoided throughout the studies included in this thesis. Patient privacy has therefore in no way been infringed by our research.

Animal welfare

Animal experiments in *Study I* were approved by the Norwegian Food Safety Authority, under approval numbers 4385 and 10038. All experiments were performed in accordance with local and national regulations.

When designing animal experiments, the *Three Rs* were used as guiding principles: *Replace, Reduce, Refine*. The primary goal of *Study I* was to characterize an animal model, and replacement with non-animal models was therefore, by definition, not possible. The findings of *Studies I* and *II* emphasized the importance of immune and stromal infiltration in claudin-low breast cancer. Immunocompetent models, with representative tumor microenvironment, are therefore likely required in order to accurately portray the tumor group. As a result, replacement of animal models (e.g. with cell lines or organoids) may be challenging for functional studies of claudin-low tumors. We made all relevant data from animal experiments publicly available, enabling other researchers to replace some animal trials with *in silico* analyses of our data. Tumors were profiled using genome-wide analyses, which enable exploration into innumerable biological characteristics using relatively few samples (i.e. reduction). Our findings improve the understanding of the tumor model, thereby enabling other researchers, in future, to design more well-informed studies, ideally using fewer animals. Animal welfare was optimized by following established best practice guidelines, and ongoing refinement of these. Mice were inspected daily ensuring that their welfare was adequately monitored. They were anaesthetized when subject to gavage and when MPA pellets were implanted subcutaneously. Mice were euthanized when a single tumor exceeded 1000mm^3 or total tumor volume exceeded 2000mm^3.

Reproducibility

Data generated in *Study I* were deposited in publicly available repositories in order to ensure reproducibility. Sequencing data were deposited to the European Nucleotide Archive and microarray data were deposited to ArrayExpress. No raw data were generated in *Studies II* and *III*, and deposition of data was not relevant for these studies.

All code used in *Studies II* and *III* were made available through GitHub and linked to in the respective manuscripts. These studies consisted entirely of analyses of publicly available data, and the studies can be replicated using only a computer. In retrospect, the code used in *Study I* should also have been made available and linked to upon publication. The analyses used in *Study I* are, however, described in sufficient detail that a capable bioinformatician should be able to replicate the results with relative ease.

The findings in *Study I* highlighted that MPA/DMBA-induced mouse mammary tumors show marked inter-tumor heterogeneity. This has been observed in previous studies[58,101,106,118]. Inter-tumor heterogeneity is a genuine feature of the model, which is likely attributable to the stochastically distributed mutagenic effect of DMBA. Therefore, if the experiments and analyses detailed in *Study I* were precisely repeated, in a similarly sized cohort, certain findings might differ from ours. In particular, the exact mutations and copy number aberrations would presumably only show moderate overlap to those observed in our study. We therefore made attempts to focus the conclusions of the study on broader ideas not related to the exact genetic aberrations observed. Reproducibility is a fundamental aspect of scientific discovery. However, it is also the nature of cancer that every tumor is unique, and therefore that the characteristics found in a one tumor can never be exactly replicated in another.

Open access and rapid dissemination

All contemporary science builds on preceding research. It follows that the more rapidly research becomes available, the faster other researchers can build on it. Peer review serves a vital purpose for scientific discourse, but it also significantly slows the distribution of findings. Pre-prints are one solution to accelerating scientific dissemination, while ultimately retaining the function of peer review. All studies in this thesis were posted to the pre-print server *bioRxiv* concurrently with journal submission. As a result, *Studies I and II* were made publicly available five and seven months prior to formal publication, respectively. *Study III* has not yet been accepted for formal publication.

Scientific findings are only useful insofar as they are accessible. Many academic journals require payment in order to access their articles, despite not having funded the research described in them. This places a financial burden on academic institutions, and blocks the general public (including patients and advocates) from accessing contemporary

research. *Studies I* and *II* were therefore published in open access journals. Pre-prints from all studies are freely accessible.

Discussion

Claudin-low and the molecular classification of breast carcinomas

The first two studies in this thesis focused on the etiology and classification of transcriptomic phenotypes in breast cancer. In *Study I*, we investigated mutations, copy number aberrations, and gene expression patterns in claudin-low-like tumors arising in a chemically induced mouse model of breast cancer. Certain aspects of human claudin-low breast tumors were investigated in parallel. In *Study II*, we used genomic, transcriptomic, and clinical data to critically re-evaluate the established interpretation of *claudin-low* as a gene expression subtype in human breast cancer. Together, these studies provide a shift in the understanding of a breast cancer phenotype, and insight into how it can be modeled.

This thesis was initiated with an aim of evaluating the validity of MPA/DMBA-induced mouse mammary tumors as models for human breast cancer. Using an established mouse mammary tumor classifier[101], we found that half of tumors induced by MPA/DMBA were assigned to claudin-low-like gene expression subtypes. The remaining tumors showed extensive phenotypic heterogeneity, and we therefore focused investigation on claudin-low-like tumors. The individual transcriptomic characteristics of claudin-low-like mouse mammary tumors were in line with those previously described in human claudin-low breast cancer[46,58,59], and histological features corroborated transcriptomic classifications. Thus, at a phenotypic level, claudin-low-like MPA/DMBA-induced mouse mammary tumors were representative of human counterparts.

The genomic features of human and murine claudin-low tumors were highly dissimilar. In general, mutations found in MPA/DMBA-induced tumors were rarely observed in human breast cancers. Also, MPA/DMBA-induced tumors showed high mutational burden, contrasting with the relatively low mutational burden in human breast cancers[76] and in claudin-low tumors in particular. At a genomic level, MPA/DMBA-induced tumors (claudin-low-like and others) can therefore not be considered accurate models of human breast cancer. However, a central observation in *Study I* was a disjunction between genetic and phenotypic characteristics in analyzed tumors. We speculated that cell-of-origin or microenvironmental influences might be the key factors determining whether or not a tumor became claudin-low-like. This also appears to be the case in human claudin-low breast cancer (evidenced by *Study II*). Disparate genomic landscapes in humans and murine claudin-low tumors might therefore not necessarily be a major concern. Given the discordance between genetic features in human and murine claudin-low tumors, it is difficult to argue that MPA/DMBA-induced mouse mammary tumors are *accurate* models of human disease. However, given the concordance in transcriptomic features, and possible non-genetic etiology of claudin-low phenotypes, it may nonetheless be argued that they can be *useful* models of human disease.

The considerable intertumor heterogeneity can act as a limitation for the model, as it is impossible to predict which tumor characteristics will arise in any given mouse. One solution might be to serially transplant MPA/DMBA-induced tumors for experiments, rather than chemically inducing primary tumors. This would allow for studies in which genetic drivers and phenotypic patterns can be pre-selected. We performed pilot experiments, and found that MPA/DMBA-induced tumors could be propagated in immunocompetent FVB/N mice (data not shown). Orthotopically transplanted tumors showed short generation times, growing to maximally permitted size within 2-4 weeks. Material generated in *Study I* could therefore serve as a valuable biobank of transplantable tumors, with known genomic and transcriptomic features, and intact tumor-immune response. Naturally, allogenic transplantation of MPA/DMBA-induced tumors would not be suitable for studies of tumor initiation.

In *Study II*, we showed that in human breast cancer, *claudin-low* should be considered a phenotype that permeates the intrinsic subtypes, rather than a *bona fide* intrinsic subtype. How should the findings in *Study I* be interpreted in light of the findings in *Study II*? It is important to note that the murine subtyping system proposed by Pfefferle *et al.*[101] identifies 17 distinct murine subtypes. This stands in stark contrast to the five intrinsic subtypes in human breast cancer. Further, one of the most fundamental divides between human breast

cancers lies in whether or not they express estrogen receptor. However, the significance of estrogen receptor signaling in mouse mammary tumors is unclear[109,110]. Therefore, mappings of murine subtypes to human subtypes should, in general, be interpreted with some caution. Overinterpreting such mappings could lead to murine gene expression patterns essentially being forced into a framework that does not necessarily apply to them. For example, it is possible to make inferences about whether a murine tumor is more basal-like or luminal-like (e.g. based on gene expression patterns, or immunohistochemical staining for cytokeratins or hormone receptors). It is, however, uncertain whether it is meaningful to interpret these features in the same way as they are interpreted in humans. We found that claudin-low-like mouse mammary tumors showed similar phenotypic patterns to human claudin-low breast tumors, but this does not prove that they are functionally, or etiologically, equivalent entities. It is therefore challenging to evaluate whether *claudin-low* in mice should be considered a *bona fide* subtype or a cancer phenotype. There is considerable room for improvement in the understanding of subtypes in mouse mammary tumors, but given current frameworks, we do not find reason to not treat *claudin-low* as a subtype in mice.

In *Study II*, we found that, in human breast cancer, surprisingly few tumor characteristics are in fact governed by claudin-low status. The majority of characteristics are accurately reflected in a tumor's intrinsic subtype. It is therefore reasonable to ask if future studies of general breast cancer populations need to take claudin-low status into consideration, or if it can, essentially, be disregarded. While the intrinsic subtypes do, to some extent, delineate distinct diseases[48,52,55,56], this does not seem to be the case for claudin-low status. Many investigations can presumably omit claudin-low stratification without missing information that would significantly alter conclusions of the study. Additionally, breast cancer cohorts, not filtered for tumor purity, show claudin-low prevalence of approximately 5-7%[167]. Claudin-low tumors must further be stratified by intrinsic subtype. Large cohorts, numbering in the hundreds, are therefore required in order to attain sufficient sample sizes for meaningful analyses stratified by claudin-low status. Claudin-low stratification may therefore, in practice, be difficult in many studies.

Despite the above considerations, claudin-low tumors showed several characteristics that may be of relevance for focused investigation. In particular, we observed many features that may affect efficacy of immunotherapeutic interventions. Most prominently, claudin-low tumors showed high levels of immune infiltration. The quality of immune infiltration in claudin-low tumors remains inadequately characterized, but initial findings indicate immunosuppressive features[46,59,168–170]. Claudin-low tumors showed low mutational burden,

which may limit the efficacy of immune checkpoint inhibitors[42,171]. Beyond immune-related aspects, EMT may have relevance for chemoresistance and may be a target for therapy[172].

In sum, *claudin-low* should no longer be viewed as a prognostically relevant intrinsic subtype, and may therefore not be important for the reasons originally proposed. Instead, *Study II* in conjunction with related studies[66,67,169,170], shifts the understanding of why *claudin-low* may be relevant for further investigation. While claudin-low prevalence is relatively low, identifying treatments which could improve prognosis for even five percent of breast cancer patients would represent meaningful advance.

It is interesting to speculate how breast cancer classification could be re-imagined if the approach from *Study II* was applied to all intrinsic subtypes. First, the distinction between subtypes and phenotypes, as used in this thesis, should be clarified. In a coherent classification system, subtypes must be mutually exclusive: A tumor cannot be both basal-like and luminal-like at once. In contrast, a phenotype does not necessarily imply mutual exclusivity: A tumor can have a claudin-low phenotype and an immunosuppressive phenotype at the same time (while also belonging to the basal-like subtype). Informally, different subtypes might reasonably be considered distinct diseases. Phenotypes should rather be thought of as characteristics representing heterogeneity within the same disease.

There is clearly a divide between basal-like and luminal-like breast tumors[37,52,55,56,120,173,174]. These subtypes are thought to have distinct cells-of-origin[175], and one could argue that analyzing them together is as meaningful as analyzing, e.g. pancreatic cancer and lung cancer together. They are also functionally demarcated by estrogen receptor signaling as a primary driver (although differences between basal-like and luminal-like tumors persist after estrogen receptor status is taken into account[176]). Therefore, delineating between basal-like and luminal-like tumors is a reasonable starting point for breast cancer classification.

The HER2-enriched subtype was named after the frequent overexpression of *ERBB2*/HER2 – and concomitant overexpression of *ERBB2*-associated genes – observed in the subtype[44,45]. The hypothesized association between *ERBB2*-amplification and the HER2-enriched subtype was confirmed in later studies[56,177]. Recently, Daemen & Manning[178] carried out a focused analysis of the subtype, and found that only 54% of HER2-enriched breast tumors in fact carried *ERBB2*-amplification. Of all breast tumors carrying *ERBB2*-amplification, only 47% were classified as HER2-enriched. These observations led the authors to conclude that *ERBB2*-amplification is not a subtype marker, but a "discrete event on top of a luminal or basal transcriptional and mutational state"[178]. They further

investigated the gene expression characteristics in HER2-enriched tumors, and found that despite mostly being ER-negative, several aspects of their transcriptional program (not directly related to ER-signaling) showed resemblance to that of ER-positive tumors. This appeared to be partially mediated by androgen receptor signaling, which essentially replaced ER-signaling as an analogous driver on top of a luminal transcriptional program. A cursory analysis of subtypes in TCGA[56,120], using PAM50[53,179], shows that at least 80% of HER2-enriched tumors display greater likeness to luminal subtypes than the basal-like subtype. The similarities between HER2-enriched tumors and luminal-like tumors is also supported by other studies[173,174,180]. In sum, there is evidence for the HER2-enriched transcriptional program representing a *bona fide* subtype[44,45,178,181]. However, it might also reasonably be viewed as a phenotype, predominantly observed on top of a luminal transcriptional program.

Luminal breast tumors are split into the luminal A and luminal B subtypes[44,45,55]. These tumors share putative cell-of-origin[175], and show considerable transcriptomic and genomic similarity[44,56]. The major molecular distinction between these subtypes lies in their proliferation levels. Luminal B tumors show substantially higher proliferation levels, and are associated with worse prognosis. Some other differences have also been noted, such as a gradient in estrogen receptor expression (higher in luminal A than in luminal B) and certain associations to genomic aberrations (e.g. *TP53* mutations in luminal B tumors)[55]. However, based on these relatively limited differences, it is difficult to argue that *luminal A* and *luminal B* truly represent distinct diseases, rather than heterogeneity within the same disease. Could these two subtypes instead be considered a single subtype, in which a group of tumors (luminal B) additionally show a proliferative phenotype?

Finally, claudin-low tumors and normal-like tumors must be considered. In *Study II*, we showed that *claudin-low* is not a subtype, but rather an EMT-like phenotype. Despite numerous studies of normal-like breast tumors, their etiology remains unclear. It has been proposed that *normal-like* is simply an artefact of normal-cell infiltration[44,45,52], and it is uncertain whether the subtype's characteristics are in fact derived from tumor cells. Single cell analyses might, in future, shed light on this issue[182].

Taken together, it is tempting to ask if *luminal-like* and *basal-like* might be the only true intrinsic breast cancer subtypes. The remaining characteristics (proliferation levels, HER2-signaling, androgen receptor signaling, EMT/claudin-lowness, normal cell infiltration, etc.) are undoubtedly important features, representing clinically relevant tumor biology. But do these characteristics truly demarcate distinct disease subtypes, or could they rather be viewed as phenotypes seen in addition to an underlying subtype? This notion is supported by Anderson *et al.*[173], in which the authors propose that there are only two

etiological breast cancer subtypes (ER positive vs. negative, or alternatively basal vs. luminal) based on age distribution data.

A further consideration is that the intrinsic subtypes might be forcing continuous features into categorical representations. For example, do luminal tumors show binary high or low proliferation levels – luminal A and B – or are proliferation levels in luminal tumors a continuum? For clinical decision making, firm cut-offs are necessary, and it can therefore be practical to discretely represent continuous features. However, for research purposes, treating the noted characteristics as phenotypes, on top of an underlying subtype, might allow for greater nuance as to whether tumor characteristics are categorical or continuous.

Finally, there is varying practice in how breast cancer cohorts are stratified. A clearer distinction between subtypes and phenotypes might standardize which stratification should be considered essential (i.e. the subtypes, treated as if from different organs), and which characteristics might instead be treated as heterogeneity within the same disease.

The ideas presented in this section are highly speculative, and not yet fully developed. From a clinical perspective, the intrinsic subtypes are undoubtedly of additional utility beyond a simple basal-like versus luminal-like split. However, for the purpose of conceptually understanding breast cancer, a clearer distinction between subtypes and phenotypes might be of value.

Determinants of cancer phenotypes

The common thread throughout studies in this thesis was understanding the interplay between etiological – genetic and epigenetic – factors, and associated gene expression patterns. In *Study I*, we were surprised to find a lack of concordance between mutations, copy number aberrations and gene expression subtypes in MPA/DMBA-induced tumors. Certain subtler associations would likely have emerged provided greater sample numbers, but there was no indication of pathognomonic genetic aberrations leading to observed phenotypes. In *Study II*, we observed previously identified associations between intrinsic breast cancer subtypes and certain genetic aberrations, e.g. high frequency of *TP53* mutations in basal-like tumors and high frequency of *PIK3CA* mutations in luminal A tumors[56]. Interestingly though, claudin-low tumors were not associated with any specific genetic aberrations, but were rather characterized by an absence of them. The associations between genetic aberrations and gene expression subtypes in breast cancer[37,44,45,56,58,101] highlight the importance of genotypes in determining cancer phenotypes. However, our

investigations of *claudin-low* also highlighted the importance of features not encoded in a tumor's DNA sequence, such as cell-of-origin or tumor microenvironment.

In *Studies I* and *II*, we correlated individual categorical DNA features (e.g. wildtype vs. mutant) to compound gene expression signatures. Throughout the investigations, it became apparent that this approach could not sufficiently explain cancer phenotypes. This motivated *Study III*, in which we attempted to more precisely enumerate the effects of genetic and epigenetic factors on gene expression in cancer.

Copy number, methylation, gene expression and whole exome sequence data were available from The Cancer Genome Atlas[120]. We were able to represent methylation, copy number and gene expression in numeric forms suitable for statistical modeling. Inclusion of DNA sequences in models was, however, avoided due to the categorical nature of data, and the vast number of features (i.e. nucleotides) which would be required to represent genes. We also chose to avoid investigating *trans* associations, in order to reduce the risk of identifying spurious correlations. Finally, there are countless etiological factors which could not be fully captured by sampling of bulk tumors (e.g. effects of paracrine and endocrine signaling), or from the analytical methods employed (e.g. three-dimensional structure of DNA and chromatin, chromosomal architecture, epigenetic modifications beyond DNA methylation). Taken together, we believe our analyses effectively used the data available in order to answer the research questions at hand. However, there are myriad biological aspects, relevant to transcriptional regulation, which could not be incorporated in our analyses. It is therefore reasonable that only a limited proportion of transcriptional variation in cancer could be accurately captured by our models (which were solely based on methylation and/or copy number).

Considerable effort was made to identify the optimal methods for modeling expression-methylation and expression-copy number associations. However, some decisions could be viewed as arbitrary, such as the number of methylation signatures used for modeling. While we believe our modeling parameters to be optimal for pan-cancer, genome-wide analyses, individual genes may be more appropriately modeled with different parameters. One major motivation for building the web application associated with *Study III* was, therefore, to avoid insinuation that our parameters were the *correct* ones. Instead, we wished to enable researchers to model E-M/E-C associations with parameters of their own choosing, modified to suit their individual study. We envision that use of our tool should act as an initial exploratory analysis in studies investigating genes transcriptionally dysregulated in cancer. Previously, such analyses could require researchers to download over 40 gigabytes of data (if using TCGA[120] – which would likely need to be analyzed on server-

grade computer hardware) and perform comprehensive pre-processing of data. This might culminate in modeling using suboptimal methods (e.g. linear regression used to model non-linear associations). Using our tool, E-M/E-C associations can be explored in a matter of minutes, within individual tumor types or pan-cancer, and with no bioinformatical knowledge required.

While DNA methylation and copy number are known to affect gene expression, causality cannot necessarily be inferred from E-M/E-C associations. For the most part, copy number aberrations are independent of one another. This increases the likelihood of a copy number aberration being causally implicated if it is correlated with gene expression. Chromosomally adjacent genes are an exception to this, due to co-amplification/co-deletion (i.e. copy number in a gene can be non-causally associated with gene expression in neighboring genes). While it cannot be formally proven, it is therefore reasonable to assume that copy number is causally implicated in most *cis* E-C associations. In contrast, cells gain coordinated methylation patterns as they progress along the differentiation hierarchy. As methylation patterns in different genes are not independent, there is a substantial probability of finding non-causal E-M associations. Testicular germ cell tumors are an instructive example. The tumor type is histologically subdivided into seminomas, which are undifferentiated and globally hypomethylated, and non-seminomas, which are more differentiated and have higher degree of methylation[183]. Therefore, a gene which is differentially expressed in seminomas and non-seminomas could show significant E-M association using methylation data from almost any chromosomal region (due to the global, coordinated, nature of differences in methylation between the two histological subtypes). Only considering E-M associations in *cis* reduces the likelihood of identifying non-causal relationships, but there is still a considerable possibility of identifying correlations without causation.

Despite limitations to our approach, several interesting trends emerged in the patterns of E-M/E-C associations. In particular, the correlations between E-M/E-C associations and variance in expression, methylation and copy number (Supplementary Figures 4, 5 and 6 in *Study III*) warrant additional discussion. At the level of individual genes, there was positive correlation between the frequency of copy number aberration in a gene (i.e. variance in copy number data) and the strength of E-C association for that gene. It is unsurprising that frequently copy number aberrant genes have stronger E-C associations than genes in which copy number aberrations are rarely found. More surprising was the relative weakness of this association: Per-gene variance in copy number data could only account for approximately ten percent of strength in E-C association, and many genes had high copy number variance

but low E-C association. This indicates that a considerable proportion of copy number aberrations might be functionally insignificant. This could be related to the mutual exclusivity in high E-M association and high E-C association: Genes with high E-M association are epigenetically silenced, leading to low E-C association in those genes irrespective of frequency of copy number aberrations. However, when the correlation between copy number variance and E-C association was stratified by E-M association, the trend persisted, indicating a limited ability for methylation to negate transcriptional effects of copy number aberrations (a conclusion also reached by Sun *et al.*[7]). One outstanding question in this regard, is why some copy number aberrations affect transcription while others don't. Whole genome sequencing, which could elucidate the genomic architecture of copy number aberrations, may shed light on the mechanisms behind copy number aberrations without associated gene expression changes. Functional analyses will also be necessary.

For methylation, there was also a positive correlation between variance in MethSigs and E-M association, at a per-gene level. Also here, it is unsurprising that genes showing more extensive differential methylation would show greater E-M association. MethSig variance could, however, only account for approximately four percent of E-M association strength, and there was an abundance of genes with high variance in methylation without associated gene expression changes. This indicates that the bulk of variation in methylation, in cancer, might be functionally insignificant, and that the quality of variation in methylation (e.g. location relative to the gene, interaction with histones) is more important than the quantity. Understanding mechanisms determining the transcriptional effects of methylation, at a genome-wide scale, will be complex and will require analyses stratified by CpG location relative to coding regions and transcription start sites, integration with other omics technologies (e.g. ChIP-Seq) and functional analyses.

The following relationships relevant to copy number were positively, and significantly correlated: Mean E-C association in a tumor type versus mean copy number variance, mean E-C association in a tumor type versus mean gene expression variance, and mean copy number variance versus mean gene expression variance. In conjunction with the gene-level findings discussed above, these correlations are strongly suggestive of copy number being a major driver of phenotypic heterogeneity within tumor types. The IntClust subtypes in breast cancer are an excellent example in support of this notion[37,69,184]. Analogous relationships for E-M associations were however more convoluted. There was a positive and significant correlation between mean MethSig variance in a tumor type and mean E-M association. However, this was entirely driven by one tumor type, and the relationship

disappears if testicular germ cell tumors are treated as an outlier tumor type. There was no correlation between mean E-M association and mean gene expression variance in a tumor type. If aberrant methylation is a driver of gross transcriptional dysregulation in cancer, it seems counter-intuitive that these relationships should not show positive correlations (as is the case for copy number). There was, however, a positive correlation between mean MethSig variance in a tumor type and mean gene expression variance. In isolation, this trend is indicative of aberrant methylation shaping cancer transcriptomes at a genome-wide scale. Viewing this trend together with the two aforementioned correlations, however, raises questions about causality. There may be underlying processes in cancer which act as common drivers of variation in gene expression, methylation and copy number. For example, copy number aberrations and structural rearrangements are associated with aberrant methylation[7,185], and hypomethylation may induce chromosomal instability[186] (although it is unlikely that the mentioned associations fully explain the observed correlation between gene expression variance and methylation variance in tumor types). Other findings relevant to the role of methylation and copy number in shaping tumor transcriptomes included the mutual exclusivity in high levels of E-M and E-C association, and the distinct functions enriched for in methylation-associated genes and copy number-associated genes.

Taken together, our findings indicated that the majority of E-M associations in cancer might primarily be a reflection of normal-cell infiltration and cell-of-origin, rather than a reflection of oncogenic aberration. This would imply that at a genome-wide scale, methylation may not be a major, direct driver of cancer-specific transcriptional heterogeneity within tumor types. Other have posited that methylation is mostly a maintainer of transcriptional states, rather than an inducer[6,7]; this notion seems congruent with our observations. Aberrant methylation can certainly have an oncogenic effect, mediated by epigenetic aberrations affecting specific genes (e.g. *MLH1*, *BRCA1*) and through effects related to genomic instability[186,187]. Aberrant methylation does, however, not seem to be a direct mechanism causing genome-wide, gross transcriptional dysregulation, as is the case for copy number aberrations. The direct oncogenic role of methylation may therefore be conceptually more analogous to mutations in in individual genes. These inferences from *Study III* are thought-provoking, and by no means conclusive. Functional experiments will be required in order to more clearly understand causality in expression-methylation associations. Single cell analyses of methylation will also be imperative to understanding the role of normal-cell admixture in E-M associations, and for accurately delineating normal and aberrant methylation states in cancer.

The three studies included in this thesis illustrate the enormous complexity of transcriptional dysregulation in cancer. Certain genetic aberrations are pathognomonic for specific forms of cancer (i.e. *BCR-ABL* translocation in chronic myelogenous leukaemia[188]). It is, however, increasingly clear that in most cases, cancer phenotypes (gene-level or compound signatures) cannot simply be explained by individual genetic or epigenetic features. We identify four key methodological issues which must be addressed in order to gain a more comprehensive understanding of the genetic determinants of cancer:

1. The observational nature of most genome-wide datasets, precluding a real understanding of causality
2. An incomplete depiction of the genome and epigenome (e.g. chromosomal architecture, epigenetic features beyond methylation)
3. Heterogeneity within bulk tumor samples, which acts as a confounding factor throughout all analyses
4. Limited sample sizes and vast numbers of features, precluding more complex modeling (e.g. machine learning)

Future perspectives

The research detailed in this thesis has highlighted numerous exciting avenues for future investigation:

- What are the mutational landscapes of non-mammary tissues in mice systemically exposed to DMBA? Several studies, in recent years, have revealed surprisingly frequent oncogenic mutations in histologically non-cancerous tissue[13–15]. Might this also be the case in mice systemically exposed to DMBA, and could this be used to study different tissues' predisposition to oncogenic transformation?
- What are the characteristics of immune cell infiltration in claudin-low breast tumors? Could immune cell profiling improve understanding of immunotherapeutic possibilities in claudin-low breast cancer?
- Could immune checkpoint inhibitors and cyclooxygenase inhibitors be used, alone or in combination, to treat claudin-low breast cancer? Serially transplanted MPA/DMBA-induced mouse mammary tumors could be an appropriate model for such a study.
- What are the characteristics of claudin-low and normal-like breast tumors at a single cell level? To what extent are their features a reflection of non-tumor cell infiltration?
- Following the thought process from *Study II*, could the entire breast cancer classification system be re-evaluated? How would breast cancer subtypes, and

phenotypes, be defined if the entire system was re-built from first principles, incorporating the past twenty years of breast cancer research?

- How do the transcriptional effects of copy number aberrations vary depending on whether they are broad (e.g. whole arm amplifications) or focal?
- Which other data could be used to extend the analyses from *Study III*? Are expression-methylation associations materially different if data from whole genome bisulphite sequencing are used? Could the newly released Pan-Cancer Analysis of Whole Genomes[18] dataset be used to extend analyses, e.g. by incorporating structural rearrangements? Could mutations or SNPs (eQTLs) be incorporated? Could data from multiple studies be pooled to increase sample size?
- How could genetic, epigenetic, and transcriptomic characteristics of tumors be altered and measured over multiple timepoints, in order to improve the understanding of causality in E-M/E-C associations?

Concluding remarks

In this thesis, we have investigated genetic and epigenetic determinants of cancer phenotypes. We carried out comprehensive analyses in two species, and in over twenty types of cancer. Phenotypes were investigated at the level of individual genes, compound molecular signatures, and clinical outcomes. Our studies necessitated the development of novel methods, resulting in two computational tools that are readily available to other researchers. We generated novel datasets, also publicly available, and a biobank of pre-characterized and serially transplantable mouse mammary tumors. Our studies yielded numerous interesting results. We observed a lack of concordance between genetic and transcriptomic features in a murine model of breast cancer, and identified immunosuppression as a potentially actionable characteristic in a subgroup of tumors. We found that the conceptual understanding of an established breast cancer subtype was oversimplified, and that previous characterizations of the tumor group were undermined by inadequate stratification. Finally, we discovered several curious pan-cancer trends in expression-methylation and expression-copy number associations, thereby raising questions around how DNA methylation and copy number shape transcriptomic heterogeneity within tumor types. Our findings open multiple paths for future study, and may correct the trajectory of certain pre-existing lines of inquiry.

References

1. Alberts, B. *et al. Molecular biology of the cell*. (Garland Science, 2015).
2. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561 (1970).
3. Myhre, S. *et al.* Influence of DNA copy number and mRNA levels on the expression of breast cancer related proteins. *Mol. Oncol.* **7**, 704–718 (2013).
4. Gonçalves, E. *et al.* Widespread post-transcriptional attenuation of genomic copy-number variation in cancer. *Cell Syst.* **5**, 386–398 (2017).
5. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.* **11**, 191–203 (2010).
6. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
7. Sun, W. *et al.* The association between copy number aberration, DNA methylation and gene expression in tumor samples. *Nucleic Acids Res.* **46**, 3009–3018 (2018).
8. Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650–665 (2018).
9. Pollack, J. R. *et al.* Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci.* **99**, 12963–8 (2002).
10. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168**, 613–628 (2017).
11. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153 (2007).
12. Puisieux, A., Pommier, R. M., Morel, A.-P. & Lavial, F. Cellular pliancy and the multistep process of tumorigenesis. *Cancer Cell* **33**, 164–172 (2018).
13. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (80-.).* **348**, 880–886 (2015).
14. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science (80-.).* **362**, 911–917 (2018).
15. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
16. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).

17. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
18. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
19. Global Burden of Disease Cancer Collaboration. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol.* **5**, 1749–1768 (2019).
20. Welch, H. G., Kramer, B. S. & Black, W. C. Epidemiologic signatures in cancer. *N. Engl. J. Med.* **381**, 1378–1386 (2019).
21. Berry, D. A. *et al.* Effect of screening and adjuvant therapy on mortality from breast cancer. *N. Engl. J. Med.* **353**, 1784–1792 (2005).
22. Welch, H. G. & Black, W. C. Overdiagnosis in Cancer. *J. Natl. Cancer Inst.* **102**, 605–613 (2010).
23. Welch, H. G., Prorok, P. C., O'Malley, A. J. & Kramer, B. S. Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. *N. Engl. J. Med.* **375**, 1438–1447 (2016).
24. Esserman, L. J. *et al.* Addressing overdiagnosis and overtreatment in cancer: a prescription for change. *Lancet Oncol.* **15**, e234–e242 (2014).
25. Trichopoulos, D., Adami, H., Ekbom, A., Hsieh, C. & Lagiou, P. Early life events and conditions and breast cancer risk: from epidemiology to etiology. *Int. J. Cancer* **122**, 481–485 (2008).
26. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).
27. Baker, S. G., Lichtenstein, P., Kaprio, J. & Holm, N. Genetic susceptibility to prostate, breast, and colorectal cancer among Nordic twins. *Biometrics* **61**, 55–63 (2005).
28. Mavaddat, N., Antoniou, A. C., Easton, D. F. & Garcia-Closas, M. Genetic susceptibility to breast cancer. *Mol. Oncol.* **4**, 174–191 (2010).
29. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
30. Rojas, K. & Stuckey, A. Breast cancer epidemiology and risk factors. *Clin. Obstet. Gynecol.* **59**, 651–672 (2016).
31. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
32. Hassiotou, F. & Geddes, D. Anatomy of the human mammary gland: Current status of knowledge. *Clin. Anat.* **26**, 29–48 (2013).
33. Bombonati, A. & Sgroi, D. C. The molecular pathology of breast cancer progression. *J. Pathol.* **223**, 308–318 (2011).
34. 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).
35. Helsedirektoratet. *Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft*. (2019).
36. Jatoi, I., Chen, B. E., Anderson, W. F. & Rosenberg, P. S. Breast cancer mortality trends in the United States according to estrogen receptor status and age at diagnosis. *J. Clin. Oncol.* **25**, 1683–1690 (2007).
37. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346 (2012).

38. Johnston, S. R. D. New strategies in estrogen receptor–positive breast cancer. *Clin. Cancer Res.* **16**, 1979–1987 (2010).
39. Hudis, C. A. Trastuzumab—mechanism of action and use in clinical practice. *N. Engl. J. Med.* **357**, 39–51 (2007).
40. Kummar, S. *et al.* Advances in using PARP inhibitors to treat cancer. *BMC Med.* **10**, 25 (2012).
41. Polk, A., Svane, I.-M., Andersson, M. & Nielsen, D. Checkpoint inhibitors in breast cancer–current status. *Cancer Treat. Rev.* **63**, 122–134 (2018).
42. Yarchoan, M., Hopkins, A. & Jaffee, E. M. Tumor mutational burden and response rate to PD-1 inhibition. *N. Engl. J. Med.* **377**, 2500–2501 (2017).
43. Schmid, P. *et al.* Atezolizumab and nab-paclitaxel in advanced triple-negative breast cancer. *N. Engl. J. Med.* **379**, 2108–2121 (2018).
44. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
45. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* **98**, 10869–10874 (2001).
46. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **12**, R68 (2010).
47. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (80-.).* **270**, 467–470 (1995).
48. Perou, C. M. & Børresen-Dale, A.-L. Systems biology and genomics of breast cancer. *Cold Spring Harb. Perspect. Biol.* **3**, a003293 (2011).
49. Sørlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci.* **100**, 8418–8423 (2003).
50. Naume, B. *et al.* Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer. *Mol. Oncol.* **1**, 160–171 (2007).
51. 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).
52. Sørlie, T. Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. *Eur. J. Cancer* **40**, 2667–2675 (2004).
53. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160 (2009).
54. 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.* **19**, 120 (2017).
55. Norum, J. H., Andersen, K. & Sørlie, T. Lessons learned from the intrinsic subtypes of breast cancer in the quest for precision therapy. *Br. J. Surg.* **101**, 925–938 (2014).
56. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61 (2012).
57. Wallden, B. *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genomics* **8**, 54 (2015).
58. 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).
59. Sabatier, R. *et al.* Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization. *Mol. Cancer* **13**, 228 (2014).

60. Dias, K. *et al.* Claudin-low breast cancer; clinical & pathological characteristics. *PLoS One* **12**, e0168669 (2017).
61. Williams, E. D., Gao, D., Redfern, A. & Thompson, E. W. Controversies around epithelial–mesenchymal plasticity in cancer metastasis. *Nat. Rev. Cancer* 1–17 (2019).
62. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).
63. Bruna, A. *et al.* TGFβ induces the formation of tumour-initiating cells in claudin low breast cancer. *Nat. Commun.* **3**, 1055 (2012).
64. Damrauer, J. S. *et al.* Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. *Proc. Natl. Acad. Sci.* **111**, 3110–3115 (2014).
65. Kardos, J. *et al.* Claudin-low bladder tumors are immune infiltrated and actively immune suppressed. *JCI Insight* **1**, e85902 (2016).
66. Morel, A.-P. *et al.* EMT inducers catalyze malignant transformation of mammary epithelial cells and drive tumorigenesis towards claudin-low tumors in transgenic mice. *PLoS Genet.* **8**, e1002723 (2012).
67. Morel, A. P. *et al.* A stemness-related ZEB1-MSRB3 axis governs cellular pliancy and breast cancer genome stability. *Nat. Med.* **23**, 568–578 (2017).
68. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).
69. Russnes, H. G., Lingjaerde, 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).
70. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
71. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
72. Ciriello, G. *et al.* Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–519 (2015).
73. Staaf, J. *et al.* Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* **25**, 1526–1533 (2019).
74. Bertucci, F. *et al.* Genomic characterization of metastatic breast cancers. *Nature* **569**, 560–564 (2019).
75. Angus, L. *et al.* The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. *Nat. Genet.* **51**, 1450–1458 (2019).
76. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
77. Jiao, W. *et al.* A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* **11**, 728 (2020).
78. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149**, 979–993 (2012).
79. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).
80. Russnes, H. G. *et al.* Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci. Transl. Med.* **2**, 38ra47-38ra47 (2010).
81. Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16**, 1465–1479 (2006).

82. Pladsen, A. V *et al.* DNA copy number motifs are strong and independent predictors of survival in breast cancer. *Commun. Biol.* **3**, 153 (2020).
83. Dawson, S., Rueda, O. M., Aparicio, S. & Caldas, C. A new genome-driven integrated classification of breast cancer and its implications. *EMBO J.* **32**, 617–628 (2013).
84. Fleischer, T. *et al.* Genome-wide DNA methylation profiles in progression to. *Genome Biol.* **15**, 435 (2014).
85. Bediaga, N. G. *et al.* DNA methylation epigenotypes in breast cancer molecular subtypes. *Breast Cancer Res.* **12**, R77 (2010).
86. Holm, K. *et al.* Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Res.* **12**, R36 (2010).
87. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520 (2002).
88. Perlman, R. L. Mouse models of human disease: An evolutionary perspective. *Evol. Med. Public Heal.* **2016**, 170–176 (2016).
89. Mestas, J. & Hughes, C. C. W. Of mice and not men: differences between mouse and human immunology. *J. Immunol.* **172**, 2731–2738 (2004).
90. Davis, M. M. A prescription for human immunology. *Immunity* **29**, 835–838 (2008).
91. Jaenisch, R. & Mintz, B. Simian virus 40 DNA sequences in DNA of healthy adult mice derived from preimplantation blastocysts injected with viral DNA. *Proc. Natl. Acad. Sci.* **71**, 1250–1254 (1974).
92. Kersten, K., de Visser, K. E., van Miltenburg, M. H. & Jonkers, J. Genetically engineered mouse models in oncology research and cancer medicine. *EMBO Mol. Med.* **9**, 137–153 (2017).
93. Rennhack, J. P. *et al.* Integrated analyses of murine breast cancer models reveal critical parallels with human disease. *Nat. Commun.* **10**, 3261 (2019).
94. Westcott, P. M. K. *et al.* The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. *Nature* **517**, 489–492 (2015).
95. 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.* **21**, 946 (2015).
96. McFadden, D. G. *et al.* Mutational landscape of EGFR-, MYC-, and Kras-driven genetically engineered mouse models of lung adenocarcinoma. *Proc. Natl. Acad. Sci.* **113**, E6409–E6417 (2016).
97. Francis, J. C. *et al.* Whole-exome DNA sequence analysis of *Brca2* - and *Trp53* - deficient mouse mammary gland tumours. *J. Pathol.* **236**, 186–200 (2015).
98. Pfefferle, A. D. *et al.* Genomic profiling of murine mammary tumors identifies potential personalized drug targets for p53-deficient mammary cancers. *Dis. Model. Mech.* **9**, (2016).
99. Hollern, D. P. & Andrechek, E. R. A genomic analysis of mouse models of breast cancer reveals molecular features of mouse models and relationships to human breast cancer. *Breast Cancer Res.* **16**, R59 (2014).
100. Hollern, D. P., Swiatnicki, M. R. & Andrechek, E. R. Histological subtypes of mouse mammary tumors reveal conserved relationships to human cancers. *PLoS Genet.* **14**, e1007135 (2018).
101. Pfefferle, A. D. *et al.* Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. *Genome Biol.* **14**, R125 (2013).

102. Medina, D. Mammary Tumorigenesis in Chemical Carcinogen-Treated Mice. I. Incidence in BALB/c and C57BL Mice2. *J. Natl. Cancer Inst.* **53**, 213–221 (1974).
103. Medina, D., Butel, J. S., Socher, S. H. & Miller, F. L. Mammary tumorigenesis in 7, 12-dimethybenzanthracene-treated C57BL× DBA/2f F1 mice. *Cancer Res.* **40**, 368–373 (1980).
104. 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).
105. McCreery, M. Q. *et al.* Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. *Nat. Med.* **21**, 1514 (2015).
106. Abba, M. C. *et al.* DMBA induced mouse mammary tumors display high incidence of activating Pik3caH1047 and loss of function Pten mutations. *Oncotarget* **5**, (2016).
107. Byrne, A. T. *et al.* Interrogating open issues in cancer precision medicine with patient-derived xenografts. *Nat. Rev. Cancer* **17**, 254 (2017).
108. 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).
109. Cornelissen, L. M. *et al.* Exogenous ERα Expression in the Mammary Epithelium Decreases Over Time and Does Not Contribute to p53-Deficient Mammary Tumor Formation in Mice. *J. Mammary Gland Biol. Neoplasia* **24**, 305–321 (2019).
110. 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).
111. Prat, A. & Perou, C. M. Mammary development meets cancer genomics. *Nat. Med.* **15**, 842 (2009).
112. Baird, W. M., Hooven, L. A. & Mahadevan, B. Carcinogenic polycyclic aromatic hydrocarbon-DNA adducts and mechanism of action. *Environ. Mol. Mutagen.* **45**, 106–114 (2005).
113. Phillips, D. H. Polycyclic aromatic hydrocarbons in the diet. *Mutat. Res. Toxicol. Environ. Mutagen.* **443**, 139–147 (1999).
114. Miyata, M., Furukawa, M., Takahashi, K., Gonzalez, F. J. & Yamazoe, Y. Mechanism of 7, 12-dimethylbenz [a] anthracene-induced immunotoxicity: role of metabolic activation at the target organ. *Jpn. J. Pharmacol.* **86**, 302–309 (2001).
115. Bláha, L., Kapplová, P., Vondráček, J., Upham, B. & Machala, M. Inhibition of gap-junctional intercellular communication by environmentally occurring polycyclic aromatic hydrocarbons. *Toxicol. Sci.* **65**, 43–51 (2002).
116. Huggins, C., Grand, L. C. & Brillantes, F. P. Mammary cancer induced by a single feeding of polynuclear hydrocarbons and its suppression. *Nature* **189**, 204–207 (1961).
117. Gonzalez-Suarez, E. *et al.* RANK ligand mediates progestin-induced mammary epithelial proliferation and carcinogenesis. *Nature* **468**, 103 (2010).
118. Yin, Y. *et al.* Characterization of medroxyprogesterone and DMBA-induced multilineage mammary tumors by gene expression profiling. *Mol. Carcinog.* **44**, 42–50 (2005).
119. Weigelt, B. *et al.* Metaplastic breast carcinomas display genomic and transcriptomic heterogeneity. *Mod. Pathol.* **28**, 340 (2015).

120. Hoadley, K. A. *et al.* Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304 (2018).
121. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
122. Consortium, G. O. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
123. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
124. Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
125. Mukherjee, A. *et al.* Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort. *NPJ Breast Cancer* **4**, 5 (2018).
126. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135 (2008).
127. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
128. Mantione, K. J. *et al.* Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Med. Sci. Monit. Basic Res.* **20**, 138 (2014).
129. Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science (80-.).* **364**, eaaw0726 (2019).
130. Ali, H. R. *et al.* Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol.* **15**, 431 (2014).
131. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
132. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107**, 16910–5 (2010).
133. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131–e131 (2016).
134. Taylor, A. M. *et al.* Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* **33**, 676–689 (2018).
135. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413 (2012).
136. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
137. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
138. D'Aurizio, R. *et al.* Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. *Nucleic Acids Res.* **44**, e154–e154 (2016).
139. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153 (2008).
140. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).

141. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).
142. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
143. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
144. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
145. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–303 (2010).
146. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).
147. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *BioRxiv* 861054 (2019).
148. Manam, S. *et al.* Activation of the Ha-, Ki-, and N-ras genes in chemically induced liver tumors from CD-1 mice. *Cancer Res.* **52**, 3347–3352 (1992).
149. Buchmann, A., Karcier, Z., Schmid, B., Strathmann, J. & Schwarz, M. Differential selection for B-raf and Ha-ras mutated liver tumors in mice with high and low susceptibility to hepatocarcinogenesis. *Mutat. Res. Mol. Mech. Mutagen.* **638**, 66–74 (2008).
150. Wong, K. *et al.* Sequencing and characterization of the FVB/NJ mouse genome. *Genome Biol.* **13**, 1–12 (2012).
151. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* **6**, 80–92 (2012).
152. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* 1 (2018).
153. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2016).
154. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213 (2013).
155. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).
156. Yong, W.-S., Hsu, F.-M. & Chen, P.-Y. Profiling genome-wide DNA methylation. *Epigenetics Chromatin* **9**, 26 (2016).
157. Bibikova, M. *et al.* Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics* **1**, 177–200 (2009).
158. Mansell, G. *et al.* Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC Genomics* **20**, 366 (2019).
159. Jolliffe, I. *Principal component analysis.* (Springer, 2011).
160. Fougner, C. *et al.* A pan-cancer atlas of transcriptional dependence on DNA methylation and copy number aberrations. *bioRxiv* 2020.05.04.076901 (2020).
161. Goeman, J. J. & Solari, A. Multiple hypothesis testing in genomics. *Stat. Med.* **33**, 1946–1978 (2014).
162. Amrhein, V., Greenland, S. & McShane, B. Scientists rise up against statistical significance. *Nature* **567**, 305–307 (2019).
163. Hastie, T. & Tibshirani, R. Generalized Additive Models. *Stat. Sci.* 297–310 (1986).

164. Burnham, K. P. & Anderson, D. R. Practical use of the information-theoretic approach. in *Model selection and inference* 75–117 (Springer, 1998).
165. Erlich, Y., Shor, T., Pe'er, I. & Carmi, S. Identity inference of genomic data using long-range familial searches. *Science (80-.).* **362**, 690–694 (2018).
166. Grishin, D., Obbad, K. & Church, G. M. Data privacy in the age of personal genomics. *Nat. Biotechnol.* **37**, 1115–1117 (2019).
167. Fougner, C., Bergholtz, H., Norum, J. H. & Sørlie, T. Re-definition of claudin-low as a breast cancer phenotype. *Nat. Commun.* **11**, 1787 (2020).
168. Fougner, C., Bergholtz, H., Kuiper, R., Norum, J. H. & Sørlie, T. Claudin-low-like mouse mammary tumors show distinct transcriptomic patterns uncoupled from genomic drivers. *Breast Cancer Res.* **21**, 85 (2019).
169. Alsuliman, A. *et al.* Bidirectional crosstalk between PD-L1 expression and epithelial to mesenchymal transition: significance in claudin-low breast cancer cells. *Mol. Cancer* **14**, 149 (2015).
170. Taylor, N. A. *et al.* Treg depletion potentiates checkpoint inhibition in claudin-low breast cancer. *J. Clin. Invest.* **127**, 3472–3483 (2017).
171. Cristescu, R. *et al.* Pan-tumor genomic biomarkers for PD-1 checkpoint blockade–based immunotherapy. *Science (80-.).* **362**, eaar3593 (2018).
172. Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
173. Anderson, W. F., Rosenberg, P. S., Prat, A., Perou, C. M. & Sherman, M. E. How many etiological subtypes of breast cancer: two, three, four, or more? *J. Natl. Cancer Inst.* **106**, dju165 (2014).
174. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
175. Nguyen, Q. H. *et al.* Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* **9**, 2028 (2018).
176. Bertucci, F., Finetti, P., Goncalves, A. & Birnbaum, D. The therapeutic response of ER+/HER2− breast cancers differs according to the molecular Basal or Luminal subtype. *NPJ Breast Cancer* **6**, 8 (2020).
177. 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, Chromosom. Cancer* **45**, 1033–1040 (2006).
178. Daemen, A. & Manning, G. HER2 is not a cancer subtype but rather a pan-cancer event and is highly enriched in AR-driven breast tumors. *Breast Cancer Res.* **20**, 8 (2018).
179. Gendoo, D. M. A. *et al.* Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics* **32**, 1097–1099 (2015).
180. Prat, A. *et al.* Genomic analyses across six cancer types identify basal-like breast cancer as a unique molecular entity. *Sci. Rep.* **3**, 3544 (2013).
181. Prat, A. *et al.* Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast* **24**, S26–S35 (2015).
182. Ali, H. R. *et al.* Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat. Cancer* **1**, 163–175 (2020).
183. Shen, H. *et al.* Integrated molecular characterization of testicular germ cell tumors. *Cell Rep.* **23**, 3392–3406 (2018).

184. Rueda, O. M. *et al.* Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature* **567**, 399–404 (2019).
185. Zhang, Y. *et al.* Global impact of somatic structural variation on the DNA methylome of human cancers. *Genome Biol.* **20**, 209 (2019).
186. Esteller, M. Epigenetics in Cancer. *N. Engl. J. Med.* **358**, 1148–1159 (2008).
187. Rodríguez-Paredes, M. & Esteller, M. Cancer epigenetics reaches mainstream oncology. *Nat. Med.* **17**, 330 (2011).
188. Ren, R. Mechanisms of BCR–ABL in the pathogenesis of chronic myelogenous leukaemia. *Nat. Rev. Cancer* **5**, 172–183 (2005).

Study I

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 Sørlie.
Breast Cancer Research 21, 85 (2019).

Fougner et al. Breast Cancer Research (2019) 21:85
https://doi.org/10.1186/s13058-019-1170-8

Breast Cancer Research

RESEARCH ARTICLE Open Access

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

Christian Fougner[1], Helga Bergholtz[1], Raoul Kuiper[2], Jens Henrik Norum[1] and Therese Sørlie[1,3,4*]

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
[1]Department of Cancer Genetics, Oslo University Hospital, Oslo, Norway
[3]Centre for Cancer Biomarkers CCBIO, University of Bergen, Bergen, Norway
Full list of author information is available at the end of the article

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^2 × length)/2. When the tumors reached the maximum allowed size of 1000 mm^3, 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 Etex 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 Pfefferle 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 Etex 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^{Ex} and squamous-like^{Ex} 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 different subtypes in all cases and are presumed to be distinct primary tumors. All normal mammary gland samples were classified as normal-like^{Ex} and clustered separately from the tumors.

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

No. of samples	Murine subtype	Cluster
6	Claudin-low^{Ex}	Claudin-low-like
2	Squamous-like^{Ex}	Claudin-low-like
3	PyMT^{Ex}	Mixed
1	Class3^{Ex}	Mixed
1	Class8^{Ex}	Mixed
1	Class14^{Ex}	Mixed
1	Erbb2-like^{Ex}	Mixed
1	Wnt1-Early^{Ex}	Mixed
1	Wnt1-Late^{Ex}	Mixed
5 (normal mammary)	Normal^{Ex}	Normal

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^{Ex} 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 spindloid 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^{Ex} tumors (Fig. 2a).

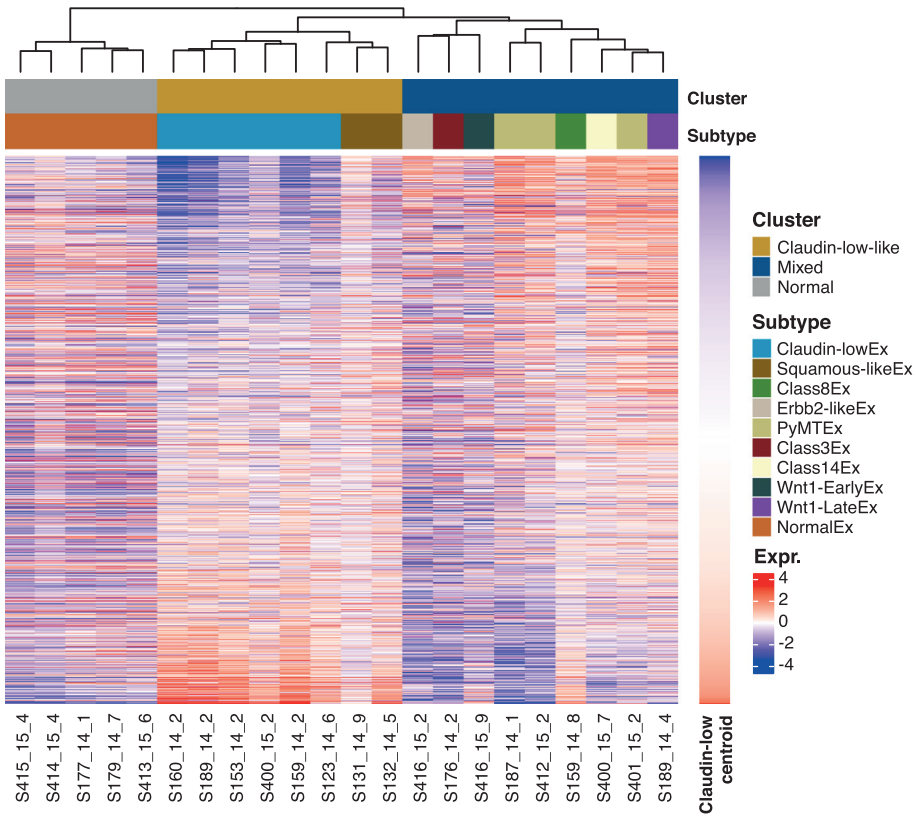


Fig. 1 Gene expression-based subtypes in the MPA/DMBA-induced tumor cohort. Using the murine intrinsic gene list [11], hierarchical clustering of gene expression data revealed two distinct tumor clusters ($p = 0.044$, SigClust [22]), one containing claudin-low-like tumors and the other containing a transcriptomically heterogeneous set of tumors. Normal mouse mammary gland samples formed a separate cluster. Genes are ordered according to correlation to the claudin-low^{Ex} 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 *Zfhx3* 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^{Ex} 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. 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].

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

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

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

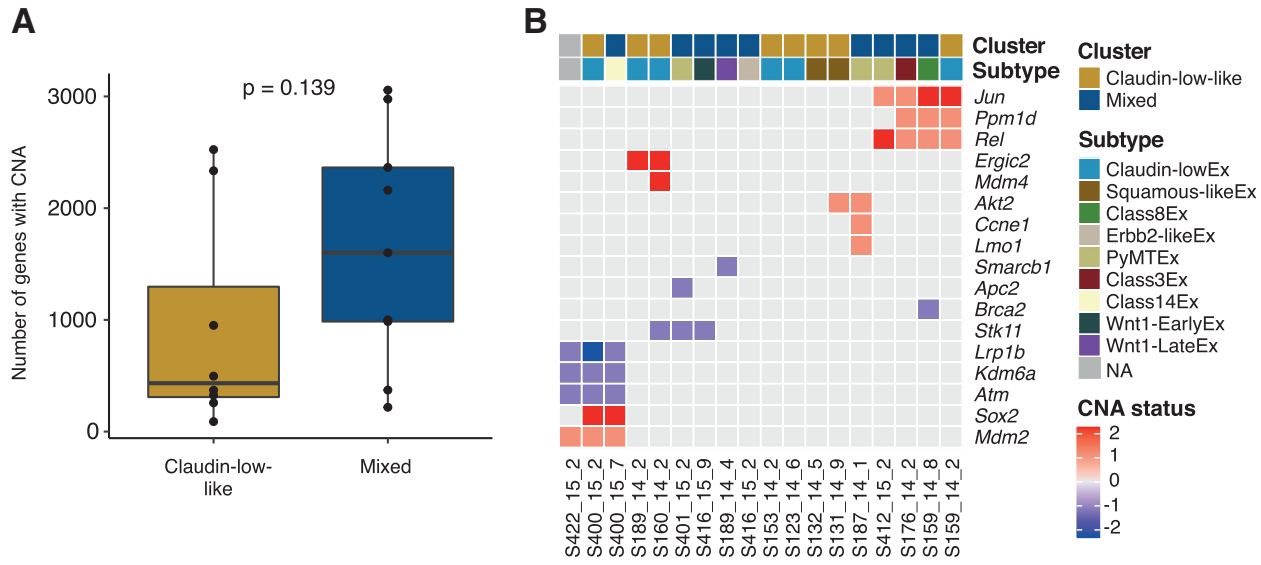


Fig. 3 Copy number aberrations in MPA/DMBA-induced mouse mammary tumors. **a** There was a trend toward a lower number of genes with copy number aberrations in the claudin-low-like cluster. **b** Copy number aberrations implicated in cancer were found in 14 of 18 MPA/DMBA-induced tumors. Two tumor sets (*S422_15_2*, *S400_15_2*, and *S400_15_7*, and *S412_15_2*, *S176_14_2*, *S159_14_8*, and *S159_14_2*) showed remarkably similar CNA profiles, but displayed different gene expression subtypes. CNA status of − 2 is a homozygous deletion, CNA status of − 1 is a heterozygous deletion, CNA status of 0 is copy number neutral, CNA status of 1 is a single copy amplification, and CNA status of 2 is a multi-copy amplification

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

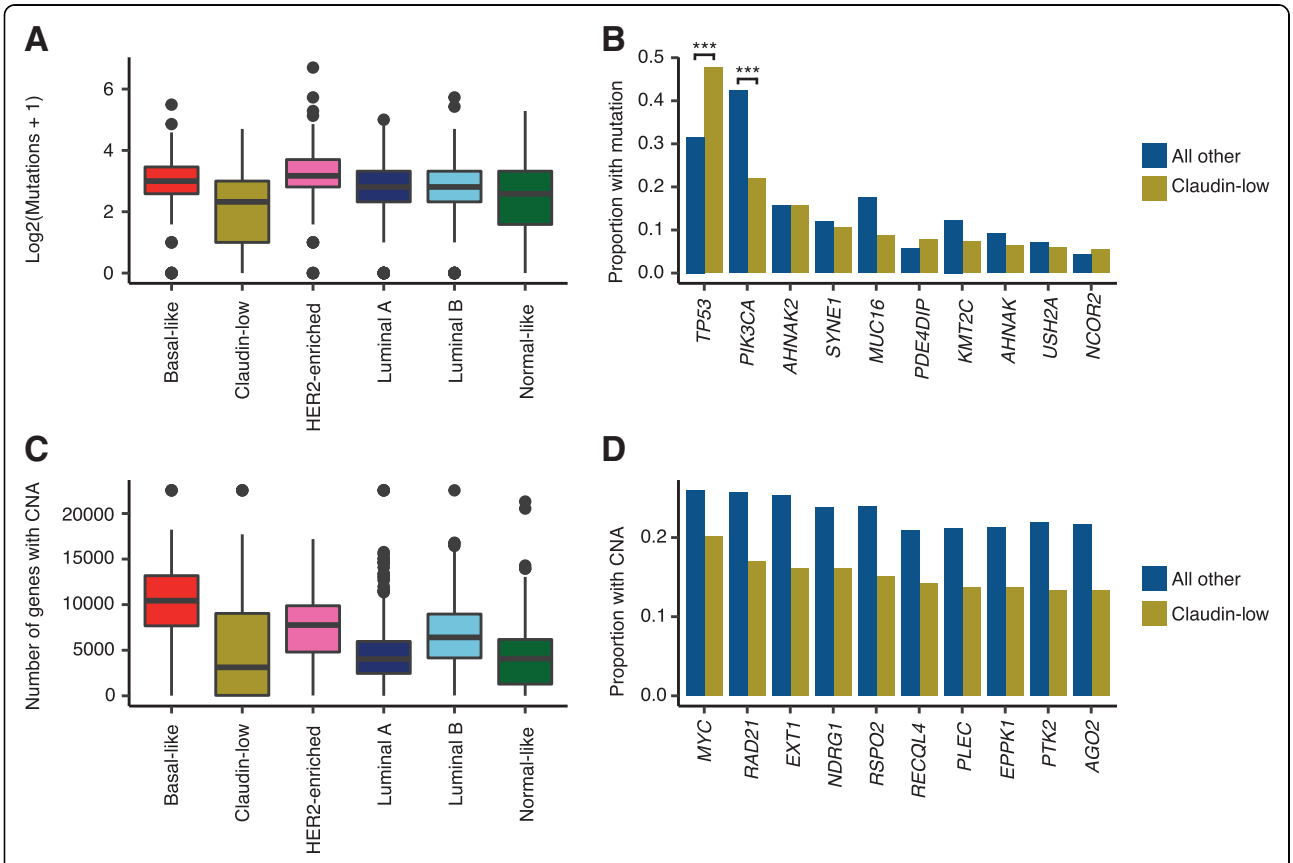


Fig. 4 Somatic mutations and copy number aberrations in human claudin-low breast cancer. **a** Claudin-low breast cancer was the subtype with the lowest mutational burden. Number of mutations displayed as $\log_2$(mutations + 1). **b** *TP53* and *PIK3CA* were the most frequently mutated genes in human breast cancer. Claudin-low tumors carried *TP53* and *PIK3CA* mutations at significantly higher and lower rates, respectively, compared to non-claudin-low breast tumors. ***$p < 0.001$. **c** Claudin-low tumors carried relatively few CNAs compared to non-claudin-low tumors. **d** The ten genes which were most frequently affected by CNA in claudin-low tumors were all found to be copy number aberrant at a higher frequency in non-claudin-low tumors. *MYC* amplification is the most common CNA event in claudin-low breast cancer

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

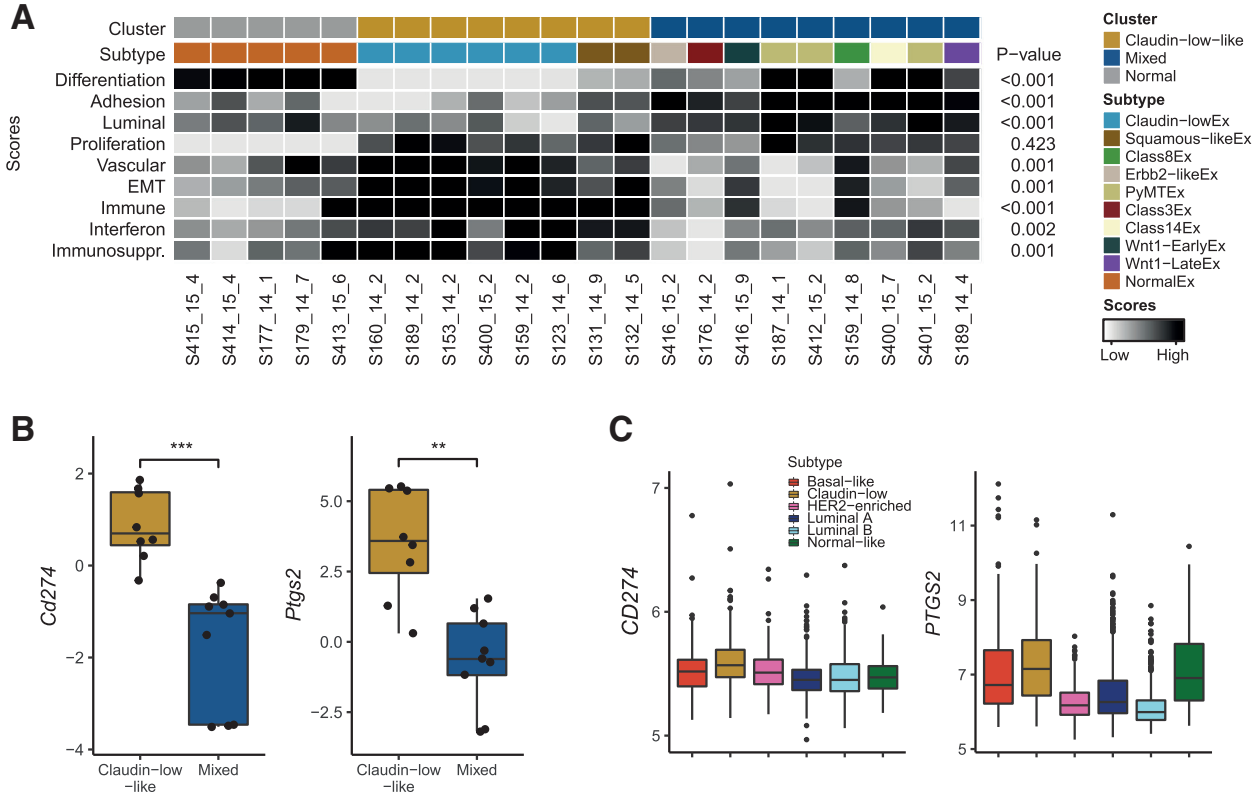


Fig. 5 Gene expression characteristics of claudin-low-like MPA/DMBA-induced tumors and human claudin-low breast cancers. **a** MPA/DMBA-induced claudin-low-like tumors recapitulated the gene expression characteristics of the claudin-low subtype as evidenced by the expression levels of relevant gene signatures. *p* values are calculated for the claudin-low-like tumors versus tumors in the mixed cluster. **b** *Cd274* and *Ptgs2* are expressed at significantly higher levels in the claudin-low-like tumors than in the mixed cluster tumors. **c** Claudin-low is the breast cancer subtype with the highest expression of *CD274* and *PTGS2*

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-l-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 transcriptomic 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 transcriptomically 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.forskningsradet.no/) (250459 to TS), South-Eastern Norway Regional Health Authority (www.helse-sorost.no/) (2012056 to TS), and the Medical Student Research Program at the University of Oslo (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

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

Received: 5 March 2019 Accepted: 17 July 2019
Published online: 31 July 2019

References

1. 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(5):R76.
2. 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.
3. 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.
4. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. Mol Oncol. 2011;5(1):5–23.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. Gonzalez-Suarez E, Jacob AP, Jones J, Miller R, Roudier-Meyer MP, Erwert R, et al. RANK ligand mediates progestin-induced mammary epithelial proliferation and carcinogenesis. Nature. 2010;468(7320):103.
14. 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.
15. Frenkel K. 7,12-dimethylbenz[a]anthracene induces oxidative DNA modification in vivo. Free Radic Biol Med. 1995;19(3):373–80.
16. Dean JH, Ward EC, Murray MJ, Lauer LD, House RV. Mechanisms of dimethylbenzanthracene-induced immunotoxicity. Clin Physiol Biochem. 1985;3(2–3):98–110.
17. 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.
18. Trichopoulos D, Adami H, Ekbom 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.
19. 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.
20. Norum JH, Bergström Å, Andersson AB, Kuiper RV, Hoelzl MA, Sørlie T, et al. A conditional transgenic mouse line for targeted expression of the stem cell marker LGR5. Dev Biol. 2015;404(2):35–48.
21. Team RC, Computing RF, for S. R. A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
29. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10-2.
30. Joshi NAFJN. Sickle: a sliding-window, adaptive, quality-based tool for FastQ files; 2011.
31. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997; 2013.
32. 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, Kernytsky 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: AACR; 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 where, the when, 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

Supplementary information

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

Fougner *et al.*

Additional file 1: Mouse characteristics and histopathological data. (.xlsx)

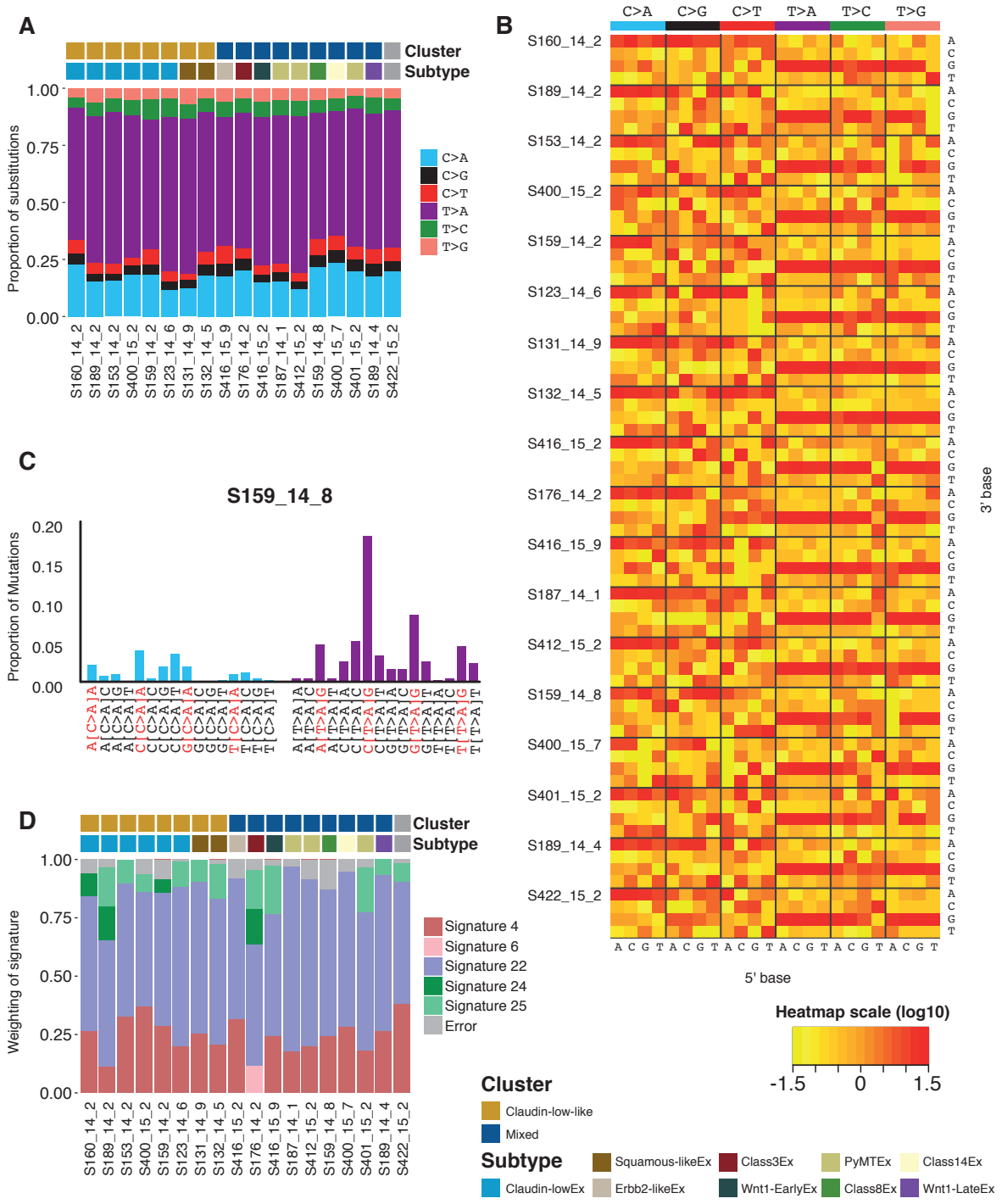
Additional file 2: Gene lists used for gene expression scores. (xlsx)

Additional file 3: Subtype correlations for MPA/DMBA-induced tumors. (.xlsx)

Additional file 4: Mutations observed in MPA/DMBA-induced tumors.

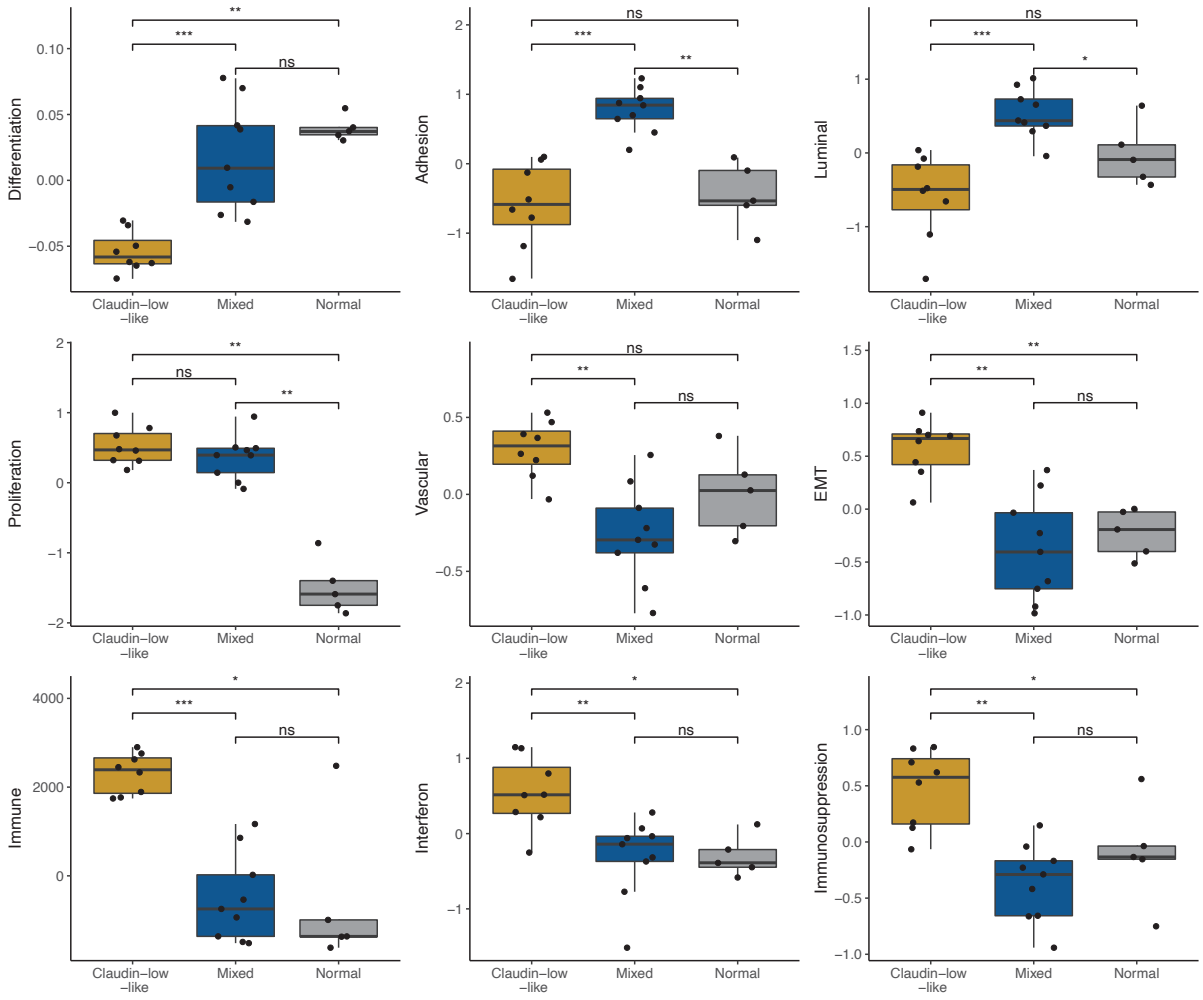
Additional file 5: Driver gene mutations in MPA/DMBA-induced tumors observed in the COSMIC database. (.xlsx)

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

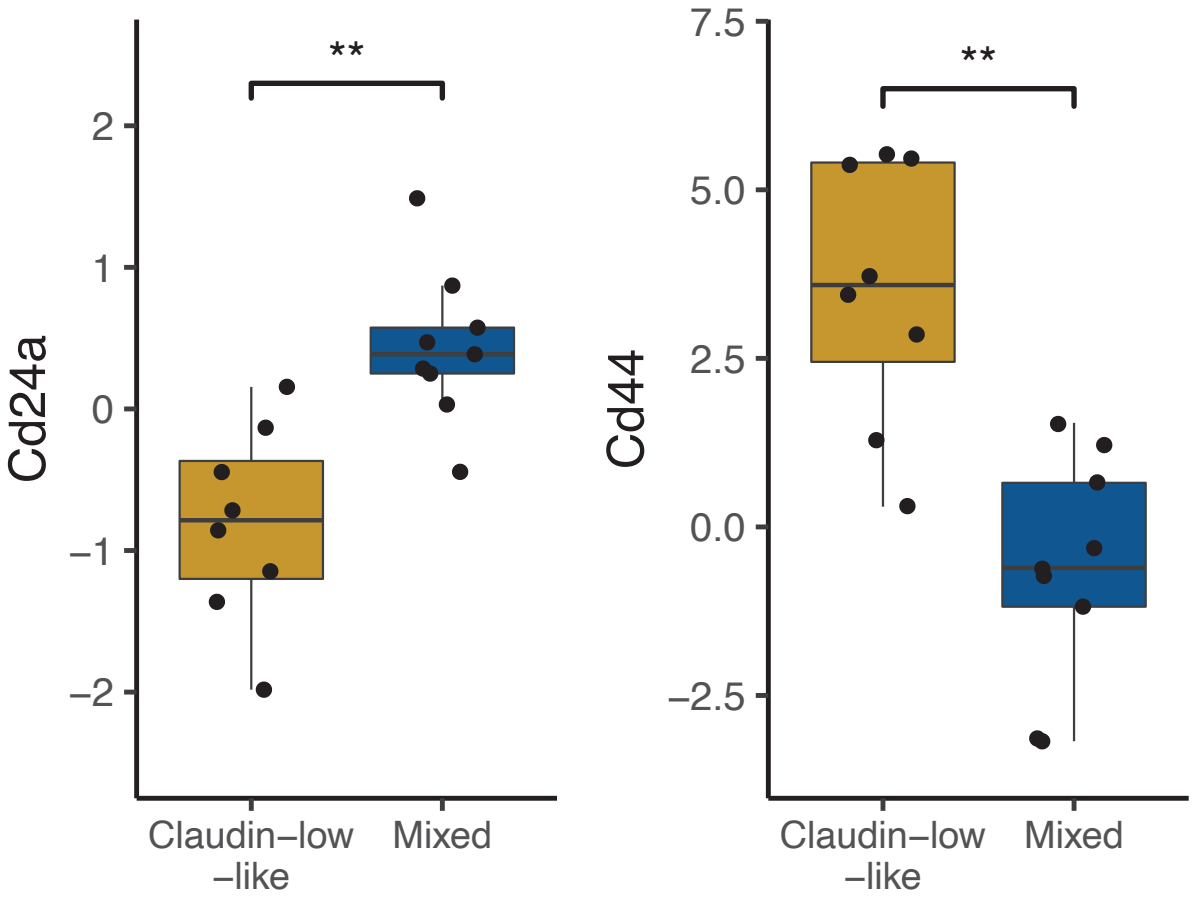


Additional file 7: The mutational spectra and mutational signatures of MPA/DMBA-induced mammary tumors. a T > A transversions were the most frequent mutation type in MPA/DMBA-induced tumors, followed by C > A transversions. **b** Heatmap of mutational frequencies by trinucleotide context. There was an 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.

Additional file 8: Mutational signatures for all MPA/DMBA-induced tumors (.zip)



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, $p > 0.05$. *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.



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.

Study II

Re-definition of *claudin-low* as a breast cancer phenotype

Christian Fougner, Helga Bergholtz, Jens Henrik Norum and Therese Sørlie.
Nature Communications 11, 1787 (2020).

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-15574-5 OPEN

Re-definition of claudin-low as a breast cancer phenotype

Christian Fougner [1,2], Helga Bergholtz [1,2], Jens Henrik Norum [1] & Therese Sørlie[1,2]✉

The claudin-low breast cancer subtype is defined by gene expression characteristics and encompasses a remarkably diverse range of breast tumors. Here, we investigate genomic, transcriptomic, and clinical features of claudin-low breast tumors. We show that claudin-low is not simply a subtype analogous to the intrinsic subtypes (basal-like, HER2-enriched, luminal A, luminal B and normal-like) as previously portrayed, but is a complex additional phenotype which may permeate breast tumors of various intrinsic subtypes. Claudin-low tumors are distinguished by low genomic instability, mutational burden and proliferation levels, and high levels of immune and stromal cell infiltration. In other aspects, claudin-low tumors reflect characteristics of their intrinsic subtype. Finally, we explore an alternative method for identifying claudin-low tumors and thereby uncover potential weaknesses in the established claudin-low classifier. In sum, these findings elucidate the heterogeneity in claudin-low breast tumors, and substantiate a re-definition of claudin-low as a cancer phenotype.

[1] Department of Cancer Genetics, Oslo University Hospital, Oslo, Norway. [2] Institute for Clinical Medicine, University of Oslo, Oslo, Norway. ✉email: therese.sorlie@rr-research.no

The five breast cancer intrinsic subtypes were initially identified by hierarchical clustering of genes with significantly greater variation in expression between different breast tumors than between paired tumor samples pre- and post-chemotherapy[1,2]. Claudin-low breast tumors did not emerge as an independent group in this analysis. The claudin-low breast cancer subtype was discovered 7 years later in an integrated analysis of human and murine mammary tumors[3]. The existence of this subtype has later been observed in several independent breast cancer cohorts[4–9], and an analogous claudin-low subtype has been identified in bladder cancer[10,11].

The claudin-low breast cancer subtype is defined by gene expression characteristics, most prominently: Low expression of cell–cell adhesion genes, high expression of epithelial–mesenchymal transition (EMT) genes, and stem cell-like/less differentiated gene expression patterns[12]. Beyond these gene expression features, claudin-low tumors have marked immune and stromal cell infiltration[9,12], but are in many other aspects remarkably heterogeneous. No specific genomic aberrations accurately delineate claudin-low tumors, and there is a greater variation in mutational burden and degree of copy number aberration (CNA) than in the other breast cancer subtypes[13]. Claudin-low tumors are, however, often genomically stable, potentially due to their less differentiated state and a protective effect mediated by the EMT-related transcription factor ZEB1[14,15]. Claudin-low breast tumors are reported to be mostly estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and human epidermal growth factor receptor 2 (HER2)-negative (triple negative), and are associated with poor prognosis[12,16]. The prevalence of claudin-low breast cancer shows striking variability, ranging from 1.5% to 14% of tumors in breast cancer cohorts[5,7,8,12].

An algorithm (predictor) for identifying claudin-low tumors was described with the original characterization of the subtype[12]. Briefly, nine claudin-low cell lines were identified by hierarchical clustering of gene expression values of 1906 breast cancer intrinsic genes[17] in a cohort of 52 cell lines. Cell lines were used to build the claudin-low predictor, rather than bulk tumor samples, to minimize immune and stromal infiltration as confounding factors[12]. Two centroids were then defined: one for the cell lines with claudin-low gene expression features and one for all other breast cancer cell lines. Claudin-low tumors are identified by correlating a tumor's gene expression values to the two centroids and defining a tumor as claudin-low if it has stronger correlation to the claudin-low centroid than the other centroid. Importantly, the intrinsic subtypes (basal-like, HER2-enriched, luminal A, luminal B and normal-like) are first identified using the PAM50 predictor[17], and claudin-low subtyping is subsequently performed as an isolated second step[12]. In published studies, claudin-low is treated as a sixth intrinsic subtype, and the subtype assigned by PAM50 is therefore overwritten in claudin-low tumors[5,8,9,12]. As a consequence, claudin-low breast tumors have, thus far, been characterized as a single group, without regard for the distribution of the underlying intrinsic subtypes in the given set of claudin-low tumors[8,9,12,13].

In this study, we aim to elucidate the heterogeneity observed in claudin-low breast cancer. By stratifying claudin-low tumors according to intrinsic subtype, we show that the characteristics of claudin-low tumors reflect the intrinsic subtype to which they are initially assigned. Further, we explore an alternative method for identifying claudin-low tumors, and demonstrate that the nine-cell line claudin-low predictor[12] may be overly inclusive in classifying tumors with marked immune and stromal infiltration as claudin-low.

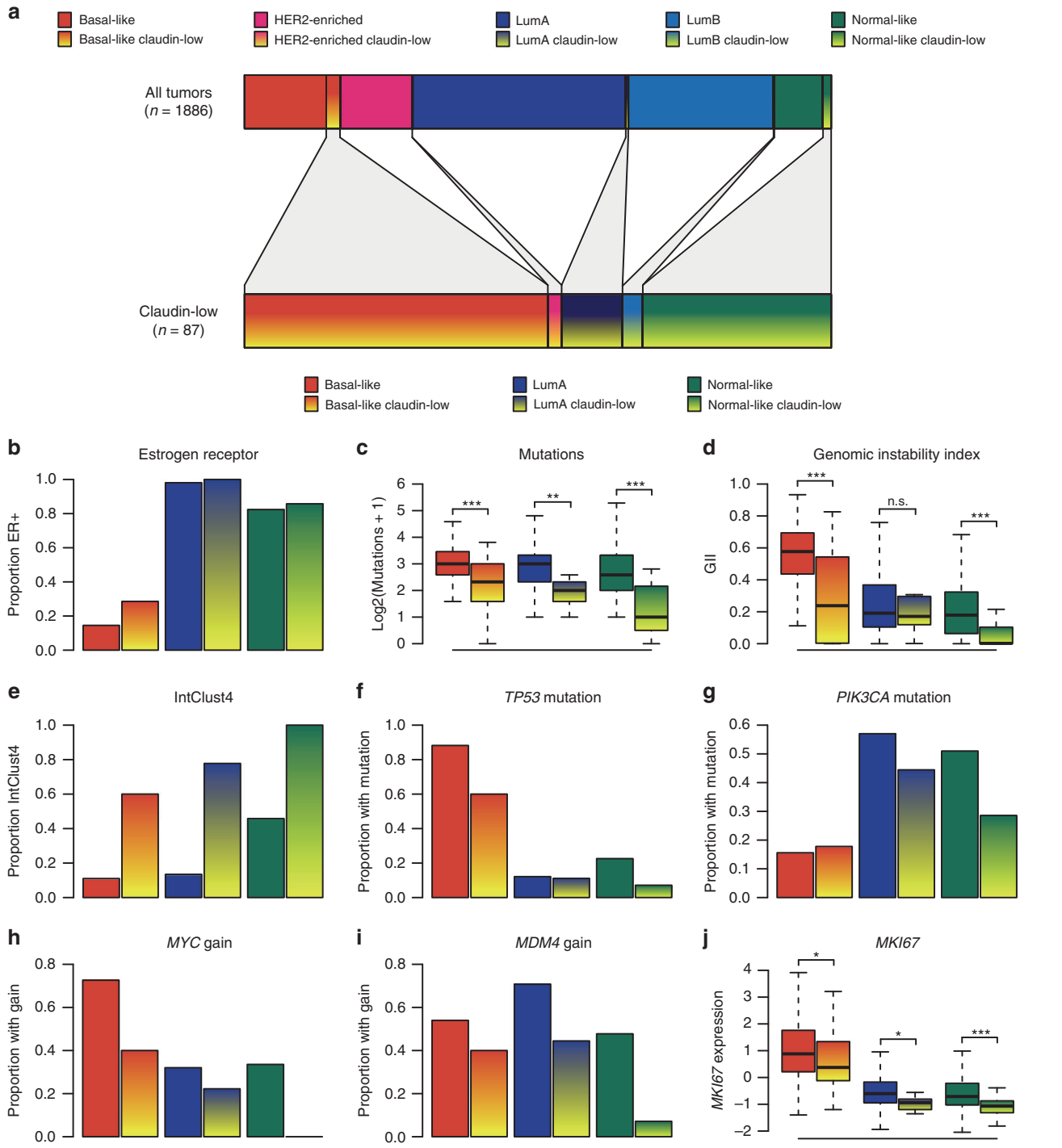
Results

Claudin-low breast tumors are delineated by intrinsic subtype. We identified 87 claudin-low tumors (4.6%) in the METABRIC cohort[4,5] using the nine-cell line claudin-low predictor[12,18] (Supplementary Data 1). By intrinsic subtype, the majority of these were classified either as basal-like (51.7%, $n = 45$), normal-like (32.2%, $n = 28$) or luminal A (LumA; 10.3%, $n = 9$) (Fig. 1a, Table 1). 14.6% and 15.3% of all basal-like and normal-like tumors, respectively, were identified as claudin-low. All three remaining subtypes were represented in the set of claudin-low tumors, but with a lower prevalence, representing 0.6 - 1.3% of tumors from each subtype. The distribution of intrinsic subtypes within the set of claudin-low tumors differed significantly from the distribution of intrinsic subtypes in non-claudin-low tumors ($P < 0.001$, χ^2-test). Basal-like and normal-like tumors were significantly overrepresented in the set of claudin-low tumors, while the remaining intrinsic subtypes were significantly underrepresented ($P = 0.001$ for HER2-enriched, $P < 0.001$ for all other, Fisher's exact test). Only two HER2-enriched and three luminal B (LumB) tumors were classified as claudin-low. These two subtypes were not analyzed further due to low sample numbers. Claudin-low tumors broadly showed similar histology to non-claudin-low tumors (Supplementary Data 1), with 70% of tumors being classified as no special type (NST). One metaplastic tumor was found in the cohort, which was classified as claudin-low.

There were significant differences in the proportion of tumors expressing estrogen receptor when claudin-low tumors were stratified by intrinsic subtype (Fig. 1b; $P < 0.001$, χ^2-test). 28.6%, 100% and 85.7% of basal-like, LumA, and normal-like claudin-low tumors, respectively, were ER-positive, closely reflecting the pattern seen in non-claudin low tumors (Fig. 1b). These findings indicate that the expression of ER in claudin-low tumors is reflected in their intrinsic subtype, and that characterizing claudin-low tumors as a triple negative subgroup of breast cancer[9,12] is an oversimplification.

Claudin-low tumors, as a whole, have previously been reported to have a low mutational burden and low level of genomic instability compared to the other subtypes[13,14]. Whole genome copy number data and sequence data from a panel of 173 cancer-associated genes were available for the METABRIC cohort[4,5]. When claudin-low tumors were stratified by intrinsic subtype, they consistently showed lower mutational burden and genomic instability compared to their non-claudin-low counterparts (Fig. 1c, d), with the exception of genomic instability in LumA tumors. There were, however, also significant differences in mutational burden ($P = 0.002$, Kruskal–Wallis test) and genomic instability ($P < 0.001$, Kruskal–Wallis test) between claudin-low tumors of the different intrinsic subtypes. Despite a degree of subtype specific variations, these findings point toward lower mutational rate and lower levels of genomic instability as bona fide claudin-low characteristics.

Curtis et al.[4] introduced breast cancer subtypes (IntClust) defined by patterns of CNA with *cis* correlation to gene expression. The genomically stable IntClust4 subtype showed overlap with claudin-low tumors[4,5,14]. In our analyses, 75% of all claudin-low tumors in the METABRIC cohort were classified as IntClust4. Stratified by intrinsic subtype, claudin-low tumors were consistently more likely to be classified as IntClust4 compared to non-claudin-low tumors of the same subtype (Fig. 1e). There were however significant variations in the proportion of claudin-low tumors classified as IntClust4 ($P < 0.001$, χ^2-test), ranging from 60% of basal-like claudin-low tumors to 100% of normal-like claudin-low tumors. Further, IntClust4 tumors have been separated into ER-positive and ER-negative groups due to major differences in their biological and clinical characteristics, despite strong similarities in gene expression patterns and associated low levels of CNA[4,5,19]. Claudin-low tumors classified as IntClust4ER+ were predominantly LumA and normal-like, whereas claudin-low tumors classified as IntClust4ER− were predominantly basal-like (Supplementary Fig. 1a, b).



The high frequency of claudin-low tumors classified as IntClust4 supports the association between claudin-low gene expression characteristics and genomic stability. However, only 21% of all IntClust4 tumors in the METABRIC cohort were classified as claudin-low, and genomic instability index (GII) did not accurately predict correlation to the claudin-low centroid, as determined by the nine-cell line predictor[12] (Supplementary Fig. 2). Thus, while most claudin-low tumors were genomically stable, only a subset of genomically stable tumors were claudin-low.

No putative driver[20] mutations or CNAs were found at a significantly higher rate in claudin-low tumors, stratified by intrinsic subtype, than in non-claudin-low tumors of the same subtype (Fisher's exact test, Bonferroni corrected; Supplementary Data 2). Rather, claudin-low tumors tended to exhibit patterns of mutation/CNA associated with their intrinsic subtype. Reflecting the lower levels of genomic instability and mutational burden, claudin-low tumors generally had lower incidences of potential driver aberrations compared to their non-claudin-low counterparts. To illustrate the relative frequencies of driver aberrations in claudin-low and non-claudin-low tumors, we selected four early genomic driver aberrations for further analysis: *TP53* mutation, *PIK3CA* mutation, *MYC* gain (located on 8q24), and *MDM4* gain (located on 1q32). Similar to the pattern observed for ER-positivity, the incidence of *TP53* mutations in claudin-low tumors largely followed the incidence seen in the tumors' intrinsic subtype (Fig. 1f). The differences in *TP53* mutation rates between

Fig. 1 Claudin-low tumors are delineated by intrinsic subtype. a Distribution of intrinsic subtypes in the METABRIC cohort for all tumors (top bar, $n = 1886$) and for claudin-low tumors only (bottom bar, $n = 87$). **b** Estrogen receptor status. 58% of claudin-low tumors were ER-positive. ER prevalence differed between claudin-low tumors stratified by intrinsic subtype ($P < 0.001$, χ^2-test). **c** Number of mutations in the panel of 173 sequenced genes. Claudin-low tumors showed lower mutational rates than non-claudin-low tumors of the same subtype (basal-like $P < 0.001$, LumA $P = 0.004$, normal-like $P < 0.001$, two-tailed Wilcoxon rank-sum test). **d** Genomic instability index (GII). Basal-like and normal-like claudin-low tumors showed lower levels of genomic instability than non-claudin-low tumors of the same subtype (basal-like $P < 0.001$, LumA $P = 0.83$, normal-like $P < 0.001$, two-tailed Wilcoxon rank-sum test). **e** IntClust4. 75% of claudin-low tumors were classified as IntClust4. IntClust4 classification differed between claudin-low tumors stratified by intrinsic subtype ($P < 0.001$, χ^2-test). **f** *TP53* mutation. 38% of claudin-low tumors carried *TP53* mutations. Rate of *TP53* mutation differed between claudin-low tumors stratified by intrinsic subtype ($P < 0.001$, χ^2-test). **g** *PIK3CA* mutation. 24% of claudin-low tumors carried *PIK3CA* mutations. Differences between claudin-low tumors stratified by intrinsic subtype were not statistically significant ($P = 0.19$, χ^2-test) **h** *MYC* gain. 26% of claudin-low tumors showed gain of *MYC*. Rate of *MYC* gain differed between claudin-low tumors stratified by intrinsic subtype ($P < 0.001$, χ^2-test). **i** *MDM4* gain. 30% of claudin-low tumors showed gain of *MDM4*. Rate of *MDM4* gain differed between claudin-low tumors stratified by intrinsic subtype ($P = 0.006$, χ^2-test). **j** *MKI67* gene expression (log2). Claudin-low tumors consistently expressed lower levels of *MKI67* compared to non-claudin-low counterparts (basal-like $P = 0.01$, LumA $P = 0.03$, normal-like $P < 0.001$, two-tailed Wilcoxon rank-sum test). **All** n.s. $P > 0.05$, $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Sample sizes provided in Table 1. Boxplot elements: center line = median, box limits = upper and lower quartiles, whiskers = 1.5 × interquartile range. Source data are provided as a Source Data file.

Table 1 Distribution of claudin-low tumors by intrinsic subtype in the METABRIC cohort.

Intrinsic subtype	Claudin-low (*n*)	Non- claudin-low (*n*)	Proportion claudin-low in subtype
Basal-like	45	263	14.6%
HER2-enriched	2	231	0.9%
LumA	9	684	1.3%
LumB	3	466	0.6%
Normal-like	28	155	15.3%

Source data are provided as a Source Data file.

claudin-low tumors stratified by intrinsic subtype were statistically significant ($P < 0.001$, χ^2-test). There were similar trends for the other three aberrations analyzed (Fig. 1g–i). Claudin-low tumors stratified by intrinsic subtype showed significantly different rates of *MYC* and *MDM4* gain ($P < 0.001$ and $P = 0.006$, χ^2-test), but not *PIK3CA* mutation ($P = 0.19$, χ^2-test).

Claudin-low tumors have previously been characterized as slower cycling, with proliferation levels lower than in basal-like tumors, but higher than in LumA and normal-like tumors[8,12]. Ki-67, encoded by the *MKI67* gene, is a commonly used proliferation marker. When claudin-low tumors were stratified by intrinsic subtype, there were significantly different levels of *MKI67* expression between subtypes (Fig. 1j; $P < 0.001$, Kruskal–Wallis test), with basal-like claudin-low tumors showing significantly higher levels of *MKI67* expression than LumA claudin-low tumors and normal-like claudin-low tumors ($P < 0.001$ for both, Wilcoxon rank-sum test). Claudin-low tumors did, however, also show significantly lower levels of *MKI67* expression than non-claudin-low counterparts in all intrinsic subtypes (Fig. 1j; $P = 0.01$, 0.03 and <0.001 claudin-low compared to non-claudin-low in basal-like, LumA, and normal-like tumors, respectively, Wilcoxon rank-sum test). Thus, *MKI67* gene expression levels indicate that claudin-low tumors reflect the proliferation levels of their intrinsic subtype but are also slower cycling than non-claudin-low counterparts.

Claudin-low tumors have previously been associated with poor prognosis[8,12]. This characterization was accurate when claudin-low tumors were viewed as a single group (Supplementary Fig. 1c). However, when the survival of patients with claudin-low tumors was stratified by intrinsic subtype, the survival patterns generally observed in non-claudin-low breast cancer[2] re-emerged (Fig. 2a). Further, there were no significant differences in survival between patients with claudin-low and non-claudin-low tumors within each intrinsic subtype (Fig. 2b–d). Thus, we did not find evidence indicating that claudin-low status affects survival in breast cancer patients.

Claudin-low tumors have been reported to mostly occur in younger patients, with age at diagnosis slightly higher than in basal-like tumors, but lower than in the remaining subtypes[8,9]. When claudin-low tumors were stratified by intrinsic subtype, there were, however, significant differences in the average age at diagnosis ($P = 0.01$, Kruskal–Wallis test; Supplementary Fig. 1d), with basal-like claudin-low tumors diagnosed at a significantly lower age than LumA claudin-low and normal-like claudin-low tumors ($P = 0.03$ and 0.01 respectively, Wilcoxon rank-sum test). Claudin-low and non-claudin-low tumors of the same intrinsic subtype showed similar age at diagnosis (basal-like $P = 0.67$, LumA $P = 0.53$, normal-like $P = 0.052$, two-tailed Wilcoxon rank-sum test).

A condensed gene list refines claudin-low classification. Claudin-low tumors have been shown to exhibit high degrees of immune and stromal infiltration[9,12]. Also when stratified by intrinsic subtype, claudin-low tumors in the METABRIC cohort consistently had higher infiltration of immune and stromal cells compared to non-claudin-low tumors (as determined by ESTIMATE, a gene expression-based tool for inferring normal-cell infiltration in tumors[21]) (Supplementary Fig. 3a, b). The nine-cell line claudin-low predictor uses 807 genes, and Prat et al. acknowledge that it may inappropriately identify some tumors as claudin-low solely due to stromal infiltration[12]. This statement is supported by a strong correlation between a tumor's inferred[21] stromal infiltration and closeness to the nine-cell line claudin-low centroid ($R^2 = 0.76$, linear regression; Supplementary Fig. 3c, d). A similar, but weaker trend was also observed for inferred[21] immune cell infiltration ($R^2 = 0.27$, linear regression; Supplementary Fig. 3e, f). We therefore considered whether a more targeted gene list could be used for claudin-low classification, in order to more accurately isolate features intrinsic to claudin-low tumors.

We created a condensed claudin-low gene list (Supplementary Table 1), consisting of 19 genes representing only the pathognomonic gene expression characteristics of claudin-low tumors: Low expression of cell–cell adhesion genes, high expression of EMT genes, and gene expression patterns typical of stem cell-like/less differentiated cells[3,8,9,12,14]. In the METABRIC cohort, hierarchical clustering of gene expression values, using the condensed gene list, revealed a tumor cluster with gene expression characteristics in line with those previously described in

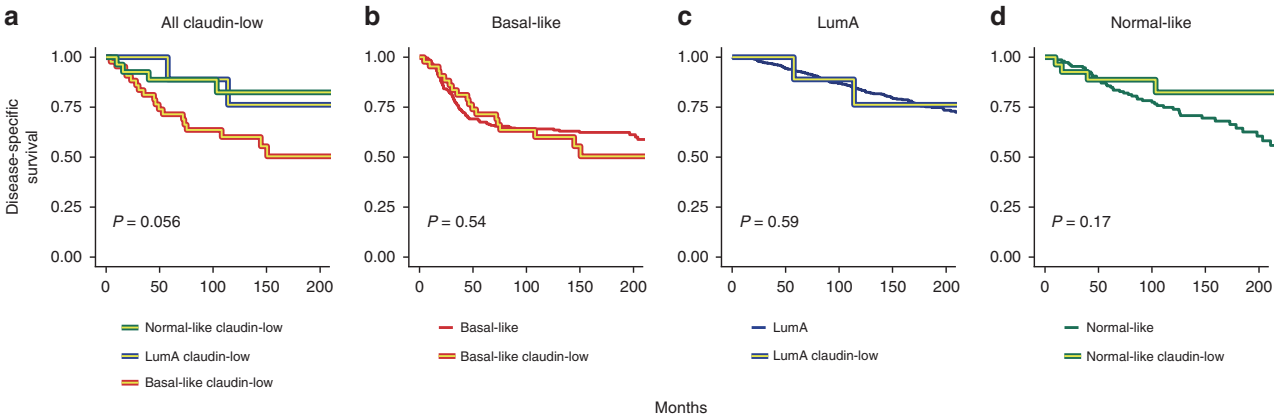


Fig. 2 No evidence of claudin-low status as an indicator of poor prognosis in the METABRIC cohort. a Disease-specific survival in basal-like claudin-low, LumA claudin-low, and normal-like claudin-low tumors in the METABRIC cohort. Survival trends recapitulated the patterns seen in non-claudin-low tumors. $P = 0.056$ when testing for difference between claudin-low tumors stratified by intrinsic subtype. **b**–**d** Disease-specific survival in claudin-low and non-claudin-low basal-like (**b**), LumA (**c**) and normal-like (**d**) tumors. Significant differences between claudin-low and non-claudin-low tumors were not found (basal-like $P = 0.54$, LumA $P = 0.59$, normal-like $P = 0.17$). **All** Two-tailed log-rank test used for significance testing. Disease-specific deaths and sample sizes: basal-like claudin-low $n = 19$ of 45, basal-like non-claudin-low $n = 98$ of 263, LumA claudin-low $n = 3$ of 9, LumA non-claudin-low $n = 144$ of 684, normal-like claudin-low $n = 5$ of 28, normal-like non-claudin-low $n = 50$ of 155. Source data are provided as a Source Data file.

claudin-low tumors (Fig. 3; $P = 0.006$, SigClust[22]). We refer to tumors in this cluster as core claudin-low (CoreCL), while claudin-low tumors (as defined by the nine-cell line predictor) outside the CoreCL cluster are referred to as other claudin-low (OtherCL). Individual inspection of gene expression values showed that OtherCL tumors displayed certain claudin-low characteristics, albeit to a lesser degree than CoreCL tumors (Supplementary Fig. 4).

The CoreCL cluster consisted of 79 tumors (4.2% of tumors in the cohort), of which 57 (72.2%) were identified as claudin-low by the nine-cell line predictor (Supplementary Data 1). While several intrinsic subtypes were prominently represented in the group of CoreCL tumors, the OtherCL ($n = 30$) tumors were predominantly basal-like ($n = 23$; Fig. 4a). Thus, our method for identifying claudin-low tumors primarily differed from the nine-cell line predictor by filtering out a set of basal-like tumors with high levels of stromal and immune infiltration (Supplementary Fig. 5a, b), but without pathognomonic claudin-low gene expression characteristics (Supplementary Fig. 6).

There were marked contrasts between the characteristics of basal-like CoreCL tumors ($n = 25$), basal-like OtherCL-tumors ($n = 23$), and non-claudin-low basal-like tumors ($n = 260$). Basal-like CoreCL tumors carried significantly fewer mutations than basal-like OtherCL tumors and non-claudin-low basal-like tumors (Fig. 4b; $P = 0.015$ & $P < 0.001$ respectively, Wilcoxon rank-sum test). Basal-like CoreCL tumors also displayed significantly lower levels of genomic instability than basal-like OtherCL tumors and non-claudin-low basal-like tumors (Fig. 4c; $P < 0.001$ for both, Wilcoxon rank-sum test). There were no significant differences in GII between basal-like OtherCL tumors and non-claudin-low basal-like tumors (Fig. 4c, $P = 0.082$, Wilcoxon rank-sum test). There was also a greater proportion of basal-like CoreCL tumors in IntClust4, than basal-like OtherCL and non-claudin-low basal-like tumors (Fig. 4d, Supplementary Fig. 5c, d). In total, 80% of basal-like CoreCL tumors were classified as IntClust4, in contrast to 43% of basal-like OtherCL tumors and 10% of basal-like non-claudin-low tumors. There were also lower rates of *TP53* mutation, *MYC* gain and *MDM4* gain, in basal-like CoreCL tumors compared to basal-like OtherCL and basal-like non-claudin-low tumors, reflecting the lower mutational burden and GII (Supplementary Fig. 5e, g). This trend was, however, not evident for *PIK3CA* (Supplementary Fig. 5h). Basal-like CoreCL tumors expressed significantly lower levels of *MKI67* than basal-like OtherCL and basal-like non-claudin-low tumors (Fig. 4e; $P < 0.001$ for both, Wilcoxon rank-sum test). There were no significant differences in *MKI67* expression between basal-like OtherCL and basal-like non-claudin-low tumors ($P = 0.63$, Wilcoxon rank-sum test). In sum, the characteristics of basal-like OtherCL tumors show weaker concordance with the characteristics of claudin-low tumors, compared to basal-like CoreCL tumors. It is therefore likely that OtherCL tumors are classified as claudin-low by the nine-cell line predictor due to their stromal infiltration (Supplementary Figs. 3c, 5b), and that the classification of these tumors as claudin-low may be dubious.

Despite differences in genomic and transcriptomic features, as well as in immune and stromal infiltration, there were no significant differences in survival between basal-like CoreCL, basal-like OtherCL and non-claudin-low basal-like tumors (Fig. 4f). These findings reinforce our observations indicating that claudin-low status is not a major determinant of survival in breast cancer patients.

There were few OtherCL samples not classified as basal-like ($n = 1$, 3, and 3 for LumA, LumB and normal-like tumors, respectively; Fig. 4a). The characteristics of normal-like CoreCL ($n = 39$) and LumA CoreCL ($n = 13$) tumors were similar to the characteristics of normal-like claudin-low and LumA claudin-low tumors identified by the nine-cell line predictor (Supplementary Figs. 7, 8). These findings indicate that the nine-cell line predictor's promiscuous classification of stromally infiltrated tumors as claudin-low may mostly be of concern in basal-like tumors.

Validation cohorts reinforce key claudin-low characteristics. To validate our findings, we queried the Oslo2 cohort[23], for which gene expression data and ER/HER2 status were available. There were 29 claudin-low tumors, as defined by the nine-cell line predictor, in the cohort (7.6%), of which most were classified as basal-like, LumA or normal-like ($n = 7$, 5 and 11, respectively; Supplementary Fig. 9a, Supplementary Data 3). When clustering using the condensed claudin-low gene list, there was a cluster with claudin-low gene expression characteristics and high levels of immune and stromal cell infiltration (Supplementary Fig. 9b;

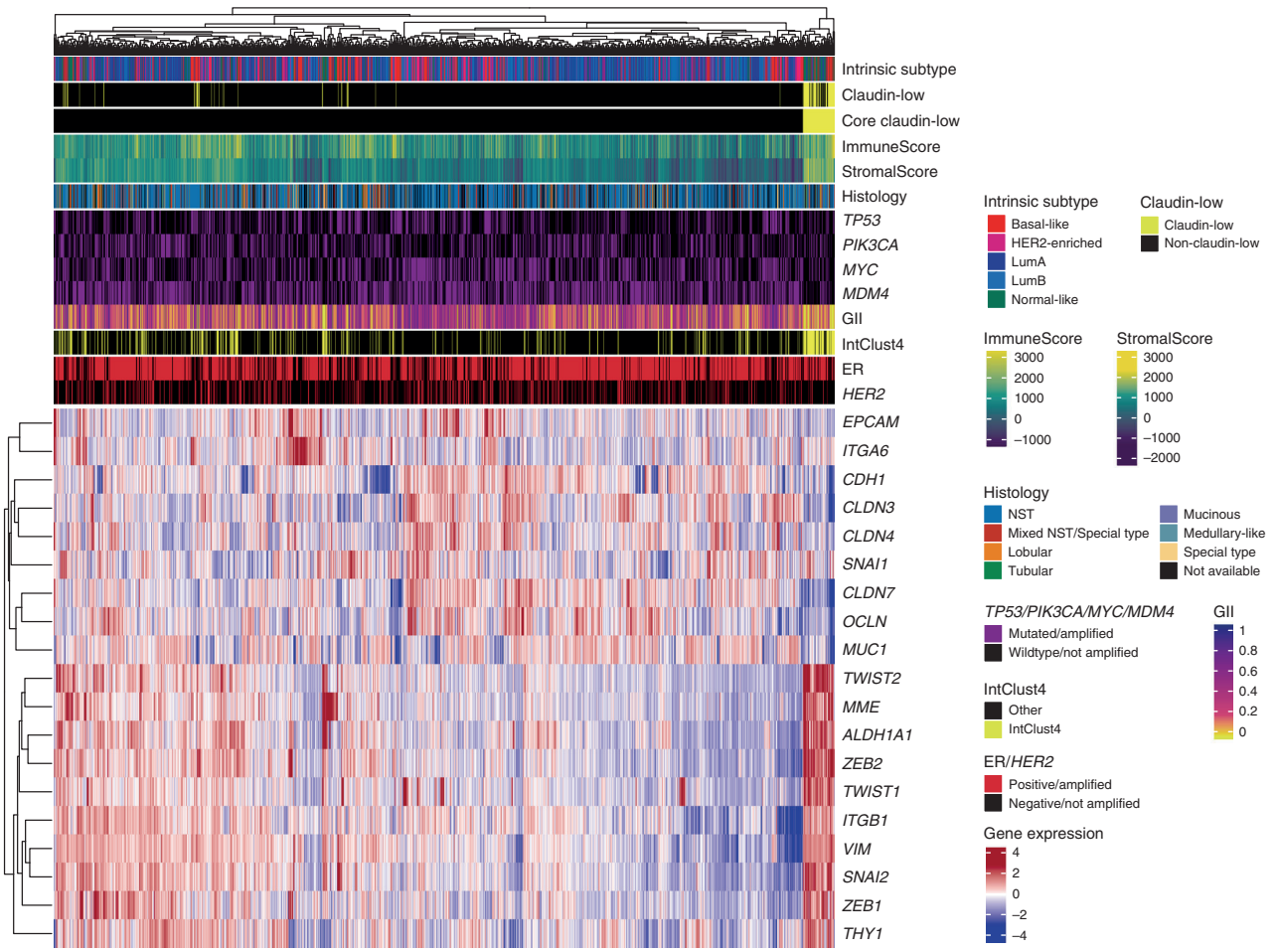


Fig. 3 A condensed claudin-low gene list identifies a set of core claudin-low tumors. Heatmap of gene expression values (log2) for a condensed claudin-low gene list in the METABRIC cohort ($n = 1886$ biologically independent samples). A cluster, marked Core claudin-low ($n = 79$), emerged with transcriptomic and genomic claudin-low characteristics ($P = 0.006$, SigClust[22]). Source data are provided as a Source Data file.

$P < 0.001$, SigClust[22]). 28 tumors in the cohort (7.3%) were located in the core claudin-low cluster (Supplementary Fig. 9c), of which 16 (57%) were identified as claudin-low by the nine-cell line predictor. Seven basal-like tumors were classified as claudin-low by the nine-cell line predictor; two of these were CoreCL, both of which were IntClust4, and the remaining five were OtherCL, none of which were IntClust4. Using IntClust4 as a surrogate marker for low levels of genomic instability[4,19,24], these findings emphasize that the nine-cell line predictor may be overly inclusive in identifying basal-like tumors as claudin-low. The OtherCL tumors in the Oslo2 cohort were, however, more diverse than in the METABRIC cohort, with 7 of 12 OtherCL tumors being non-basal-like ($n = 1$, 4 and 2 for HER2-enriched, LumA, and LumB, respectively). In total, 89% of CoreCL tumors in the Oslo2 cohort were classified as IntClust4, compared to 38% of OtherCL tumors and 20% of non-claudin-low tumors. Thus, the characteristics of claudin-low tumors in the Oslo2 cohort were mostly consistent with those observed in the METABRIC cohort.

Finally, we explored the TCGA breast cancer cohort[7,25]. 32 of 1082 tumors (3.0%) were classified as claudin-low by the nine cell-line predictor (Supplementary Data 4); however, no core claudin-low cluster emerged (Supplementary Fig. 10). As previously noted, non-tumor cell infiltration is a central characteristic of claudin-low tumors. An inclusion criterion in the TCGA protocol is a tumor cellularity over 60%[7]. The METABRIC cohort was originally divided into a discovery cohort with a cellularity cut-off of 40%, which had a claudin-low prevalence of 3.6%, and a validation cohort with no cellularity cut-off[4], which had a claudin-low prevalence of 5.6%. There was no cut-off for cellularity in the Oslo2 cohort[23], which had a claudin-low prevalence of 7.6%. Thus, there may be an association between cellularity cut-off in a cohort and claudin-low prevalence (Fig. 5). This strengthens the observation of non-tumor cell infiltration as a fundamental claudin-low characteristic and may explain the absence of a core-claudin-low cluster in the TCGA-BRCA cohort.

Claudin-low tumors in the TCGA breast cancer cohort mostly showed histological features in line with those of non-claudin-low tumors (Supplementary Data 4). There were eight metaplastic tumors in the cohort, of which six were classified as claudin-low, confirming that most metaplastic tumors are claudin-low[26].

Discussion

Here, we have re-evaluated the characteristics of claudin-low breast tumors, from the perspective of claudin-low as a phenotype that may permeate the intrinsic subtypes. Through analyses of genomic, transcriptomic and clinical data, we have shown that the characteristics of claudin-low tumors reflect their intrinsic subtype. Characteristics that are associated with claudin-low status include marked immune and stromal cell infiltration, low levels of genomic instability and mutational burden, and reduced proliferation levels. Finally, we explored an alternative method for identifying claudin-low tumors, and thereby showed that a subset of tumors with pronounced immune and stromal infiltration may

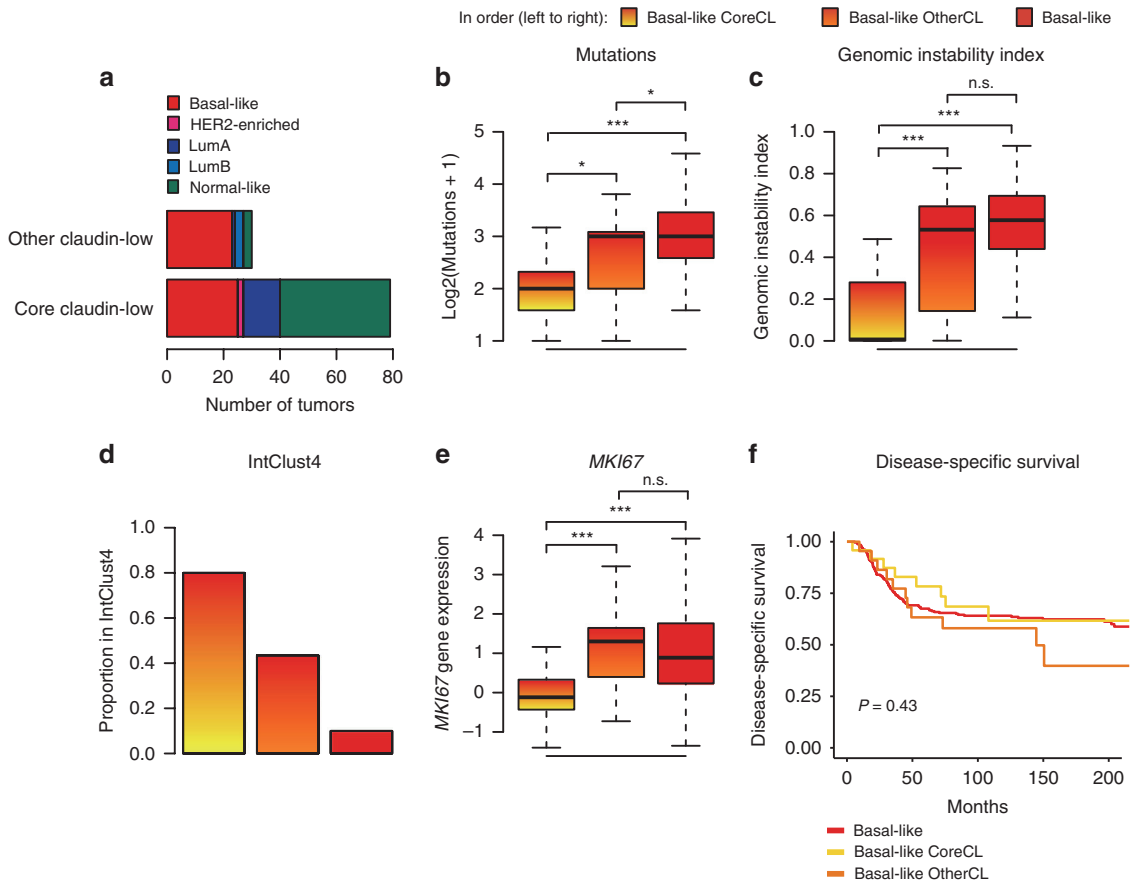


Fig. 4 Basal-like OtherCL tumors may be inappropriately classified as claudin-low. a Distribution of subtypes in CoreCL and OtherCL tumors in the METABRIC cohort. Hierarchical clustering with the condensed claudin-low gene list filtered out a subset of basal-like claudin-low tumors (as defined by the nine-cell line predictor) with weak claudin-low characteristics. **b** Number of mutated genes in the panel of 173 sequenced genes. Basal-like CoreCL tumors carried significantly fewer mutations than basal-like OtherCL tumors and basal-like non-claudin-low tumors ($P = 0.02$ and $P < 0.001$ respectively, two-tailed Wilcoxon rank-sum test). **c** Distribution of genomic instability index. Basal-like CoreCL tumors showed significantly lower levels of genomic instability than basal-like OtherCL tumors and non-claudin-low basal-like tumors ($P < 0.001$ for both, two-tailed Wilcoxon rank-sum test). **d** Proportion of tumors in IntClust4. 80% of basal-like CoreCL tumors were classified as IntClust4. **e** *MKI67* gene expression (log2). Basal-like CoreCL tumors expressed significantly lower levels of *MKI67* than basal-like OtherCL and non-claudin-low tumors ($P < 0.001$ for both, two-tailed Wilcoxon rank-sum test). **f** Disease-specific survival in basal-like CoreCL, basal-like OtherCL and non-claudin-low basal-like tumors. Disease-specific survival in basal-like breast tumors did not significantly differ when stratified by claudin-low status (two-tailed log-rank test). Disease-specific deaths: basal-like CoreCL $n = 8$ of 25, basal-like OtherCL $n = 12$ of 23, basal-like non-claudin-low $n = 97$ of 260. **All** n.s. $P > 0.05$, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Boxplot elements: center line = median, box limits = upper and lower quartiles, whiskers = 1.5 × interquartile range. Basal-like CoreCL $n = 25$, basal-like OtherCL $n = 23$, basal-like non-claudin-low $n = 260$, HER2-enriched CoreCL $n = 2$, LumA CoreCL $n = 13$, LumA OtherCL $n = 1$, LumB OtherCL $n = 3$, normal-like CoreCL $n = 39$, normal-like OtherCL $n = 3$ biologically independent samples. Source data are provided as a Source Data file.

be inappropriately classified as claudin-low by the established claudin-low predictor[12].

We stratified claudin-low tumors by intrinsic subtype and found differences between claudin-low tumors of different intrinsic subtypes in almost all aspects analyzed. Perhaps most surprisingly, we found no evidence indicating that claudin-low status affects disease-specific survival, contrasting with previous reports of claudin-low as a poor prognosis subtype[8,12]. These findings imply that a large subset of previously reported characteristics of claudin-low tumors are not bona fide claudin-low characteristics but are rather an average of the characteristics of several intrinsic subtypes. Thus, the established practice of analyzing claudin-low tumors as a single entity, without taking intrinsic subtype into consideration, may obscure the features that are attributable to claudin-low status.

Claudin-low breast cancer has previously been considered a single disease entity, analogous to the intrinsic breast cancer subtypes[8,9,12,13] (Fig. 6a). Our findings, however, imply that breast tumors are not claudin-low instead of the intrinsic subtype to which they are assigned by the PAM50 predictor, rather that they can carry a claudin-low phenotype in addition to their intrinsic subtype (Fig. 6b). According to this interpretation, claudin-low is a measure of a set of biological features which is distinct from the set of biological features measured by the intrinsic subtypes.

We explored a method of identifying claudin-low tumors using a condensed gene list. The claudin-low tumors identified using this method (CoreCL) showed more consistent traits than the claudin-low tumors identified by the nine-cell line predictor. OtherCL tumors can be interpreted to not be genuine claudin-low tumors. OtherCL tumors did, however, display some genomic and transcriptomic traits which were consistent with the claudin-low phenotype, though to a lesser degree than CoreCL tumors. A compelling interpretation may instead be that claudin-low is a continuum (degree of "claudin-lowness", Fig. 6c), rather than a binary feature (claudin-low vs. non-claudin-low, Fig. 6b).

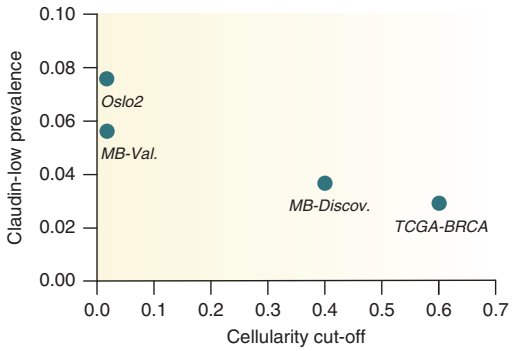


Fig. 5 Cut-offs for tumor cellularity may affect the prevalence of claudin-low tumors in breast cancer cohorts. Relationship between cut-offs for tumor cellularity in a cohort and the prevalence of claudin-low tumors (as identified by the nine-cell line predictor). Cohorts which excluded tumors due to low cellularity had a lower claudin-low prevalence than cohorts without a cellularity exclusion criterion. *MB-Discov.*: METABRIC discovery cohort, *MB-Val.*: METABRIC validation cohort. MB-Discov. $n = 957$, MB-Val. $n = 929$, Oslo2 $n = 381$, TCGA-BRCA $n = 1082$ biologically independent samples. Source data are provided as a Source Data file.

According to this interpretation, breast tumors might exist along a spectrum of claudin-lowness, in which they lie somewhere between: (1) non-claudin-low, fully concordant with an intrinsic subtype, (2) moderately claudin-low with marked imprint of an intrinsic subtype (exemplified by the average claudin-low tumor identified by the nine-cell line predictor), (3) extensively claudin-low, with limited imprint of an intrinsic subtype (exemplified by the average CoreCL tumor), or (4) purely claudin-low, with no imprint of intrinsic subtype (perhaps exemplified by special histological subtypes[26,27]). This model would be consistent with recent descriptions of partial EMT phenotypes in cancer[28,29] and cellular pliancy as an etiological explanation for the claudin-low phenotype[14,15].

Claudin-low tumors had high levels of non-tumor cell infiltration, and there was a lower prevalence of claudin-low tumors in the cohorts with a cut-off for tumor cellularity. It is also known that EMT-like gene expression features in tumors are similar to the gene expression characteristics of stromal tissue[29], and a subset of normal breast tissue samples show marked similarities to claudin-low-like gene expression patterns[30,31]. In the context of these observations, it is pertinent to ask: How much of the claudin-low phenotype is a result of stromal infiltration, and could the claudin-low phenotype simply be an artifact of stromal infiltration? If the claudin-low phenotype were only a sampling artifact, irrelevant to a tumor's biology, one would expect claudin-low tumors to be similarly distributed among the intrinsic subtypes. Claudin-low tumors were, however, overrepresented in basal-like and normal-like tumors, and underrepresented in the remaining intrinsic subtypes. Further, if claudin-lowness were only mediated by stromal infiltration, it should be possible to accurately identify claudin-low tumors solely on the basis of stromal infiltration. However, while almost all claudin-low tumors had high levels of stromal infiltration, only a minority of tumors with high levels of stromal infiltration were in fact classified as claudin-low (Supplementary Data 1, 3, and 4). Finally, numerous studies have identified features in claudin-low tumors (human, murine and cell-line), which are directly attributable to claudin-low tumor cells[6,9,12,14,32]. Therefore, non-tumor cell infiltration is undoubtedly an important feature of the claudin-low tumor microenvironment[33–36], and may even be the feature that induces EMT in claudin-low tumor initiating cells[15]. However, the characteristics observed in claudin-low tumors cannot solely be attributed to non-tumor-cell infiltration.

While we did not find evidence that claudin-low status affects survival, certain claudin-low characteristics may nonetheless be clinically relevant and/or actionable. For example, claudin-low tumors show high levels of immune cell infiltration[8,12], express high levels of *PD-L1*[13], are immunosuppressed by T-regulatory cells[33], and carry low mutational burden;[13,14] these factors may all be relevant for the efficacy of immunotherapeutics in claudin-low tumors. The EMT phenotype in claudin-low tumors may in itself be a therapeutic target, and may also have implications for chemoresistance[37]. Due to the major influence of non-tumor-cell infiltration, it is likely that immunocompetent animal models will be of particular importance for functionally evaluating how these features can be therapeutically targeted[3,13,38,39].

Several limitations of this study should be noted. Despite analyzing over three thousand breast tumors, we identified relatively few claudin-low tumors. The findings presented in this article must therefore be interpreted with a degree of caution. While the Kaplan–Meier curves for the METABRIC cohort (Fig. 2, Supplementary Fig. 8) show clear resemblance to those observed in non-claudin-low tumors[2], the claudin-low cohort was not powered to detect statistically significant differences between groups. Further, it is difficult to ascertain the extent to which the observations from bulk tumor samples represent the characteristics of tumor cells or non-tumor cells[29]. It is therefore likely that single-cell transcriptomic analyses will be required in order to effectively disentangle the features of tumor cells and infiltrating immune and stromal cells. Finally, it must be highlighted that we deliberately chose a biased approach to building the condensed claudin-low gene list. This choice was motivated by our findings (Supplementary Fig. 3) and informed by contemporary studies of claudin-low tumors[6,9,14,15,32]. We therefore stress that there is no gold standard for identifying claudin-low tumors, and that the method presented here may lack external generalizability. Additional approaches to refining claudin-low classification, which could be used in conjunction with the nine-cell line predictor or the method presented here, might include: Immunohistochemical staining of EMT-related protein markers, implementing a cut-off for maximum permitted GII, or checking for overlap with IntClust4 status.

In summary, we have comprehensively analyzed claudin-low breast tumors, and through these analyses substantiated a re-definition of claudin-low as a breast cancer phenotype. Our findings explain the large degree of heterogeneity observed in claudin-low breast tumors, thereby enabling more accurate and nuanced investigations into this poorly understood form of cancer.

Methods

Cohorts. The METABRIC[4,5], Oslo2[23], and TCGA-BRCA[7,25] cohorts were analyzed in this study. Processed data from the METABRIC cohort were downloaded from cBioportal;[40,41] queried data include hormone receptor status, IntClust subtype, disease-specific survival, mutation status in a panel of 173 sequenced genes[5], gene-centric copy number status, and normalized gene expression values. Intrinsic subtypes (identified using the PAM50 predictor[17]) for the METABRIC cohort were retrieved from supplementary files in Curtis et al.[4]. Copy number segments and tumor ploidy were retrieved from the repository associated with Pereira et al.[5]. There were 1886 tumors in the METABRIC cohort with aforementioned data available. Centrally reviewed histological classifications were kindly made available by Dr. Elena Provenzano[42]. Histological classification was available for 1575 tumors in the dataset.

For the Oslo2 cohort, normalized gene expression values, intrinsic subtypes (identified using PAM50) and hormone receptor status were downloaded from the Gene Expression Omnibus (GEO), accession GSE80999. All 381 samples from the cohort were included in the analyses. Analyses were carried out using GEOquery[43] and Biobase[44]. Copy number data were only available for seven claudin-low tumors and was therefore not used in the analyses.

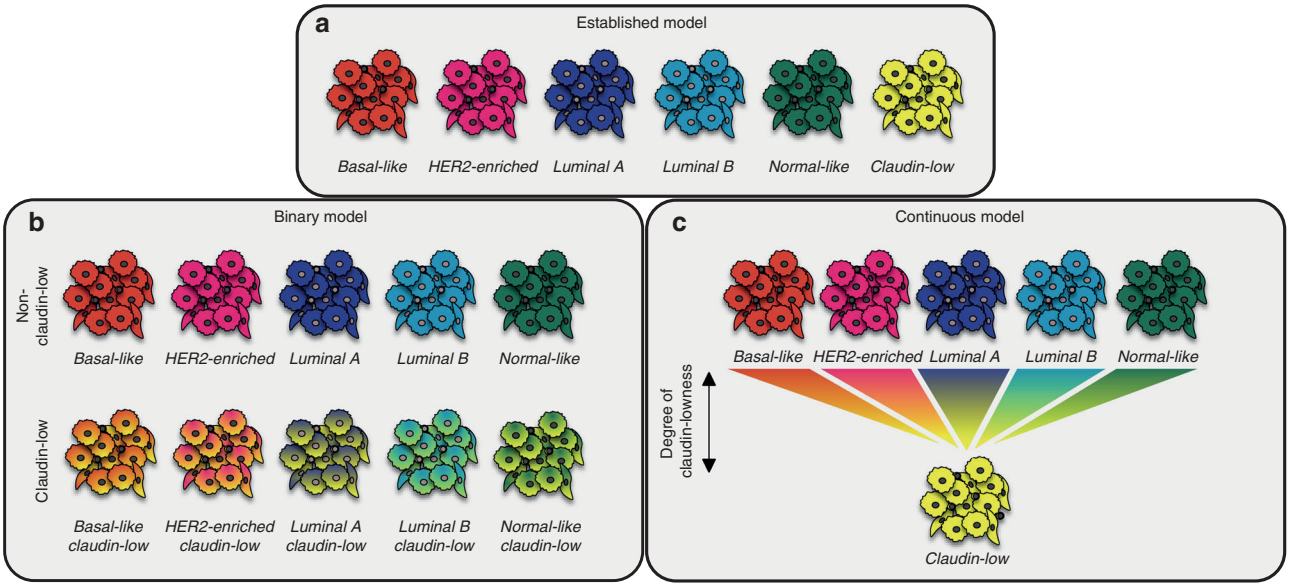


Fig. 6 Re-definition of claudin-low as a breast cancer phenotype. a In the established model, claudin-low is interpreted as a sixth subtype, analogous to the intrinsic subtypes. **b** When stratified by intrinsic subtype, claudin-low tumors, however, show characteristics associated with their intrinsic subtype. This implies that tumors are not claudin-low instead of their intrinsic subtype, rather that tumors can carry a claudin-low phenotype in addition to their intrinsic subtype. In the binary model, a tumor is either classified as claudin-low, or non-claudin-low. **c** The comparative analysis of CoreCL tumors and claudin-low tumors identified by the nine-cell line predictor, indicates that claudin-low may in fact be a continuous feature. Thus, individual tumors may show varying degrees of claudin-lowness, rather than simply being claudin-low or non-claudin-low. In this model, CoreCL tumors, on average, have a higher degree of claudin-lowness than claudin-low tumors identified by the nine-cell line predictor. The continuous model opens for the possibility of pure claudin-low tumors, uncoupled from the intrinsic subtypes.

Normalized gene expression values, intrinsic subtype (identified using PAM50) and histological classification from tumors in the TCGA-BRCA cohort were downloaded from cBioportal[7,25,40,41]. All 1082 tumors from the TCGA-BRCA cohort were analyzed.

Transcriptomic analyses. The generation and pre-processing of gene expression data are described in the cohorts' respective publications[4,5,7,23,25]. Gene expression values were mean centered and scaled (z-score). In the Oslo2 cohort, genes represented by multiple probes were reduced to a single gene expression value by finding the mean of all probes representing the given gene.

Claudin-low tumors were identified using the implementation of the nine-cell line claudin-low predictor[12] in the Genefu[18] package for R[45]. Euclidean distance was used as the distance metric for claudin-low classification. IntClust subtypes in the Oslo2 and TCGA-BRCA cohorts were determined using a gene-expression-based IntClust-classifier[24] implemented in Genefu[18]. Immune and stromal infiltration was inferred from gene expression data using *ImmuneScore* and *StromalScore* derived by ESTIMATE[21].

We observed that the nine-cell line claudin-low predictor was heavily influenced by the effect of non-tumor cell infiltration (Supplementary Fig. 3). This can be related to the marked stromal and immune infiltration in claudin-low tumors[12], and to the partial overlap in gene expression features between stromal tissue and tumors with an EMT phenotype[29–31]. Given these challenges which arose from the unbiased approach used by Prat et al.[12], and the progress made in the understanding of claudin-low tumors[6,14,15,32], we chose to explore a biased approach to identifying claudin-low tumors. The reduced gene set used to identify core claudin-low tumors (Supplementary Table 1) was manually selected on the basis of published characterizations of claudin-low gene expression features and advances in understanding the etiological basis of claudin-low tumors[3,6,8,9,12,14,15,32]. We reasoned that the genes should capture the characteristics unique to claudin-low tumors: Low expression of cell–cell adhesion genes, high expression of EMT genes, and stem-cell like/undifferentiated gene expression pattern. Further, we reasoned that the gene list should not include characteristics that are not unique to claudin-low tumors, such as a low expression of luminal epithelium-related genes. Inclusion of such genes would risk recapitulating the intrinsic subtypes. Hierarchical clustering using the reduced gene list was performed by complete linkage with Euclidean distance as the distance metric. Clustering and visualization were performed using the ComplexHeatmap package[46] for R. The significance of the core claudin-low cluster was evaluated using SigClust[22].

Genomic analyses. GII was derived by dividing the number of copy number aberrant nucleotides by the total number of nucleotides in the genome. GII was ploidy-corrected by defining a segment as copy number aberrant if the copy number state deviated from the nearest integer value for ploidy. All GII values were ploidy-corrected.

Individually analyzed genomic aberrations were chosen according to the following criteria: (1) known function as early driver events;[20,47] (2) among the most frequently observed aberrations in breast cancer;[4,5] (3) significantly different incidence between intrinsic subtypes (χ^2-test $P < 0.05$);[4,5] (4) non-overlap with other selected events (i.e. only one CNA located on 8q24). *TP53* mutation, *PIK3CA* mutation, *MYC* amplification (8q24), and *MDM4* amplification (1q32) were selected for further analysis on the basis of these criteria.

Survival analyses. Survival analyses were performed using the Survival package[48] for R, and Kaplan–Meier plots were generated using the Survminer package.

Statistical analyses. All significance tests (where applicable) were two-tailed. For continuous variables, Wilcoxon rank-sum test and Kruskal–Wallis test were used to test for differences between two or more than two groups, respectively. For categorical variables, Fisher's exact test and χ^2-test were used to test for differences between two or more than two groups, respectively. Significance in survival analyses was determined by log-rank tests. Adjustments were made for multiple hypothesis testing in the analyses detailed in Supplementary Data 2 (Bonferroni-correction); no other corrections were made for multiple hypothesis testing. Whiskers in box-and-whisker plots were generated using the Tukey method; individual data points were not plotted, as the imbalance in sample numbers between groups (Table 1) tended to obscure overall trends.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data used in this study are available through cBioportal[40,41] (METABRIC[4,5], TCGA[7,25]), GSE80999[23], supplementary tables 2 and 3 in Curtis et al.[4] and the repository associated with Pereira et al.[5]. Histological classification of the METABRIC cohort may be available upon request to Mukherjee et al.[42]. Detailed instructions for gathering data can be found in the repository associated with this study. The source data underlying each figure are provided as a Source Data file. All other data are available within the

Article, Supplementary Information files or available from the author upon reasonable request.

Code availability

All code used in the described analyses is available at https://github.com/clfougner/ClaudinLow.

Received: 24 September 2019; Accepted: 16 March 2020;
Published online: 14 April 2020

References

1. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
2. Sørlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci.* **98**, 10869–10874 (2001).
3. 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).
4. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346 (2012).
5. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).
6. Bruna, A. et al. TGFβ induces the formation of tumour-initiating cells in claudin low breast cancer. *Nat. Commun.* **3**, 1055 (2012).
7. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61 (2012).
8. Sabatier, R. et al. Claudin-low breast cancers: clinical, pathological, molecular and prognostic characterization. *Mol. Cancer* **13**, 228 (2014).
9. Dias, K. et al. Claudin-low breast cancer; clinical & pathological characteristics. *PLoS ONE* **12**, e0168669 (2017).
10. Damrauer, J. S. et al. Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. *Proc. Natl Acad. Sci.* **111**, 3110–3115 (2014).
11. Kardos, J. et al. Claudin-low bladder tumors are immune infiltrated and actively immune suppressed. *JCI Insight* **1**, e85902 (2016).
12. Prat, A. et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **12**, R68 (2010).
13. Fougner, C., Bergholtz, H., Kuiper, R., Norum, J. H. & Sørlie, T. Claudin-low-like mouse mammary tumors show distinct transcriptomic patterns uncoupled from genomic drivers. *Breast Cancer Res.* **21**, 85 (2019).
14. Morel, A. P. et al. A stemness-related ZEB1-MSRB3 axis governs cellular pliancy and breast cancer genome stability. *Nat. Med.* **23**, 568–578 (2017).
15. Puisieux, A., Pommier, R. M., Morel, A.-P. & Lavial, F. Cellular pliancy and the multistep process of tumorigenesis. *Cancer Cell* **33**, 164–172 (2018).
16. Prat, A. et al. Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. *Oncologist* **18**, 123–133 (2013).
17. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160 (2009).
18. Gendoo, D. M. A. et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics* **32**, 1097–1099 (2015).
19. Russnes, H. G., Lingjaerde, 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).
20. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
21. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
22. 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).
23. 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).
24. Ali, H. R. et al. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol.* **15**, 431 (2014).
25. Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304 (2018).
26. Weigelt, B. et al. Metaplastic breast carcinomas display genomic and transcriptomic heterogeneity. *Mod. Pathol.* **28**, 340 (2015).
27. Vidal, M. et al. Gene expression-based classifications of fibroadenomas and phyllodes tumours of the breast. *Mol. Oncol.* **9**, 1081–1090 (2015).
28. McFaline-Figueroa, J. L. et al. A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nat. Genet.* **51**, 1389–1398 (2019).
29. Williams, E. D., Gao, D., Redfern, A. & Thompson, E. W. Controversies around epithelial–mesenchymal plasticity in cancer metastasis. *Nat. Rev. Cancer* **19**, 716–732 (2019).
30. 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).
31. 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).
32. Morel, A.-P. et al. EMT inducers catalyze malignant transformation of mammary epithelial cells and drive tumorigenesis towards claudin-low tumors in transgenic mice. *PLoS Genet.* **8**, e1002723 (2012).
33. Taylor, N. A. et al. Treg depletion potentiates checkpoint inhibition in claudin-low breast cancer. *J. Clin. Invest.* **127**, 3472–3483 (2017).
34. Alsuliman, A. et al. Bidirectional crosstalk between PD-L1 expression and epithelial to mesenchymal transition: significance in claudin-low breast cancer cells. *Mol. Cancer* **14**, 149 (2015).
35. Chuck Harrell, J. et al. Endothelial-like properties of claudin-low breast cancer cells promote tumor vascular permeability and metastasis. *Clin. Exp. Metastasis* **31**, 33–45 (2014).
36. Hanahan, D. & Coussens, L. M. Accessories to the crime: functions of cells recruited to the tumor microenvironment. *Cancer Cell* **21**, 309–322 (2012).
37. Nieto, M. A., Huang, R. Y.-J., Jackson, R. A. & Thiery, J. P. EMT: 2016. *Cell* **166**, 21–45 (2016).
38. Pfefferle, A. D. et al. Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. *Genome Biol.* **14**, R125 (2013).
39. Norum, J. H. et al. GLI1 induced mammary gland tumours are transplantable and maintain major molecular features. *Int. J. Cancer* **146**, 1125–1138 (2019).
40. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2**, 401–404 (2012).
41. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
42. Mukherjee, A. et al. Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort. *NPJ breast cancer* **4**, 5 (2018).
43. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846–1847 (2007).
44. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115 (2015).
45. R Core Team. *R: A Language and Environment for Statistical Computing* ISBN 3-900051-07-0 (R Foundation for Statistical Computing, 2017).
46. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
47. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
48. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model* (Springer, 2000).

Acknowledgements

The authors thank Aleix Prat and Ole Christian Lingjærde for insightful discussions and critical reading of the manuscript. We are grateful to Elena Provenzano for providing us with centrally reviewed histological classifications of tumors in the METABRIC cohort, Hege G. Russnes for histopathological support, and the Oslo Breast Cancer Research Consortium (OSBREAC) for access to data from the Oslo2 cohort. C.F., H.B., and J.H.N. are supported by grants from the Norwegian Research Council (163027) and South-Eastern Norway Regional Health Authority (2012056) to T.S.

Author contributions

C.F. conceptualized and designed the study, and performed all analyses. C.F., H.B., J.H.N., and T.S. interpreted the results. J.H.N. and T.S. provided supervision. T.S. acquired funding. C.F. wrote the original manuscript draft. C.F., H.B., J.H.N., and T.S. reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-15574-5.

Correspondence and requests for materials should be addressed to T.Sør.

Peer review information Nature Communications thanks Stephen Chanock and Brian Lehmann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

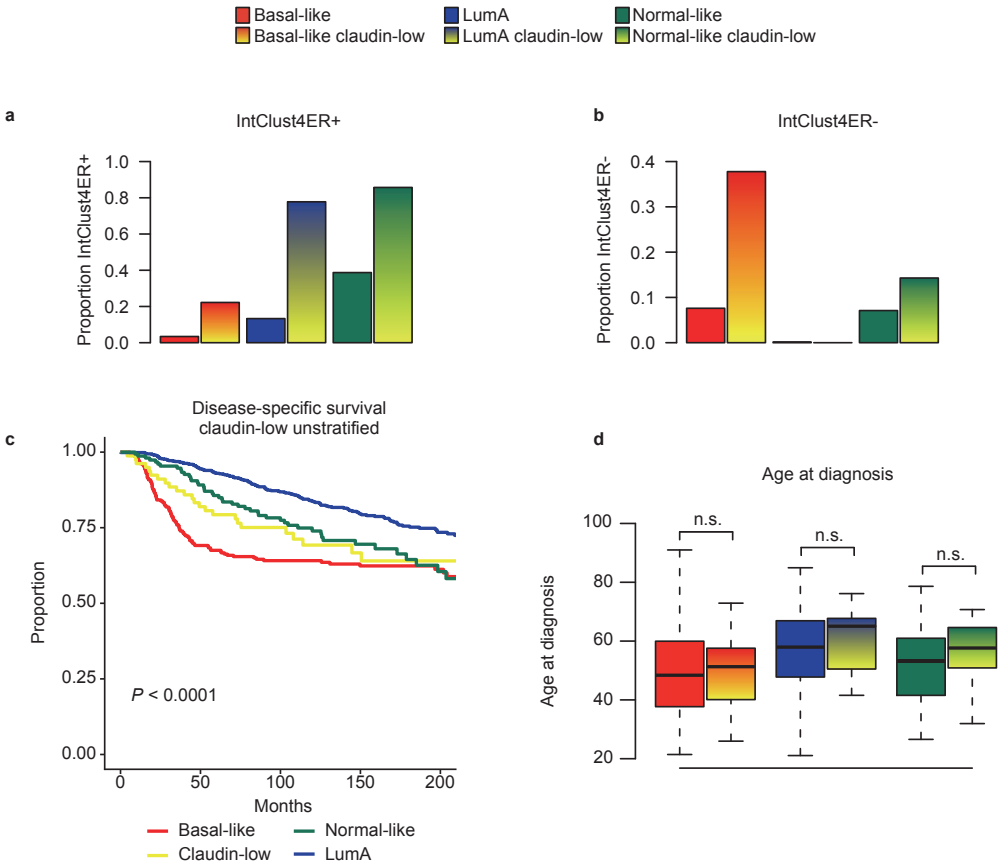
Reprints and permission information is available at http://www.nature.com/reprints

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

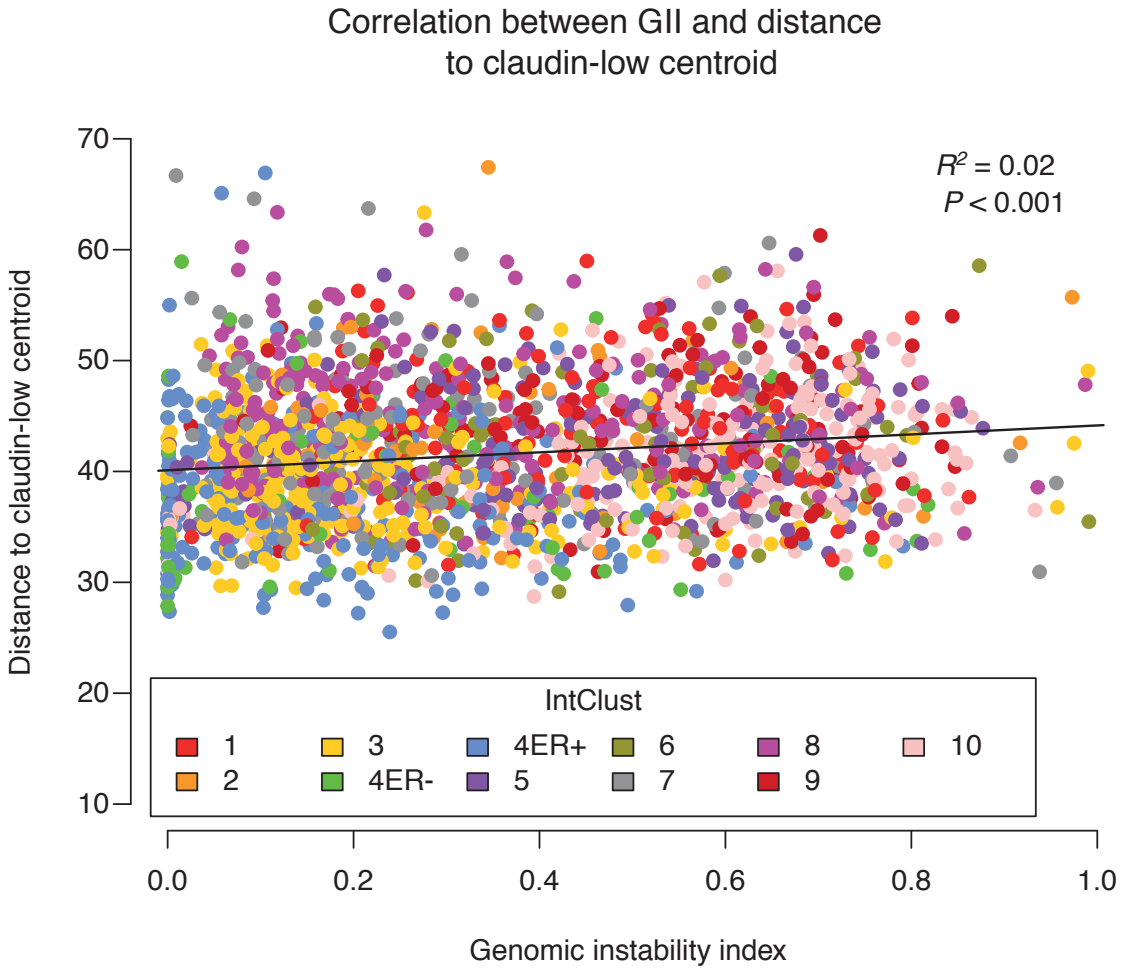
Supplementary information

Re-definition of *claudin-low* as a breast cancer phenotype

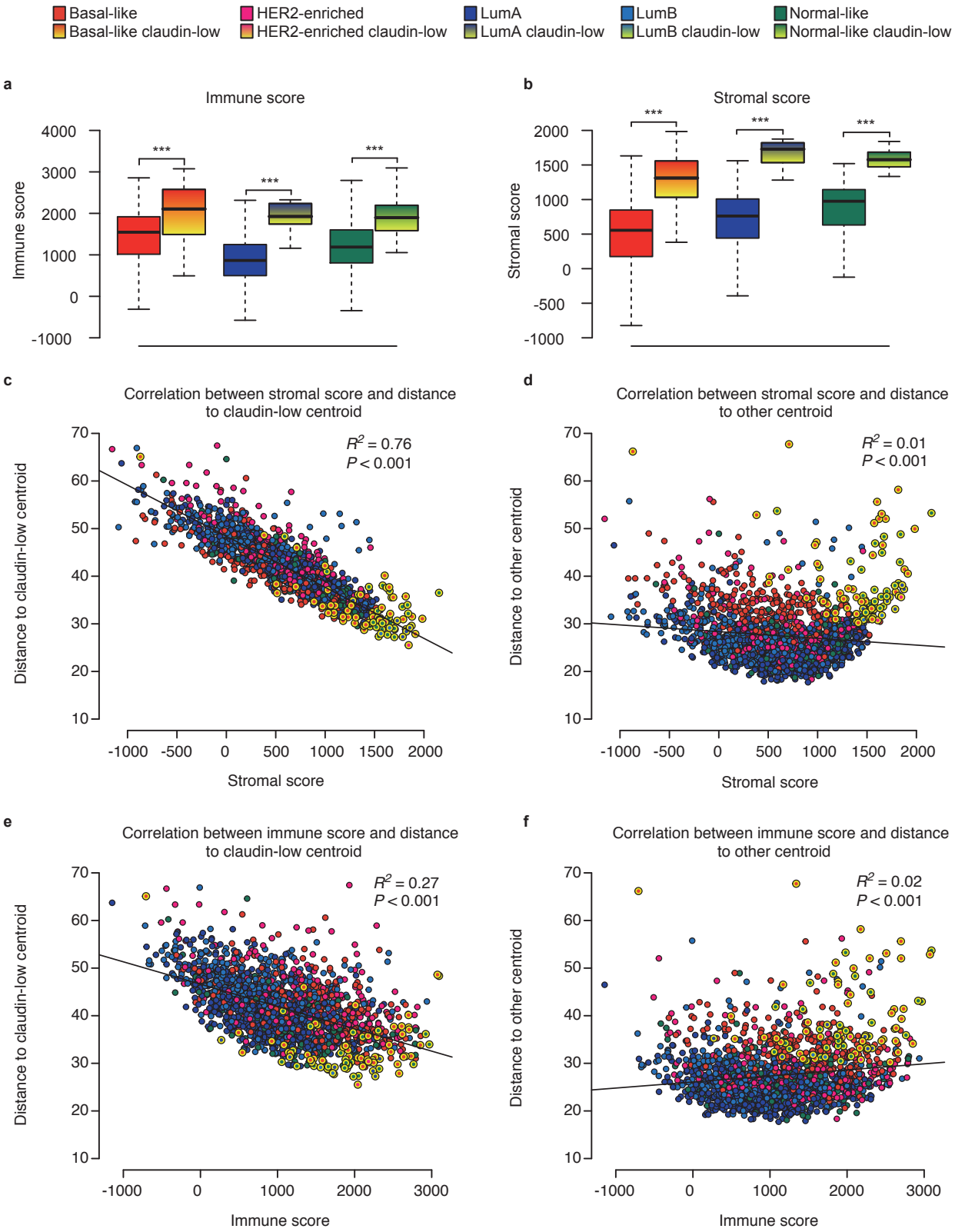
Fougner *et al.*



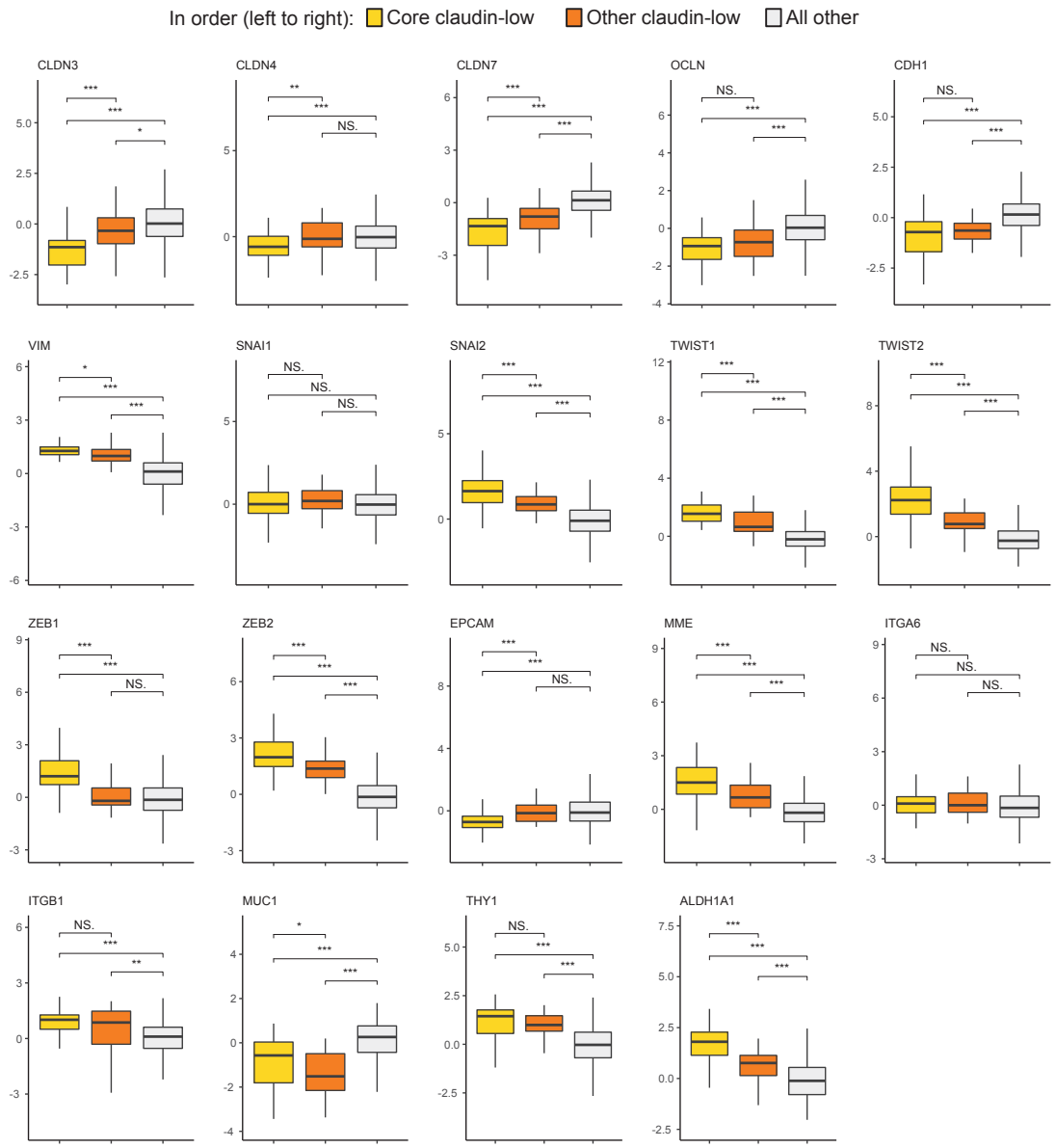
Supplementary Figure 1: Claudin-low tumors are delineated by intrinsic subtype, continued. **a - b** Proportion of tumors in IntClust4ER+ (a) and IntClust4ER- (b) by intrinsic subtype and claudin-low status, in the METABRIC cohort. Basal-like claudin-low tumors tended to be IntClust4ER-, whereas normal-like claudin-low and LumA claudin-low tumors tended to be IntClust4ER+. **c** Disease-specific survival in the METABRIC cohort. Patients with claudin-low tumors are here treated as a single group. The disease-specific survival of patients with claudin-low tumors was superior to that of patients with basal-like tumors, but inferior to that of patients with normal-like and LumA tumors. Disease-specific deaths and sample sizes: Basal-like n = 98 of 263, claudin-low n = 30 of 87, LumA n = 144 of 684, normal-like n = 50 of 155. Difference between groups: $P < 0.0001$, two-tailed log-rank test. **d** Age at diagnosis in claudin-low and non-claudin-low tumors. Claudin-low tumors stratified by intrinsic subtype were diagnosed at significantly different ages ($P = 0.01$, Kruskal-Wallis test), with basal-like claudin-low tumors being diagnosed at a significantly lower age than LumA claudin-low and normal-like claudin-low tumors ($P = 0.01$ and $P = 0.03$ respectively, two-tailed Wilcoxon rank-sum test). Claudin-low and non-claudin-low tumors of the same intrinsic subtype showed similar age at diagnosis (basal-like P = 0.67, LumA P = 0.53, normal-like P = 0.052, two-tailed Wilcoxon-rank-sum test). n.s. $P > 0.05$. Boxplot elements: center line = median, box limits = upper and lower quartiles, whiskers = 1.5 × interquartile range. **All** Sample sizes provided in Table 1. Source data are provided as a Source Data file.



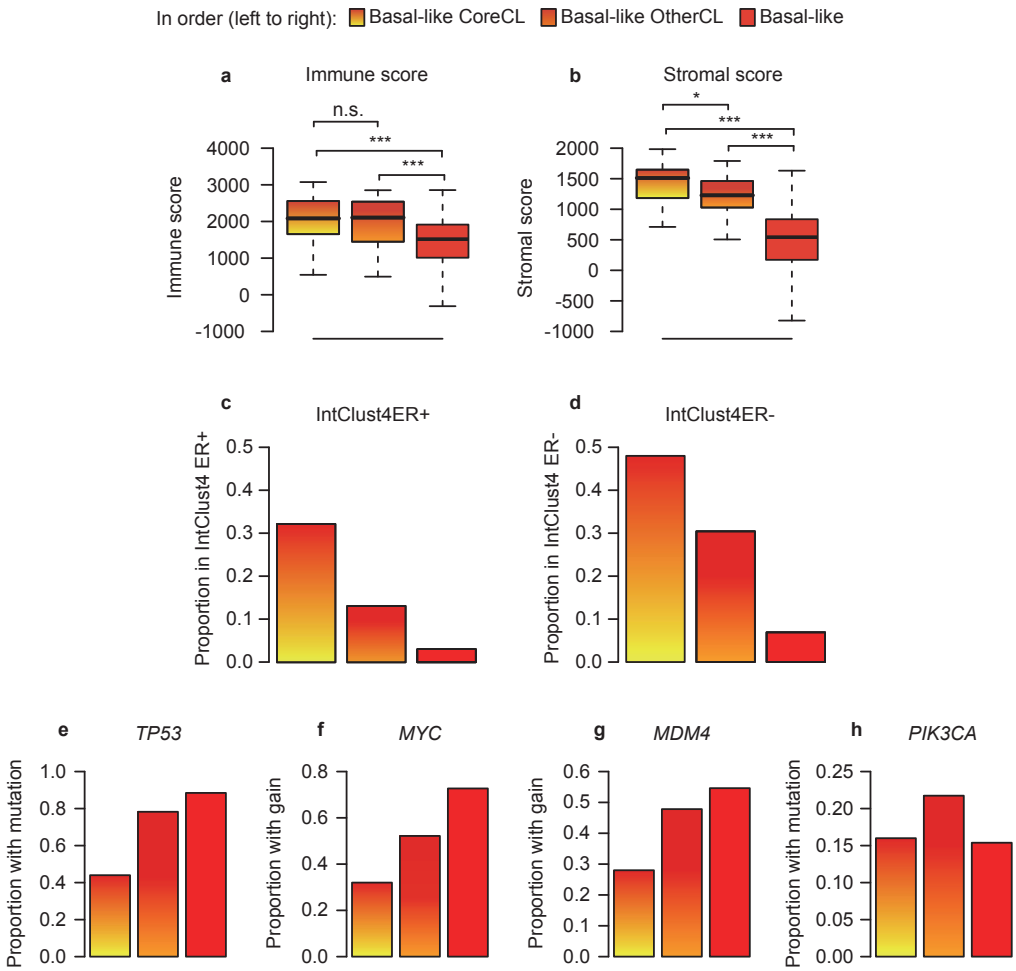
Supplementary Figure 2: Genomic instability index alone does not accurately predict correlation to the claudin-low centroid. Correlation between genomic instability index (GII) and distance to the claudin-low centroid from the nine-cell line claudin-low predictor in the METABRIC cohort. While the correlation between GII and distance to the claudin-low centroid was statistically significant, GII only accounted for 2% of the variance ($P < 0.001$, linear regression). Tumors from all intrinsic subtypes, including HER2-enriched and LumB, are included in the figure. n = 1886 biologically independent samples. Source data are provided as a Source Data file.


Basal-like
Basal-like claudin-low
HER2-enriched
HER2-enriched claudin-low
LumA
LumA claudin-low
LumB
LumB claudin-low
Normal-like
Normal-like claudin-low
a
Immune score
Immune score
b
Stromal score
Stromal score
c
Correlation between stromal score and distance to claudin-low centroid
$R^2 = 0.76$
$P < 0.001$
Distance to claudin-low centroid
Stromal score
d
Correlation between stromal score and distance to other centroid
$R^2 = 0.01$
$P < 0.001$
Distance to other centroid
Stromal score
e
Correlation between immune score and distance to claudin-low centroid
$R^2 = 0.27$
$P < 0.001$
Distance to claudin-low centroid
Immune score
f
Correlation between immune score and distance to other centroid
$R^2 = 0.02$
$P < 0.001$
Distance to other centroid
Immune score

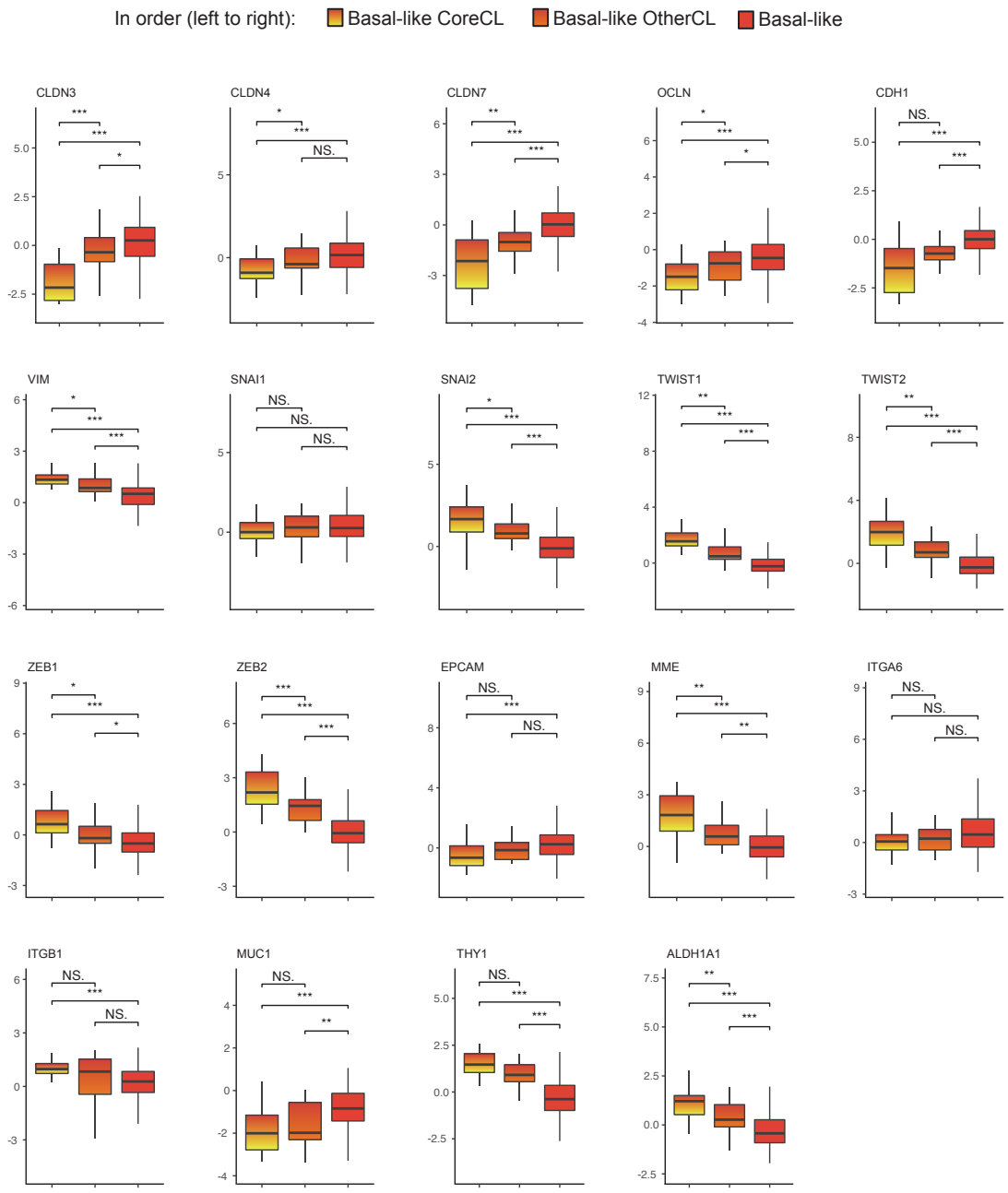
Supplementary Figure 3: Claudin-low tumors show high levels of immune and stromal infiltration. a - b Immune and stromal score, from ESTIMATE, in claudin-low and non-claudin-low tumors in the METABRIC cohort. Claudin-low tumors of all subtypes had higher levels of immune and stromal infiltration than non-claudin-low tumors of the same subtype ($P < 0.001$ for all, two-tailed Wilcoxon rank-sum test). **c - d** Relationship between stromal score and Euclidean distance to the claudin-low centroid (**c**) and the other centroid (**d**) from the nine-cell line predictor (linear regression). An inverse correlation between stromal score and distance to the claudin-low centroid was observed ($R^2 = 0.76$). **e - f** Relationship between immune score and Euclidean distance to the claudin-low centroid (**e**) and the other centroid (**f**) from the nine-cell line predictor (linear regression). An inverse correlation between immune score and distance to the claudin-low centroid was observed ($R^2 = 0.27$). **c – f** $P < 0.001$ for all linear regressions. **All** *** $P < 0.001$. Boxplot elements: center line = median, box limits = upper and lower quartiles, whiskers = 1.5 × interquartile range. Sample sizes provided in Table 1 (HER2-enriched and LumB tumors are included in panels **c** - **f**). Source data are provided as a Source Data file.



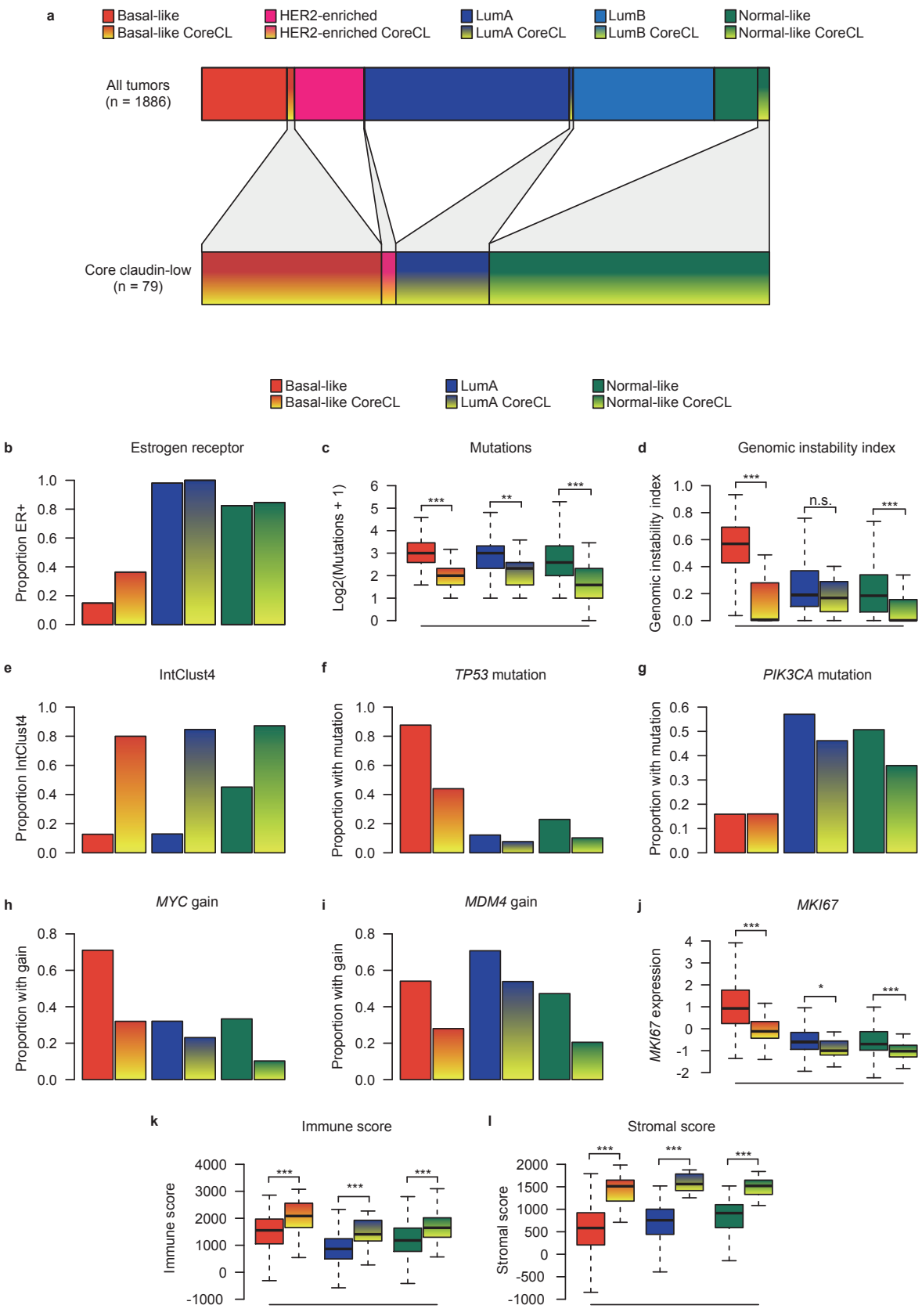
Supplementary Figure 4: CoreCL tumors show gene expression features consistent with the claudin-low phenotype. Gene expression (*log2*) for all 19 genes in the condensed claudin-low gene list, in the METABRIC cohort (n = 1886 biologically independent samples), separated into CoreCL, OtherCL and all other tumors. CoreCL tumors showed gene expression characteristics in line with those previously described for claudin-low tumors. OtherCL tumors showed some claudin-low characteristics, albeit to a lesser degree than CoreCL tumors. Two-tailed Wilcoxon rank-sum test used for significance testing. NS. $P > 0.05$, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Boxplot elements: center line = median, box limits = upper and lower quartiles, whiskers = 1.5 × interquartile range. CoreCL n = 79, OtherCL n = 30, non-claudin-low n = 1777 biologically independent samples. Source data are provided as a Source Data file.



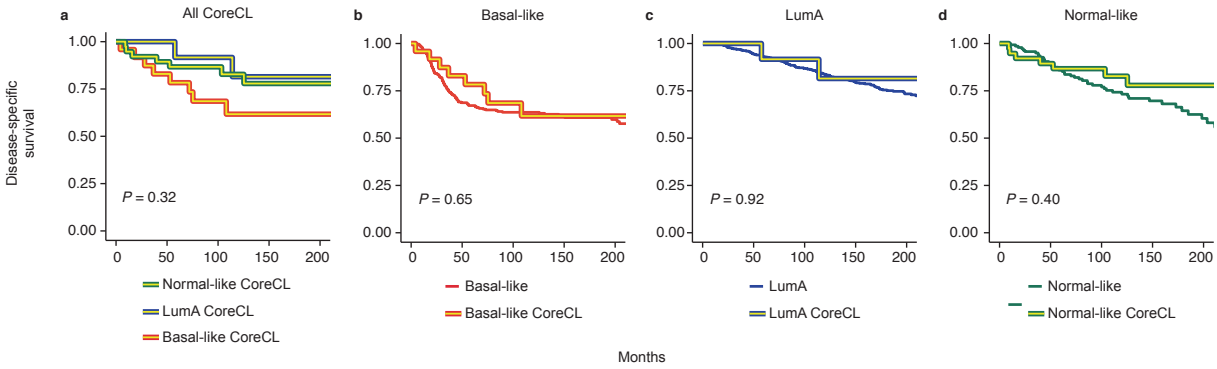
Supplementary Figure 5: Basal-like OtherCL tumors may be inappropriately classified as claudin-low, continued. a - b Immune and stromal score in basal-like core claudin-low tumors, basal-like other claudin-low tumors, and basal-like non-claudin-low tumors in the METABRIC cohort. Basal-like CoreCL and OtherCL tumors showed higher immune and stromal infiltration than basal-like non-claudin-low tumors ($P < 0.001$ for all, two-tailed Wilcoxon rank-sum test). **c - d** Proportion of basal-like tumors in IntClust4ER+ (**c**) and IntClust4ER- (**d**) by claudin-low status. The majority of basal-like CoreCL tumors were classified as IntClust4, with an overweight of tumors not expressing ER. **e - h** Proportion of tumors with mutation or copy number gain in key genes. Basal-like CoreCL tumors showed lower rates of *TP53* mutation (**e**), *MYC* gain (**f**) and *MDM4* gain (**g**) than basal-like OtherCL and non-claudin-low basal-like tumors. This trend was however not evident in the distribution of *PIK3CA* mutations (**h**). **All** n.s. $P > 0.05$, * $P < 0.05$, *** $P < 0.001$. Boxplot elements: center line = median, box limits = upper and lower quartiles, whiskers = 1.5 × interquartile range. Basal-like CoreCL $n = 25$, basal-like OtherCL $n = 23$, basal-like non-claudin-low $n = 260$ biologically independent samples. Source data are provided as a Source Data file.



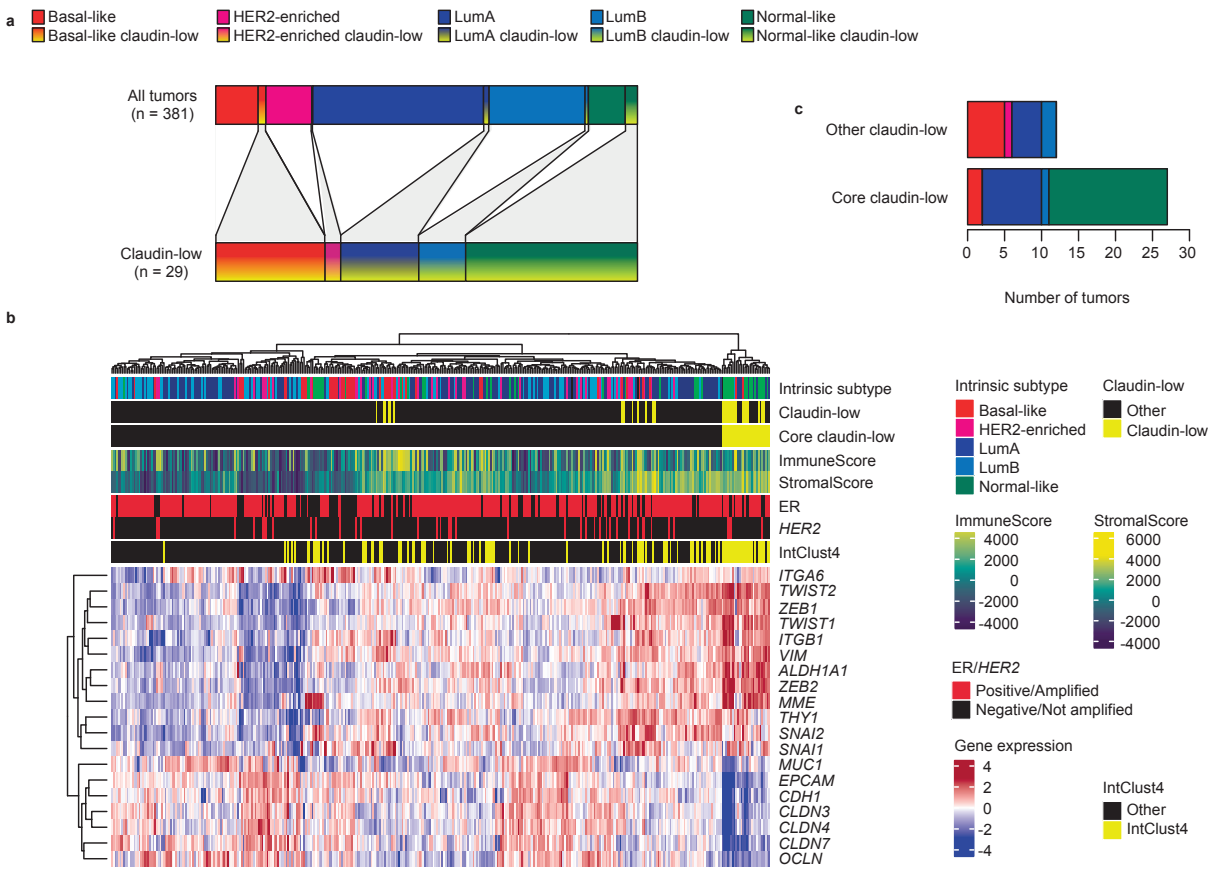
Supplementary Figure 6: Basal-like CoreCL tumors show claudin-low gene expression characteristics. Gene expression (*log2*) for all 19 genes in the condensed claudin-low gene list for basal-like core claudin-low tumors, basal-like other claudin-low tumors and non-claudin-low basal-like tumors in the METABRIC cohort. Two-tailed Wilcoxon rank-sum test used for significance testing. NS. $P > 0.05$, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Boxplot elements: center line = median, box limits = upper and lower quartiles, whiskers = 1.5 × interquartile range. Basal-like CoreCL $n = 25$, basal-like OtherCL $n = 23$, basal-like non-claudin-low $n = 260$ biologically independent samples. Source data are provided as a Source Data file.


a
Basal-like
Basal-like CoreCL
HER2-enriched
HER2-enriched CoreCL
LumA
LumA CoreCL
LumB
LumB CoreCL
Normal-like
Normal-like CoreCL
All tumors
(n = 1886)
Core claudin-low
(n = 79)
Basal-like
Basal-like CoreCL
LumA
LumA CoreCL
Normal-like
Normal-like CoreCL
b
Estrogen receptor
Proportion ER+
c
Mutations
Log2(Mutations + 1)
d
Genomic instability index
Genomic instability index
n.s.
e
IntClust4
Proportion IntClust4
f
TP53 mutation
Proportion with mutation
g
PIK3CA mutation
Proportion with mutation
h
MYC gain
Proportion with gain
i
MDM4 gain
Proportion with gain
j
MKI67
MKI67 expression
k
Immune score
Immune score
l
Stromal score
Stromal score

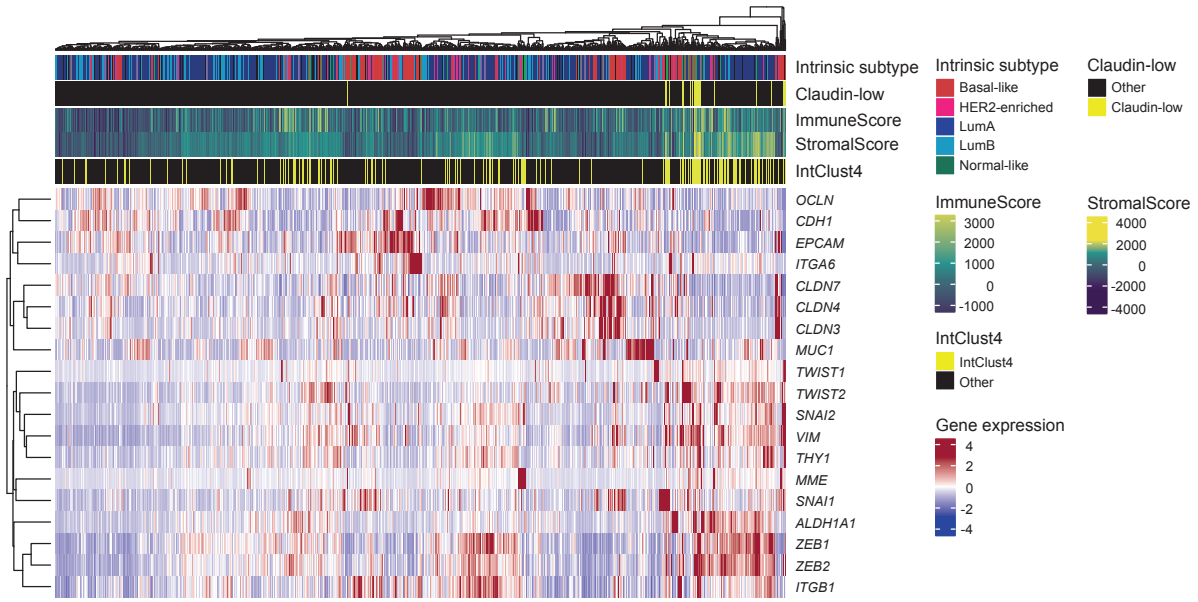
Supplementary Figure 7: CoreCL tumors are more homogeneous than claudin-low tumors identified by the nine-cell line predictor. **a** Distribution of intrinsic subtypes in the METABRIC cohort for all tumors (top bar, $n = 1886$) and for CoreCL tumors only (bottom bar, $n = 79$). **b** Distribution of estrogen receptor-positivity. **c** Number of mutations in the panel of 173 sequenced genes. CoreCL tumors showed lower mutational rates than non-claudin-low tumors of the same subtype (basal-like $P < 0.001$, LumA $P = 0.009$, normal-like $P < 0.001$, two-tailed Wilcoxon rank-sum test). **d** Distribution of genomic instability index (GII). Basal-like and normal-like CoreCL tumors showed lower levels of genomic instability than non-claudin-low tumors of the same subtype (basal-like $P < 0.001$, LumA $P = 0.17$, normal-like $P < 0.001$, two-tailed Wilcoxon rank-sum test). **e** Distribution of allocation to the IntClust4 subtype. **f** Distribution of *TP53* mutations. **g** Distribution of *PIK3CA* mutations. **h** Distribution of *MYC* gain. **i** *MDM4* gain. **j** Distribution of *MKI67* gene expression (*log2*). CoreCL tumors consistently expressed lower levels of *MKI67* compared to non-claudin-low counterparts (basal-like $P < 0.001$, LumA $P = 0.01$, normal-like $P < 0.001$, two-tailed Wilcoxon rank-sum test). **k - l** Distribution of immune and stromal score from ESTIMATE. CoreCL tumors had higher immune and stromal score than non-claudin-low tumors ($P < 0.001$ for all, two-tailed Wilcoxon-rank sum test). **All** There was a reduced variability in the characteristics of basal-like CoreCL tumors compared to basal-like claudin-low tumors as classified by the nine-cell line predictor. The characteristics of LumA CoreCL and normal-like CoreCL tumors were similar to the characteristics of LumA claudin-low and normal-like claudin-low tumors as classified by the nine-cell line predictor. n.s. $P > 0.05$, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Boxplot elements: center line = median, box limits = upper and lower quartiles, whiskers = 1.5 × interquartile range. Here, OtherCL tumors are treated as non-claudin-low. Basal-like CoreCL $n = 25$, basal-like non-claudin-low $n = 283$, HER2-enriched CoreCL $n = 2$, HER2-enriched non-claudin-low $n = 231$, LumA CoreCL $n = 13$, LumA non-claudin-low $n = 680$, LumB CoreCL $n = 0$, LumB non-claudin-low $n = 469$, normal-like CoreCL $n = 39$, normal-like non-claudin-low $n = 144$ biologically independent samples. Source data are provided as a Source Data file.



Supplementary Figure 8: No evidence of CoreCL-status as an indicator of poor prognosis in the METABRIC cohort. a Disease-specific survival in basal-like CoreCL, LumA CoreCL, and normal-like CoreCL tumors in the METABRIC cohort. Survival trends recapitulated the patterns seen in non-claudin-low tumors, although differences in disease-specific survival between CoreCL tumors stratified by intrinsic subtype did not approach statistical significance (P = 0.32, two-tailed log-rank test). **b - d** Disease specific survival in CoreCL and non-claudin-low basal-like (**b**), LumA (**c**) and normal-like (**d**) tumors. Significant differences between CoreCL and non-claudin-low tumors were not found (basal-like P = 0.65, LumA P = 0.92, normal-like P = 0.40, two-tailed log-rank test). **All** Here, OtherCL tumors are treated as non-claudin-low. Basal-like CoreCL n = *25,* basal-like non-claudin-low n = 283, LumA CoreCL n = 13, LumA non-claudin-low n = 680, normal-like CoreCL n = 39, normal-like non-claudin-low n = 144 biologically independent samples. Disease-specific deaths: Basal-like CoreCL n = 8 of 25, basal-like non-claudin-low n = 109 of 283, LumA CoreCL n = 3 of 13, LumA non-claudin-low n = 144 of 680, normal-like CoreCL n = 9 of 39, normal-like non-claudin-low n = 46 of 144. Source data are provided as a Source Data file.



Supplementary Figure 9: Claudin-low tumors in the Oslo2 cohort recapitulate characteristics observed in the METABRIC cohort. a Distribution of intrinsic subtypes in the Oslo2 cohort for all tumors (top bar, $n = 381$ biologically independent samples) and for claudin-low tumors, as defined by the nine-cell line predictor (bottom bar, $n = 29$ biologically independent samples). Most claudin-low tumors were either basal-like, LumA, or normal-like. **b** Heatmap of gene expression values (*log2*) for the condensed claudin-low gene list in the Oslo2 cohort ($n = 381$ biologically independent samples). Hierarchical clustering identified a core claudin-low cluster ($P < 0.001$, SigClust) with similar characteristics to those observed in the METABRIC cohort. Copy number data was not available, however, the representation of IntClust4 in the core claudin-low cluster implies genomic stability in the group. **c** Distribution of subtypes in core and other claudin-low tumors in the Oslo2 cohort. The distribution of subtypes was similar to that seen in the METABRIC cohort, with a slightly larger variation in the intrinsic subtypes of OtherCL tumors. Basal-like CoreCL $n = 2$, basal-like OtherCL $n = 5$, HER2-enriched OtherCL $n = 1$, LumA CoreCL $n = 8$, LumA OtherCL $n = 4$, LumB CoreCL $n = 1$, LumB OtherCL $n = 2$, normal-like CoreCL $n = 16$ biologically independent samples. $n = 1$ CoreCL and $n = 1$ OtherCL sample without available intrinsic subtype. **All** Source data are provided as a Source Data file.



Supplementary Figure 10: No core claudin-low cluster in the TCGA-BRCA cohort. Heatmap of gene expression values (*log2*) for the condensed claudin-low gene list in the TCGA-BRCA cohort (n = 1082 biologically independent samples). No core claudin-low cluster emerged. Source data are provided as a Source Data file.

Gene	EntrezID	Characteristic	Expected regulation in claudin-low
CLDN3	1365	Cell-cell adhesion & tight junction	Down
CLDN4	1364	Cell-cell adhesion & tight junction	Down
CLDN7	1366	Cell-cell adhesion & tight junction	Down
OCLN	100506658	Cell-cell adhesion & tight junction	Down
CDH1	999	Cell-cell adhesion & tight junction	Down
VIM	7431	EMT	Up
SNAI1	6615	EMT	Up
SNAI2	6591	EMT	Up
TWIST1	7291	EMT	Up
TWIST2	117581	EMT	Up
ZEB1	6935	EMT	Up
ZEB2	9839	EMT	Up
EPCAM	4072	Stem cell & epithelial differentiation	Down
MUC1	4582	Stem cell & epithelial differentiation	Down
MME	4311	Stem cell & epithelial differentiation	Up
ITGA6	3655	Stem cell & epithelial differentiation	Up
ITGB1	3688	Stem cell & epithelial differentiation	Up
THY1	7070	Stem cell & epithelial differentiation	Up
ALDH1A1	216	Stem cell & epithelial differentiation	Up

Supplementary Table 1: A condensed claudin-low gene list. A list of 19 genes pathognomonic to the claudin-low phenotype. Characteristics/functions listed in the "Characteristic" column are guiding; listed genes may be representative of multiple functions.

Supplementary Data 1: METABRIC cohort sample information. Table containing relevant annotations for samples and patients in the METABRIC cohort. (.xlsx)

Supplementary Data 2: Comparative rates of mutations and CNAs in claudin-low and non-claudin-low tumors in the METABRIC cohort. No mutations were found at a significantly higher rate in claudin-low tumors, stratified by intrinsic subtype, than in non-claudin-low tumors of the same subtype (two-tailed Fisher's exact test, Bonferroni corrected). In analyses of core claudin-low tumors (see subheading "A condensed gene list refines claudin-low classification"), OtherCL tumors are treated as non-claudin-low. (.xlsx)

Supplementary Data 3: Oslo2 cohort sample information. Table containing relevant annotations for samples and patients in the Oslo2 cohort. (.xlsx)

Supplementary Data 4: TCGA-BRCA cohort sample information. Table containing relevant annotations for samples and patients in the TCGA-BRCA cohort. (.xlsx)

nature research

Corresponding author(s): Therese Sørlie

Last updated by author(s): Feb 20, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No custom code was used to collect data in this study.
Data analysis	All analyses performed in this study are described at www.github.com/clfougner/ClaudinLow All analyses were run in R Version 3.6.0. The R packages used in the analyses are: - Genefu v2.16.0 - ESTIMATE v1.0.13 - ComplexHeatmap v2.0.0 - SigClust v1.1.0 - Biobase v2.44.0 - GEOquery v2.52.0 - Survival v2.44-1.1 - Survminer v0.4.5 - ggsci v2.9 - circlize v0.4.7 - ggplot2 v3.2.1 - ggsignif v0.6.0 - gtools v3.8.1 - gridExtra v2.3

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data used in this study are available through cBioportal (METABRIC [1], TCGA [2]), GSE80999 [3], supplementary tables 2 and 3 in Curtis et al. [4] and the repository [5] associated with Pereira et al. [6] Histological classification of the METABRIC cohort may be available upon request to Mukherjee et al. [7] Detailed instructions for gathering data can be found in the repository [8] associated with this study. The source data underlying each figure are provided as a Source Data file.

[1] https://www.cbioportal.org/study/summary?id=brca_metabric
[2] https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018]
[3] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80999
[4] Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346 (2012).
[5] https://github.com/cclab-brca/mutationalProfiles/tree/master/Data
[6] Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat. Commun. 7, 11479 (2016).
[7] Mukherjee, A. et al. Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort. NPJ breast cancer 4, 5 (2018).
[8] https://github.com/clfougner/ClaudinLow

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All samples with required data publicly available from the METABRIC, Oslo2, and TCGA-BRCA cohorts were included.
Data exclusions	HER2-enriched and Luminal B tumors from the METABRIC cohort were not studied in depth as there were only 2 and 3 claudin-low tumors from each group, respectively (not pre-established).
Replication	N/A as no experiments were performed in this study (only deterministic analyses of publicly available datasets).
Randomization	N/A as this study contained no experimental intervention, and there was consequently no allocation to experimental groups.
Blinding	N/A as there were no experimental interventions to which investigators could be blinded.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Breast cancer patients from three publicly available cohorts were included. These cohorts primarily include female patients sampled from European and North American populations. The detailed characteristics of each cohort are described in their respective publications [1-4]. [1] Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346 (2012). [2] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 490, 61 (2012). [3] Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell 173, 291–304 (2018). [4] 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).
Recruitment	No patients were recruited as part of this study.
Ethics oversight	This study is an analysis of de-identified publicly available data, and no ethical approvals were therefore required. The ethical approvals for the cohorts used in this study are available in their respective publications [1-4]. [1] Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346 (2012). [2] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 490, 61 (2012). [3] Hoadley, K. A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell 173, 291–304 (2018). [4] 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).

Note that full information on the approval of the study protocol must also be provided in the manuscript.