Precision Oncology in Breast Cancer: Multi-Omics Molecular Profiling and DNA Methylation-Based Prognostics

Hohmann, Lennart

2026

Document Version:
Publisher's PDF, also known as Version of record

Precision Oncology in Breast Cancer

Multi-Omics Molecular Profiling and DNA Methylation-Based Prognostics

LENNART HOHMANN
DEPARTMENT OF LABORATORY MEDICINE | FACULTY OF MEDICINE | LUND UNIVERSITY

Precision Oncology in Breast Cancer

Multi-Omics Molecular Profiling and DNA Methylation-Based Prognostics

Lennart Hohmann

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD)
at the Faculty of Medicine at Lund University
to be publicly defended on the 27th of May 2026 at 09.00
in Belfragesalen, BMC, Lund, Sweden

Faculty opponent
Katherine Hoadley, PhD
Associate Professor, Department of Genetics
The University of North Carolina at Chapel Hill, USA

Organization: Lund University

Document name: Doctoral Dissertation **Date of issue:** 2026-05-27

Author: Lennart Hohmann

Title: Precision Oncology in Breast Cancer: Multi-Omics Molecular Profiling and DNA Methylation-Based Prognostics

Abstract: Breast cancer is a biologically heterogeneous disease in which molecular profiling is playing an increasingly important role in treatment decisions. In particular, the PAM50 classifier has been widely adopted to define intrinsic subtypes based on gene expression patterns, yet the biological interpretation of these classifications and their clinical relevance within established clinical groups, such as ER-positive/HER2-negative disease, remain incompletely understood. This thesis aimed to refine the understanding of molecular tumor profiling for biological characterization and risk stratification in early breast cancer by leveraging large, population-representative cohorts and exploring complementary molecular layers. A central finding of this work is that PAM50 subtyping is better understood as a continuum rather than a set of discrete categories. Tumors are positioned across multiple transcriptional programs, where proliferation, steroid hormone signaling, and basal keratin expression, alongside genes selected to define subtype specific features, jointly shape subtype assignment. This challenges the conventional view of PAM50 subtypes as distinct entities and instead supports a model in which subtype labels reflect the relative balance of underlying biological processes. Within this framework, tumors classified as HER2-enriched or Basal-like in ER-positive/HER2-negative breast cancer emerge as clinically relevant subgroups rather than classification artifacts. These tumors display molecular characteristics consistent with their counterparts in HER2-positive disease and triple-negative breast cancer, underscoring that intrinsic subtyping captures biology transcending clinical receptor status. Clinically, they are associated with significantly poorer outcomes following endocrine therapy alone, highlighting an unmet need for alternative treatment strategies. Their molecular profiles further point toward potentially actionable features, including immune infiltration and homologous recombination deficiency. Beyond providing insights into gene expression-based subtyping, this thesis also demonstrates that tumor DNA methylation may provide additional prognostic information in ER-positive/HER2-negative breast cancer, particularly in capturing the risk of late recurrence. This suggests a complementary role for methylation profiling in a clinical context where traditional prognostic factors lose discriminative power over time. Taken together, these findings support a more nuanced understanding of molecular tumor profiling in early breast cancer, where interpretation in the context of clinical subgroup and integration of complementary molecular features capturing different aspects of tumor biology enhance both biological characterization and risk stratification.

Language: English **Number of pages:** 96

ISSN: 1652-8220 **ISBN:** 978-91-8021-884-9

Precision Oncology in Breast Cancer

Multi-Omics Molecular Profiling and DNA Methylation-Based Prognostics

Lennart Hohmann

LUND
UNIVERSITY

Cover illustration front by Christine Noll

Cover photo back by Iñaki Sasiain Casado

Cover text back by Lennart Hohmann

Faculty of Medicine

Department of Laboratory Medicine

ISSN 1652-8220

ISBN 978-91-8021-884-9

Lund University, Faculty of Medicine Doctoral Dissertation Series 2026:86

Printed in Sweden by Media-Tryck, Lund University

Lund, 2026

Table of Contents

Abstract

Breast cancer is a biologically heterogeneous disease in which molecular profiling is playing an increasingly important role in treatment decisions. In particular, the PAM50 classifier has been widely adopted to define intrinsic subtypes based on gene expression patterns, yet the biological interpretation of these classifications and their clinical relevance within established clinical groups, such as ER-positive/HER2-negative disease, remain incompletely understood. This thesis aimed to refine the understanding of molecular tumor profiling for biological characterization and risk stratification in early breast cancer by leveraging large, population-representative cohorts and exploring complementary molecular layers. A central finding of this work is that PAM50 subtyping is better understood as a continuum rather than a set of discrete categories. Tumors are positioned across multiple transcriptional programs, where proliferation, steroid hormone signaling, and basal keratin expression, alongside genes selected to define subtype specific features, jointly shape subtype assignment. This challenges the conventional view of PAM50 subtypes as distinct entities and instead supports a model in which subtype labels reflect the relative balance of underlying biological processes. Within this framework, tumors classified as HER2-enriched or Basal-like in ER-positive/HER2-negative breast cancer emerge as clinically relevant subgroups rather than classification artifacts. These tumors display molecular characteristics consistent with their counterparts in HER2-positive disease and triple-negative breast cancer, underscoring that intrinsic subtyping captures biology transcending clinical receptor status. Clinically, they are associated with significantly poorer outcomes following endocrine therapy alone, highlighting an unmet need for alternative treatment strategies. Their molecular profiles further point toward potentially actionable features, including immune infiltration and homologous recombination deficiency. Beyond providing insights into gene expression-based subtyping, this thesis also demonstrates that tumor DNA methylation may provide additional prognostic information in ER-positive/HER2-negative breast cancer, particularly in capturing the risk of late recurrence. This suggests a complementary role for methylation profiling in a clinical context where traditional prognostic factors lose discriminative power over time. Taken together, these findings support a more nuanced understanding of molecular tumor profiling in early breast cancer, where interpretation in the context of clinical subgroup and integration of complementary molecular features capturing different aspects of tumor biology enhance both biological characterization and risk stratification.

Popular Summary

Breast cancer is the most common cancer among women worldwide. Although all breast tumors arise in the same tissue, they can differ substantially in how they respond to treatment, and whether they return after therapy. Understanding these differences requires looking beyond the clinical characteristics of a tumor, such as its size and grade, and into its underlying biology. Tumor biology can be studied through multiple molecular layers. These include changes to the DNA sequence itself, modifications that regulate how the DNA sequence is read, and the resulting patterns of gene expression, which reflect the activity of genes within the tumor cells.

A key goal in breast cancer research is to use this molecular information to classify patients into groups that are likely to respond differently to treatment, allowing therapy to be better tailored to the individual. In clinical practice, breast tumors are routinely characterized into clinical subgroups by the presence or absence of specific proteins on the tumor cell surface, most importantly the estrogen receptor (ER) and the human epidermal growth factor receptor 2 (HER2). These receptor-status markers guide treatment decisions but do not fully capture the biological heterogeneity of breast tumors, meaning that tumors within the same clinical subgroup can still differ substantially in their underlying biology. In this context, molecular subtyping provides a more refined characterization of tumor biology. The well-established PAM50 classifier assigns a tumor to a molecular subtype based on its expression of 50 specifically selected genes. Yet important questions remain about what these subtypes truly reflect biologically, and what they mean for patients within specific clinical subgroups.

This thesis aimed to refine the understanding of molecular tumor profiling for biological characterization and for assessing the risk of cancer recurrence. A central finding is that PAM50 subtypes should not be understood as clearly distinct biological groups. Although a tumor is assigned to a specific subtype, this classification represents only a simplified approximation. In reality, tumors can show similarity to multiple subtypes at the same time and lie along a spectrum with gradual transitions between them. The assignment of a tumor to a particular subtype therefore reflects the interplay of different biological characteristics, such as how quickly tumor cells divide or how they respond to hormones, rather than a clearly defined tumor type. Within the clinical subgroup of ER-positive/HER2-negative breast cancer, which generally carries a relatively favorable prognosis, this thesis

identifies a subset of tumors with a molecular profile more typical of other clinical subgroups. This thesis shows that these are not misclassified outliers but clinically meaningful subsets, as patients with these tumors have significantly poorer clinical outcomes than the broader group they are assigned to. Their molecular profiles also point toward potentially actionable features, including immune cell infiltration into the tumor and defects in the DNA repair machinery. Finally, this thesis suggests that tumor DNA methylation, chemical modifications of the DNA sequence, can provide prognostic information beyond what clinical characteristics alone can capture. In ER-positive/HER2-negative breast cancer, where the risk of cancer returning many years after treatment remains difficult to predict, methylation profiling shows promise as a complementary tool, particularly for identifying patients at risk of late recurrence.

Taken together, these findings support a more complete approach to molecular tumor profiling in early breast cancer. Different molecular layers capture different aspects of tumor biology and interpreting them in combination and in the right clinical context improves both the biological characterization of tumors and the ability to predict outcomes for individual patients.

Populärwissenschaftliche Zusammenfassung

Brustkrebs ist die häufigste Krebserkrankung bei Frauen weltweit. Obwohl alle Brusttumoren im selben Gewebe entstehen, können sie sich erheblich darin unterscheiden, wie sie auf eine Behandlung ansprechen und ob sie nach einer Therapie zurückkehren. Um diese Unterschiede zu verstehen, muss man über die klinischen Merkmale eines Tumors, wie seine Größe und seinen Differenzierungsgrad, hinausblicken und seine zugrunde liegende molekulare Biologie betrachten. Die Biologie eines Tumors lässt sich auf mehreren molekularen Ebenen untersuchen. Dazu gehören Veränderungen der DNA-Sequenz selbst, Modifikationen, die regulieren, wie die DNA-Sequenz abgelesen wird, sowie die daraus resultierenden Muster der Genexpression, die die Aktivität der Gene in den Tumorzellen widerspiegeln.

Ein zentrales Ziel der Brustkrebsforschung ist es, diese molekularen Informationen zu nutzen, um Patientinnen in Gruppen einzuteilen, die unterschiedlich auf eine Behandlung ansprechen, sodass die Therapie besser auf die einzelne Patientin abgestimmt werden kann. In der klinischen Praxis werden Brusttumore routinemäßig anhand des Vorhandenseins oder Fehlens bestimmter Proteine auf der Tumorzelloberfläche charakterisiert, allen voran des Östrogenrezeptors (ER) und des humanen epidermalen Wachstumsfaktorrezeptors 2 (HER2). Diese Rezeptorstatus-Marker leiten zwar Therapieentscheidungen, erfassen jedoch nicht die gesamte biologische Heterogenität von Brusttumoren. Das bedeutet, dass Tumore innerhalb derselben Rezeptorstatus-definierten Gruppe sich stets noch in ihrer zugrunde liegenden Biologie erheblich unterscheiden können. In diesem Zusammenhang spielt die Klassifizierung von Tumoren anhand von molekularen Eigenschaften eine wichtige Rolle. Die etablierte PAM50-Klassifikation von Brustkrebs, basierend auf der Aktivität von 50 spezifisch ausgewählten Genen, ordnet einen Tumor einem von fünf molekularen Subtypen zu. Trotz der regelmäßigen Anwendung bleiben wichtige Fragen offen: Was spiegeln diese Subtypen biologisch tatsächlich wider und welche Bedeutung haben sie für Patientinnen innerhalb spezifischer, durch den Rezeptorstatus-definierter Untergruppen.

Ziel dieser Arbeit war es, besser zu verstehen, wie Tumore anhand ihrer molekularen Eigenschaften biologisch eingeordnet und hinsichtlich ihres Risikos

für ein Wiederauftreten der Erkrankung bewertet werden können. Ein zentrales Ergebnis ist, dass PAM50-Subtypen nicht als klar biologisch voneinander abgegrenzte Gruppen verstanden werden sollten. Zwar wird ein Tumor einem bestimmten Subtyp zugeordnet, diese Zuordnung stellt jedoch nur eine vereinfachte Annäherung dar. In Wirklichkeit können Tumoren mehreren Subtypen gleichzeitig ähneln und liegen auf einem Spektrum mit fließenden Übergängen. Die Einteilung eines Tumors in einen bestimmten Subtyp spiegelt daher das Zusammenspiel verschiedener biologischer Eigenschaften wider, zum Beispiel wie schnell sich die Tumorzellen teilen oder wie sie auf Hormone reagieren, und nicht einen klar abgegrenzten Tumortyp. Weiterhin zeigt diese Arbeit, dass der ER-positive/HER2-negative Brustkrebs, obwohl er im Allgemeinen eine relativ gute Prognose hat, zwei PAM50-definierte Untergruppen umfasst, die mit einem deutlich ungünstigeren Krankheitsverlauf verbunden sind. Die molekularen Profile dieser Tumore deuten zudem auf potenziell therapeutisch nutzbare Eigenschaften hin, darunter Immunzellinfiltration und Defekte in DNA-Reparaturmechanismen. Schließlich legt diese Arbeit nahe, dass die DNA-Methylierung von Tumoren, also chemische Modifikationen der DNA-Sequenz, prognostische Informationen liefern kann, die über das hinausgehen, was klinische Merkmale allein erfassen können. Bei ER-positiven/HER2-negativen Brustkrebs ist das Risiko eines Krankheitsrückfalls viele Jahre nach der Behandlung schwer vorherzusagen. Hier zeigt die Erfassung von Mustern der DNA-Methylierung das Potenzial, Patientinnen mit einem Risiko für ein Spätrezidiv zu identifizieren.

Zusammengenommen unterstützen diese Erkenntnisse einen umfassenderen Ansatz zur molekularen Tumorprofilierung bei frühem Brustkrebs. Verschiedene molekulare Ebenen erfassen unterschiedliche Aspekte der Tumorbiologie, und ihre kombinierte Interpretation im richtigen klinischen Kontext verbessert sowohl die biologische Charakterisierung von Tumoren als auch die Fähigkeit, Behandlungsergebnisse für einzelne Patientinnen vorherzusagen.

Scientific Publications

Papers Included in this Thesis

Paper I

Perturbation and stability of PAM50 subtyping in population-based primary invasive breast cancer

Srinivas Veerla, **Lennart Hohmann**, Deborah F. Nacer, Johan Vallon-Christersson, Johan Staaf

npj Breast Cancer 2023; 9:83

Paper II

Genomic characterization of the HER2-enriched intrinsic molecular subtype in primary ER-positive HER2-negative breast cancer

Lennart Hohmann, Kristin Sigurjonsdottir, Ana Bosch Campos, Deborah F. Nacer, Srinivas Veerla, Frida Rosengren, Poojaswini T. Reddy, Jari Häkkinen, Nicklas Nordborg, Johan Vallon-Christersson, Yasin Memari, Daniella Black, Ramsay Bowden, Helen R. Davies, Åke Borg, Serena Nik-Zainal, Johan Staaf

Nature Communications 2025; 16:2208

Paper III

Molecular profiling of the Basal-like intrinsic molecular subtype in primary ER-positive HER2-negative breast cancer

Lennart Hohmann, Deborah F. Nacer, Mattias Aine, Yasin Memari, Daniella Black, Ramsay Bowden, Helen R. Davies, Åke Borg, Johan Vallon-Christersson, Serena Nik-Zainal, Johan Staaf

Genome Medicine 2025; 17:146

Paper IV

DNA methylation for recurrence risk stratification in early-stage breast cancer

Lennart Hohmann, Johan Vallon-Christersson, Åke Borg, Aurélien Latouche, Johan Staaf

Manuscript

Author's Contributions to Papers

Listed following the CRediT (Contributor Roles Taxonomy) system.

Paper I

Methodology, Investigation, Formal analysis, Data curation, Visualization, Writing - original draft, Writing - review & editing.

Paper II

Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Visualization, Writing - original draft (lead), Writing - review & editing.

Paper III

Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Visualization, Writing - original draft (lead), Writing - review & editing.

Paper IV

Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Visualization, Project administration, Writing - original draft (lead), Writing - review & editing.

Papers Not Included in this Thesis

Homologous recombination deficiency in primary ER-positive and HER2-negative breast cancer

Helen R. Davies , Daniella Black, Anders Kvist, Kristín Sigurjónsdóttir, Ana Bosch, Ramsay Bowden, Yasin Memari, Ziqian Chen, Giuseppe Rinaldi, Frida Rosengren, Deborah F. Nacer, Srinivas Veerla, **Lennart Hohmann**, Nicklas Nordborg, Jari Häkkinen, Johan Vallon-Christersson, Åke Borg, Serena Nik-Zainal, Johan Staaf

Communications Medicine 2026; 6:118

Abbreviations

ADC	Antibody Drug Conjugate
ASCAT	Allele-Specific Copy Number Analysis of Tumors
ATAC	Assay for Transposase-Accessible Chromatin
AUC	Area under the Curve
Basal	Basal-like (PAM50 Subtype)
CDK4/6	Cyclin-Dependent Kinases 4 and 6
CpG	Cytosine-Phosphate-Guanine
DNA	Deoxyribonucleic Acid
DRFI	Distant Recurrence-Free Interval
DWR	Death Without Recurrence
ER	Estrogen Receptor
HER2	Human Epidermal Growth Factor Receptor 2
HER2E	HER2-Enriched (PAM50 Subtype)
HR	Hormone Receptor
HRD	Homologous Recombination Deficiency
IBS	Integrated Brier Score
ICI	Immune Checkpoint Inhibitor
IDFS	Invasive Disease-Free Survival
IHC	Immunohistochemistry
Indel	Insertion/Deletion
LumA	Luminal A (PAM50 Subtype)
LumB	Luminal B (PAM50 Subtype)
MeRS	Methylation Risk Score
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium

mRNA	Messenger RNA
OS	Overall Survival
PAM50	Prediction Analysis of Microarray 50
PARP	Poly(ADP-ribose) Polymerase
PD-1	Programmed Cell Death Protein 1
PD-L1	Programmed Death-Ligand 1
PR	Progesterone Receptor
RFI	Recurrence-Free Interval
RNA	Ribonucleic Acid
SCAN-B	Sweden Cancerome Analysis Network – Breast
SNV	Single Nucleotide Variant
SNP	Single Nucleotide Polymorphism
TCGA	The Cancer Genome Atlas
TILs	Tumor-Infiltrating Lymphocytes
TMB	Tumor Mutational Burden
TNBC	Triple-Negative Breast Cancer
TNM	Tumor, Node, Metastasis
WGS	Whole-Genome Sequencing

Declaration of Generative AI Use

In compliance with the guidelines on the use of generative AI issued by the Research Studies Board at the Faculty of Medicine, Lund University, I report the limited use of generative artificial intelligence tools for proofreading and improving the clarity of scientific text, as well as to support searching for relevant literature. The scientific content, interpretations, and conclusions were developed independently. I critically reviewed and revised AI-generated suggestions and take full responsibility for the content of this thesis.

Introduction

Cancer

Almost everyone is touched by cancer, either through personal experience or through its impact on others. The fear of a cancer diagnosis is a shadow that modern medicine has never fully dispelled, shaped by centuries of uncertainty about the disease and its outcomes. Cancer has accompanied humanity throughout history, first recorded in ancient Egyptian texts and later described by Greek physicians [1]. From early efforts to remove visible tumors to the rise of cellular pathology and the continued unraveling of tumor biology at the molecular level, each era has deepened our understanding [2-4]. Yet despite this long history of inquiry and the scientific progress it has generated, the disease persists.

Cancer arises from our own cells, subverting the biological processes that sustain life in the first place. The capacity to grow, to adapt, and to survive, the very strengths that make our cells resilient, are the same ones cancer exploits. These properties make the disease exceptionally difficult to defeat and relate to the largely hard-won and incremental nature of progress, with only occasional breakthroughs.

Cancer is not a single disease, but a complex group of disorders unified by one defining characteristic: the uncontrolled proliferation of cells that have escaped the regulatory mechanisms governing normal tissue, giving rise to a tumor. What distinguishes malignant from benign tumors, and ultimately defines cancer, is the ability to invade surrounding structures and, in advanced disease, to disseminate to distant organs through metastasis.

Virtually every cell type capable of proliferation in the body can undergo malignant transformation, with the cell of origin determining the type of cancer that arises. This transformation, known as tumorigenesis, does not occur as a single event but through a multi-step process in which normal cells progressively acquire alterations that collectively drive malignant behavior [5].

To bring conceptual order to this complexity, Hanahan and Weinberg proposed the hallmarks of cancer framework, defining a set of functional capabilities that cells acquire during tumor development (**Figure 1**) [6-8]. One of the most fundamental capabilities is sustaining proliferative signaling, in which cancer cells dysregulate cell division by chronically activating growth-promoting signaling pathways while evading growth suppressors that normally limit proliferation. Malignant cells

further escape normal constraints on survival by resisting cell death and by enabling replicative immortality. Another hallmark is the deregulation of cellular metabolism, through which cancer cells adapt metabolic pathways to support the biosynthetic and energetic demands of rapid proliferation. Tumors must also exploit host systems, inducing or accessing vasculature to secure sufficient oxygen and nutrients while avoiding immune destruction by evading or suppressing immune surveillance. By unlocking phenotypic plasticity cancer cells are able to escape normal differentiation programs and adopt alternative cellular states. Finally, cancer cells acquire the capacity to activate invasion and metastasis, enabling dissemination from the primary tumor and colonization of distant organs. The acquisition of these capabilities is facilitated by enabling characteristics. Genome instability and mutation generate the genetic diversity that increases the likelihood of acquiring advantageous alterations, while tumor-promoting inflammation can shape a microenvironment rich in cytokines, growth factors, and proteases that further support tumor progression. Non-mutational epigenetic reprogramming enables rapid transcriptional adaptation without changes in DNA sequence and at the microenvironmental level, interactions with microbiomes and the secretory activity of senescent cells have also been proposed to influence tumor development and progression.

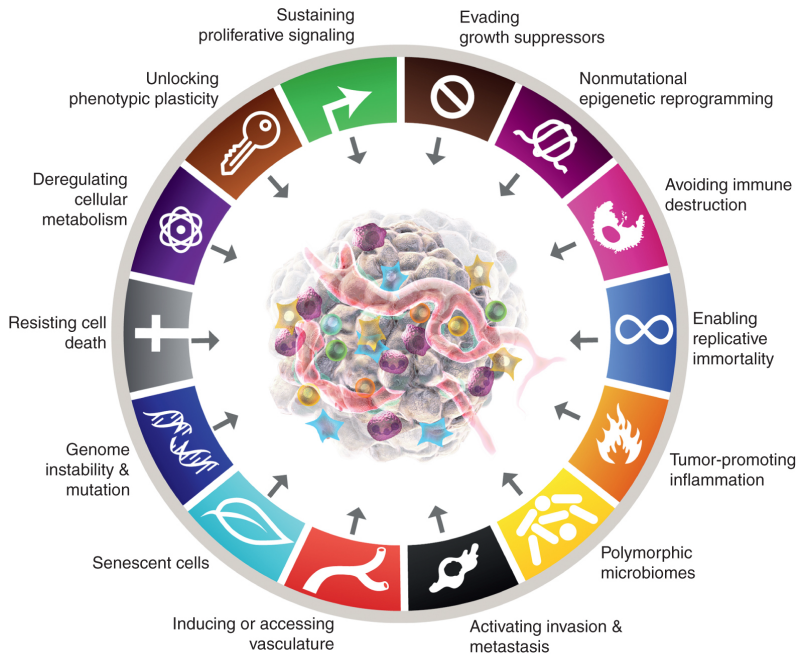

Figure 1. The hallmarks of cancer.

Source: Reproduced with permission from Hanahan et al. [8].

But even within a shared set of hallmark capabilities, cancer is characterized by profound heterogeneity, both between patients with the same cancer type and within a single tumor itself. Inter-tumor heterogeneity reflects the fact that different patients accumulate distinct combinations of alterations during and after tumorigenesis, producing tumors with different molecular profiles, with implications for clinical behavior, and responses to treatment [9]. Intra-tumor heterogeneity arises because tumor cells continue to evolve after initiation. As a tumor grows, individual cells acquire new alterations, giving rise to genetically distinct subpopulations, or clones, that coexist within the same tumor mass [10]. Through clonal expansion, those subpopulations with the greatest growth advantage outcompete others, driving tumor progression and adaptation. This process is governed by principles analogous to Darwinian evolution: heritable alterations arise continuously in individual cells, and natural selection acts on the resulting diversity, favoring those with the greatest capacity to proliferate and survive [11]. This continuous evolutionary process means that by the time a tumor is diagnosed it is rarely a uniform entity but a dynamic ecosystem of competing clones.

This ecosystem extends beyond the cancer cells themselves. The tumor microenvironment, comprising the stromal, immune, and vascular cells that surround and infiltrate the tumor, is an active participant in tumor biology. Far from being passive bystanders, these cells communicate continuously with the tumor, shaping its growth, its capacity to spread, and its relationship with the immune system. The result is a highly complex and dynamic tissue, one in which the behavior of the cancer is influenced not only by its own molecular alterations but by the environment it has created around itself [12].

Understanding cancer at this level of complexity requires looking to the molecular alterations that initiate and sustain malignant transformation. At the heart of every cancer is a disrupted molecular program, one written across multiple layers of cellular information. These include alterations to the DNA sequence and its structural organization, changes to the epigenetic marks that govern how that sequence is read, and the downstream consequences reflected in gene expression, protein production, and cellular metabolism. Together, these layers constitute the molecular architecture of cancer and deciphering them is fundamental to understanding why tumors behave the way they do and how patients might be better diagnosed, prognosticated, and treated.

Breast Cancer

This thesis is focused on breast cancer, which arises from the malignant transformation of epithelial cells of the mammary gland, primarily from the ductal and lobular epithelium (**Figure 2**). Although arising from the same tissue of origin, tumors can vary markedly in their molecular characteristics [13]. As a result, they may differ in their clinical behavior, including sensitivity or resistance to specific treatments and overall patient prognosis.

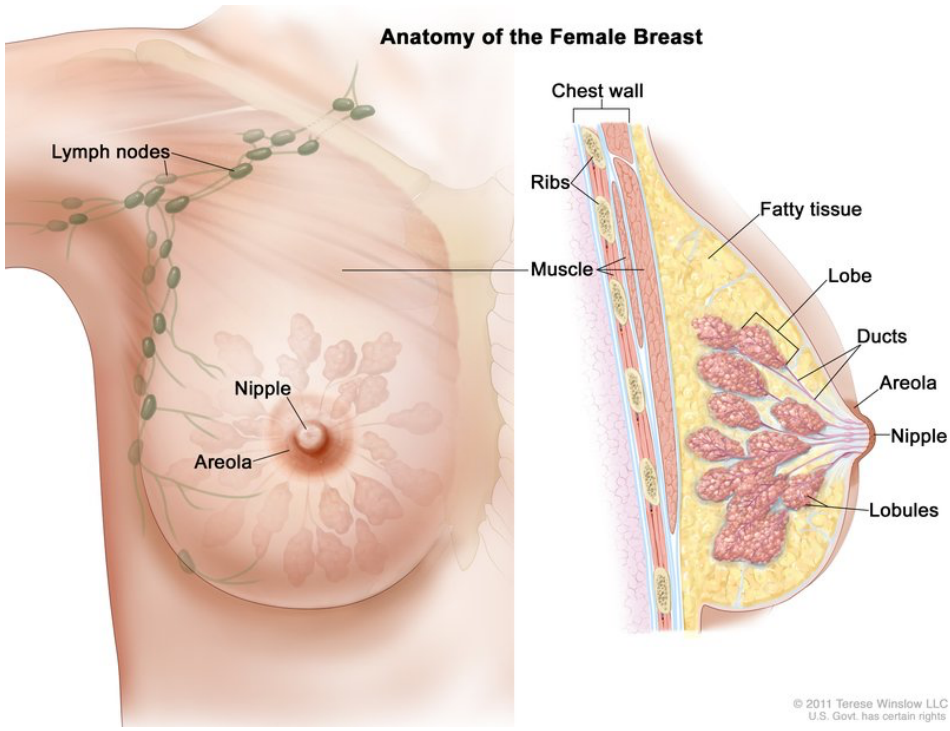

Figure 2. The anatomy of the female breast.

Source: The National Cancer Institute © 2011 Terese Winslow LLC, U.S. Govt. has certain rights. Reproduced with permission from the copyright holder.

Epidemiology

Breast cancer occurs in both sexes, but it almost exclusively affects women. Over her lifetime, a woman faces an estimated risk of about one in eight of developing breast cancer, making it the most diagnosed cancer in women worldwide (**Figure 3**) [14]. Despite substantial advances in early detection and systemic therapies over recent decades, breast cancer remains the leading cause of cancer-related mortality in women, reflecting its high incidence and the persistent clinical challenges posed by aggressive forms of the disease [14, 15].

Survival outcomes in breast cancer are generally encouraging, with approximately 91% of patients alive at 5 years and 86% at 10 years after diagnosis [14]. However, unlike many cancers where long-term survival is considered a functional cure, breast cancer carries a persistent risk of recurrence for decades after initial treatment. Patients who appear disease-free years after diagnosis remain at risk, with recurrence rates between 5 and 20 years after diagnosis ranging from 10% to 41% depending on tumor characteristics [16]. This pattern of late recurrence is one of the most clinically challenging aspects of the disease.

Breast cancer poses a global health issue, though its incidence and outcomes vary substantially between countries due to differences in lifestyle, healthcare access, early detection programs, and treatment options [17]. In high-income countries, such as Sweden, comprehensive screening initiatives and state-of-the-art treatment protocols have contributed to earlier detection and a steady decline in breast cancer mortality over the past few decades, where it currently ranks as the second leading cause of cancer-related mortality in women [18]. In lower-income countries however, limited access to early detection and treatment often results in later-stage diagnoses and higher mortality rates [17, 19].

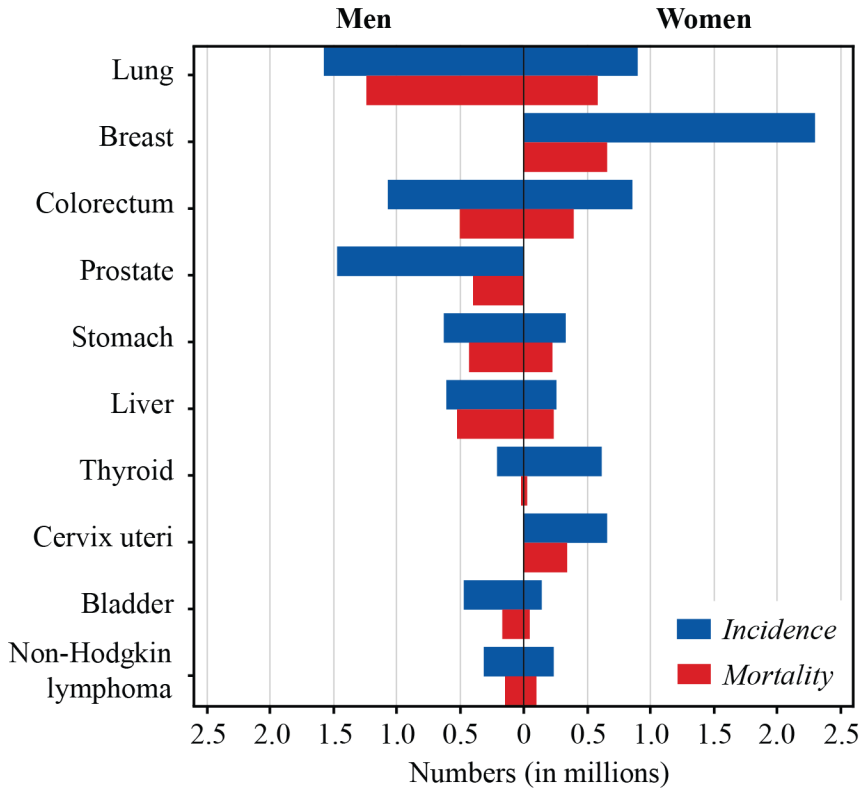

Figure 3. Incidence and mortality of the top ten most common cancers worldwide based on absolute numbers for the year 2022.

Source: World Health Organization Global Cancer Observatory (https://gco.iarc.fr) [20].

Risk Factors

Several factors influence a woman's likelihood of developing breast cancer. Increasing age is one of the most important determinants of breast cancer risk. Another major group of risk factors relates to prolonged exposure to endogenous estrogen and progesterone. These include an early onset of menstruation, late menopause, and the use of hormone replacement therapy, all of which extend the duration of hormone exposure over a lifetime [21]. By contrast, factors that reduce cumulative hormone exposure, such as pregnancy and breastfeeding, are associated with a lower risk of developing breast cancer [22]. Factors related to lifestyle that increase the likelihood of developing breast cancer include alcohol consumption, physical inactivity, and obesity [23].

A woman's risk of developing breast cancer can be strongly influenced by hereditary factors as well. Family history of the disease often reflects an underlying

genetic predisposition, in which alterations in genes associated with tumorigenesis are passed from parent to child [24]. These are known as pathogenic germline variants, and they are responsible for approximately 5-10 % of all breast cancers [25, 26]. Many of the genes implicated in hereditary breast cancer are tumor suppressor genes, such as *BRCA1* and *BRCA2* that help maintain genomic stability by repairing DNA damage. According to the two-hit hypothesis, a cell in which one copy of a tumor suppressor gene is already altered by inheritance needs a second pathogenic alteration in the remaining normal copy before the gene's protective function is lost. In addition to increasing susceptibility, these inherited gene alterations can influence the effectiveness of certain treatments and necessitate more rigorous monitoring and preventive strategies for those affected [27].

Diagnosis

The initial symptoms a breast cancer patient experiences are commonly varied, reflecting the diversity of the disease. Breast cancer may not present any noticeable symptoms at all, which underscores the importance of regular screening. When symptoms do arise, they typically include a noticeable lump or changes in the size or shape of the breast. In advanced stages, symptoms may become more severe and include persistent pain and swelling in nearby lymph nodes, e.g., in the armpit [28].

Diagnostic imaging techniques are important in the identification and initial assessment of breast cancer. For screening, mammography is the primary tool of choice, which is based on using low-dose X-rays to detect abnormalities or tumors before they become palpable [29]. In Sweden, routine mammography check-ups are recommended every two years for women over 40 [30]. Ultrasound is used alongside mammography to further investigate suspicious findings and distinguish between solid tumors and fluid-filled cysts. For more complex cases, magnetic resonance imaging helps to assess the extent of the disease and evaluate dense breast tissues which are less visible on mammograms.

Following the initial imaging, biopsy techniques confirm the presence of breast cancer but also provide information about tumor characteristics, which are vital for determining the most effective treatment plan. A biopsy involves the removal of a small sample of breast tissue for examination by a pathologist. The most common types are fine needle aspiration, which uses a thin needle to remove cells from a suspicious area, and core needle biopsy, which uses a larger needle to collect a bigger sample of tissue. The biopsy technique is chosen based on factors such as the size and location of the suspicious area, the patient's medical history, and the need for detailed tissue analysis.

In patients eligible for surgical treatment, final diagnostics are run on the resected tissue to further evaluate tumor characteristics and help guide post-surgical treatment options [31]. These surgically resected specimens provide the most

comprehensive assessment of the primary tumor and constitute the source material for the molecular data analyzed in this thesis.

Breast Cancer Classification

Breast cancer classification aims to stratify tumors into clinically meaningful subgroups to guide prognosis and treatment decisions. This structured approach integrates multiple levels of assessment, ranging from the anatomical extent of disease to microscopic pathological evaluation and molecular characterization. Together, these layers of classification provide a structured framework for approaching the inherent heterogeneity of breast cancer, enabling a more precise understanding of tumor behavior and more individualized patient management.

Clinicopathological Assessment

Stage

An important initial classification of breast cancer involves determining how far the cancer has spread within the breast and to other parts of the body. The TNM staging system is the most widely used method for classifying the extent of breast cancer and provides essential information for understanding the seriousness of the disease and planning treatment. "TNM" stands for Tumor, Node, and Metastasis, which form the three criteria pillars of the staging system. The tumor criterion describes the size and extent of the main tumor, which is commonly referred to as the primary tumor, and includes whether the tumor has invaded adjacent structures such as the chest wall or skin. The node criterion indicates whether cancer has spread to nearby lymph nodes, located in the armpit (axilla), near the clavicle and breastbone [32]. The lymph nodes that receive lymphatic drainage directly from the tumor site are the first ones that cancer spreads to. Therefore, they are particularly important for assessing the extent of cancer spread and are commonly referred to as the sentinel lymph nodes [33]. The metastasis criterion describes whether cancer has spread to other parts of the body, with common metastasis sites being the bones, liver, lungs, and brain [34]. The TNM categories are combined to assign stages from I to IV, reflecting increasing severity and spread of the disease. The tumor stage at diagnosis helps clinicians to determine the prognosis and tailor treatment plans to a specific patient’s situation. In this thesis, the research is centered on early breast cancer at diagnosis, defined as localized, non-metastatic disease treated with primary surgical resection.

Grade

To provide a pathological diagnosis and basic prognostic information, biopsy tissue is examined by a pathologist under the microscope examining cellular and tissue structures. Grading a tumor involves evaluating how much the cancer cells differ from normal cells and helps to predict its growth rate and potential to spread. The pathologist examines the tissue for features, such as the arrangement of cells, whether they form tubules, their resemblance to normal breast cells, and the mitotic count [35]. The widely used Nottingham Grading System incorporates these features to provide a standardized tumor grade [36]. Tumors are graded on a scale from 1 to 3, with grade 1 tumors resembling normal cells and grade 3 tumors being highly abnormal and aggressive. The grade helps clinicians to predict how the cancer might behave and tailor treatment plans accordingly. Low-grade cancers generally grow more slowly and are less likely to recur after treatment, whereas high-grade cancers grow and spread quickly and are more likely to recur.

Histopathology

Histological subtyping categorizes tumors according to their growth pattern and site of origin within the breast. Breast carcinomas are broadly divided into in situ and invasive forms. In situ carcinomas remain confined to the ducts or lobules without breaching the basement membrane, while invasive carcinomas are characterized by malignant cells that have infiltrated the surrounding breast tissue, extending beyond their site of origin [37]. At the time of diagnosis, invasive disease is far more common than non-invasive forms. The most common invasive subtype, accounting for 70% of cases, is invasive ductal carcinoma, originating from the ductal structures of the breast. Invasive lobular carcinoma arises from the lobules and constitutes about 10-15% of invasive cases [38]. Non-invasive subtypes include ductal carcinoma in situ and lobular carcinoma in situ. The studies in this thesis focus on invasive breast cancer and do not further differentiate between histological subtypes.

Molecular Assessment

Breast cancer is a highly heterogeneous disease, with tumors differing not only in their stage and histology but also in their underlying molecular characteristics. Traditional pathological assessments provide important information, but they do not fully capture the biological identity of the tumor, which ultimately determines how it behaves, its likely clinical course, and how it will respond to treatment. Molecular subtyping tries to organize this complexity by grouping tumors with similar molecular profiles, providing a structured framework to better understand tumor behavior and guide more precise, individualized treatment decisions.

Hormone receptor and HER2 status

Determining the most appropriate treatment strategy for patients with invasive breast cancer involves the routine assessment of three key biomarkers. The estrogen receptor (ER) and progesterone receptor (PR), collectively referred to as hormone receptors (HR), act as transcription factors regulating the expression of specific genes implicated in cell proliferation and survival [39]. ER represents the primary determinant of endocrine responsiveness, whereas PR expression further reflects functional activity of the ER signaling pathway and provides additional prognostic refinement within ER-positive disease [40, 41]. Pathologists assess the presence of these proteins by evaluating the percentage of tumor cells that bind to specific antibodies in immunohistochemistry (IHC) assays. While international guidelines commonly define ER-positivity as staining in more than 1% of tumor cells, Swedish guidelines apply a higher cutoff of more than 10% [42, 43]. This distinction reflects evidence that tumors with low ER expression (1 to 9%) show unreliable response to hormone therapy and are therefore not considered clinically hormone receptor-positive in the Swedish setting [44].

The third key biomarker is the human epidermal growth factor receptor 2 (HER2), a transmembrane receptor tyrosine kinase that plays a key role in regulating cell growth and differentiation [45]. Its overexpression or the amplification of the encoding *ERBB2* gene acts as the primary oncogenic driver in a subset of breast cancers. The dependence of these tumors on HER2 signaling, often referred to as oncogene addiction, provides a therapeutic vulnerability that can be exploited using targeted treatments. As for hormone receptors, pathologists assess HER2 status using IHC, which scores protein expression across a defined range. Intermediate scores require further testing using fluorescence in situ hybridization to assess *ERBB2* gene amplification and confirm HER2 status. Tumors with low protein expression and no *ERBB2* amplification are classified as HER2-low, a category of clinical relevance due to the emergence of antibody-drug conjugate therapies [46].

Beyond receptor status, Ki67 is a nuclear protein whose expression is linked to greater proliferative activity and assessed by IHC in the same manner as the receptors. Together, the receptor status of ER, PR, and HER2 forms the foundation for clinically relevant subgroup classification and guides systemic treatment decisions, while Ki67 provides additional information regarding tumor aggressiveness (**Figure 4**).

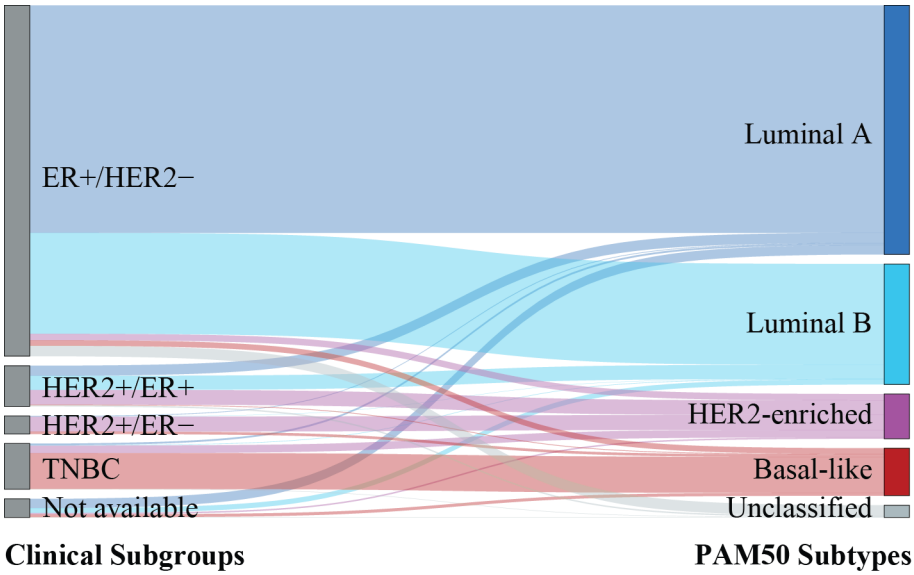

Figure 4. Distribution of PAM50 molecular subtypes across clinical subgroups.

Created based on data from a population-based cohort of early invasive breast cancer obtained from Staaf et al. [47].

Intrinsic molecular subtypes

The development of high-throughput gene expression profiling enabled systematic analysis of thousands of genes simultaneously, providing new insight into the molecular heterogeneity of breast cancer. This technological advance allowed researchers to move beyond single-marker assessment and instead characterize tumors based on global transcriptional patterns. Using hierarchical clustering of gene expression data, Perou et al. demonstrated that breast cancers could be grouped into distinct subtypes [48]. These intrinsic molecular subtypes were characterized by stable and reproducible gene expression signatures and were found to differ in their underlying biological features as well as in clinical outcomes [49].

The initial gene set defining these subtypes was later refined into a streamlined 50-gene classifier, known as PAM50 [50]. This reduced gene panel retained the ability to distinguish intrinsic subtypes while improving reproducibility and clinical applicability. PAM50-based classification is now widely used in research and clinical practice to stratify patients according to tumor biology and associated prognosis. Classification is performed using a nearest-centroid method, in which each subtype is represented by a centroid, which are vectors of mean expression values across the 50 genes, derived from a labelled training cohort. A new sample is then assigned to the subtype whose centroid its expression profile most closely correlates with, producing a categorization that reflects the tumor's underlying biology. The four intrinsic molecular subtypes that tumors can be classified as are

Luminal A (LumA), Luminal B (LumB), HER2-enriched (HER2E), and Basal-like (Basal) (**Figure 4**). A fifth "Normal-like" subtype was originally identified, however, later studies indicated that this group most likely represents samples with a high proportion of normal breast tissue [50].

While each PAM50 subtype has distinct characteristics they exist on a biological continuum rather than as completely separate entities [51]. This is a fundamental tension with the nearest-centroid classification approach, which by design forces every sample into a discrete category. Moreover, the centroids are derived from a specific reference cohort, and classification relies on relative gene expression and appropriate normalization to that reference. Inadequate normalization or differences in cohort composition may therefore influence subtype assignment, and failure to properly account for these factors can lead to erroneous classification [47]. The stability of PAM50 subtype assignment, the relationship between nearest and second-nearest centroids, and the influence of underlying biological gene modules on classification were investigated in **Paper I**.

Gene expression signatures for prognosis

Beyond molecular subtyping, gene expression data has been leveraged to develop prognostic signatures, defined as sets of genes whose combined expression patterns are used to predict the risk of disease recurrence. Their clinical value stems from patients with similar tumor characteristics often experiencing markedly different outcomes, with some relapsing early, others many years after diagnosis, and some remaining recurrence-free long-term. As treatment decisions are partially guided by recurrence risk, imprecision in prognostic classification can lead to both overtreatment and undertreatment, with important consequences for patient quality of life and long-term outcomes.

Several multigene prognostic tests have been developed for this purpose. Oncotype DX, MammaPrint, and the PAM50-based Prosigna assay are used in early ER-positive/HER2-negative breast cancer to assess recurrence risk and guide adjuvant therapy decisions, particularly regarding the administration of chemotherapy [50, 52-54]. Despite their clinical utility, outcome heterogeneity persists even within risk categories defined by these assays, and their validated value remains restricted to specific clinical subgroups [31].

This has motivated research into whether incorporating additional layers of biological information including genomic instability measures, epigenomic alterations, and features of the tumor microenvironment, could improve the resolution of risk stratification beyond what clinicopathological characteristics and gene expression can provide [55-58]. The potential of epigenomic features for recurrence risk prediction is explored in **Paper IV**, where a DNA methylation-based signature is developed and evaluated for its prognostic value across different breast cancer subgroups.

Clinical Outcome Patterns and Molecular Characteristics

The clinical breast cancer subgroups defined by hormone receptor and HER2 status exhibit distinct outcome trajectories that reflect fundamental differences in oncogenic drivers, proliferative dynamics, and tumor-microenvironment interactions, as well as differences in the availability and effectiveness of targeted therapies (**Figure 5**).

ER-positive/HER2-negative (ER+/HER2−) tumors are primarily driven by estrogen-dependent transcriptional programs and generally display lower proliferative activity and less genomic instability than other subgroups [59]. Clinically, ER+/HER2− disease is associated with a favorable short-term prognosis, however, the risk of recurrence may persist for many years, reflecting a pattern of late relapse that distinguishes this group [16].

HER2-positive (HER2+) tumors are biologically characterized by constitutive activation of HER2-mediated growth signaling, driven by amplification of the *ERBB2* oncogene [60, 61]. They frequently exhibit high proliferative rates, elevated Ki67 expression, and greater genomic instability than ER+/HER2− tumors. Historically, these cancers were associated with early relapse and poor survival, however the introduction of HER2-targeted therapies has led to substantial improvements in clinical outcomes [62].

Triple-negative breast cancer (TNBC) represents the most biologically and clinically aggressive subgroup. These tumors are commonly associated with high proliferation rates and marked genomic instability [63]. A considerable proportion exhibit features of homologous recombination deficiency (HRD), reflecting impairment of the cellular pathway responsible for repairing DNA double-strand breaks. This defect is most commonly caused by germline or somatic *BRCA1/2* alterations or by epigenetic silencing of key DNA repair genes [64]. Impaired DNA repair further contributes to genomic instability and an elevated mutational burden. In addition, a subset of TNBCs displays a high presence of immune cells within the tumor microenvironment, referred to as tumor-infiltrating lymphocytes (TILs) which are associated with improved survival in this subgroup [58, 65]. Recent epigenetic profiling has also identified basal and non-basal tumor groups within TNBC, underscoring biological heterogeneity beyond traditional receptor-based classification [66]. Clinically, TNBC is characterized by early recurrence, typically within the first few years following diagnosis, and its aggressive biology, combined with the lack of targeted therapies, contributes to its status as the subgroup with the poorest overall outcomes [67].

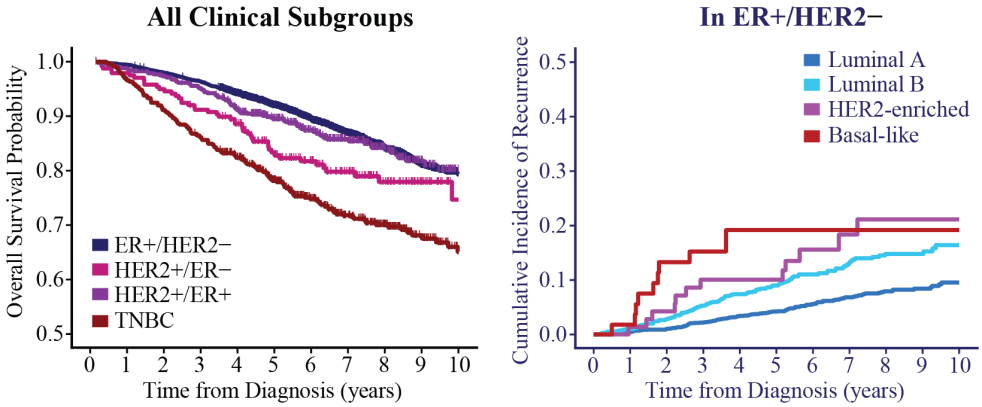

Figure 5. Survival across clinical subgroups and recurrence risk by PAM50 subtype within ER+/HER2− disease.

Created based on data from a population-based cohort of early invasive breast cancer obtained from Staaf et al. [47].

Molecular stratification within clinical subgroups

PAM50-defined intrinsic molecular subtypes further refine clinical classifications by capturing underlying biological heterogeneity within receptor-defined groups. In ER+/HER2− tumors, PAM50 predominantly identifies the two luminal intrinsic subtypes: LumA tumors, characterized by low proliferative activity and favorable prognosis, and LumB tumors, which exhibit higher proliferation and an increased risk of recurrence. ER+/HER2− disease provides a clear example of how intrinsic molecular subtypes capture additional clinically relevant heterogeneity and enable stratification of recurrence risk beyond receptor-defined classification (**Figure 5**). HER2+ tumors largely correspond to the HER2E subtype, whereas TNBC most frequently maps to the Basal subtype [68]. Overall, intrinsic molecular subtyping helps explain why patients with similar receptor profiles may experience divergent clinical outcomes, although biological heterogeneity persists even within each subtype.

Importantly, analyses of large population-representative cohorts demonstrate that all intrinsic subtypes can be observed across clinical groups, albeit in smaller proportions (**Figure 4**) [69]. This includes the presence of HER2E and Basal intrinsic subtypes within ER+/HER2− disease. The biological and clinical significance of these "atypical" subtype assignments remains incompletely understood and raises questions regarding whether they represent biologically meaningful entities with therapeutic implications or merely reflect methodological aspects of nearest-centroid classification. In this thesis, **Paper II** and **Paper III** systematically addressed this question by comprehensively characterizing the HER2E and Basal intrinsic subtypes within ER+/HER2− breast cancer, analyzing their clinical outcomes and underlying transcriptomic and genomic characteristics.

Early Breast Cancer Treatment

The integration of molecular biomarkers and subtypes into clinical practice has enabled increasingly personalized therapeutic strategies and risk stratification. Importantly, early breast cancer is considered curable, with current multimodal treatment approaches offering long-term disease control and high cure rates in the majority of patients [13]. The primary aim of treatment in early breast cancer is therefore to eradicate locoregional disease and prevent metastatic recurrence.

Locoregional Treatment

For patients with early breast cancer, surgical excision of the primary tumor remains the cornerstone of curative treatment. In recent years, breast-conserving surgery has largely replaced mastectomy, aiming to remove the tumor while preserving as much of the breast tissue as possible [70]. To assess whether the cancer has spread through lymph nodes, sentinel lymph node biopsies are performed as the standard staging approach. The procedure removes the initial lymph nodes draining from the tumor, helping to decide if additional axillary surgery or radiation is necessary. Following breast-conserving surgery, radiation therapy may help to eliminate residual cancer cells, reducing the risk of locoregional recurrence and improving overall survival [31, 70].

Systemic Therapy

Most patients however also require systemic therapy to reduce the risk of disease recurrence and improve long-term survival by targeting tumor cells that may have spread beyond the breast and regional lymph nodes. Systemic treatment may be administered in the adjuvant setting, following surgery to eradicate residual microscopic disease, or in the neoadjuvant setting, prior to surgery. Neoadjuvant therapy aims to reduce tumor burden, increase the likelihood of breast-conserving surgery, and allow assessment of treatment response [71].

Endocrine therapy

Endocrine therapy is the principal approach to treating ER-positive breast cancer aiming to disrupt estrogen signaling, the principal driver of tumor growth within this subgroup [59]. Disruption can be achieved in different ways by using selective modulators that bind and block estrogen receptors, or aromatase inhibitors that inhibit an enzyme catalyzing estrogen production. Selective degraders not only block but also degrade the estrogen receptor, offering an alternative mechanism for suppressing estrogen receptor signaling. The benefit of combining these therapies with CDK4/6 inhibitors, which inhibit cyclin-dependent kinases involved in cell

cycle progression, has been shown in selected patient groups with intermediate- and high-risk disease [72, 73]. In advanced disease, additional therapeutic strategies are being developed to overcome endocrine resistance. These include next-generation selective estrogen receptor degraders and agents targeting *ESR1* mutations, which are associated with acquired resistance to endocrine therapy [74, 75].

Chemotherapy

Despite its toxicity, chemotherapy remains essential in the treatment of more aggressive breast cancer subtypes, including TNBC and HER2+ cases. In ER+/HER2− breast cancer, chemotherapy is typically reserved for patients with high-risk features due to its substantial side effects [31]. Selecting patients for chemotherapy in this context is challenging, as overtreatment can lead to unnecessary toxicity, underscoring the need for precise biomarkers such as molecular recurrence risk signatures to guide treatment decisions. Chemotherapy agents target rapidly dividing cancer cells through various mechanisms. Anthracyclines insert into DNA and lead to cell death, while taxanes prevent cell division by stabilizing microtubules. Alkylating agents cause DNA crosslinking, thereby hindering replication, and antimetabolites function by mimicking cellular components and disrupting DNA synthesis. Therapy regimens commonly combine multiple agents to maximize cancer cell elimination and minimize resistance.

Targeted HER2 therapies and antibody-drug conjugates

In HER2+ breast cancer, the cornerstone of systemic therapy is anti-HER2 targeted treatment, most prominently with monoclonal antibodies such as trastuzumab, which bind the HER2 receptor and inhibit downstream signaling while mediating antibody-dependent cellular cytotoxicity [62]. Antibody-drug conjugates (ADCs) represent an additional therapeutic class in which an anti-HER2 antibody is linked to a cytotoxic payload, enabling targeted delivery of chemotherapy to HER2-expressing tumor cells while limiting systemic exposure. The activity of HER2-targeting ADCs has also established HER2-low tumors as a therapeutically relevant category, as tumors with low levels of HER2 expression can derive benefit despite lacking *ERBB2* gene amplification or classical HER2-driven biology [76]. While initially focused on HER2, these therapies are now being developed and applied to target a broader range of tumor-associated antigens [77].

Immune checkpoint inhibitors

Immunotherapy has emerged as an important addition to the treatment landscape of breast cancer. Immune checkpoint inhibitors (ICIs) target inhibitory pathways such as PD-1/PD-L1, which tumors exploit to suppress anti-tumor immune responses. By blocking these signals, ICIs restore T-cell mediated immune recognition and tumor cell elimination. In early breast cancer, ICIs are primarily used in high-risk TNBC, where they are administered in combination with chemotherapy in the neoadjuvant

setting and may be continued in the adjuvant phase [78]. Their use in TNBC is supported by the relatively high immunogenicity of this subtype, which frequently exhibits increased levels of tumor-infiltrating lymphocytes (TILs), a feature associated with improved response to immune checkpoint inhibition and providing a biological rationale for exploring immune-based treatment strategies also in other clinical subgroups [58, 79-81]. As the understanding of the tumor microenvironment and tumor-immune interactions continues to improve, there is increasing potential for immunotherapeutic approaches to be extended beyond TNBC to additional clinical subgroups [82].

PARP inhibitors

PARP inhibitors represent a targeted therapeutic strategy for HRD-positive patients. By inhibiting poly(ADP-ribose) polymerase (PARP), these agents interfere with DNA repair mechanisms, leading to accumulation of DNA damage and selective tumor cell death [83]. In early breast cancer, PARP inhibitors are used in patients with high-risk HER2− disease and germline *BRCA1/2* mutations as part of adjuvant systemic therapy [84]. As homologous recombination deficiency extends beyond patients with germline *BRCA1/2* mutations, alternative methods for its assessment may facilitate the identification of a broader population eligible for DNA repair-targeted therapies [55, 83].

Aims

Overall Aim

To refine the application of molecular tumor profiling for biological characterization and risk stratification in early breast cancer by improving its interpretation and clinical relevance within established clinical subgroups, particularly ER+/HER2− disease.

Specific Aims

Paper I

To investigate the biological determinants and robustness of PAM50 subtyping across clinically defined breast cancer subgroups in a population-representative cohort.

Paper II

To comprehensively characterize the clinicopathological, transcriptomic, and genomic features of ER-positive/HER2-negative breast cancers classified as PAM50 HER2-enriched and to determine their clinical significance and biological similarity to other ER-positive/HER2-negative tumors and HER2-positive tumors.

Paper III

To comprehensively characterize the clinicopathological, transcriptomic, and genomic features of ER-positive/HER2-negative breast cancers classified as PAM50 Basal-like and to determine their clinical significance and biological similarity to other ER-positive/HER2-negative tumors and triple-negative tumors.

Paper IV

To evaluate the prognostic utility of genome-wide tumor DNA methylation profiles for recurrence risk prediction across clinical subgroups in early-stage breast cancer and to assess whether they provide additional information beyond established clinicopathological variables.

Materials and Methods

Patient Cohorts

SCAN-B

The Sweden Cancerome Analysis Network - Breast (SCAN-B) is a prospective, population-based study launched in 2009 in the South Sweden healthcare region (ClinicalTrials.gov identifier NCT02306096) [85, 86]. All eligible patients at participating hospitals are offered inclusion, resulting in an unselected cohort of patients diagnosed with invasive breast cancer that is representative of cases within its geographic catchment area [69]. Since SCAN-B is not designed as a randomized clinical trial, participation does not influence treatment decisions.

As part of the initiative, peripheral blood samples and fresh tumor tissue are collected either at surgery or, for patients receiving neoadjuvant therapy, prior to surgery. Sampling is performed only after routine pathological assessment to ensure that standard diagnostic procedures are not disrupted. Clinicopathological variables and longitudinal follow-up data are obtained from the Swedish National Breast Cancer Quality Registry. Whole-transcriptome RNA sequencing is routinely performed for primary tumors to enable transcriptomic characterization.

The SCAN-B cases included in this thesis were drawn from a population-based RNA sequencing cohort of 6660 primary invasive breast cancer cases with prospective clinical follow-up, as described by Staaf et al. [47]. Intrinsic molecular subtypes were assigned by repeated classification against reference sets designed to mimic the composition of the original PAM50 training cohort, with the final subtype robustly determined by majority vote.

Paper I included SCAN-B cases with RNA sequencing data and linked clinicopathological information. **Paper II** and **Paper III** were primarily based on the ER+/HER2- subset of the SCAN-B cohort with RNA sequencing data. Within this subset, whole-genome sequencing and genome-wide DNA methylation profiling were performed for selected cases. Additionally, analyses providing context across clinical subgroups also included HER2+ and TNBC cases. **Paper IV** included SCAN-B cases with genome-wide DNA methylation data and clinical outcome information.

METABRIC

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort is a large, retrospectively assembled breast cancer dataset comprising primary invasive tumors diagnosed between 1977 and 2005. Tumors were collected from centers in the United Kingdom and Canada and represent a well-characterized historical breast cancer population with mature survival data [87]. The cohort includes extensive clinicopathological annotation with long-term follow-up and molecular profiling. Available molecular data include genome-wide gene expression and copy number profiles, along with targeted sequencing data for selected genes. Within this thesis, METABRIC was primarily used as an independent validation cohort in **Paper II** and **Paper III**.

BASIS

The BASIS cohort is derived from a large international whole-genome sequencing study of breast cancer in which primary breast tumors were analyzed to characterize the somatic genetic basis of the disease [88]. Tumor samples were collected through collaborative cancer research initiatives and represent a diverse set of primary breast cancers subjected to high-coverage whole-genome sequencing. The study systematically identified driver mutations and defined mutational processes underlying breast cancer at genome-wide resolution. In this thesis, ER+/HER2− tumors from the BASIS cohort were used to complement whole-genome sequencing performed in selected SCAN-B subgroups. They enabled comparative analyses of genomic alterations in rare ER+/HER2− intrinsic subtypes investigated in **Paper II** and **Paper III**.

TCGA

The Cancer Genome Atlas (TCGA) is a large-scale, multi-institutional research program launched in 2006 by the National Cancer Institute and the National Human Genome Research Institute to characterize the molecular landscape of human cancers. The TCGA breast cancer (BRCA) cohort comprises primary breast tumors collected across multiple centers in the United States [68]. Available data include gene expression profiles based on RNA sequencing, genome-wide DNA methylation data, copy number data, somatic mutation data derived from exome sequencing, and associated clinical annotation. These datasets are publicly accessible and widely used in cancer genomics research. TCGA-BRCA represents a retrospectively assembled research cohort and does not constitute a population-based series. Within this thesis, TCGA-BRCA served as an external reference dataset. In **Paper II** and **Paper III**, transcriptomic and DNA methylation data of the TCGA cohort were used to validate subgroup-specific gene expression and

methylation patterns identified in the SCAN-B cohort. In **Paper I**, TCGA gene expression data were accessed via the online cBioPortal platform for independent gene-gene correlation analyses.

Statistical Hypothesis Testing

Statistical hypothesis testing is used for evaluating whether observed differences or associations in data are likely to reflect true biological signals or arise by chance. A null hypothesis is formulated based on the question being tested, and the appropriate statistical test is then chosen based on the nature of the data and the assumptions that can reasonably be made. Parametric tests assume that the data follow a specific distribution, most commonly the normal distribution, whereas non-parametric alternatives make fewer distributional assumptions and often operate on ranks of the observations.

To evaluate the null hypothesis, a test statistic is computed from the data, summarizing how much the observed data deviate from what would be expected under the null hypothesis. The p-value is then derived by comparing this test statistic to its theoretical distribution under the null hypothesis, representing the probability of obtaining the observed result or a more extreme one if the null hypothesis were true. A p-value below a predefined significance threshold, conventionally set at 0.05, is taken as sufficient evidence to reject the null hypothesis. This threshold reflects an accepted 5% risk of incorrectly rejecting a true null hypothesis, known as a type I error or false positive. The complementary error, a type II error or false negative, occurs when a true difference fails to be detected. These two error types represent a fundamental trade-off in hypothesis testing, as lowering the significance threshold to reduce false positives increases the risk of false negatives. When multiple hypotheses are tested, the probability of false positive findings increases substantially, as each individual test carries a chance of incorrectly rejecting the null hypothesis, accumulating across tests. This is particularly relevant in high-dimensional molecular analyses where thousands of features are tested at once. To control for this, multiple testing correction methods are applied, which differ in how stringently they control false positive findings. More stringent methods minimize the risk of any false positive occurring but at the cost of missing true findings, while less stringent approaches accept a proportion of false positives among significant results, making them better suited for exploratory high-dimensional analyses.

Survival Analysis

Survival analysis refers to a class of statistical methods for time-to-event data, where the primary outcome is the time elapsed from a defined starting point to the occurrence of a specific event of interest. Because the event may not occur within the study period for all individuals, observations can be right-censored, meaning that follow-up ends before the event is observed.

In this thesis, the following clinical endpoints were used. Overall survival (OS) was defined as the time from diagnosis to death from any cause. Recurrence-free interval (RFI) was defined as the time from diagnosis to the first documented disease recurrence. Distant recurrence-free interval (DRFI) was defined as the time from diagnosis to the occurrence of distant metastatic recurrence. Invasive disease-free survival (IDFS) was defined as the time from diagnosis to the first invasive disease event, including locoregional recurrence, distant recurrence, second primary invasive cancer, or death from any cause.

A fundamental tool for summarizing survival data is the Kaplan Meier estimator, a non-parametric method that estimates the survival function, defined as the probability of remaining event free beyond a given time point [89]. Survival probabilities are calculated directly from the observed event times, resulting in a stepwise survival curve that reflects the proportion of individuals who remain event free over time. To compare survival between groups, the log rank test is commonly used to evaluate whether survival distributions differ.

To quantify the association between predictors and time-to-event outcomes, regression modelling approaches are used. A widely applied method is the Cox proportional hazards model, a semi parametric model that relates predictors to the hazard function, defined as the instantaneous event rate at a given time among individuals who have not yet experienced the event [90]. The Cox model assumes that covariate effects act multiplicatively on the hazard and remain constant over time, known as the proportional hazards assumption. The model estimates regression coefficients that quantify the association between each predictor and the hazard. Exponentiating these coefficients yields hazard ratios, which represent the relative change in hazard associated with a one unit increase in the predictor, holding other variables constant. In a univariate Cox model, a single predictor is included. The model can be extended to include multiple predictors simultaneously, allowing estimation of adjusted effects while accounting for potential confounding. This framework enables evaluation of the independent contribution of several interrelated variables.

In **Paper I**, Kaplan Meier estimation and Cox proportional hazards models were used to evaluate the prognostic relevance of PAM50 subtype classifications and their stability in relation to OS and DRFI. In **Paper II** and **Paper III**, these methods

were used to compare RFI, IDFS, and OS between intrinsic subtypes within ER+/HER2− breast cancer, including analyses stratified by treatment regimen.

Competing Risks

Standard survival analysis methods assume that only one type of event can occur. In many clinical settings, however, individuals are at risk of multiple mutually exclusive events. Studies of cancer recurrence represent a clear example, as a patient who dies cannot subsequently experience a tumor recurrence. This is particularly relevant in breast cancer, which predominantly affects older women, many of whom have substantial comorbidity and a non-negligible risk of death from causes unrelated to cancer. When recurrence-free interval (RFI) is the outcome of interest, death without prior recurrence (DWR), defined as the time from diagnosis to death from any cause before a documented recurrence event, constitutes a competing risk. If competing events are treated as simple censoring, individuals who die are incorrectly assumed to remain at risk of recurrence beyond their time of death. This leads to overestimation of the absolute risk of recurrence, especially when mortality from other causes is substantial [91].

In such settings, the relevant quantity is the cumulative incidence function, which represents the probability of experiencing a specific event by a given time while accounting for competing events. Regression approaches adapted to competing risks, such as the Fine and Gray subdistribution hazard model, relate covariates directly to the cumulative incidence [92]. These methods are therefore more appropriate than standard Cox models when the objective is to estimate or predict the absolute risk of recurrence in the presence of death as a competing event. In **Paper IV**, Fine and Gray models were used to estimate the cumulative incidence of RFI, with DWR treated as a competing event.

Limitations

First, survival analyses are inherently cohort dependent. Hazard ratios and absolute risk estimates reflect the characteristics of the specific patient population under study, and their interpretation depends on how well this population represents the broader clinical setting. A major strength of this thesis is the use of the SCAN-B cohort, which has repeatedly been reported to be population-representative. Nevertheless, differences in healthcare systems, screening practices, demographic structure, and treatment strategies may influence how well the findings from one cohort translate to other populations.

Second, the treatment landscape changes over time as national guidelines are updated. The SCAN-B patients included in this thesis were enrolled between 2010 and 2018, a period during which treatment recommendations were revised,

including guidelines regarding adjuvant chemotherapy in ER+/HER2- disease. As a result, patients with similar tumor characteristics may have received different treatments depending on the year of diagnosis.

Third, survival analyses depend on the quality and completeness of follow-up data. Recurrence information used in this thesis originated primarily from data reported to the Swedish National Quality Registry for Breast Cancer, where reporting of recurrence events may be incomplete or delayed. In addition, detailed treatment information is not typically available from the registry, meaning that heterogeneity within treatment groups may remain unaccounted for and influence outcome comparisons.

Predictive Modelling in the High-Dimensional Setting

When the objective extends beyond assessing associations between individual molecular features and outcome to constructing models that can predict outcomes for new patients, predictive modelling provides the appropriate methodological framework. These methods use observed data to learn relationships between predictors and outcome to generate individualized risk estimates [93].

When many molecular features are measured simultaneously, the number of predictors is large relative to the sample size and, more importantly in survival analysis, relative to the number of observed events. In such high-dimensional, low event settings, standard multivariable survival models are prone to overfitting and unstable effect estimates [94]. As a result, they may perform well in the training data but fail to generalize to independent samples. The concepts discussed in this section provide the methodological context for the predictive modelling framework implemented in **Paper IV**.

Penalized Regression

Prediction can be approached using a range of modelling strategies. Regression based methods are commonly chosen because they provide interpretable risk estimates, can incorporate competing risks within established statistical frameworks, and allow for intrinsic feature selection [94]. To enable stable estimation and principled feature selection in a high-dimensional setting, penalized regression methods introduce regularization by constraining the size of regression coefficients. This reduces estimation variance, improves numerical stability, and limits overfitting. Penalized regression methods differ in how they constrain the regression coefficients and therefore in how they behave. Ridge regression shrinks all coefficients toward zero, which stabilizes estimation, especially when predictors are highly correlated, but generally keeps all predictors in the model. In contrast,

the lasso penalty can shrink some coefficients exactly to zero, effectively performing feature selection by excluding less informative predictors. The elastic net incorporates both penalties, enabling simultaneous shrinkage and variable selection, and can appropriately handle correlated predictors by retaining groups of correlated variables that jointly contribute to the outcome.

Hyperparameter Tuning

Penalized regression models include hyperparameters that control aspects such as the overall strength of coefficient shrinkage and, in the case of the elastic net, the balance between ridge and lasso penalties. These hyperparameters are not estimated directly during model fitting but must instead be chosen based on model performance within the training data. Because their values determine the degree of regularization applied during model fitting, directly influencing the number of features retained in the model and its predictive performance, their selection is a critical step in model development.

Overfitting and Generalization

When features or hyperparameters are chosen to optimize performance on the available data, there is a risk that the model adapts too closely to random variation in the training data. This phenomenon, known as overfitting, leads to overly optimistic estimates of predictive performance. A model that performs well in the training data may therefore perform substantially worse when applied to new independent cases.

High-dimensional omics data further increase this risk, as the number of molecular features is typically large relative to the number of patients and observed events. Controlling model complexity through penalization and rigorous validation is therefore essential to ensure generalizable performance.

Data Splitting and Cross-Validation

To obtain unbiased estimates of predictive performance, model development and model evaluation must be clearly separated. Data are therefore partitioned into subsets used for training and subsets reserved for evaluation. Performance assessed on data that were not used for model fitting provides a more realistic estimate of generalization ability.

Cross-validation provides an efficient resampling strategy for internal performance estimation, as it allows all observations to contribute to both model development and evaluation. In k-fold cross-validation, the dataset is partitioned into k approximately equal subsets, or folds. In each iteration, one fold is held out as

evaluation data, while the remaining k-1 folds are used for model training. This process is repeated k times so that each fold serves once as evaluation data, and predictive performance is averaged across iterations.

When model development involves steps such as feature selection and hyperparameter tuning, all such steps must be performed exclusively within the training data of each iteration. If information from the evaluation data influences any stage of model development, information leakage occurs, potentially leading to optimistically biased performance estimates. Nested cross-validation addresses this by providing an explicit framework that separates model development from performance estimation. The inner loop is used to optimize model development steps, including feature selection and hyperparameter tuning, while the outer loop is used solely to estimate predictive performance (**Figure 6**).

However, internal estimates remain an approximation of how a model will perform in independent data, and external validation on an independent dataset represents the gold standard for evaluating generalizability. When such data are available, the inner loop hyperparameter selection procedure is repeated on the complete training dataset without holding out data for evaluation, and the refitted model is then applied to the external dataset.

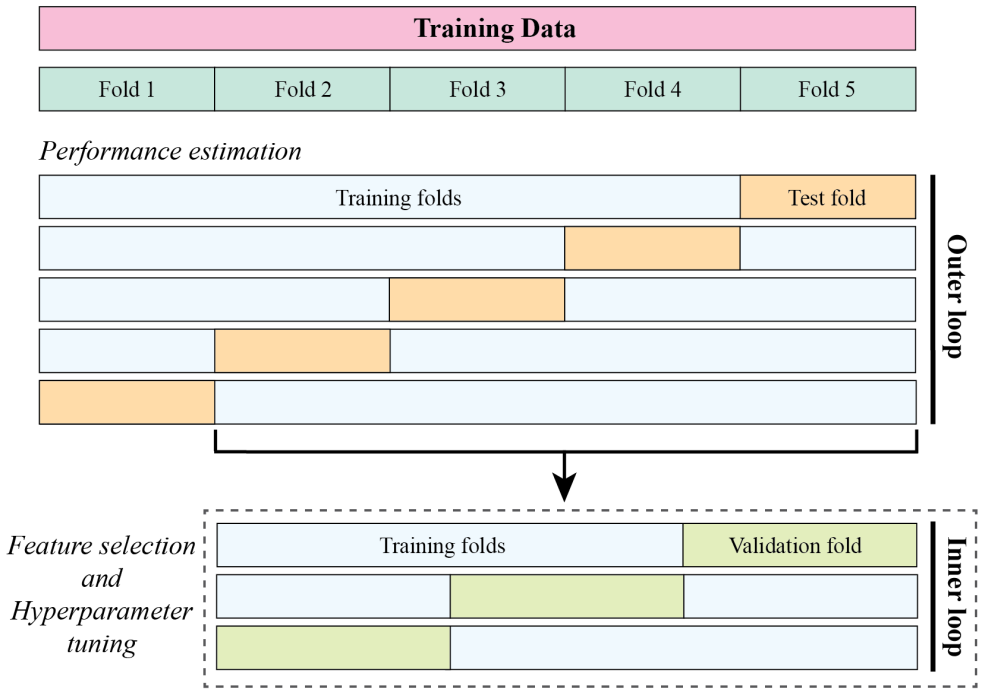

Figure 6. Nested cross-validation framework for model development and unbiased performance estimation.

Performance Metrics for Predictive Models

The performance of predictive models is typically evaluated in terms of discrimination and calibration. Discrimination refers to the ability of a model to distinguish between patients with different outcomes. In time-to-event settings, the concordance index is a commonly used global measure of discrimination, assessing whether patients with higher predicted risk experience the event earlier than those with lower predicted risk among comparable patient pairs. When prediction at a specific time point is of interest, time dependent measures such as the time dependent area under the receiver operating characteristic curve (AUC(t)) assess how well the model separates patients who experience the event by that time from those who do not.

Calibration refers to the agreement between predicted risks and observed outcome frequencies. A well calibrated model provides absolute risk estimates that correspond closely to what is observed in the data. Calibration and overall prediction accuracy can be quantified using the Brier score, which measures the mean squared difference between predicted probabilities and observed outcomes at a given time

point. The integrated Brier score (IBS) summarizes prediction error across follow-up.

Discrimination and calibration capture complementary aspects of model performance: a model may rank patients correctly yet systematically overestimate or underestimate risk. Comprehensive evaluation therefore requires assessment of both [95].

Limitations

In addition to the limitations discussed for survival analysis, the predictive modelling framework has further methodological considerations. Despite the use of penalized regression to mitigate the challenges of high-dimensional survival modelling, the relatively favorable prognosis of breast cancer means that the number of recurrence events remains limited relative to the number of predictors. Under these conditions, the selected predictors and their estimated effects may exhibit some instability, and the resulting risk estimates should be interpreted with caution until validated in independent cohorts.

Dimensionality reduction methods could in principle be applied prior to model fitting to address the high-dimensional nature of the data. However, such approaches transform the original features into composite variables that are difficult to interpret biologically and clinically. In prognostic modelling, the ability to identify which specific features drive risk predictions is important both for understanding the underlying biology and for supporting clinical trust in model decisions.

Sample size and event count also constrain the number of cross-validation folds that can be used during model development. Although increasing the number of folds allows a larger proportion of the data to contribute to model training and can reduce bias in performance estimation, it may result in hold-out evaluation sets containing too few events for reliable performance assessment.

Lastly, both Cox proportional hazards models and Fine-Gray subdistribution hazard models assume that covariate effects act multiplicatively on the hazard and remain proportional over time. On the log-hazard scale, this corresponds to a linear combination of predictors. Consequently, these regression approaches may fail to represent more complex effects in high-dimensional molecular data.

Genomic Alteration Profiling

The human genome comprises approximately three billion base pairs that define the DNA sequence, of which only around 1-2% encode the approximately 20,000 protein-coding genes, with the remainder comprising regulatory elements and non-

coding sequences [96, 97]. This sequence is organized across 23 pairs of chromosomes, with every cell in the body carrying two copies of each, one inherited from each parent, constituting the diploid genome. While this inherited genome contains germline variants that are present in all cells of an individual, cancer development is characterized by the accumulation of acquired somatic alterations that arise specifically within the tumor cell population. Characterizing these genomic alterations plays an important role in understanding tumor biology and was performed in **Paper II** and **Paper III**.

Profiling Technologies

Whole-genome sequencing (WGS) provides a comprehensive assessment of the tumor genome, allowing detection of somatic single nucleotide variants (SNVs), small insertions and deletions (indels), copy number alterations, and structural rearrangements across the entire genome [98]. During sequencing, DNA is fragmented and each fragment is sequenced, generating short nucleotide sequences referred to as reads. These reads are aligned to a reference genome to reconstruct the tumor sequence. By sequencing both tumor and matched normal DNA, somatic alterations can be distinguished from germline variation. The sensitivity and reliability of detecting such alterations depends on the sequencing depth, defined as the average number of reads covering each genomic position. In contrast to WGS, targeted next-generation sequencing panels restrict analysis to predefined genomic regions and are typically used to detect mutations in clinically relevant genes. Genome-wide copy number alterations can alternatively be assessed using single nucleotide polymorphism (SNP) arrays, which are microarray-based platforms designed to detect genomic gains and losses across the genome [99, 100].

To accurately infer copy number alterations from sequencing or SNP array data, allele-specific analytical approaches such as ASCAT (Allele-Specific Copy number Analysis of Tumors) can be applied [101]. Since raw copy number signals from tumor samples are confounded by the presence of non-tumor cells in the sample and the chromosomal ploidy of the tumor, ASCAT explicitly estimates tumor purity and ploidy to correct for these effects. This yields allele-specific copy number profiles, enabling the calling of genomic gains, losses, and loss of heterozygosity.

Genomic Characterization of Tumors

Depending on their location, SNVs and indels may alter coding sequences or regulatory elements and thereby impact gene function. Not all detected mutations contribute to tumorigenesis. Mutations affecting established driver genes are of particular interest, as somatic alterations in these genes are known to drive cancer development. A straightforward way to compare tumors genomically is to evaluate the frequency of alterations in these driver genes. The overall number of somatic

mutations per tumor, often summarized as tumor mutational burden (TMB), provides an additional measure of genomic instability and can be compared across tumor subgroups.

Copy number alterations reflect gains or losses of genomic material and may result in oncogene amplification or tumor suppressor loss. Tumors can be compared by assessing copy number gains and losses as well as the proportion of the genome affected by such alterations, providing insight into chromosomal instability and genomic architecture. Structural rearrangements include deletions, tandem duplications, inversions, and translocations, which may disrupt genes or arise from defective DNA repair mechanisms.

Beyond cataloguing individual variants, whole-genome sequencing enables characterization of the mutational processes that shape tumor genomes. Distinct endogenous and exogenous mutational mechanisms leave characteristic patterns across the genome. These patterns are captured through mutational and rearrangement signatures inferred from genomic context and combinations of SNVs, indels, copy number changes, and structural rearrangements [102, 103]. Rearrangement signatures specifically describe recurrent patterns of structural variation across the genome and can reflect underlying defects in DNA repair mechanisms. The utility of this framework is exemplified by HRDetect, which leverages specific genome-wide patterns associated with homologous recombination deficiency to generate a probabilistic score of HRD-positivity [55]. This approach captures the cumulative genomic consequences of impaired homologous recombination repair rather than relying on direct detection of mutations in individual genes. In addition to genome-wide patterns, localized clusters of hypermutation may occur. This phenomenon, termed kataegis, is characterized by closely spaced base substitutions and can co-occur with structural rearrangements, reflecting localized genomic instability [102].

Limitations

All molecular profiling data in this thesis were derived from bulk tissue samples, where a small portion of a tumor biopsy, comprising a mixture of cells, is used for analysis. As different regions of a tumor can harbor distinct cell populations, a single biopsy may not capture the full diversity of the tumor mass and therefore not be fully representative of the entire tumor with respect to both intrinsic and microenvironment characteristics [10]. Molecular measurements from such samples reflect an average signal across all constituent cells, meaning alterations or signals of minority cell populations may be diluted or obscured. This is an inherent limitation of bulk tissue profiling that applies throughout the analyses described in this thesis. For the WGS-based analyses in this thesis, this means that the spatial organization of genomic alterations or potential subclonal populations within the tumor cannot be determined.

Furthermore, as tumor purity and sequencing depth affect the sensitivity with which somatic alterations can be detected, low-frequency variants may remain undetected [104]. The tumor WGS data analyzed in this thesis had an average sequencing depth of 37×, which is sufficient for the detection of most somatic alterations but limits the ability to reliably infer subclonal tumor populations and assess intra-tumor heterogeneity.

DNA Methylation Profiling

DNA methylation is a chemical modification of the DNA in which a methyl group is added to the cytosine base within a cytosine-phosphate-guanine (CpG) dinucleotide, known as a CpG site. Unlike genetic mutations, methylation does not alter the DNA sequence but regulates gene expression through effects on chromatin structure and transcriptional accessibility [105]. CpG sites are unevenly distributed across the genome and occur in distinct genomic contexts. Genomic regions with a high density of CpG sites are known as CpG islands, which are typically located at or near gene promoters. Promoter-associated methylation is commonly linked to transcriptional repression, whereas methylation in other genomic contexts, such as gene bodies or distal regulatory elements, exhibits more context-dependent regulatory effects [106]. CpG sites can also be evaluated in the context of chromatin accessibility by assessing overlap with open chromatin regions identified by ATAC sequencing, where methylation changes are more likely to have regulatory relevance [107]. In cancer, DNA methylation patterns are frequently altered, disrupting normal gene regulation and genomic stability [108]. In this thesis, DNA methylation data were used in **Paper II** and **Paper III** to investigate epigenetic patterns in molecular subtypes and in **Paper IV** for prognostic modeling and risk stratification.

Profiling Technologies

Genome-wide methylation profiling is commonly performed using array-based technologies, such as the Illumina 450K and EPIC platforms, which interrogate hundreds of thousands of predefined CpG sites across promoters, gene bodies, CpG islands, and distal regulatory regions. Methylation levels are commonly represented as beta values ranging from 0 to 1, reflecting the proportion of methylated signal at each CpG site. For statistical analyses, methylation levels can be converted to M-values, defined as the log2 ratio of methylated to unmethylated signal intensities [109]. This transformation stabilizes variance across the methylation range and improves suitability for linear and regression-based modeling. When methylation measurements are derived from bulk tumor tissue, adjustment for tumor cell content can be performed to better approximate tumor-specific methylation patterns, for example by correcting beta values based on estimated tumor purity [110].

Characterization of Tumors Based on DNA Methylation Patterns

Genome-wide DNA methylation profiling enables systematic characterization of tumors based on their epigenetic landscape. By quantifying methylation levels across large numbers of CpG sites, tumors can be described according to global and regional patterns, as well as gene associated features. Unsupervised clustering of tumors based on similarity in DNA methylation patterns can be used to define epigenetic subtypes, also referred to as epitypes. These epitypes may partially align with known biological phenotypes but also capture heterogeneity beyond established molecular classifications [66].

Differential DNA methylation analysis identifies CpG sites that differ in methylation levels between predefined groups of samples, such as molecular subtypes or clinical categories. The resulting CpGs can then be mapped to their genomic context and analyzed using functional enrichment approaches to provide insight into associated regulatory mechanisms and biological processes [111].

Beyond genome-wide analyses, epigenetic alterations can also be examined at the level of individual genes to assess potential regulatory effects. CpG methylation within promoter regions in proximity to transcription start sites (TSS) is of particular interest because methylation in these regions is closely linked to transcriptional control [106]. Visualization of CpG island and shore methylation patterns within promoter and TSS regions can therefore reveal specific hypermethylation or hypomethylation profiles. Integration of promoter region methylation data with matched gene expression measurements further enables assessment of whether observed epigenetic alterations are associated with transcriptional changes.

Limitations

In the context of DNA methylation analyses, the use of bulk tumor tissue means that measured DNA methylation levels reflect a composite signal arising from the methylation states of different cell types within the tumor and surrounding microenvironment, and variation in a samples cellular composition may therefore influence the observed methylation patterns. Consequently, cell type-specific methylation patterns, spatial methylation heterogeneity, and potential subclonal methylation alterations within the tumor cannot be assessed.

Transcriptomic Profiling

Gene expression refers to the process by which information encoded in DNA is transcribed into RNA and, for protein-coding genes, subsequently translated into functional proteins. The collection of RNA molecules produced by a cell at any

given time, known as the transcriptome, reflects which genes are active and at what levels. Gene expression is tightly regulated and varies between cell types, tissues, and physiological conditions. This regulation takes place primarily at the transcriptional level through transcription factors that bind regulatory DNA elements and is further shaped by epigenetic mechanisms such as DNA methylation and histone modification, which influence chromatin accessibility and transcriptional output [112]. In cancer, gene expression patterns are frequently altered as a consequence of genetic and epigenetic changes, driving key aspects of tumor behavior. In this thesis, transcriptomic data were used in **Paper I** to investigate the stability and biological drivers of PAM50 gene expression subtype classification, and in **Paper II** and **Paper III** to characterize intrinsic molecular subtypes.

Profiling Technologies

RNA sequencing is the predominant technology for profiling the transcriptome and forms the basis of the transcriptomic analyses that are part of this thesis. Briefly, RNA molecules extracted from a biological sample are converted into complementary DNA, sequenced, and the resulting reads are aligned to a reference genome. Reads mapping to each gene are then counted and summarized into a gene-level expression matrix, providing a quantitative measure of transcriptional activity.

Because raw read counts are influenced by both gene length and the total sequencing depth of a sample, normalization is commonly applied before expression values can be meaningfully compared. Several normalization methods exist, including Fragments per Kilobase Million (FPKM), which adjusts read counts for both gene length and sequencing depth. Prior to downstream analysis, expression values are commonly log2-transformed and scaled to compresses the wide dynamic range of gene expression data into a more symmetric distribution, which better satisfies the assumptions of many statistical methods.

Transcriptomic Characterization of Tumors

Transcriptomic data can be interrogated at multiple levels to characterize the gene expression landscape of tumors. At the most fundamental level, the expression of individual genes of interest can be examined directly, allowing targeted comparison across tumor subgroups. To systematically identify genes that differ in expression between defined groups, supervised differential expression analysis can be applied, testing all genes across the genome simultaneously. The resulting set of differentially expressed genes can then be interpreted in a biological context through functional enrichment analysis, which tests whether genes associated with particular biological processes, molecular pathways, or transcription factor targets are overrepresented [113].

Beyond individual genes, transcriptional programs can be summarized using metagene signatures, which aggregate the expression of co-regulated gene sets into a single activity score [114]. This enables quantification of broad biological programs such as proliferation, immune activation, or hormone response signaling across tumors, capturing coordinated transcriptional behavior.

To explore broader transcriptomic structure without imposing prior assumptions, unsupervised dimensionality reduction methods, such as Uniform Manifold Approximation and Projection (UMAP) can be applied [115]. These approaches reduce high-dimensional gene expression data into a low-dimensional representation, in which tumors are positioned according to the overall similarity of their transcriptional profiles. Clusters emerging from such analyses reflect shared patterns of gene expression.

Because bulk RNA sequencing captures the combined transcriptional output of all cells within a tumor sample, computational deconvolution methods can be used to estimate the relative abundance of distinct cell types within, providing an indirect measure of tumor microenvironment composition [116].

Limitations

In the context of gene expression analyses, the use of bulk tumor tissue means that the analyses performed in this thesis cannot determine spatial or cell population specific gene expression patterns as addressing these questions would require single-cell and spatial transcriptomics approaches.

Another consideration is that the SCAN-B transcriptomic data analyzed in this thesis were FPKM-normalized expression values from RNA sequencing. Because FPKM values within a sample do not sum to a constant, differences in transcriptional composition between samples can introduce systematic bias when comparing the absolute expression of a given gene across samples [117].

Statistical Software and Programming Environment

The analyses presented in this thesis were conducted using programming, statistical modelling, and large-scale data processing approaches typical of modern bioinformatics research. Most data processing, statistical analyses, and visualization were implemented using the R programming language, which provides an extensive ecosystem of community developed libraries. These libraries implement a wide range of analytical methods that can be readily integrated into custom analysis workflows. Python was used in selected contexts for data manipulation and machine learning approaches using isolated Conda environments to manage software

dependencies. All analyses were performed within a Unix based computing environment using the Bash shell, enabling scripting and automation of analysis workflows. Code development and interactive data exploration were primarily done in RStudio and Visual Studio Code. Version control of analysis scripts was managed using Git, and code repositories were maintained on GitHub to facilitate version tracking and reproducible computational workflows.

Ethical Considerations

This thesis included molecular and clinical data derived from previously collected patient samples, with ethical considerations relating to how such data and biological material are collected, handled, stored, and analyzed. The underlying studies from which data were sourced were conducted under ethical approval and written informed consent. For instance, the SCAN-B study was approved by the Regional Ethical Review Board in Lund, Sweden, and the Swedish Ethical Review Authority. For the TCGA and METABRIC cohorts, samples were collected under institutional review board approval and informed consent, and access to TCGA data is additionally governed by data use policies designed to protect participant privacy. Tumor tissue was taken during routine clinical procedures, and research analyses used leftover material after diagnostic tests were done. Because clinical needs always come first, this process can introduce sampling bias, as tumors with more leftover tissue may be overrepresented, which could mean larger tumors or those with higher cellularity are more common in the samples.

All patient data were pseudonymized before analysis to protect privacy, following ethical and legal rules. In Sweden, raw DNA sequence data are classified as sensitive personal data under applicable data protection regulations, precluding their public disclosure. Consequently, only processed or aggregated data were made publicly available as part of the publications included in this thesis.

Lastly, all molecular analyses were restricted to somatic tumor alterations, and germline variants were not analyzed. This distinction is ethically important, as germline variants carry implications beyond the individual patient, potentially revealing heritable information requiring broader consent considerations than those governing somatic tumor research.

Results and Discussion

Papers in the Thesis Context

The first part of the thesis elucidates the nature of PAM50 subtyping by examining the stability of subtype assignment and the influence of transcriptional programs on classification across tumors and clinical subgroups (**Paper I**).

Building on this, the thesis then focuses on the interpretation of atypical subtype assignments within clinical subgroups by comprehensively characterizing PAM50 HER2E and Basal tumors within ER+/HER2− breast cancer, comparing them both to other tumors within this clinical subgroup and to tumors sharing the same subtype in their typical clinical contexts (**Paper II** and **Paper III**).

The final part of the thesis moves beyond molecular subtyping toward prognostic modeling, shifting the objective from biological characterization to clinical risk stratification, and investigates the utility of genome-wide DNA methylation profiles for recurrence risk prediction (**Paper IV**).

Paper I

PAM50 subtyping is based on nearest-centroid classification, where a tumor's gene expression profile is correlated with each of the five PAM50 subtype centroids and assigned to the subtype with the highest correlation. These centroids reflect characteristic transcriptional patterns that can be interpreted as underlying biological processes [114]. Although the Normal-like subtype has been suggested to reflect samples with a high proportion of normal breast tissue rather than a distinct tumor-intrinsic subtype, it remains part of the PAM50 classification framework and was therefore included in the analyses that are part of this study [50, 69, 118].

When applied to large, population-representative cohorts, all PAM50 subtypes are observed across clinically defined subgroups [69]. While expected from the relative nature of nearest-centroid classification, this highlights that subtype assignment is not intrinsically tied to clinical subgroup and makes it important to understand how these assignments arise in order to interpret them appropriately.

We therefore aimed to better understand the mechanisms underlying PAM50 subtype assignment by characterizing the correlation relationships inherent to nearest-centroid classification and determining which biological processes within the PAM50 gene set drive subtype assignment across clinically defined breast cancer subgroups. The analyses performed as part of this study were conducted using a population-representative patient cohort of 6233 invasive breast cancer cases and combined with a rigorous nearest-centroid classification strategy using multiple reference sets for normalization and gene centering, balanced to mimic the original cohort composition of Parker et al. [47, 50, 69].

PAM50 Subtype Distinctiveness and Second-best Relationships

To better understand how tumors relate to the PAM50 centroids beyond the assigned subtype, we examined both the nearest (best) and second-nearest (second-best) centroid for each tumor based on their correlation profiles. Across the cohort, consistent patterns were observed in the relationship between the best and second-best subtype. For PAM50 Basal tumors, the second-best subtype was most often HER2E or Normal-like (Normal), for HER2E tumors most often LumB or Basal, for LumA tumors most often LumB or Normal, and for LumB tumors most often LumA or HER2E (**Figure 7A**). These patterns corresponded closely to the correlation structure between PAM50 centroids, indicating that second-best subtype assignments arise as an inherent consequence of how the centroids are constructed and inter-correlated (**Figure 7B**).

To quantify how distinctly tumors align with a given subtype, we evaluated for each tumor the difference in correlation between the best and second-best centroid. This difference provides a measure of subtype distinctiveness, reflecting how much more

strongly a tumor correlates with the assigned centroid compared to the next closest alternative. Basal tumors showed the largest difference in correlation between the best and second-best centroid, indicating more distinct subtype assignments, whereas Normal tumors showed the smallest difference (**Figure 7C**). This distinctiveness also varied across clinical subgroups. For example, HER2E-classified tumors in ER+/HER2− breast cancer were less distinctly correlated to the HER2E centroid than in other clinical subgroup contexts (**Figure 7D**). We further observed that in ER+/HER2−, lymph node-negative disease, the second-best subtype was associated with overall survival, demonstrating that proximity to alternative subtype centroids captures biologically and clinically meaningful variation beyond the primary subtype label.

Section discussion

These findings highlight aspects of PAM50 subtype classification that are not considered by a simplified interpretation as a set of discrete tumor classes or biological entities. Rather than forming strictly separated categories, tumors are positioned relative to multiple centroids, and their assigned subtype reflects the closest match within this framework. However, this assignment represents only an approximation of intrinsic tumor biology, as some cases show a strong and distinct alignment to a single centroid, whereas others exhibit only weak or comparable similarity across multiple centroids. In such cases, discrete calling of a single PAM50 subtype can be close to arbitrary, a notion further supported by the fact that key expression patterns underlying subtype calls, such as proliferation-related signals, are inherently continuous [119].

We observed that the distinctiveness of the best centroid match varies depending on the PAM50 subtype and the clinical subgroup context. In particular, the distinctiveness of subtype calls in typically disparate molecular subgroups can be limited, such as HER2E classifications in ER+/HER2− tumors. In these settings, the weak support for subtype assignment may call into question the biological and clinical relevance of such classifications, particularly when technical artifacts or misclassification by conventional pathology markers can be reasonably excluded.

Together, the observed subtype relationships, differences in distinctiveness, and the prognostic relevance of the second-best subtype underline that tumor PAM50 subtypes may at times be more appropriately viewed as a combination of centroid correlations rather than a single nearest-centroid label, supporting the notion that one-class PAM50 subtype assignment is a conceptual oversimplification and that tumors are more accurately positioned along a continuum between subtypes.

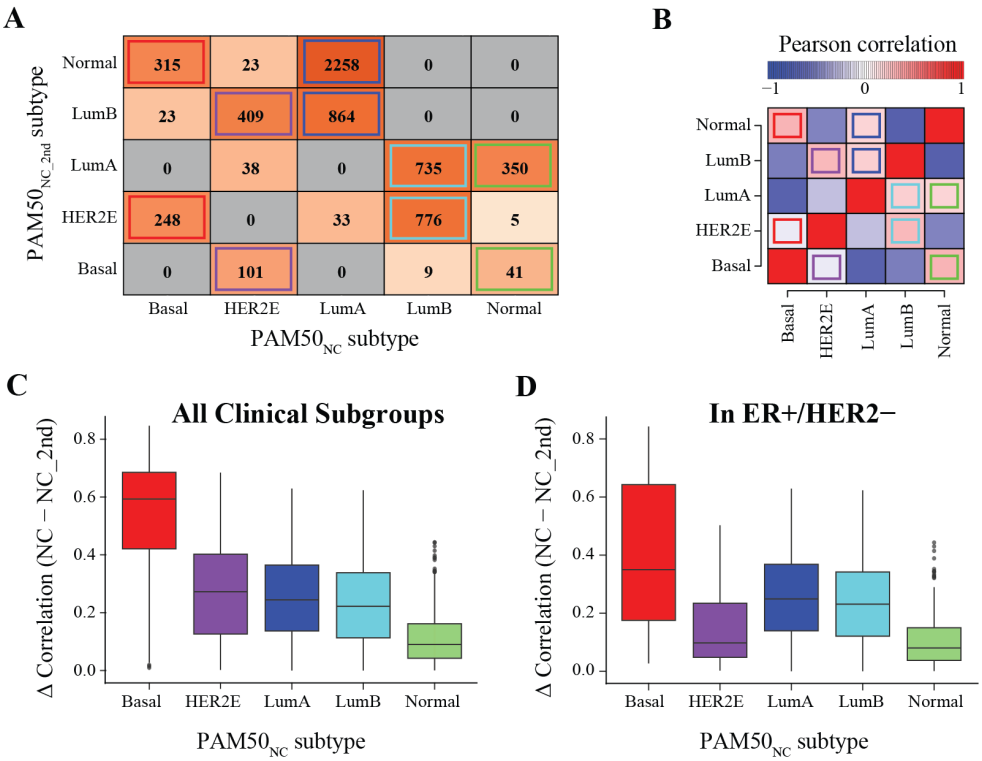

Figure 7. PAM50 subtype relationships.

A Cross-tabulation of PAM50 nearest-centroid ($PAM50_{NC}$) versus second-best ($PAM50_{NC\_2nd}$) subtype assignments across all tumors. Colored boxes highlight consistent subtype pairings between primary and second-best subtype. **B** Heatmap of Pearson correlations between PAM50 subtype centroids. Highlighted cells indicate centroid correlation patterns corresponding to the $PAM50_{NC}$ and $PAM50_{NC\_2nd}$ subtype relationships observed in panel A. **C** Subtype distinctiveness measured as the difference in correlation between the best and second-best centroid for each tumor (Δ = NC – NC_2nd) across all clinical subgroups. **D** Subtype distinctiveness measured as the difference in correlation between the best and second-best centroid for each tumor (Δ = NC – NC_2nd) in ER+/HER2− breast cancer. *Figure adapted from Paper I [51].*

Role of Transcriptional Programs in PAM50 Subtype Assignment

To investigate how specific transcriptional programs contribute to PAM50 subtype assignment, we applied a leave-one-gene-cluster-out perturbation strategy. Co-expressed groups of PAM50 genes were first identified using SRIQ clustering, resulting in seven gene clusters. Among these, three clusters captured major biological programs within the PAM50 signature, corresponding to proliferation (gene set 1), steroid response (gene set 2), and basal keratin expression (gene set 3), as supported by correlations with established metagenes and functional enrichment analyses (**Figure 8A**).

This highlights that broader transcriptional programs are naturally reflected within the PAM50 gene set, alongside genes selected to capture more subtype-specific expression patterns [50]. Notably, the selection of subtype-specific PAM50 genes is not necessarily driven by typical co-expression patterns across breast cancers, and no gene set showed a strong association with stroma-, lipid-, or immune-related metagenes. This is consistent with the original aim of defining an intrinsic gene list that primarily reflects tumor cell-intrinsic biology rather than the tumor microenvironment [48].

In the perturbation analysis, one gene cluster at a time was excluded from the PAM50 centroids and tumors were reclassified based on the remaining genes. This allowed us to assess how subtype assignment depends on specific gene expression programs. Across the full cohort, the impact of perturbation was primarily driven by exclusion of the large gene sets 1-3, representing proliferation, steroid response, and basal keratins. The extent of subtype switching differed between PAM50 subtypes and depended on the excluded gene set. Basal tumors showed the highest stability, with only a small proportion of cases changing subtype across perturbations, whereas Normal tumors were highly unstable (**Figure 8B**). LumA and LumB tumors showed intermediate stability, with LumA tumors primarily affected by exclusion of steroid response-related genes and LumB tumors by exclusion of proliferation- and basal keratin-related genes. Consistently, exclusion of gene set 1 frequently caused LumB tumors to switch to LumA, highlighting the well-established role of proliferation as a key divider between these subtypes.

Beyond overall effects, the impact of the leave-one-gene-cluster-out perturbation was different for PAM50 subtypes also within specific clinical subgroups (**Figure 8C**). Across TNBC, ER−/HER2+, and ER+/HER2+ tumors the dominant $PAM50_{NC}$ subtypes (Basal, HER2E, and HER2E, respectively) showed the highest classification stability, remaining largely unchanged across all gene set exclusions. Exclusion of smaller gene sets (gene sets 4-7) generally resulted in low proportions of tumors switching subtype, indicating a limited impact of these gene groups on classification stability. In ER+/HER2− tumors, exclusion of gene set 4 had minimal effect, whereas exclusion of gene set 7 resulted in a greater proportion of tumors switching subtype, highlighting a comparatively stronger contribution of *FGFR4*-related expression to HER2E classification in this context, a finding we explore further in **Paper II**. Across clinical and molecular subgroups, tumors that switched subtype frequently changed to their previously identified second-best subtype, reflecting the intrinsic centroid-to-centroid correlation structure and effectively illustrating these relationships (**Figure 7B**).

Section discussion

The perturbation analyses demonstrated that PAM50 classification is driven by the combined effect of multiple transcriptional programs rather than individual genes. In particular, clusters of co-expressed genes associated with proliferation, steroid

response, and basal keratin expression had the largest impact on subtype assignment, indicating that these biological processes form the core of the classification approach. Subtype assignment thus reflects the interplay between these broader transcriptional programs and prototypically selected genes included to capture subtype-specific features. Notably, even genes specifically selected to represent individual subtype biology showed only modest influence on their classification stability when excluded. For example, gene sets included in PAM50 classification to capture HER2E biology (gene sets 4 and 7, including *ERBB2*/*GRB7* and *FGFR4*) showed only limited impact on HER2E subtype assignment in clinically HER2+ disease. This suggests that HER2E classification is predominantly dictated by the interplay between other included gene sets, like those capturing proliferation, steroid response, and basal keratin expression.

Importantly, the effect of perturbation depended on the clinical context of the tumor, where typical subtype-clinical subgroup combinations showed the highest stability, whereas subtype assignments outside of these contexts were more sensitive to perturbation. In a broader context, this likely reflects that tumors classified within their typical clinical subgroup more closely resemble the prototypical samples from which the original centroid values were derived and therefore exhibit stronger and more stable correlations. This underscores the importance of considering the composition of the cohorts used to define the PAM50 centroids when applying and interpreting subtype assignments across diverse breast cancer contexts. The identification of tumors that remained stable across all perturbations further supports the existence of core subtype cases with highly consistent transcriptional profiles, which may represent prototypical examples of each subtype providing a basis for refining subtype centroids.

Taken together, these findings challenge the interpretation of PAM50 subtypes as fixed entities and instead support a model in which subtype assignment reflects the relative contribution of underlying biological processes.

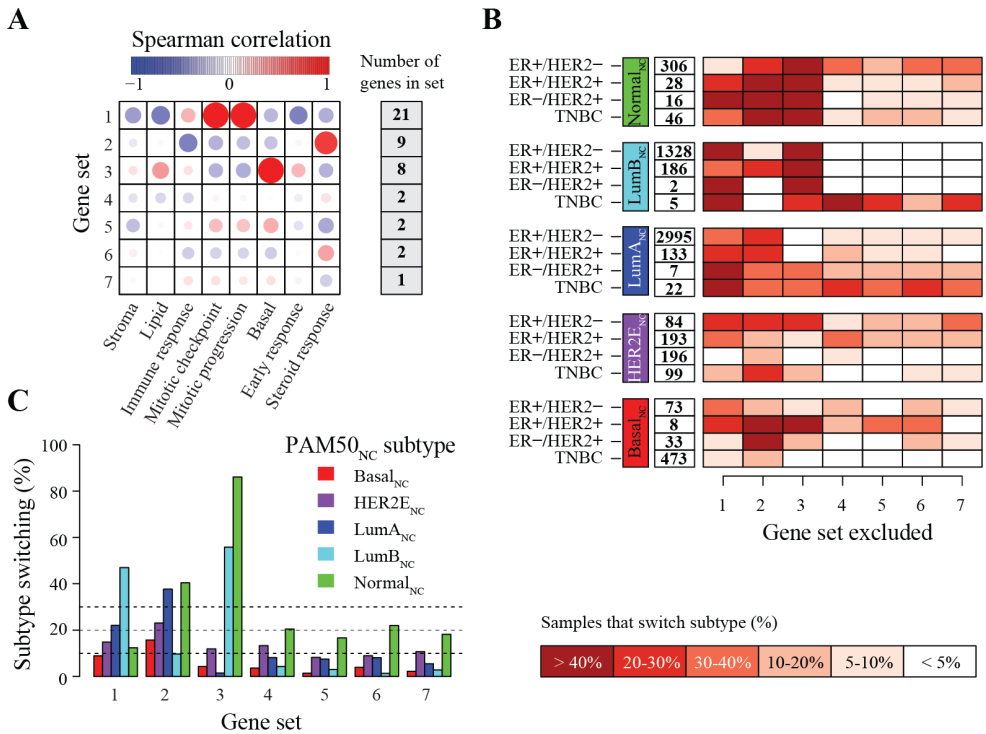

Figure 8. Effects of gene set perturbation on PAM50 classification.

A Spearman correlation matrix between expression scores of gene sets derived from SRIQ clustering and eight established biological metagenes, with the number of genes in each cluster indicated. Five PAM50 genes (*MYC*, *MMP11*, *BAG1*, *MDM2*, and *BLVRA*) were not assigned to any core cluster by SRIQ and are therefore excluded, resulting in 45 of 50 genes. One cluster containing *ERBB2*, *GRB7*, and *FGFR4* was manually split into two based on genomic proximity and expression correlation, giving seven final gene clusters. **B** Heatmap showing the proportion of tumors that switched subtype after gene set exclusion stratified by molecular and clinical subgroup. Switching is defined as assignment to a different PAM50 subtype compared to the original $PAM50_{NC}$ classification. **C** PAM50 reclassification results stratified by molecular and clinical subgroups. Heatmap showing the proportion of tumors switching subtype after gene set exclusion across subtype-subgroup combinations, with numbers indicating total group sizes per row. *Figure adapted from Paper I [51].*

Limitations

A limitation of the current study relates to the size of the PAM50 gene set as removal of larger gene clusters inevitably increases subtype switching. In addition, perturbing the centroids alters their original definition, and correlations to modified centroids should therefore be interpreted with caution. Despite this, subtype switching following perturbation was frequently consistent with the previously identified second-best subtype, supporting that the observed patterns reflect the underlying centroid correlation structure rather than random effects. Finally, the use

of bulk mRNA expression data represents an inherent limitation, as tissue heterogeneity and variable tumor cellularity may influence subtype assignment [118, 120].

In the context of this thesis, an improved understanding of the subtyping scheme supported the interpretation of tumors with seemingly disparate classifications such as tumors classified as Basal or HER2E in ER+/HER2− breast cancer, as examined in the subsequent studies.

Paper II and Paper III

Within ER+/HER2− breast cancer, the dominant PAM50 subtypes are LumA and LumB, yet small but consistent subgroups of tumors are classified as HER2E or Basal. These tumors challenge conventional clinical categorization, as their PAM50 subtypes suggest molecular profiles more reminiscent of other clinical subgroups than the one they are clinically classified into.

Rather than dismissing these as classification artifacts, the here described studies sought to characterize whether they represent biologically coherent subgroups with distinct molecular features and clinical implications, by comparing them both to other tumors within the same clinical subgroup (ER+/HER2−) and to tumors sharing the same subtype in their typical clinical contexts, namely HER2+ disease and TNBC.

Cohorts and Study Design

Both studies were conducted using the population-representative SCAN-B cohort as the primary discovery cohort, comprising 4487 and 4474 ER+/HER2− cases for **Paper II** and **Paper III** respectively, with the METABRIC cohort serving as an independent validation cohort.

Within the ER+/HER2− subset of the SCAN-B cohort, HER2E tumors constituted 89 cases and Basal tumors 76 cases, reflecting their low prevalence in an unselected invasive breast cancer population. The population-representative nature of SCAN-B is a key strength of both studies, as given the rarity of these subtypes, a large, unselected cohort of this kind is a prerequisite for their meaningful molecular characterization.

For DNA-level analyses in ER+/HER2−, whole-genome sequencing was performed on 28 HER2E and 16 Basal tumors, with 73 LumA and 105 LumB tumors from the BASIS cohort used for comparison. DNA methylation profiling using Illumina EPIC beadchips was available for a subset of SCAN-B cases and used in both studies.

Comparisons to HER2+ disease in **Paper II** were based on 564 cases from the SCAN-B cohort for transcriptomic analyses and on 162 tumors from the METABRIC cohort for genomic analyses. TNBC comparisons in **Paper III** were based on 228 SCAN-B cases with matched WGS, RNA sequencing, and DNA methylation data.

Importantly, ER-positivity in SCAN-B was defined as ≥10% positively stained tumor cells according to Swedish guidelines and is therefore unaffected by ongoing discussions on raising the ER-positivity cutoff from the commonly used 1% threshold [44, 121, 122].

Clinical and Prognostic Characteristics

HER2E and Basal tumors each represent small but consistently observed subgroups, each comprising approximately 2% of the invasive ER+/HER2− breast cancer population. Both subtypes were associated with aggressive clinicopathological features akin to LumB tumors and were commonly classified as high-risk by gene expression-based risk models, indicating that these patients would likely be identified as high-risk in clinical practice [47].

Survival analyses of clinical outcomes demonstrated that both subtypes carried a substantially worse prognosis than both LumA and LumB tumors when treated with endocrine therapy alone, providing independent prognostic information also in multivariate Cox regression models adjusting for lymph node status, tumor size, grade, and age, consistent with observations from previous studies (**Figure 9A, 9B**) [123-126]. In contrast, these differences were less pronounced in patients receiving combined chemotherapy and endocrine therapy.

These findings demonstrate that HER2E and Basal subtype assignments in ER+/HER2− disease are not clinically inconsequential but carry relevance for treatment considerations. A deeper molecular characterization of these subgroups is therefore needed to refine patient selection for targeted treatment approaches and to identify features that could serve as potential new therapeutic targets.

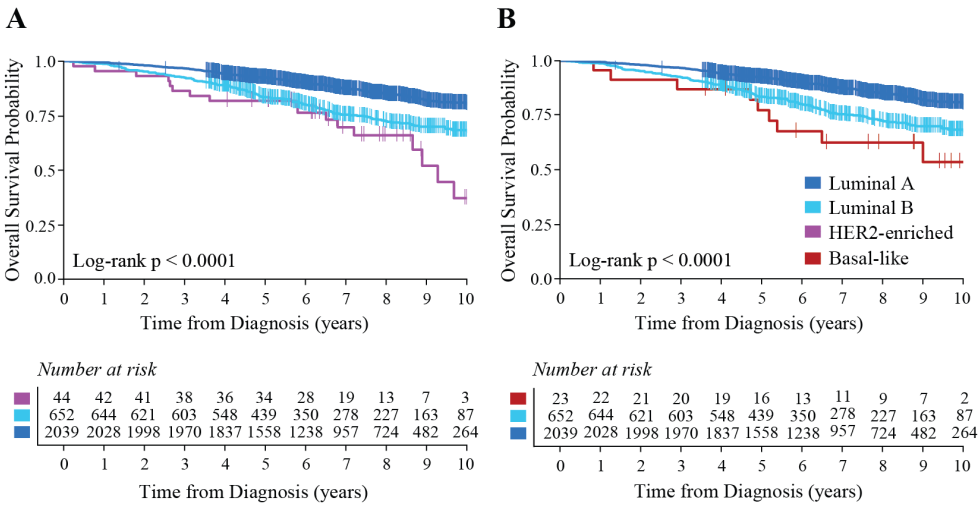

Figure 9. Overall survival in ER+/HER2− patients treated with endocrine therapy alone.

A Kaplan-Meier curves of overall survival stratified by PAM50 subtypes Luminal A, Luminal B, and HER2-enriched. **B** Kaplan-Meier curves of overall survival stratified by PAM50 subtypes Luminal A, Luminal B, and Basal-like. *Figure adapted from Paper II and Paper III [127, 128].*

Molecular Characteristics of PAM50 HER2E Tumors in ER+/HER2− Breast Cancer

Within ER+/HER2− breast cancer, HER2E tumors were characterized by high proliferation akin to the proliferative LumB subtype, clearly distinguishing them from LumA tumors (**Figure 10A**).

Furthermore, HER2E tumors showed features of a more immune-infiltrated tumor microenvironment compared to other luminal subtypes, collectively supported by higher immune metagene scores, elevated *CD274* expression, and estimated immune cell proportions from in silico cell type deconvolution (**Figure 10B**). In line with this observation, Griguolo et al. reported higher immune infiltration in non-luminal subtypes within HR+/HER2− breast cancer [129].

At the same time, steroid response pathway activity was markedly reduced, further reflected in significantly lower *ESR1* expression compared to both LumA and LumB tumors (**Figure 10C, 10D**). This reduced ER signaling activity likely contributes to the poor outcome on endocrine therapy, suggesting that these tumors are less dependent on estrogen-driven proliferation than other luminal subtypes, consistent with a hypothesis that acquisition of the HER2E subtype in metastatic disease is linked to estrogen independence [130]. Notably, no difference in *ESR1* mutation frequency between PAM50 subtypes was observed, confirming that poor endocrine

therapy outcomes in HER2E tumors are not explained by resistance through *ESR1* mutations, consistent with previous observations [131].

At the DNA level, HER2E tumors were predominantly characterized by a high frequency of *TP53* mutations, consistent with its general association with poor outcomes in luminal disease and with previously reported elevated *TP53* mutation frequencies in HER2E metastatic disease (**Figure 10G**) [131-133]. While *ERBB2* mutations were more frequent than in other luminal subtypes, their overall frequency remained too low to represent a group-defining trait, an observation similarly reported in the metastatic setting [131].

Despite these characterizing features, HER2E tumors do not appear to represent a discrete biological entity within ER+/HER2− disease. Rather, they form a subgroup that is distinct from LumA but tends to resemble LumB, primarily driven by shared high proliferation. This was reflected in both genome-wide expression patterns and the copy number landscape, where no region emerged as a defining trait of HER2E. Consistently, PAM50 centroid correlations of HER2E tumors indicated low subtype distinctiveness in the ER+/HER2− clinical subgroup, with LumB typically representing the second-best subtype, as also observed in **Paper I**.

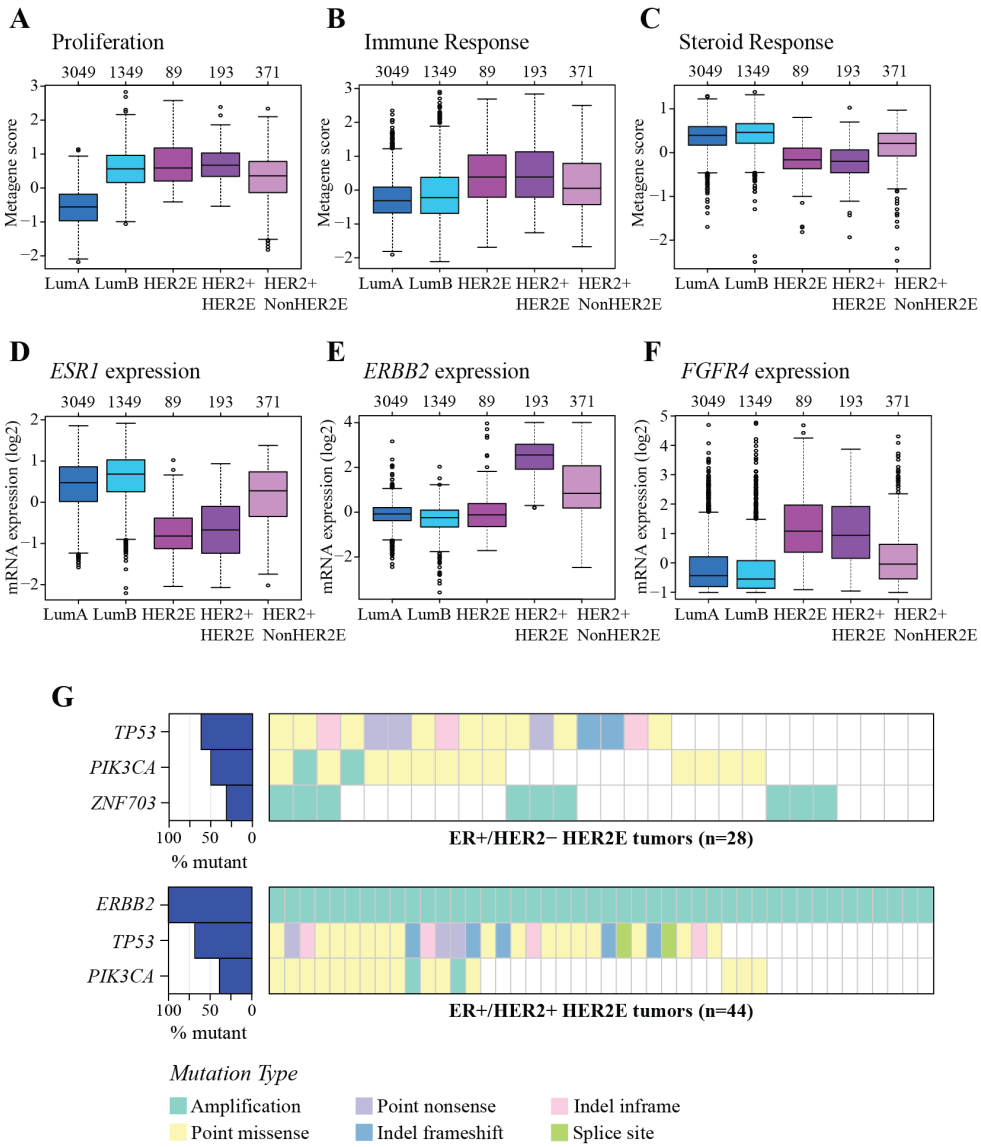

Figure 10. Transcriptional characteristics and driver gene alterations in PAM50 HER2E tumors.

A-C Metagenes scores capturing activity of proliferation, immune response, and steroid response transcriptomic programs. **D-F** mRNA expression of *ESR1*, *ERBB2*, and *FGFR4* scaled across all clinical subgroups. **G** Top three most frequently altered driver genes in PAM50 HER2-enriched tumors in ER+/HER2− and ER+/HER2+ disease. *Subtypes denoted by intrinsic names only (LumA, LumB, HER2E) correspond to ER+/HER2− disease. Figure adapted from Paper II [127].*

Comparison to HER2+ disease

To further contextualize the observed phenotype, ER+/HER2− HER2E tumors were compared to ER+/HER2+ tumors stratified by PAM50 subtype into HER2E (HER2+/HER2E) and non-HER2E groups.

This comparison revealed that the molecular features characterizing HER2E in ER+/HER2− disease were consistent with those of HER2+/HER2E tumors, with the expected exception of *ERBB2* expression which was higher in both HER2+ groups (**Figure 10E**). None of the differentially expressed genes distinguishing HER2E in ER+/HER2− disease showed differential expression when compared to HER2+/HER2E, and biological mRNA metagene scores and mutational features were all comparable between them.

Together these findings demonstrate that the molecular features defining HER2E tumors in ER+/HER2− disease are consistent with the HER2E subtype in its typical clinical context, and that its transcriptional phenotype is not dependent on *ERBB2* gene status, as further discussed in the following section.

Determinants of subtype classification

A central question addressed in **Paper II** was whether HER2E tumors in the ER+/HER2− subgroup represent clinically misclassified HER2+ cases. Assessment of *ERBB2* mRNA expression and gene amplification status indicated that this is not the case, as these tumors generally did not exhibit elevated *ERBB2* expression compared to other luminal subtypes. This is consistent with perturbation analyses in **Paper I**, which demonstrated that exclusion of the *ERBB2/GRB7* gene set had limited impact on HER2E classification in ER+/HER2− disease, meaning that cases can be classified as HER2E even without high expression of these genes.

Notably, **Paper I** further showed that exclusion of *FGFR4* caused a greater proportion of ER+/HER2− HER2E tumors to switch subtype than exclusion of *ERBB2/GRB7*, highlighting *FGFR4* as a comparatively stronger contributor to HER2E classification in this clinical context, consistent with its identification as a top differentially expressed gene (**Figure 10F**). The mechanism underlying variable *FGFR4* expression was further elucidated through DNA methylation profiling, which identified hypomethylation of CpG shore regions in the *FGFR4* promoter as a potential primary driver, explaining heterogeneity in *FGFR4* levels both across PAM50 subtypes and within HER2E tumors.

HER2E classification in ER+/HER2− disease thus appears to be primarily driven by a transcriptional state characterized by high proliferation, reduced ER signaling, and high *FGFR4* expression rather than by *ERBB2* biology.

Therapeutic opportunities

The aggressive clinical behavior and poor outcomes of HER2E patients in ER+/HER2− disease indicate a need for additional systemic treatment options beyond endocrine therapy. Given their high-risk and proliferative nature, chemotherapy represents a rational component of early management, while several molecularly informed therapeutic avenues may offer further opportunities to improve patient care.

While not being defining features of the HER2E subtype as a whole, a subset of HER2E tumors were predicted HRD-positive by HRDetect (18%), identifying patients who may benefit from PARP inhibitors. Similarly, while HER2-low frequency in HER2E tumors (81%) was not significantly different from other luminal subtypes, the large HER2-low patient subset within HER2E may still benefit from HER2-targeted antibody-drug conjugates, supported by demonstrated efficacy of trastuzumab deruxtecan in HER2-low hormone receptor-positive metastatic breast cancer [76, 134].

At the group level, the elevated immune response and *CD274* expression suggest potential benefit from immune checkpoint inhibition. Both the CheckMate-7FL and KEYNOTE-756 trials recently demonstrated significant improvement in pathological complete response when adding a PD-1 inhibitor to neoadjuvant chemotherapy in high-risk ER+/HER2− disease, with greater benefit observed in tumors with higher immune activation [80, 135]. Given these findings, the HER2E subtype may serve as a surrogate marker for luminal tumors with a higher likelihood of ICI benefit, warranting further investigation. The high-risk nature of these tumors also makes them a relevant subgroup for evaluation of CDK4/6 inhibitors in the early setting, supported by observations of benefit in HER2E metastatic disease [125, 136].

Finally, high *FGFR4* expression represents a distinguishing feature of HER2E tumors and a promising therapeutic target, supported by preclinical evidence of tumor growth inhibition with FGFR4 inhibition and several FGFR4-specific inhibitors currently in clinical trials across solid tumors [137, 138]. The pan-breast cancer nature of promoter hypomethylation as the primary driver of *FGFR4* expression is particularly relevant in this context, and since expression is not driven by mutation or amplification, epigenetic status at the promoter shore region may represent a potentially more appropriate biomarker for patient selection in future trials of FGFR4-targeted agents.

Molecular Characteristics of PAM50 Basal Tumors in ER+/HER2− Breast Cancer

Within ER+/HER2− breast cancer, Basal tumors displayed a molecular profile that set them apart not only from LumA but also distinctly from LumB, with biological

features not typically expected in the clinically ER-positive subgroup. As observed for the HER2E subtype, Basal tumors were similarly characterized by elevated proliferative activity, resembling the LumB subtype (**Figure 11A**). Their markedly low steroid response pathway activity, further reflected in significantly reduced expression of *ESR1* and *PGR*, was consistent with the higher frequency of PR negativity and likely contributed the poor outcomes under endocrine therapy (**Figure 11B, 11C**).

An important characteristic of Basal tumors, which is further discussed in the comparison to TNBC, was the presence of features consistent with a more immune-infiltrated tumor microenvironment, reflected in higher immune response metagene scores, significantly elevated *CD274* mRNA expression, and enrichment of immune response-associated pathways among genes differentially expressed relative to both LumA and LumB tumors (**Figure 11D-F**).

Two transcription factors with key roles in breast cancer biology showed divergent expression in Basal tumors. *FOXA1*, which facilitates ER binding to chromatin and is typically highly expressed in luminal breast cancer, was significantly downregulated (**Figure 11G**) [139]. In contrast, *FOXC1*, associated with Basal phenotypes and more aggressive tumor behavior, was significantly upregulated (**Figure 11H**) [140]. DNA methylation profiling revealed that these transcriptional alterations were reflected at the epigenetic level, with hypermethylation of *FOXA1* promoter shore regions and hypomethylation of *FOXC1* promoter regions, suggesting epigenetic regulation of their expression irrespective of ER status.

At the DNA level, the landscape of driver alterations in Basal tumors was characterized by a high frequency of *TP53* mutations and *MYC* amplifications (**Figure 11J**). Strikingly, 44% of Basal tumors were predicted as HRD-positive by HRDetect, a substantially higher proportion than observed in LumA and LumB tumors, consistent with increased exposure to mutational signatures associated with homologous recombination deficiency (**Figure 11I**). Overall, HRD occurs only in a small minority of ER+/HER2− breast cancers, and the marked enrichment observed here therefore highlights HRD as a distinguishing feature of the Basal subgroup within ER+/HER2− disease [141]. At the broader genomic level, Basal tumors also showed a substantially different copy number alteration landscape compared to both LumA and LumB tumors, with large differences in alteration frequencies across the genome, further underlining their molecular distinctiveness within ER+/HER2− disease.

Some considerations underlying the HER2E subtype are shared with Basal tumors, such as the combination of poor outcomes under endocrine therapy together with the absence of activating *ESR1* mutations, indicating that these tumors are intrinsically less dependent on estrogen signaling compared to canonical luminal subtypes, and the high frequency of *TP53* mutations consistent with its established association with poor outcomes in luminal disease.

Taken together, these molecular analyses demonstrated marked differences between Basal and the canonical ER+/HER2− subtypes tumors across all investigated molecular layers. The characteristics of Basal tumors collectively indicate a biologically distinct profile within ER+/HER2− disease, consistent with the high PAM50 centroid distinctiveness and classification stability of the Basal subtype observed across clinical subgroups in **Paper I**.

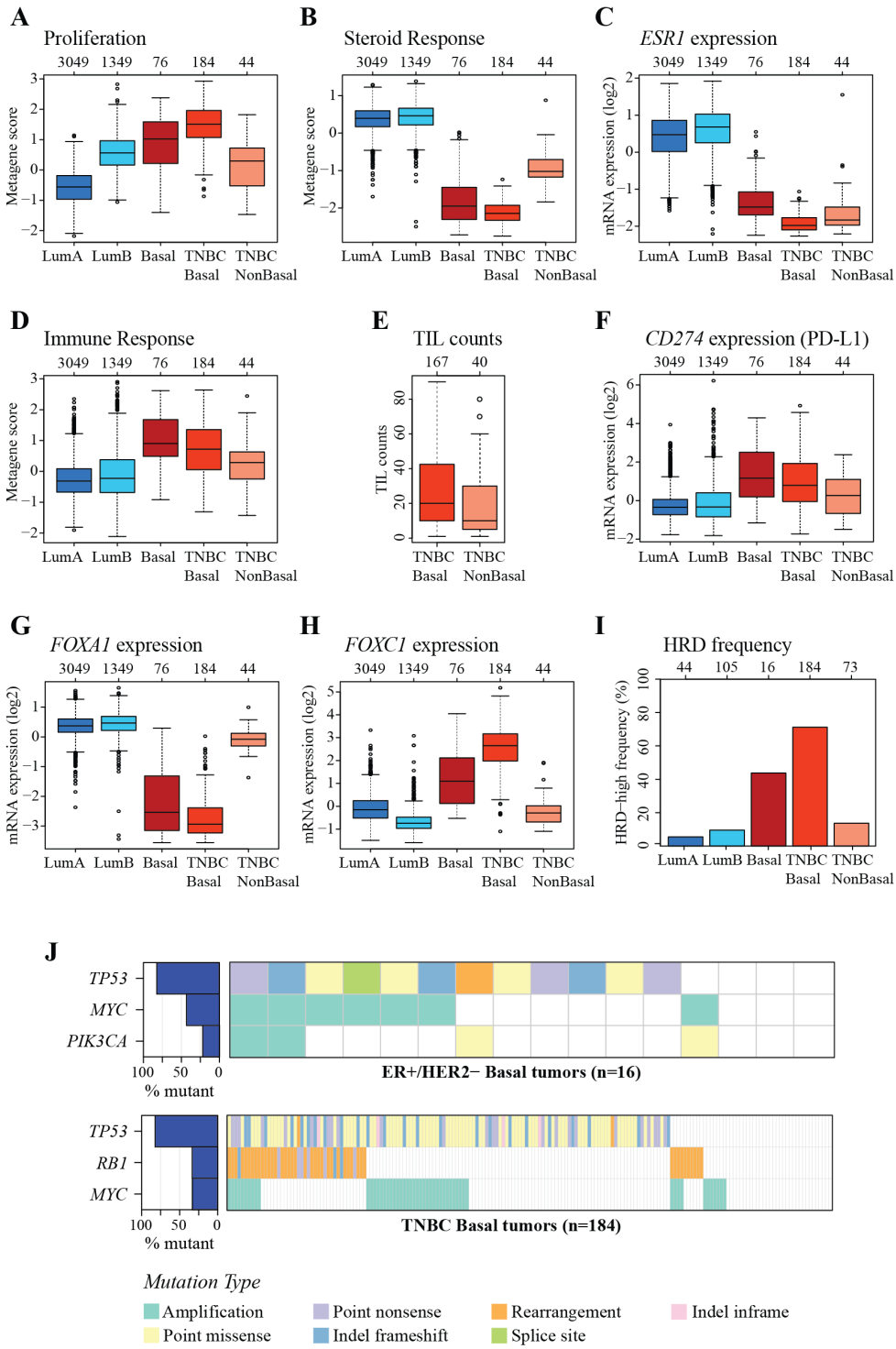
A
Proliferation
3049 1349 76 184 44
Metagene score
LumA LumB Basal TNBC Basal TNBC NonBasal
B
Steroid Response
3049 1349 76 184 44
Metagene score
LumA LumB Basal TNBC Basal TNBC NonBasal
C
ESR1 expression
3049 1349 76 184 44
mRNA expression (log2)
LumA LumB Basal TNBC Basal TNBC NonBasal
D
Immune Response
3049 1349 76 184 44
Metagene score
LumA LumB Basal TNBC Basal TNBC NonBasal
E
TIL counts
167 40
TIL counts
TNBC Basal TNBC NonBasal
F
CD274 expression (PD-L1)
3049 1349 76 184 44
mRNA expression (log2)
LumA LumB Basal TNBC Basal TNBC NonBasal
G
FOXA1 expression
3049 1349 76 184 44
mRNA expression (log2)
LumA LumB Basal TNBC Basal TNBC NonBasal
H
FOXC1 expression
3049 1349 76 184 44
mRNA expression (log2)
LumA LumB Basal TNBC Basal TNBC NonBasal
I
HRD frequency
44 105 16 184 73
HRD−high frequency (%)
LumA LumB Basal TNBC Basal TNBC NonBasal
J
TP53
MYC
PIK3CA
100 50 0
% mutant
ER+/HER2− Basal tumors (n=16)
TP53
RB1
MYC
100 50 0
% mutant
TNBC Basal tumors (n=184)
Mutation Type
Amplification
Point nonsense
Rearrangement
Indel inframe
Point missense
Indel frameshift
Splice site

Figure 11. Transcriptional characteristics, HRD frequency and driver gene alterations in PAM50 Basal tumors.

A Proliferation metagene scores. **B** Steroid response metagene scores. **C** *ESR1* mRNA expression scaled across all clinical subgroups. **D** Immune response metagene scores. **E** Tumor infiltrating lymphocyte counts. **F** *CD274* mRNA expression (encoding PD-L1) scaled across all clinical subgroups. **G-H** mRNA expression of *FOXA1* and *FOXC1* scaled across all clinical subgroups. **I** Predicted HRD frequency using HRDetect. **J** Top three most frequently altered driver genes in PAM50 Basal-like tumors in ER+/HER2− and TNBC disease. *Subtypes denoted by intrinsic names only (LumA, LumB, Basal) correspond to ER+/HER2− disease. Figure adapted from Paper III [128].*

Comparison to TNBC

Given the molecular features described above, which in several aspects deviate from what would be expected in clinically ER-positive tumors, Basal tumors were compared to TNBC disease stratified by PAM50 subtype into Basal and non-Basal groups to assess whether the observed phenotype reflects a biological state consistent with PAM50 Basal tumors in their typical clinical context. Transcriptionally, Basal tumors were characterized by high proliferation and elevated immune response across both ER+/HER2− and TNBC. The presence of a more immune-infiltrated tumor microenvironment in ER+/HER2− Basal tumors was further supported by TIL counts available for TNBC cases, which corroborated observed immune response metagene scores (**Figure 11E**).

Despite their ER-positive classification, the steroid response and *ESR1* expression of ER+/HER2− Basal tumors was positioned closer to TNBC than to ER+/HER2− groups. Expression patterns of *FOXA1* and *FOXC1* showed the same direction of dysregulation in both Basal groups. This similarity was further reflected at the epigenetic level through shared patterns of *FOXA1* promoter hypermethylation and *FOXC1* promoter hypomethylation, shown by comparing DNA methylation patterns of ER+/HER2− Basal tumors to TNBC tumors stratified into epiBasal and epiNonBasal groups. These epitypes were defined by Aine et al., based on a set of differentially methylated CpG sites which were used to assess whether epigenetic similarity extends beyond individual genes to the global level [66]. Within these CpGs, DNA methylation patterns demonstrated a close resemblance between ER+/HER2− Basal tumors and TNBC epiBasal tumors, indicating the Basal phenotype to be preserved at the DNA methylation level.

At the genomic level, both ER+/HER2− Basal and TNBC Basal tumors were characterized by a similar landscape of copy number alteration frequencies and driver gene alterations. Notably, homologous recombination deficiency represented a key feature, with both ER+/HER2− Basal and TNBC Basal tumors showing markedly elevated HRD frequencies compared to non-Basal subtypes.

Taken together, these analyses revealed a strong and consistent molecular similarity between ER+/HER2− Basal and TNBC Basal tumors, demonstrating that the Basal phenotype transcends clinical receptor classification. Furthermore, these findings support the hypothesis that all Basal tumors share a cellular origin within the luminal

progenitor compartment, consistent with the concept of luminal-to-basal plasticity and observed elevated *ELF5* expression in Basal tumors, given its role in promoting an ER-negative fate by suppressing *ESR1* and *FOXA1* expression [142-144].

Determinants of subtype classification

The high centroid distinctiveness and classification stability of Basal tumors across clinical subgroups observed in **Paper I** is consistent with the transcriptionally coherent profile of ER+/HER2− Basal tumors and with the striking molecular similarity to TNBC Basal tumors described above. This stability likely reflects that Basal tumors have a robust core transcriptional state that resists perturbation regardless of which biological programs are excluded and retain a strong transcriptional similarity to the prototypical Basal samples from which the centroid was originally derived. The consistency of this stability across clinical subgroups, suggests that Basal classification captures a deeply rooted biological identity rather than a context-dependent classification artifact.

Together, these observations suggest that Basal classification in ER+/HER2− disease is not an artifact of the classification framework but reflects a genuinely Basal transcriptional state that persists irrespective of clinical receptor status.

Therapeutic opportunities

The high-risk and proliferative nature of ER+/HER2− Basal tumors supports similar treatment considerations as discussed for HER2E tumors, including chemotherapy and CDK4/6 inhibitors in the early setting. Beyond these, the striking molecular similarity to TNBC Basal tumors raises the important question of whether patients with ER+/HER2− Basal tumors could additionally benefit from treatment approaches established in TNBC.

Two features of particular therapeutic relevance emerging from this similarity are the high predicted HRD-positive frequency and the immune-infiltrated tumor microenvironment. The high frequency of HRD-positive Basal tumors (44%) suggests potential benefit from PARP inhibitors, an approach with demonstrated benefit in high-risk HER2− early breast cancer with germline *BRCA1/2* pathogenic variants [84]. Given that a proportionally high rate of germline *BRCA1/2* carriers is likely within this subgroup, PAM50 Basal classification in ER+/HER2− disease could be considered as an indication for germline testing.

Finally, as already mentioned previously, immune checkpoint inhibition is clinical routine in neoadjuvant therapy of TNBC and recent trials demonstrated promising results also in high-risk ER+/HER2− disease, with greater benefit observed in tumors with higher immune activation [80, 135]. Therefore, the immune-infiltrated profile of ER+/HER2− Basal tumors and their overall molecular similarity to TNBC Basal disease, makes them a particularly relevant subgroup in this context, and PAM50 Basal classification may represent a useful tool for identifying ER+/HER2−

patients most likely to benefit from immune checkpoint inhibition. Notably, the GIADA trial demonstrated that the combination of a Basal intrinsic subtype and high TILs predicted pathological complete response in HR+/HER2− patients treated with neoadjuvant chemotherapy and immune checkpoint inhibition, providing direct clinical evidence supporting this notion [145].

Limitations

The primary limitation shared across both studies is the low prevalence of the investigated subgroups. In ER+/HER2− breast cancer both HER2E and Basal tumors each constitute approximately 2% of the population, resulting in small absolute group sizes that affect statistical power across analyses. In addition, the limited number of cases restricts the ability to perform more detailed comparisons of treatment effects across different therapeutic regimens.

Both studies relied on observational cohorts rather than randomized clinical trials, meaning that treatment groups are subject to potential selection bias and treatment effects. The METABRIC cohort, used as an independent validation dataset in both studies, may not fully reflect the general breast cancer population given its patient selection and diagnostic period, and the lower ER-positivity threshold applied in METABRIC compared to the Swedish guideline definition used in SCAN-B likely contributes to differences in subgroup frequency and molecular characteristics between cohorts.

For both HER2E and Basal tumors in ER+/HER2−, immune infiltration estimates were based on RNA sequencing rather than in situ quantification such as TIL scoring or multiplexed immunohistochemistry, although the specific immune response metagene used has previously been shown to correlate well with pathologist-assessed TIL counts in SCAN-B TNBC tumors [146]. Additionally, bulk mRNA expression data is subject to confounding from tissue heterogeneity and variable tumor cellularity, which may influence both subtype assignment and transcriptomic characterization. Finally, another limitation is the use of FPKM-normalized expression values for transcriptomic analyses in SCAN-B. However, the concordance of findings across two independent cohorts profiled on different platforms, with SCAN-B using RNA sequencing and METABRIC using microarray, suggests the results are not an artefact of normalization bias.

Paper IV

Current prognostic assessment of recurrence risk in breast cancer relies heavily on clinicopathological variables yet fails to fully resolve outcome heterogeneity within clinical subgroups. Gene expression-based prognostic assays in clinical use today have improved risk stratification in ER+/HER2− disease, but substantial uncertainty persists even within the risk categories they define, and their demonstrated clinical utility does not extend to other clinical breast cancer subgroups.

This has motivated the search for complementary molecular biomarkers capable of capturing additional dimensions of tumor biology such as DNA methylation. As a fundamental epigenetic regulatory mechanism, methylation alterations arise early in carcinogenesis, remain comparatively stable, and may encode regulatory features of tumor biology not fully reflected in transcriptional profiles alone. In early-stage breast cancer, where a substantial proportion of patients may die from unrelated causes before a recurrence event occurs, recurrence prediction modelling must take death as a competing risk into account.

This study sought to evaluate the prognostic utility of genome-wide tumor DNA methylation for recurrence risk prediction in early-stage breast cancer, comprising 1347 cases from the population-representative SCAN-B cohort and representing the largest study addressing this question to date.

Methodological Approach

Genome-wide methylation was profiled using Illumina EPIC arrays and filtered to approximately 190,000 CpG sites overlapping breast cancer-specific open chromatin regions defined by ATAC sequencing, enriching for sites where methylation is more likely to carry functional regulatory consequences [66, 147]. All prognostic modelling used the Fine-Gray subdistribution hazard framework, which directly models the cumulative incidence of recurrence in the presence of the competing event of death without prior recurrence (DWR), yielding absolute risk estimates appropriate for clinical use [92]. This was particularly relevant in this study cohort, as both the event of interest (recurrence) and the competing risk (DWR) occurred at similar frequencies. Models were developed separately for ER+/HER2− breast cancer, TNBC, and the full cohort, allowing the prognostic value of DNA methylation to be assessed across biologically divergent breast cancer subgroups known to differ in their recurrence patterns and underlying epigenetic landscapes.

The full model construction pipeline, comprising CpG site selection, methylation risk score construction, and final model fitting, was implemented within a 5-fold nested cross-validation framework to provide unbiased internal performance estimates.

CpG site selection combined variance filtering with cause-specific elastic-net penalized Cox regression, fit separately for RFI and DWR outcomes. This approach is appropriate for the competing risks setting, where distinct biological processes may underlie recurrence and death without recurrence, and where both cause-specific hazards jointly determine the cumulative incidence of the event of interest. Modelling each endpoint separately therefore allows identification of CpG sites associated with either outcome, capturing features that influence recurrence risk directly or indirectly through competing events.

Furthermore, when working in a high-dimensional setting with hundreds of thousands of CpG sites, enforcing sparsity in feature selection to identify a small number of prognostically relevant sites is clinically desirable, as it yields interpretable models based on individual features rather than complex combinations derived from dimensionality reduction. If a robust sparse signature can be identified and validated, it could potentially be assessed using targeted assays rather than genome-wide profiling, which may considerably lower the technical and cost barriers to routine clinical implementation.

Following their selection, the CpG sites were used to derive a single continuous methylation risk score (MeRS) per patient by fitting a ridge-penalized Fine-Gray model. Finally, three unpenalized Fine-Gray models were evaluated in each outer fold, differing in their predictor sets: (1) a methylation-only model incorporating the MeRS as the sole predictor; (2) a clinical-only model incorporating established clinical prognostic factors (Tumor Size, Lymph Node Status, Tumor Grade, and Patient Age at diagnosis); for the overall cohort, ER, PR, and HER2 status were additionally included as clinical prognostic variables; and (3) a combined model incorporating both the MeRS and clinical prognostic factors. Prediction performance was assessed by calculating time-dependent AUC and integrated Brier scores (IBS) on the held-out test folds and averaged across the five outer folds (see **Figure 6** for a conceptual overview of the nested cross-validation framework).

Prognostic Performance Across Clinical Subgroups

The prognostic value of DNA methylation differed substantially across clinical subgroups. Within the current modelling framework, the addition of methylation information improved discrimination over clinical variables in ER+/HER2− breast cancer, where the methylation-based model achieved a mean time-dependent AUC of 0.718 compared to 0.642 for the clinical model, with a small improvement in calibration. In contrast, no meaningful improvement was observed in TNBC or in the full cohort, where methylation showed lower or similar performance compared to clinical variables.

This subgroup-specific pattern may reflect underlying biological and clinical differences between breast cancer subgroups and parallels the restriction of

validated clinical utility to ER+/HER2− disease of established gene expression-based prognostic assays. In TNBC, recurrence risk is partially explained by clinicopathological variables, indicating unresolved residual heterogeneity, yet no prognostic signal was captured by the methylation risk score (mean AUC of 0.510). This may suggest that the relevant biological processes driving recurrence in TNBC are not well reflected in DNA methylation profiles or that they are not captured by the employed modelling framework. In the full cohort, the methylation model did capture some prognostic signal (mean AUC of 0.674), but the combined model did not improve upon the clinical model. This suggests that the methylation information captured in this context may largely reflect the same biological differences already encoded in receptor status, which are included as predictors in the full cohort model. If the methylation signals capturing recurrence risk are subgroup-specific, then selecting features across pooled subgroups is unlikely to identify a coherent prognostic signature. This interpretation is supported by the limited overlap between CpG sites selected in ER+/HER2− and those selected in the full cohort model.

At first glance, it may appear counterintuitive that the combined model incorporating both the MeRS and clinical variables performed intermediately between the two single-predictor models, rather than matching or exceeding the better of the two. However, this behavior likely reflects a bias-variance tradeoff inherent to the unpenalized Fine-Gray models used at the final fitting stage [94]. While unpenalized models are necessary to obtain absolute cumulative incidence estimates, the absence of regularization means that additional predictors increase estimation variance without constraint. In a low-event setting, predictors carrying limited independent information therefore introduce noise rather than signal, an effect compounded by the MeRS itself being an estimated quantity with inherent uncertainty. As a result, the combined model is more susceptible to overfitting and unstable coefficient estimates, leading to performance that falls between the two individual models rather than improving upon them.

Time-Varying Discriminative Performance in ER+/HER2− Breast Cancer

An interesting observation in ER+/HER2− breast cancer was that the performances of different model types varied over time. Clinical predictors were the stronger discriminators in the early years after diagnosis, but the methylation score progressively outperformed them over time (**Figure 12B**). This may be explained by early recurrences being driven by aggressive tumor characteristics well captured by clinicopathological variables, such as grade and nodal status, whereas late recurrences involve mechanisms such as tumor dormancy and immune evasion that may be more durably encoded in the epigenetic state of the tumor [148-153].

This time-varying pattern also indicates a potentially limited added prognostic value of DNA methylation over clinicopathological variables in TNBC, since if its utility lies specifically in capturing late recurrence risk, it would be expected to contribute little in a subgroup where recurrences cluster in the first few years [154]. Furthermore, it is worth noting that methylation data unadjusted for tumor purity may partly reflect tumor microenvironment composition and immune infiltration, signals that could in principle carry prognostic relevance in the TNBC context given the established prognostic role of tumor-infiltrating lymphocytes, yet no such signal was captured under the current modelling framework [155].

Within the ER+/HER2− subgroup, the low correlation between predicted risks from the clinical and methylation models ($r = 0.17$) further suggests that the two predictor sets capture largely distinct aspects of recurrence-relevant tumor biology. Nevertheless, the combined model was predominantly driven by the MeRS as indicated by the high correlation of predicted risks by the two models ($r = 0.98$) (**Figure 12D**). This raises the question if alternative combination strategies, such as ensembling the clinical and methylation model predictions rather than jointly fitting all predictors in a single model, could better leverage the complementary information captured by each predictor set.

Taken together, these findings suggest that the prognostic value of DNA methylation lies primarily in the assessment of late recurrence risk of patients with ER+/HER2− breast cancer.

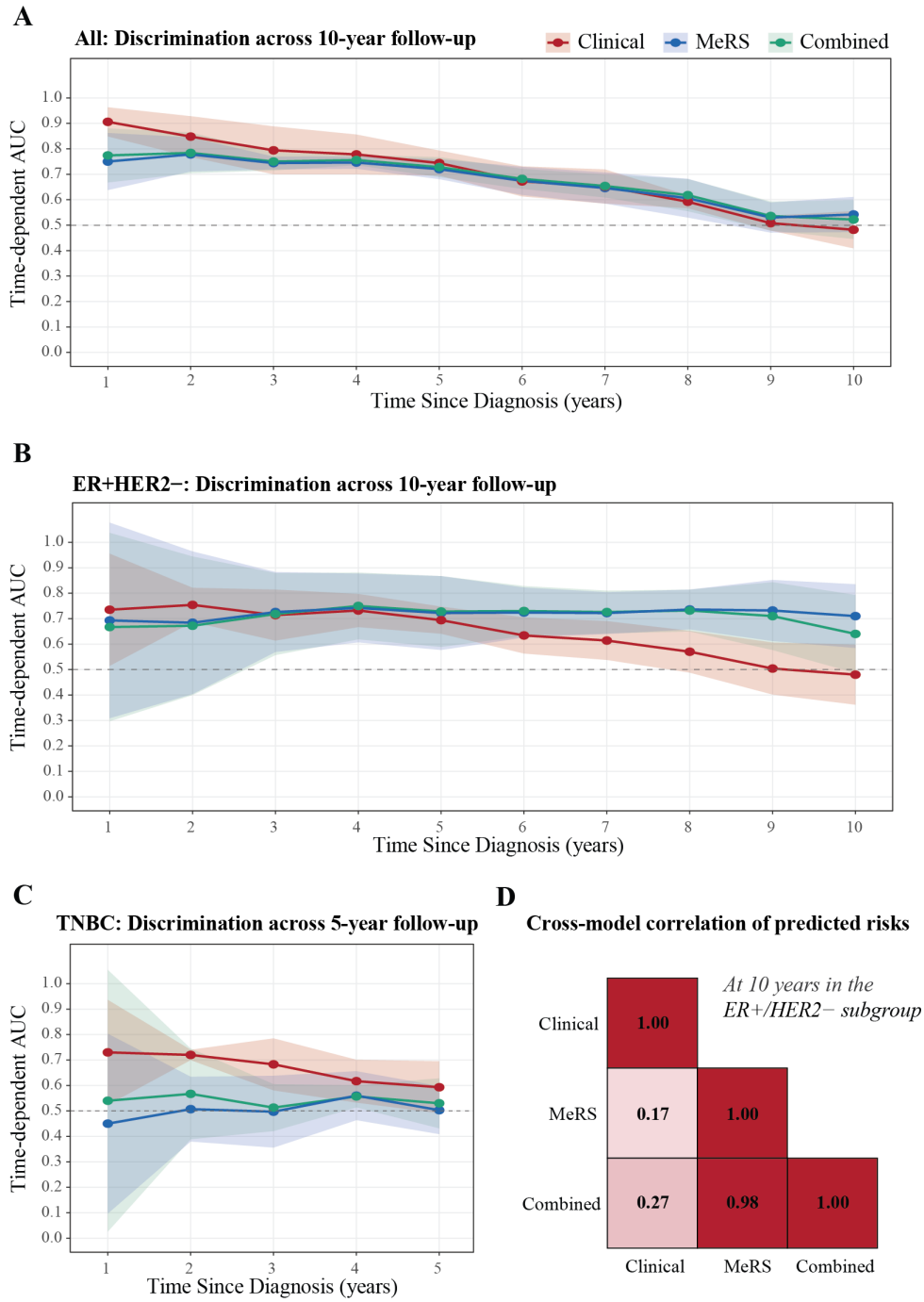
A
All: Discrimination across 10-year follow-up
Clinical
MeRS
Combined
Time-dependent AUC
Time Since Diagnosis (years)
B
ER+HER2−: Discrimination across 10-year follow-up
Time-dependent AUC
Time Since Diagnosis (years)
C
TNBC: Discrimination across 5-year follow-up
Time-dependent AUC
Time Since Diagnosis (years)
D
Cross-model correlation of predicted risks
At 10 years in the
ER+/HER2− subgroup
Clinical
MeRS
Combined
1.00
0.17
1.00
0.27
0.98
1.00
Clinical
MeRS
Combined

Figure 12. Time-dependent AUC across predictor sets and breast cancer subgroups.

A Time-dependent AUC as a function of follow-up time (up to 10 years) after diagnosis in the full cohort. **B** Time-dependent AUC as a function of follow-up time (up to 10 years) after diagnosis in the ER+/HER2− subgroup. **C** Time-dependent AUC as a function of follow-up time (up to 5 years) after diagnosis in the TNBC subgroup, with administrative censoring applied. **D** Pearson correlation of predicted risks across models in the ER+/HER2− subgroup at 10 years. Predictors included in Fine-Gray models: Clinical (red); MeRS (blue); Combined (Clinical + MeRS; green). Points represent mean AUC across five outer cross-validation folds, with shaded bands indicating 95% confidence intervals. *Figure adapted from Paper IV.*

Biological Interpretation of Selected CpG Sites

As part of the biological interpretation of selected CpG sites, each site was annotated with respect to its genomic context, and the sets of CpGs selected in each clinical subgroup were tested for enrichment of biological functions and pathways. No significant overlaps were identified, which likely reflects both the modest number of selected sites and the nature of penalized regression, which identifies features that collectively optimize predictive performance rather than sites sharing a common biological function, potentially resulting in a sparse and functionally heterogeneous feature set.

To contextualize the methylation-based risk predictions with the observations from **Paper II** and **Paper III**, the distribution of predicted recurrence risk was examined across PAM50 subtypes in the ER+/HER2− subgroup (**Figure 13A-C**). In the clinical model, LumB, HER2E and Basal tumors were predicted to have a higher risk of recurrence than LumA, consistent with their association with more aggressive clinicopathological features demonstrated in the previous studies. In the models including the MeRS however, HER2E and Basal tumors were elevated above both luminal subtypes, while LumA and LumB showed comparable predicted risk. This suggests that the MeRS captures recurrence-relevant epigenetic signal that is not aligned with proliferation-associated differences between LumA and LumB, but appears to be reflected in the classification of HER2E and Basal tumors. This is consistent with the findings in **Paper III**, where Basal tumors demonstrated a distinct epigenetic profile that was preserved across clinical contexts, suggesting that the Basal epigenetic phenotype may partially reflect recurrence-associated features.

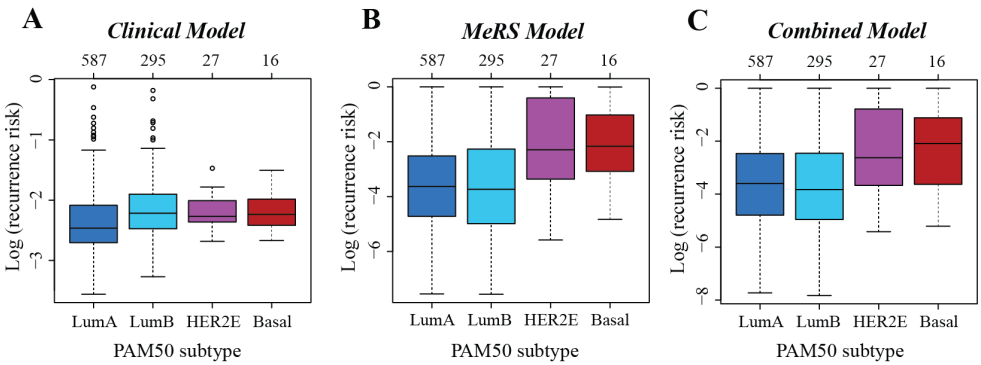

Figure 13. Predicted 10-year recurrence risk by PAM50 subtype in the ER+/HER2− subgroup.

A Predicted 10-year recurrence risk using the clinical model. **B** Predicted 10-year recurrence risk using the methylation risk score. **C** Predicted 10-year recurrence risk using the combined model. Risks correspond to cumulative incidence estimated from Fine Gray models and are shown on the log scale. *Figure adapted from Paper IV.*

Limitations

The relatively favorable outcomes in early-stage breast cancer in combination with a population-representative contemporary cohort result in low recurrence event rates, which constrains the precision of all estimated effects and limits model complexity. Furthermore, as treatment decisions in routine clinical practice are partly guided by the same clinicopathological features used as predictors in this study, patients at higher risk may have systematically received more aggressive adjuvant treatment, potentially reducing their recurrence risk and thereby attenuating the prognostic signals the models aim to capture. A further consideration is that the methylation data used for model development in this study were not adjusted for tumor purity and may therefore partly reflect tumor microenvironment composition in addition to tumor-intrinsic epigenetic states.

The employed modelling framework assumes linear effects of predictors on the log-hazard scale and does not capture non-linear or interaction-based methylation effects, though such effects could in principle be incorporated in alternative modelling approaches.

Additionally, restricting CpG sites to open chromatin regions, while biologically motivated, necessarily excludes sites in other genomic contexts that may carry prognostic signal. Similarly, the variance filter limits the analysis to the most variable CpG sites, which may reflect broader tumor biology rather than prognosis specifically, and thus may not fully capture prognostically relevant methylation. Moreover, CpGs should ideally be interpreted in a genomic context, including their genomic position, as illustrated by the presence of CpG islands in the genome. CpG context and genomic position could in principle also be incorporated in alternative

modelling approaches by e.g. using a concept of differentially methylated regions instead of single CpGs as input.

It should also be noted that clinical variables were deliberately not considered in the CpG site selection step, so that the methylation risk score entering both the methylation-only and combined models is derived independently of clinical information. While this ensures a clear assessment of the independent contribution of methylation, it may also mean that CpG sites with prognostic value specifically in combination with clinical variables were not captured during selection.

Finally, the performance estimates reported here are based on a robust nested cross-validation framework but remain internal to the training cohort. Independent external validation of the developed prognostic models will be required to establish their generalizability.

Conclusions

The aim of this thesis was to refine the use of molecular tumor profiling for stratification in early breast cancer by improving its biological interpretation and clinical relevance within established clinical subgroups, particularly ER+/HER2− disease.

Paper I demonstrates that PAM50 subtype assignment is better understood as a continuous positioning of tumors across multiple centroids rather than a discrete classification into fixed biological categories. Classification primarily reflects the combined contribution of broader transcriptional programs reflecting proliferation, steroid response, and basal keratin expression, alongside prototypically selected genes included to capture subtype-specific features. Together, these findings challenge the conventional interpretation of PAM50 subtypes as distinct and separate tumor classes, supporting instead an interpretation where subtype assignment represents the relative balance of underlying biological processes, and where a single nearest-centroid label is a conceptual oversimplification of intrinsic tumor biology.

Paper II and **Paper III** demonstrate that tumors subtyped as PAM50 HER2-enriched or PAM50 Basal-like within ER+/HER2− breast cancer are not classification artifacts but reflect molecular identities consistent with their prototypical counterparts in HER2+ disease and TNBC, respectively, highlighting that PAM50 captures biology transcending clinical receptor classification. Both subgroups carry substantially worse prognosis than canonical luminal tumors when treated with endocrine therapy alone, identifying an unmet clinical need for additional treatment options in this patient population. Their molecular characterization identifies several potentially actionable features, including immune infiltration, homologous recombination deficiency, and elevated *FGFR4* expression, with PAM50 subtype assignment representing one potential avenue for patient selection in future trials.

Paper IV demonstrates that tumor DNA methylation offers potential prognostic value beyond clinical variables specifically in ER+/HER2− breast cancer, with its utility lying primarily in capturing late recurrence risk. This positions methylation profiling as a promising complementary biomarker in this setting, addressing a clinically important window where established clinicopathological variables lose discriminative power.

Future Perspectives

The findings of this thesis raise several questions and open avenues for future investigation. The improved understanding of the PAM50 classification framework and its underlying biological processes provides a basis for refinement, potentially improving the robustness and biological relevance of subtype assignments across diverse cohorts. Beyond the current reliance on mRNA expression alone, the molecular layers characterizing tumor biology are deeply interconnected, and integrating multiple modalities for subtyping has the potential to yield a more comprehensive and biologically coherent characterization of intrinsic tumor biology. Furthermore, the application of single-cell and spatial approaches could offer important insights into intratumor heterogeneity and how distinct cell populations within the same tumor contribute to subtype assignments. A clinical consideration emerging from this thesis is whether PAM50 subtyping can guide patient selection for targeted therapies in ER+/HER2− disease, which could be explored through biomarker-enriched subgroup analyses within clinical trials. Given the suggested immune-infiltrated tumor microenvironment of both HER2-enriched and Basal-like tumors, these labels may aid in identifying ER+/HER2− patients most likely to benefit from immune checkpoint inhibition. Similarly, as the clinical development of FGFR4-targeted agents progresses, HER2-enriched tumors may represent a subgroup of particular interest, with promoter hypomethylation as a potential biomarker for patient selection. For Basal-like tumors, the high frequency of homologous recombination deficiency raises the question of whether their classification in ER+/HER2− disease could serve as an indication for germline *BRCA1/2* testing, potentially identifying patients who may benefit from PARP inhibitors. Additionally, as population-based cohorts grow over time, larger sample sizes will enable more detailed characterization and treatment-stratified analyses that could more directly inform the clinical management of rare high-risk patient groups. Finally, how to best leverage the growing availability of high-dimensional molecular data for recurrence risk prediction remains an open question, spanning methodological choices around modelling complexity and feature selection. As demonstrated in this thesis, different molecular layers capture distinct dimensions of tumor biology, and integrating multiple modalities represents a promising direction for more comprehensively resolving the outcome heterogeneity that persists across breast cancer subgroups. How such modalities are most effectively combined, whether through joint integration or through ensemble approaches that preserve the independent signal of each predictor, remains to be established.

Acknowledgements

These four years would not have been the same without a lot of people, and I would like to take a moment to thank them here.

First of all, **Johan** (**Staaf**), my supervisor for this thesis. Thank you for the confidence you placed in me and for the freedom you gave me to work in my own way, follow my own ideas, and find my own path. You were always there when I needed advice, and that combination of trust and support defined this PhD for me. I genuinely believe I could not have had better conditions to do this work, and I enjoyed these four years as much as I did largely because of that dynamic.

My co-supervisors, **Johan** (**Vallon-Christersson**), thank you for all the nice exchanges and ideas we discussed around the SCAN-B cohort, and **Mattias** (**Aine**), for the advice and the good exchanges on DNA methylation biology.

Aurélien (**Latouche**), thank you for the opportunity to join your group at Institut Curie. I felt at home from the very first day and really appreciated you taking the time to give advice whenever I needed it, as well as the welcoming and supportive atmosphere you created during my time in Paris.

The past and present members of the Lung/Breast cancer group not only for their academic support and discussions, but also for making this a truly enjoyable place to spend these years. **Deborah**, having you in the group was a blessing. You were always ready to help, always encouraging, and always in good humor. Honestly, thank you, having you in my corner made everything so much easier. **Suze**, sharing an office and talking through our PhD experiences together shaped how I found my footing. Thank you for all the advice and fun discussions. **Iñaki**, colleague and close friend. From sleeping on your couch to sharing coffee in Lund, in Paris, and wherever we end up next, it is always a pleasure, and it always will be. **Sunny**, you earn that nickname every day. Your positive energy, your good input on projects, and your encouraging words always gave me a boost. **Elsa**, it was really nice sharing an office with you at the end, I always enjoyed our chats. I would also like to thank **Anna**, **Christel**, **Daniel**, **Frida**, **Gudrun**, **Isa**, **Jari**, **Maria**, **Marija**, **Mats**, **Per Niklas** and **Sara** for the good company at Christmas dinners, group retreats, and all the moments in between.

The members of the Statistical Methods for Precision Medicine group at Institut Curie, **Antoine**, **Alexandre**, **Astrid**, **Beatriz**, **Cristina**, **Jimmy**, **Mary**, and **Xavier**,

as well as everyone else at the Institut Curie site in Saint-Cloud, thank you for the warm welcome and for all the fun and constructive conversations.

The Swedish National Graduate School in Medical Bioinformatics for the great courses, for letting me discover different universities and cities across Sweden, and for the chance to meet like-minded PhD students along the way.

The past and present PhD students I met during my time in Lund, **Dora**, **Jacob**, **Juliane**, **Margareta**, **Maria**, **Marius**, **Markus**, **Mirjam**, **Völli**, and many others I am surely forgetting, thank you for the good company during coffee breaks and lunches. **Arthur**, thank you for being such a good friend since the very beginning of my time in Lund.

Mama und Papa, danke, dass ihr mich immer und bedingungslos unterstützt habt, egal wohin es mich zum Studieren verschlagen hat. Diese Doktorarbeit ist auch euer Verdienst, denn ohne eure Unterstützung wäre das alles nicht möglich gewesen. Ihr habt mich meinen eigenen Weg gehen lassen und mich nie zurückgehalten. Das bedeutet mir mehr als ich in Worte fassen kann.

Annika, ich bewundere wie du deinen eigenen Weg gehst und dabei so viel auf dich nimmst und souverän meisterst. Und ich bin dir von Herzen dankbar, dass du dich so selbstverständlich um die Dinge zuhause kümmerst, ohne es je zu erwähnen. Nur dadurch ermöglichst du mir, überhaupt im Ausland zu leben, und das werde ich dir nie vergessen.

Sabrina, tu as été à mes côtés tout au long de cette aventure, et avant même que je ne commence mes études. Tu m'as donné rien que de l'amour et un soutien constant. Tu as soutenu ma décision de faire ce doctorat même lorsque cela signifiait que nous ne vivions pas au même endroit. Je n'aurais pas pu faire cela sans toi, et je t'en suis reconnaissant chaque jour. Je t'aime et j'ai hâte de tout ce qui nous attend.

References

1. Hajdu SI. A note from history: landmarks in history of cancer, part 1. Cancer. 2011;117(5):1097-102.
2. Hajdu SI. A note from history: landmarks in history of cancer, part 2. Cancer. 2011;117(12):2811-20.
3. Hajdu SI. A note from history: landmarks in history of cancer, part 3. Cancer. 2012;118(4):1155-68.
4. Hajdu SI. A note from history: landmarks in history of cancer, part 4. Cancer. 2012;118(20):4914-28.
5. Vogelstein B, Kinzler KW. The multistep nature of cancer. Trends in Genetics. 1993;9(4):138-41.
6. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000;100(1):57-70.
7. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
8. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discovery. 2022;12(1):31-46.
9. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013;501(7467):338-45.
10. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. New England Journal of Medicine. 2012;366(10):883-92.
11. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481(7381):306-13.
12. Anderson NM, Simon MC. The tumor microenvironment. Current Biology. 2020;30(16):R921-R5.
13. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, et al. Breast cancer. Nature Reviews Disease Primers. 2019;5(1):66.
14. Howlader N, Noone AM, Krapcho M. SEER Cancer Statistics Review, 1975–2017. Bethesda, MD: National Cancer Institute; 2020.
15. Hortobagyi GN. Breast Cancer: 45 Years of Research and Progress. Journal of Clinical Oncology. 2020;38(21):2454-62.
16. Pan H, Gray R, Braybrooke J, Davies C, Taylor C, McGale P, et al. 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. New England Journal of Medicine. 2017;377(19):1836-46.

17. Trapani D, Ginsburg O, Fadelu T, Lin NU, Hassett M, Ilbawi AM, et al. Global challenges and policy solutions in breast cancer control. Cancer Treatment Reviews. 2022;104:102339.
18. Socialstyrelsen. Statistik om bröstcancer, Accessed: 20 March 2026, https://www.socialstyrelsen.se/publikationer/statistik-om-brostcancer-2023-10-8807.
19. McCormack V, McKenzie F, Foerster M, Zietsman A, Galukande M, Adisa C, et al. Breast cancer survival and survival gap apportionment in sub-Saharan Africa (ABC-DO): a prospective cohort study. The Lancet Global Health. 2020;8(9):e1203-e12.
20. International Agency for Research on Cancer. Global Cancer Observatory: Cancer Today: International Agency for Research on Cancer, Accessed: 12 March 2026, https://gco.iarc.who.int/today.
21. Rojas K, Stuckey A. Breast Cancer Epidemiology and Risk Factors. Clinical Obstetrics and Gynecology. 2016;59(4):651-72.
22. Russo J, Moral R, Balogh GA, Mailo D, Russo IH. The protective role of pregnancy in breast cancer. Breast Cancer Research. 2005;7(3):131-42.
23. Obeagu EI, Obeagu GU. Breast cancer: A review of risk factors and diagnosis. Medicine (Baltimore). 2024;103(3):e36905.
24. Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. Nature Genetics. 2008;40(1):17-22.
25. Breast Cancer Association Consortium. Breast Cancer Risk Genes — Association Analysis in More than 113,000 Women. New England Journal of Medicine. 2021;384(5):428-39.
26. Hu C, Hart SN, Gnanaolivu R, Huang H, Lee KY, Na J, et al. A Population-Based Study of Genes Previously Implicated in Breast Cancer. New England Journal of Medicine. 2021;384(5):440-51.
27. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. New England Journal of Medicine. 2009;361(2):123-34.
28. Koo MM, von Wagner C, Abel GA, McPhail S, Rubin GP, Lyratzopoulos G. Typical and atypical presenting symptoms of breast cancer and their associations with diagnostic intervals: Evidence from a national audit of cancer diagnosis. Cancer Epidemiology. 2017;48:140-6.
29. Duffy SW, Tabár L, Yen AM, Dean PB, Smith RA, Jonsson H, et al. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. Cancer. 2020;126(13):2971-9.
30. Autier P, Koechlin A, Smans M, Vatten L, Boniol M. Mammography screening and breast cancer mortality in Sweden. Journal of the National Cancer Institute. 2012;104(14):1080-93.
31. Loibl S, André F, Bachelot T, Barrios CH, Bergh J, Burstein HJ, et al. Early breast cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Annals of Oncology. 2024;35(2):159-82.
32. Amin MB, Edge S, Greene F. AJCC Cancer Staging Manual. 8th ed 2017.

33. Giammarile F, Vidal-Sicart S, Paez D, Pellet O, Enrique EL, Mikhail-Lette M, et al. Sentinel Lymph Node Methods in Breast Cancer. Seminars in Nuclear Medicine. 2022;52(5):551-60.
34. Liang Y, Zhang H, Song X, Yang Q. Metastatic heterogeneity of breast cancer: Molecular mechanism and potential therapeutic targets. Seminars in Cancer Biology. 2020;60:14-27.
35. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991;19(5):403-10.
36. Rakha EA, El-Sayed ME, Lee AHS, Elston CW, Grainge MJ, Hodi Z, et al. Prognostic Significance of Nottingham Histologic Grade in Invasive Breast Carcinoma. Journal of Clinical Oncology. 2008;26(19):3153-8.
37. Board WHOCoTE. WHO Classification of Tumours. Volume 2: Breast Tumours. 5th ed. Lyon: International Agency for Research on Cancer; 2019.
38. U.S. National Institutes of Health NCI. SEER Training Modules: Types of Breast Histologies 2026, Accessed: 06 March 2026, https://training.seer.cancer.gov.
39. Frasor J, Danes JM, Komm B, Chang KCN, Lyttle CR, Katzenellenbogen BS. Profiling of Estrogen Up- and Down-Regulated Gene Expression in Human Breast Cancer Cells: Insights into Gene Networks and Pathways Underlying Estrogenic Control of Proliferation and Cell Phenotype. Endocrinology. 2003;144(10):4562-74.
40. Horwitz KB, McGuire WL. Estrogen control of progesterone receptor in human breast cancer. Correlation with nuclear processing of estrogen receptor. Journal of Biological Chemistry. 1978;253(7):2223-8.
41. Bardou VJ, Arpino G, Elledge RM, Osborne CK, Clark GM. Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. Journal of Clinical Oncology. 2003;21(10):1973-9.
42. Allison KH, Hammond MEH, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, et al. Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update. Journal of Clinical Oncology. 2020;38(12):1346-66.
43. Svensk Förening för Patologi. Kvalitetsdokument för bröstpatologi. 2025.
44. Acs B, Hartman J, Sönmez D, Lindman H, Johansson ALV, Fredriksson I. Real-world overall survival and characteristics of patients with ER-zero and ER-low HER2-negative breast cancer treated as triple-negative breast cancer: a Swedish population-based cohort study. The Lancet Regional Health Europe. 2024;40:100886.
45. Raghav KPS, Moasser MM. Molecular Pathways and Mechanisms of HER2 in Cancer Therapy. Clinical Cancer Research. 2023;29(13):2351-61.
46. Tarantino P, Hamilton E, Tolaney SM, Cortes J, Morganti S, Ferraro E, et al. HER2-Low Breast Cancer: Pathological and Clinical Landscape. Journal of Clinical Oncology. 2020;38(17):1951-62.

47. Staaf J, Häkkinen J, Hegardt C, Saal LH, Kimbung S, Hedenfalk I, et al. RNA sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer. npj Breast Cancer. 2022;8(1):94.
48. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.
49. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences. 2001;98(19):10869-74.
50. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. Journal of Clinical Oncology. 2009;27(8):1160-7.
51. Veerla S, Hohmann L, Nacer DF, Vallon-Christersson J, Staaf J. Perturbation and stability of PAM50 subtyping in population-based primary invasive breast cancer. npj Breast Cancer. 2023;9(1):83.
52. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. New England Journal of Medicine. 2004;351(27):2817-26.
53. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530-6.
54. Vijver MJvd, He YD, Veer LJvt, Dai H, Hart AAM, Voskuil DW, et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. New England Journal of Medicine. 2002;347(25):1999-2009.
55. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nature Medicine. 2017;23(4):517-25.
56. Yuan T, Edelmann D, Fan Z, Alwers E, Kather JN, Brenner H, Hoffmeister M. Machine learning in the identification of prognostic DNA methylation biomarkers among patients with cancer: A systematic review of epigenome-wide studies. Artificial Intelligence in Medicine. 2023;143:102589.
57. Zarean E, Li S, Southey MC, Dugué PA. A review of the use of tumour DNA methylation for breast cancer subtyping and prediction of outcomes. Clinical Epigenetics. 2025;17(1):109.
58. Loi S, Drubay D, Adams S, Pruneri G, Francis PA, Lacroix-Triki M, et al. Tumor-Infiltrating Lymphocytes and Prognosis: A Pooled Individual Patient Analysis of Early-Stage Triple-Negative Breast Cancers. Journal of Clinical Oncology. 2019;37(7):559-69.
59. Burstein HJ. Systemic Therapy for Estrogen Receptor–Positive, HER2-Negative Breast Cancer. New England Journal of Medicine. 2020;383(26):2557-70.
60. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science. 1987;235(4785):177-82.

61. Moasser MM. The oncogene HER2: its signaling and transforming functions and its role in human cancer pathogenesis. Oncogene. 2007;26(45):6469-87.
62. Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE, Davidson NE, et al. Trastuzumab plus Adjuvant Chemotherapy for Operable HER2-Positive Breast Cancer. New England Journal of Medicine. 2005;353(16):1673-84.
63. Bianchini G, Balko JM, Mayer IA, Sanders ME, Gianni L. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. Nature Reviews Clinical Oncology. 2016;13(11):674-90.
64. Staaf J, Glodzik D, Bosch A, Vallon-Christersson J, Reuterswärd C, Häkkinen J, et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nature Medicine. 2019;25(10):1526-33.
65. Denkert C, von Minckwitz G, Darb-Esfahani S, Lederer B, Heppner BI, Weber KE, et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. The Lancet Oncology. 2018;19(1):40-50.
66. Aine M, Nacer DF, Arbajian E, Veerla S, Karlsson A, Häkkinen J, et al. The DNA methylation landscape of primary triple-negative breast cancer. Nature Communications. 2025;16(1):3041.
67. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. New England Journal of Medicine. 2010;363(20):1938-48.
68. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.
69. Vallon-Christersson J, Häkkinen J, Hegardt C, Saal LH, Larsson C, Ehinger A, et al. Cross comparison and prognostic assessment of breast cancer multigene signatures in a large population-based contemporary clinical series. Scientific Reports. 2019;9(1):12184.
70. Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER, et al. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. New England Journal of Medicine. 2002;347(16):1233-41.
71. Slanetz PJ, Moy L, Baron P, diFlorio RM, Green ED, Heller SL, et al. ACR Appropriateness Criteria(®) Monitoring Response to Neoadjuvant Systemic Therapy for Breast Cancer. Journal of the American College of Radiology. 2017;14(11s):S462-S75.
72. Johnston S, Martin M, O'Shaughnessy J, Hegg R, Tolaney SM, Guarneri V, et al. Overall survival with abemaciclib in early breast cancer. Annals of Oncology. 2026;37(2):155-65.
73. Hortobagyi GN, Lacko A, Sohn J, Cruz F, Ruiz Borrego M, Manikhas A, et al. A phase III trial of adjuvant ribociclib plus endocrine therapy versus endocrine therapy alone in patients with HR-positive/HER2-negative early breast cancer: final invasive disease-free survival results from the NATALEE trial. Annals of Oncology. 2025;36(2):149-57.
74. Patel R, Klein P, Tiersten A, Sparano JA. An emerging generation of endocrine therapies in breast cancer: a clinical perspective. npj Breast Cancer. 2023;9(1):20.

75. Jeselsohn R, Buchwalter G, De Angelis C, Brown M, Schiff R. ESR1 mutations—a mechanism for acquired endocrine resistance in breast cancer. Nature Reviews Clinical Oncology. 2015;12(10):573-83.
76. Modi S, Jacot W, Yamashita T, Sohn J, Vidal M, Tokunaga E, et al. Trastuzumab Deruxtecan in Previously Treated HER2-Low Advanced Breast Cancer. New England Journal of Medicine. 2022;387(1):9-20.
77. Bardia A, Hurvitz SA, Tolaney SM, Loirat D, Punie K, Oliveira M, et al. Sacituzumab Govitecan in Metastatic Triple-Negative Breast Cancer. New England Journal of Medicine. 2021;384(16):1529-41.
78. Schmid P, Cortes J, Pusztai L, McArthur H, Kümmel S, Bergh J, et al. Pembrolizumab for Early Triple-Negative Breast Cancer. New England Journal of Medicine. 2020;382(9):810-21.
79. Rizzo A, Ricci AD. Biomarkers for breast cancer immunotherapy: PD-L1, TILs, and beyond. Expert Opinion on Investigational Drugs. 2022;31(6):549-55.
80. Cardoso F, O'Shaughnessy J, Liu Z, McArthur H, Schmid P, Cortes J, et al. Pembrolizumab and chemotherapy in high-risk, early-stage, ER+/HER2− breast cancer: a randomized phase 3 trial. Nature Medicine. 2025;31(2):442-8.
81. El Bairi K, Haynes HR, Blackley E, Fineberg S, Shear J, Turner S, et al. The tale of TILs in breast cancer: A report from The International Immuno-Oncology Biomarker Working Group. npj Breast Cancer. 2021;7(1):150.
82. Heater NK, Warrior S, Lu J. Current and future immunotherapy for breast cancer. Journal of Hematology and Oncology. 2024;17(1):131.
83. Lord CJ, Ashworth A. PARP inhibitors: Synthetic lethality in the clinic. Science. 2017;355(6330):1152-8.
84. Tutt ANJ, Garber JE, Kaufman B, Viale G, Fumagalli D, Rastogi P, et al. Adjuvant Olaparib for Patients with BRCA1- or BRCA2-Mutated Breast Cancer. New England Journal of Medicine. 2021;384(25):2394-405.
85. Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, et al. The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Medicine. 2015;7(1):20.
86. Rydén L, Loman N, Larsson C, Hegardt C, Vallon-Christersson J, Malmberg M, et al. Minimizing inequality in access to precision medicine in breast cancer by real-time population-based molecular analysis in the SCAN-B initiative. British Journal of Surgery. 2018;105(2):e158-e68.
87. 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):346-52.
88. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016;534(7605):47-54.
89. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958;53(282):457-81.

90. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological). 1972;34(2):187-220.
91. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine. 2007;26(11):2389-430.
92. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;94(446):496-509.
93. Steyerberg EW. Clinical Prediction Models - A Practical Approach to Development, Validation, and Updating. 2nd ed. Cham: Springer; 2019.
94. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction: Springer; 2009.
95. van Geloven N, Giardiello D, Bonneville EF, Teece L, Ramspek CL, van Smeden M, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. BMJ. 2022;377:e069249.
96. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
97. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45.
98. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458(7239):719-24.
99. LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Research. 2009;37(13):4181-93.
100. Shlien A, Malkin D. Copy number variations and cancer. Genome Medicine. 2009;1(6):62.
101. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences. 2010;107(39):16910-5.
102. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149(5):979-93.
103. 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.
104. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. 2013;31(3):213-9.
105. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nature Reviews Genetics. 2016;17(8):487-500.
106. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics. 2012;13(7):484-92.

107. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018;362(6413).
108. Baylin S, Bestor TH. Altered methylation patterns in cancer cell genomes: cause or consequence? Cancer Cell. 2002;1(4):299-305.
109. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.
110. Sasiain I, Nacer DF, Aine M, Veerla S, Staaf J. Tumor purity estimated from bulk DNA methylation can be used for adjusting beta values of individual samples to better reflect tumor biology. NAR Genomics and Bioinformatics. 2024;6(4):lqae146.
111. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics. 2016;32(2):286-8.
112. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nature Genetics. 2003;33(3):245-54.
113. 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. Proceedings of the National Academy of Sciences. 2005;102(43):15545-50.
114. Fredlund E, Staaf J, Rantala JK, Kallioniemi O, Borg A, Ringnér M. The gene expression landscape of breast cancer is shaped by tumor protein p53 status and epithelial-mesenchymal transition. Breast Cancer Research. 2012;14(4):R113.
115. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018.
116. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology. 2019;37(7):773-82.
117. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences. 2012;131(4):281-5.
118. Lien TG, Ohnstad HO, Lingjærde OC, Vallon-Christersson J, Aaserud M, Sveli MAT, et al. Sample Preparation Approach Influences PAM50 Risk of Recurrence Score in Early Breast Cancer. Cancers (Basel). 2021;13(23).
119. Denkert C, Budczies J, von Minckwitz G, Wienert S, Loibl S, Klauschen F. Strategies for developing Ki67 as a useful biomarker in breast cancer. Breast. 2015;24 Suppl 2:S67-72.
120. Nielsen T, Wallden B, Schaper C, Ferree S, Liu S, Gao D, et al. Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. BMC Cancer. 2014;14:177.
121. Makhlouf S, Althobiti M, Toss M, Muftah AA, Mongan NP, Lee AHS, et al. The Clinical and Biological Significance of Estrogen Receptor-Low Positive Breast Cancer. Modern Pathology. 2023;36(10):100284.

122. Fujii T, Kogawa T, Dong W, Sahin AA, Moulder S, Litton JK, et al. Revisiting the definition of estrogen receptor positivity in HER2-negative primary breast cancer. Annals of Oncology. 2017;28(10):2420-8.
123. Cejalvo JM, Pascual T, Fernández-Martínez A, Brasó-Maristany F, Gomis RR, Perou CM, et al. Clinical implications of the non-luminal intrinsic subtypes in hormone receptor-positive breast cancer. Cancer Treatment Reviews. 2018;67:63-70.
124. Prat A, Cheang MC, Galván P, Nuciforo P, Paré L, Adamo B, et al. Prognostic Value of Intrinsic Subtypes in Hormone Receptor-Positive Metastatic Breast Cancer Treated With Letrozole With or Without Lapatinib. JAMA Oncology. 2016;2(10):1287-94.
125. Prat A, Chaudhury A, Solovieff N, Paré L, Martinez D, Chic N, et al. Correlative Biomarker Analysis of Intrinsic Subtypes and Efficacy Across the MONALEESA Phase III Studies. Journal of Clinical Oncology. 2021;39(13):1458-67.
126. 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. 2020;6:8.
127. Hohmann L, Sigurjonsdottir K, Campos AB, Nacer DF, Veerla S, Rosengren F, et al. Genomic characterization of the HER2-enriched intrinsic molecular subtype in primary ER-positive HER2-negative breast cancer. Nature Communications. 2025;16(1):2208.
128. Hohmann L, Nacer DF, Aine M, Memari Y, Black D, Bowden R, et al. Molecular profiling of the Basal-like intrinsic molecular subtype in primary ER-positive HER2-negative breast cancer. Genome Medicine. 2025;17(1):146.
129. Griguolo G, Dieci MV, Paré L, Miglietta F, Generali DG, Frassoldati A, et al. Immune microenvironment and intrinsic subtyping in hormone receptor-positive/HER2-negative breast cancer. npj Breast Cancer. 2021;7(1):12.
130. Prat A, Brase JC, Cheng Y, Nuciforo P, Paré L, Pascual T, et al. Everolimus plus Exemestane for Hormone Receptor-Positive Advanced Breast Cancer: A PAM50 Intrinsic Subtype Analysis of BOLERO-2. Oncologist. 2019;24(7):893-900.
131. Prat A, Solovieff N, Su F, Bardia A, Neven P, Hortobagyi GN, et al. Abstract PD2-05: Genomic profiling of PAM50-based intrinsic subtypes in HR+/HER2-advanced breast cancer (ABC) across the MONALEESA (ML) studies. Cancer Research. 2022;82(4_Supplement):PD2-05.
132. Griffith OL, Spies NC, Anurag M, Griffith M, Luo J, Tu D, et al. The prognostic effects of somatic mutations in ER-positive breast cancer. Nature Communications. 2018;9(1):3476.
133. Berns EM, Klijn JG, Look MP, Grebenchtchikov N, Vossen R, Peters H, et al. Combined vascular endothelial growth factor and TP53 status predicts poor response to tamoxifen therapy in estrogen receptor-positive advanced breast cancer. Clinical Cancer Research. 2003;9(4):1253-8.
134. Bardia A, Hu X, Dent R, Yonemori K, Barrios CH, O'Shaughnessy JA, et al. Trastuzumab Deruxtecan after Endocrine Therapy in Metastatic Breast Cancer. New England Journal of Medicine. 2024;391(22):2110-22.

135. Loi S, Salgado R, Curigliano G, Romero Díaz RI, Delaloge S, Rojas García CI, et al. Neoadjuvant nivolumab and chemotherapy in early estrogen receptor-positive breast cancer: a randomized phase 3 trial. Nature Medicine. 2025;31(2):433-41.
136. Klocker EV, Egle D, Bartsch R, Rinnerthaler G, Gnant M. Efficacy and Safety of CDK4/6 Inhibitors: A Focus on HR+/HER2- Early Breast Cancer. Drugs. 2025;85(2):149-69.
137. Garcia-Recio S, Thennavan A, East MP, Parker JS, Cejalvo JM, Garay JP, et al. FGFR4 regulates tumor subtype differentiation in luminal breast cancer and metastatic disease. Journal of Clinical Investigation. 2020;130(9):4871-87.
138. Levine KM, Ding K, Chen L, Oesterreich S. FGFR4: A promising therapeutic target for breast cancer and other solid tumors. Pharmacology and Therapeutics. 2020;214:107590.
139. Martin EM, Orlando KA, Yokobori K, Wade PA. The estrogen receptor/GATA3/FOXA1 transcriptional network: lessons learned from breast cancer. Current Opinion in Structural Biology. 2021;71:65-70.
140. Elian FA, Yan E, Walter MA. FOXC1, the new player in the cancer sandbox. Oncotarget. 2018;9(8):8165-78.
141. Davies HR, Black D, Kvist A, Sigurjónsdóttir K, Bosch A, Bowden R, et al. Homologous recombination deficiency in primary ER-positive and HER2-negative breast cancer. Communications Medicine. 2026;6(1):118.
142. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nature Medicine. 2009;15(8):907-13.
143. Mohamed GA, Mahmood S, Ognjenovic NB, Lee MK, Wilkins OM, Christensen BC, et al. Lineage plasticity enables low-ER luminal tumors to evolve and gain basal-like traits. Breast Cancer Research. 2023;25(1):23.
144. Kalyuga M, Gallego-Ortega D, Lee HJ, Roden DL, Cowley MJ, Caldon CE, et al. ELF5 suppresses estrogen sensitivity and underpins the acquisition of antiestrogen resistance in luminal breast cancer. PLOS Biology. 2012;10(12):e1001461.
145. Dieci MV, Guarneri V, Tosi A, Bisagni G, Musolino A, Spazzapan S, et al. Neoadjuvant Chemotherapy and Immunotherapy in Luminal B-like Breast Cancer: Results of the Phase II GIADA Trial. Clinical Cancer Research. 2022;28(2):308-17.
146. Roostee S, Ehinger D, Jönsson M, Phung B, Jönsson G, Sjödahl G, et al. Tumour immune characterisation of primary triple-negative breast cancer using automated image quantification of immunohistochemistry-stained immune cells. Scientific Reports. 2024;14(1):21417.
147. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics. 2019;20(4):207-20.
148. Ahn SG, Lee HM, Cho SH, Bae SJ, Lee SA, Hwang SH, et al. The difference in prognostic factors between early recurrence and late recurrence in estrogen receptor-positive breast cancer: nodal stage differently impacts early and late recurrence. PLOS One. 2013;8(5):e63510.
149. Crea F, Nur Saidy NR, Collins CC, Wang Y. The epigenetic/noncoding origin of tumor dormancy. Trends in Molecular Medicine. 2015;21(4):206-11.

150. Jung H, Kim HS, Kim JY, Sun JM, Ahn JS, Ahn MJ, et al. DNA methylation loss promotes immune evasion of tumours with high mutation and copy number load. Nature Communications. 2019;10(1):4278.
151. Mittempergher L, Saghatchian M, Wolf DM, Michiels S, Canisius S, Dessen P, et al. A gene signature for late distant metastasis in breast cancer identifies a potential mechanism of late recurrences. Molecular Oncology. 2013;7(5):987-99.
152. Esserman LJ, Moore DH, Tsing PJ, Chu PW, Yau C, Ozanne E, et al. Biologic markers determine both the risk and the timing of recurrence in breast cancer. Breast Cancer Research and Treatment. 2011;129(2):607-16.
153. Sosa MS, Bragado P, Aguirre-Ghiso JA. Mechanisms of disseminated cancer cell dormancy: an awakening field. Nature Reviews Cancer. 2014;14(9):611-22.
154. Ma T, Hao X-m, Chen H-d, Zheng M-h, Chen X-g, Cai S-L, Zhang J. Predictive markers of rapid disease progression and chemotherapy resistance in triple-negative breast cancer patients following postoperative adjuvant therapy. Scientific Reports. 2025;15(1):386.
155. Park JH, Jonas SF, Bataillon G, Criscitiello C, Salgado R, Loi S, et al. Prognostic value of tumor-infiltrating lymphocytes in patients with early-stage triple-negative breast cancers (TNBC) who did not receive adjuvant chemotherapy. Annals of Oncology. 2019;30(12):1941-9.

Printed by Media-Tryck, Lund 2026 NORDIC SWAN ECOLABEL 3041 0903

Precision Oncology in Breast Cancer

Breast cancers can differ substantially in their underlying biology, shaping patient prognosis and response to therapy. Molecular tumor profiling can capture this variation by classifying tumors into subtypes to guide individualized treatment decisions. Drawing on population-representative patient cohorts, this thesis examines what intrinsic molecular subtypes truly reflect biologically, and how they should be interpreted within a specific clinical context. It further explores whether DNA methylation provides prognostic value for assessing the risk of cancer recurrence beyond what is captured by traditional clinical factors.

Lennart Hohmann

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Laboratory Medicine

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2026:86
ISBN 978-91-8021-884-9
ISSN 1652-8220