Tumour evolution and novel biomarkers in breast cancer

Jana Biermann

UNIVERSITY OF GOTHENBURG

Department of Oncology
Institute of Clinical Sciences
Sahlgrenska Academy, University of Gothenburg

Gothenburg 2019

Tumour evolution and novel biomarkers in breast cancer

jana.biermann@gu.se

ISBN 978-91-7833-440-7 (PRINT)
ISBN 978-91-7833-441-4 (PDF)
http://hdl.handle.net/2077/59802

Printed in Gothenburg, Sweden 2019
Printed by BrandFactory

ABSTRACT

Several gene signatures have been proposed in the past two decades to improve outcome prediction for breast cancer patients and to guide treatment decisions. Current treatment guidelines, however, primarily focus on established clinicopathological features. In **Paper I**, we identified a novel 18-marker gene expression signature predicting breast cancer-specific survival. The 18-marker signature was validated in three independent cohorts and showed increased predictive power over the clinically validated Oncotype Dx signature.

Despite increasing survival rates, about 6-23% of patients suffer from recurrences within five years of initial diagnosis indicating treatment failure. It is highly important to differentiate between clonally related recurrences and independent primary tumours due to potentially differing prognoses and treatment regimes. Currently, there is no consensus on how to define clonal relatedness between multiple tumours in the same patient. In **Paper II**, we identified the Similarity Index (SI) as the most reliable tool to classify tumour clonality.

The mammary gland is known to be highly sensitive to radiation, especially at a young age. In the years from 1920-1965, a total of 17,200 female Swedish infants were treated with ionizing radiation for skin haemangioma, resulting in an increased risk of developing breast cancer. In **Paper III**, we analysed breast tumours for genomic instability, which can be induced by ionizing radiation. Patients with higher absorbed doses to the breast exhibited increased genomic instability compared to patients exposed to lower absorbed doses. These results strongly suggest radiation-induced genomic instability as a biological link between ionizing radiation exposure at a young age and the increased breast cancer risk in subsequent decades.

In conclusion, this work highlights the importance of complementing established clinicopathological features with molecular biology and statistical models to improve breast cancer risk assessment and personalize treatment strategies.

Keywords: breast cancer, gene signature, molecular biomarkers, tumour clonality, genomic instability, Swedish haemangioma cohort

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. **Biermann J**, Nemes S, Parris TZ, Engqvist H, Werner Rönnerman E, Forssell-Aronsson E, Steineck G, Karlsson P, Helou K. A novel 18-marker panel predicting clinical outcome in breast cancer.
Cancer Epidemiology, Biomarkers & Prevention (2017)
DOI: 10.1158/1055-9965.EPI-17-0606

II. **Biermann J**, Parris TZ, Nemes S, Danielsson A, Engqvist H, Werner Rönnerman E, Forssell-Aronsson E, Kovács A, Karlsson P, Helou K. Clonal relatedness in tumour pairs of breast cancer patients.
Breast Cancer Research (2018)
DOI: 10.1186/s13058-018-1022-y

III. **Biermann J**, Langen B, Nemes S, Holmberg E, Parris TZ, Werner Rönnerman E, Engqvist H, Kovács A, Helou K, Karlsson P. Radiation-induced genomic instability in breast carcinomas of the Swedish haemangioma cohort.
Genes, Chromosomes and Cancer (2019)
DOI: 10.1002/gcc.22757

All published articles were reprinted with permission from the publishers.

The following papers are not included in the thesis but are of relevance to the field.

1. Parris TZ, Werner Rönnerman E, Engqvist H, **Biermann J**, Truvé K, Nemes S, Forssell-Aronsson E, Solinas G, Kovács A, Karlsson P, Helou K. Genome-wide multi-omics profiling of the 8p11-p12 amplicon in breast carcinoma.
   Oncotarget (2018)
   DOI: 10.18632/oncotarget.25329

2. Engqvist H, Parris TZ, Werner Rönnerman E, Söderberg EMV, **Biermann J**, Mateoiu C, Sundfeldt K, Kovács A, Karlsson P, Helou K. Transcriptomic and genomic profiling of early-stage ovarian carcinomas associated with histotype and overall survival.
   Oncotarget (2018)
   DOI: 10.18632/oncotarget.26225

3. Parris TZ, Larsson P, **Biermann J**, Engqvist H, Werner Rönnerman E, Kovács A, Karlsson P, Helou K. Optimization of the resazurin-based cell viability assay to improve reproducibility of cancer drug sensitivity screens.
   Manuscript

4. Engqvist H, Parris TZ, Kovács A, Nemes S, Werner Rönnerman E, De Lara S, **Biermann J**, Sundfeldt K, Karlsson P, Helou K. Immunohistochemical validation of COL3A1, GPR158 and PITHD1 as prognostic biomarkers in early-stage ovarian carcinomas.
   Submitted

5. **Biermann J**, Nemes S, Parris TZ, Engqvist H, Werner Rönnerman E, Kovács A, Karlsson P, Helou K. A 17-marker panel for global genomic instability in breast cancer.
   Submitted

CONTENT

ABBREVIATIONS

aCGH	Array comparative genomic hybridization
AIC	Akaike information criterion
ANOVA	Analysis of variance
AUC(t)	Time-dependent area under the ROC curve function
BAF	B allele frequency
BM	Bilateral-metachronous
BMA	Bayesian Model Averaging
BS	Bilateral-synchronous
C-index	Concordance index
CAAI	Complex arm-wise aberration index
CI	Confidence interval
CNA	Copy number alteration
CNV	Copy number variation
CTLP	Chromothripsis-like pattern
DSB	DNA double-strand break
DFS	Disease-free survival
DSS	Disease-specific survival
EAR	Excess absolute risk
ER	Oestrogen receptor
ERR	Excess relative risk
FFPE	Formalin-fixed paraffin-embedded
FGA	Fraction of the genome altered
G2I	Genomic instability index (developed by Bonnet *et al.*)
GII	Genomic instability index
Gy	Gray
η (eta)	Linear predictor
HER2	Human epidermal growth factor receptor 2
HR	Hazard ratio
IHC	Immunohistochemistry
IM	Ipsilateral-metachronous
IPA	Ingenuity Pathway Analysis
IS	Ipsilateral-synchronous
LR2	Likelihood ratio with individual comparisons
LRR	Log R ratio
MAPD	Median of the Absolute Values of all Pairwise Differences
MDS	Multidimensional scaling
ND	Not determined
ndSNPQC	SNP Quality Control of Normal Diploid Markers

NOS	Not otherwise specified
NST	No special type
OS	Overall survival
ρ (rho)	Spearman’s rho
PR	Progesterone receptor
qRT-PCR	Quantitative real-time reverse-transcriptase polymerase chain reaction
RFS	Recurrence-free survival
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
ROS	Reactive oxygen species
SI	Similarity Index
SI$_{\mathbf{met}}$	Modified SI for methylation data
SNP	Single nucleotide polymorphism
TCGA	The Cancer Genome Atlas
TMA	Tissue microarray
TNM	Tumour-node-metastasis
TSCE	Two-stage clonal expansion model
WES	Whole exome sequencing
WGS	Whole genome sequencing

1 INTRODUCTION

1.1 Cancer

Cancer defines a heterogeneous group of diseases caused by uncontrolled cell growth with the potential to spread to other parts of the body. The transformation of normal cells into tumour cells is termed carcinogenesis and typically progresses from a pre-cancerous lesion to a malignant tumour. The risk of developing cancer is increased by specific genetic factors and external agents, including physical carcinogens (e.g. ultraviolet and ionizing radiation), chemical carcinogens (e.g. tobacco smoke), and biological carcinogens, such as infections from certain viruses or bacteria [1]. The World Health Organization estimated about 18 million new cases of cancer globally in 2018 with more than 9 million cancer-related deaths, making cancer the second leading cause of death worldwide [2].

1.1.1 Cancer as a genetic disease

Cancer is a disease of the genome where each patient's tumour encompasses a unique combination of genetic and epigenetic changes, such as DNA mutations, DNA copy number alterations (CNAs) and epigenetic modifications of DNA and histone proteins. Alterations that confer selective growth advantages to cancer cells are driver mutations, which induce and promote carcinogenesis by activating proto-oncogenes, inactivating tumour suppressor genes, or altering DNA repair genes. A typical tumour contains two to eight driver mutations [3]. The remaining mutations are passenger mutations that do not provide a growth advantage, but were generated in an ancestor cancer cell during the acquisition of driver mutations [3, 4]. Consequently, tumour genomes are characterized by a high frequency of genetic alterations, where most alterations do not cause cancer but are rather a result of uncontrolled cell division [5].

1.1.2 Genomic heterogeneity

The majority of cancers accumulate sequential somatic alterations that are developed over the course of 20-30 years [3]. Driver alterations primarily affect signalling pathways that regulate cell fate determination, cell survival, and genome maintenance [3]. Specific driver and passenger mutations differ between individual tumours, but usually involve the same pathways [3]. In contrast to acquired mutations, inherited mutations play a major role in about 5-10% of all cancers and predispose individuals to develop specific types of cancer [6].

Genomic heterogeneity can be observed in tumours from different patients and multiple tumours from the same patient (intertumour heterogeneity). Even within one tumour, different cell populations may exist that harbour unique genetic alterations (intratumour heterogeneity) [7]. Intratumour heterogeneity can be identified in most cancers and is considered a major problem affecting the accuracy of tumour diagnosis, as single biopsies will not reflect the pathology of the tumour adequately [7]. Thus, intratumour heterogeneity can facilitate the expansion of drug-resistant populations and potentially affect treatment response of metastases (**Figure 1**) [3, 7].

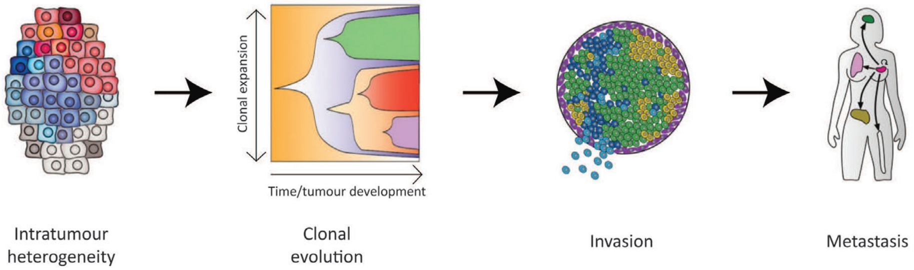

Figure 1. Intratumour heterogeneity can lead to the expansion of certain subpopulations of a tumour. Some tumour cells acquire the ability to infiltrate into the surrounding tissues and spread via blood or lymph circulation far beyond the original tumour to form distant metastases. Adapted from Navin, 2015.

1.2 Breast cancer

1.2.1 The female breast

The female breast consists of glandular, adipose and connective tissue distributed in varying amounts and proportions (**Figure 2**). When fully developed, the glandular tissue includes 15-20 lobes composed of lobules, which contain clusters of alveoli [8]. The lobes, lobules, and alveoli are linked by a network of ducts converging on the nipple [8]. Ducts and lobules are composed of luminal epithelial and myoepithelial cell layers [9]. During lactation, the inner luminal epithelial cells of the terminal ducts and the lobules produce milk [9]. The outer myoepithelial cells assist in milk ejection and play a role in maintaining the normal structure and function of the lobule and basement membrane [9]. Other components of the breast are lymph vessels, which carry the lymph fluid between lymph nodes forming a network throughout the body to filter lymph and store white blood cells [10, 11]. Clusters of lymph nodes are located near the breast in the axilla, above the collarbone, and in the chest [11]. Additionally, blood vessels and nerves can be found in the breast.

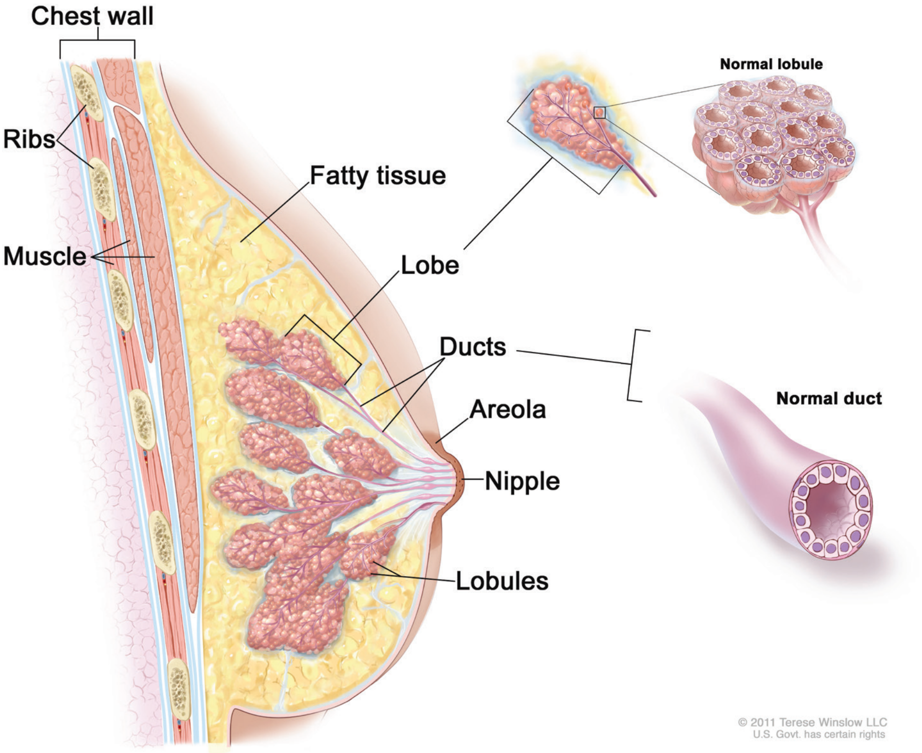

Figure 2. Anatomy of the female breast with cross sections of lobes and ducts. Adapted from https://www.teresewinslow.com.

1.2.2 Epidemiology and risk factors

Breast cancer is the most common type of cancer in women (24.2%) with an estimated number of more than 2 million new cases worldwide in 2018 [2]. According to the World Health Organization, breast cancer is the leading global cause of cancer-related death among women (15%) with approximately 627,000 breast cancer-related deaths in 2018 [2]. Since the late 1980s, mortality rates have declined in most developed countries due to improved detection, earlier diagnosis, and more effective treatments [12]. Risk factors for developing breast cancer include female gender, increased age, obesity, alcohol consumption, increased breast tissue density, prior hormone replacement therapy, exposure to ionizing radiation, prior incidence of breast cancer, changes in breast cancer susceptibility genes, increased amounts of endogenous oestrogen through menstrual history (early menarche/late menopause), nulliparity, and older age at first child birth [11]. Familial predisposition accounts for about 5-10% of breast cancer cases in which the high-risk genes *BRCA1* and *BRCA2* play a major role [12]. However, the risk of developing breast cancer depends on a combination of factors, including family history as well as reproductive and lifestyle factors [12].

1.2.3 Breast pathology

Breast cancer is a clinically, genetically and histologically heterogeneous disease. There are more than 20 histological subtypes of breast cancer. Most breast cancers arise from glandular epithelial cells (termed carcinomas) and are subdivided into *in situ* and invasive lesions. In the case of *in situ* carcinomas, intraductal malignant epithelial cells are restricted to the ducts surrounded by an intact myoepithelial cell layer and thus not invading the surrounding tissues. Approximately 70% of invasive carcinomas are categorized as “no special type” (previously known as “invasive ductal carcinoma”) comprising a heterogeneous group of tumours that show no specific morphological features [12]. The most common of the special subtypes include lobular carcinoma, tubular carcinoma, mucinous carcinoma, carcinoma with medullary and apocrine features, etc. [12].

Several histopathological tools are routinely used in the clinic to guide treatment decisions. The histological assessment of tumour grade provides powerful prognostic information by analysing how closely a tumour resembles its tissue of origin based on tubular formation, nuclear differentiation, and cell proliferation. High-grade tumours tend to be more aggressive and display an unfavourable prognosis. Another powerful and well-established pathological tool is the tumour-node-metastasis (TNM) staging system, which takes into account the size of the tumour, spread to the lymph nodes, and the occurrence of distant metastases [13]. Nonetheless, patients with a similar type, grade, or TNM stage of breast cancer still respond differently to therapy and differ in clinical outcome.

1.2.4 Biomarkers for breast cancer

The expression of the oestrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and Ki-67 is routinely evaluated to select the most appropriate treatment for breast cancer patients. The hormone oestrogen (17β-estradiol) affects proliferation, differentiation, and function of the mammary gland by binding to its receptors, ERα and ERβ, among others. The activated ER translocates into the nucleus and functions as a DNA-binding transcription factor to regulate gene transcription. Currently, only ERα is clinically measured for treatment decisions [14]. The ER induces PR expression, which is activated by the steroid hormone progesterone. Overexpression of ER and PR predicts the likelihood to benefit from endocrine therapy using adjuvant tamoxifen [15]. The *ERBB2* proto-oncogene encodes the receptor tyrosine kinase erbB-2, also known as HER2, and is amplified and/or overexpressed in approximately 15-20% of breast cancers [16, 17]. HER2 overexpression prognosticates increased tumour aggressiveness and a higher incidence of recurrence [15]. Furthermore, HER2 overexpression is a predictive factor for response to targeted therapy using the monoclonal antibody trastuzumab [15]. Ki-67 is a cellular marker for proliferation that identifies ER-positive breast cancer patients who would benefit from adjuvant chemotherapy [18]. Any new biomarker needs to contribute clinically useful information beyond that already provided by the current clinical and histopathological markers [19].

1.2.5 Molecular subtypes

Microarray-based gene expression profiling has shown that breast cancer encompasses a collection of different diseases with unique patterns of gene expression. Perou *et al*. [20] and Sorlie *et al*. [21] identified hierarchical clusters based on gene expression that revealed the existence of intrinsic breast cancer subtypes. The expression patterns of the intrinsic subtypes overlap with the routinely evaluated biomarkers and highlight that ER-positive and ER-negative breast cancers represent molecularly distinct diseases (**Table 1**). The luminal subtypes (luminal A, luminal B/HER2-negative, luminal B/HER2-amplified) encompass the hormone receptor-positive tumours (i.e. ER- and PR-positive) and can be treated with endocrine therapy resulting in good or intermediate prognoses. HER2-amplified subtypes (luminal B/HER2-amplified, HER2-positive) offer a target for treatment with trastuzumab but have more unfavourable prognoses. The triple-negative subgroup has the worst prognosis as neither the hormone receptors nor HER2 can be targeted for treatment.

Table 1. Intrinsic subtypes and associated biomarkers, treatment options and prognoses. Adapted from [22-24].

	Luminal A	**Luminal B HER2-negative**	**Luminal B HER2-amplified**	**HER2-positive**	**Triple-negative**
ER	+	+	+	-	-
PR	±	±	±	-	-
HER2	-	-	+	+	-
Ki-67	<20%	>20%	any	any	any
Treatment	Endocrine	Endocrine	Endocrine, trastuzumab	Trastuzumab	Chemo-therapy
Prognosis	Good	Intermediate	Intermediate	Poor	Poor

1.2.6 Gene signatures

A gene signature comprises a set of genes representing distinct gene expression patterns, which are associated with clinical outcome (prognostic), response to a particular therapy (predictive), or distinguish phenotypically similar conditions (diagnostic). The 70-gene expression signature (MammaPrint) was identified from gene expression profiles of 117 breast tumours and predicted the occurrence of metastasis in lymph node-negative breast cancer patients more accurately than the established clinicopathological markers [25, 26]. The MINDACT trial demonstrated the clinical utility of the 70-gene signature to categorize high- and low-risk patients [27]. The 21-gene recurrence score (Oncotype Dx) is a clinically validated assay for ER+ and node-negative breast cancer patients to assess the risk of metastasis and predict the response to adjuvant chemotherapy [28, 29].

A plethora of different gene expression signatures have been proposed for breast cancer, several of which mainly identify high-risk patients based on high expression of proliferation-related genes [30, 31]. Indeed, the identification of patients at risk for disease recurrence can guide treatment decisions to avoid adjuvant chemotherapy or, alternatively, select more aggressive therapy options. About 60% of early-stage breast cancer patients receive adjuvant chemotherapy, while only 2-15% of this patient group benefit from chemotherapy [32]. Consequently, treatment tailoring offers an opportunity to minimize the risk of toxic side effects by avoiding over-treatment.

Despite many decades of developing prognostic gene signatures, a major drawback of the field is the lack of consensus gene expression models for prognosis [33]. Prognostic gene signatures aim to categorize a common set of biological features but show surprisingly little overlap. However, interpretation of gene expression signatures is difficult due to the complexity of the underlying biological processes, as up to 30% of genes in any given signature have unknown functions [33]. This complicates the identification of affected pathways, which might consolidate single genes from different signatures. Hence, linking gene signatures to underlying molecular mechanisms of cancer is crucial to enable translation into the clinic.

1.3 Survival analysis

Survival analysis is a branch of statistics analysing the expected time to an event of interest, e.g. the death of a patient by any cause (OS; overall survival) or the time from initial diagnosis to disease-specific death (DSS; disease-specific survival). Since the event of interest is dichotomized, some parts of the data might be censored (i.e. patients that have not experienced the event by the time the study ends, patients lost to follow-up during the study period, or patients that withdrew from the study) [34]. Survival data can be described using the survival function $S(t)$, which is defined as the probability of an individual surviving from the time of origin to a specified time t [34]. These survival probabilities for different values of t describe survival of the cohort [35]. The hazard function $h(t)$ is interconnected with the survival function and gives the instantaneous potential of having an event at the time t, given that the patient has survived up to time t [35]. The hazard function focusses on the event occurring (current event rate) while the survivor function in contrast focusses on the event not happening (cumulative non-occurrence) [34].

The hazard ratio (HR) measures the relative survival experience in two groups over time, where HR = 1 means no difference in survival between the groups, while HR >1 indicates increased mortality and HR <1 decreased mortality of the group [34, 35]. HRs are usually estimated using regression modelling techniques, such as the Cox proportional hazards model (hereinafter referred to as Cox model) [34, 36]. Cox models estimate the effects of a set of covariates (i.e. known quantities potentially affecting prognosis) on survival, while the baseline hazard $h_0(t)$ remains unspecified (semiparametric model) [36, 37]. The proportional hazard assumption requires the HR to be constant over time, thus the hazard for one individual has to be proportional to the hazard for any other individual and independent of time [37]. Mathematically, the Cox model can be described using the following equation:

$$h(t, \mathrm{X}) = h_0(t) \exp\left(\beta_1 \mathrm{X}_1 + \ \dots + \beta_p \ \mathrm{X}_p\right)$$

where:

- $h(t, \mathrm{X})$ is the hazard at time t, considering covariates X
- $h_0(t)$ is the baseline hazard at time t
- p is the number of covariates
- β_p is the value of the p^{th} Cox coefficient
- X_p is the value of the p^{th} covariate.

In proportional hazard models, the HR is the exponentiated Cox coefficient $\exp(\beta_p)$. The linear predictor η (eta) is represented by the product of the covariate vector X and the Cox coefficient β, where η >0 indicates a poor prognosis (high-risk group) and η <0 a favourable outcome (low-risk group).

1.4 Predictive modelling

Generating a strong predictive model for patient survival is based on feature selection and model construction (**Figure 3**) [38]. Univariable Cox modelling can be used to estimate the utility of each probe (feature) of a gene expression microarray on an individual basis [38]. Sets of selected features can be used to build multivariable models that take the dependencies between the features into account [38, 39]. Iterative Bayesian model averaging (BMA) uses the weighted average of posterior distributions of multiple contending models and combines their effectiveness [38-40]. Hence, iterative BMA has the ability to account for model uncertainty and to select a small and parsimonious number of predictive features [38-40]. Iterative BMA represents a more accurate evaluation of feature importance than a P-value by implementing the posterior probability of each feature belonging in the model [40]. There are different ways to measure the quality of a model's fit. The C-index (concordance index) is a scalar that represents the predictive discrimination of a fitted survival model and ranges from 0.5 (random prediction) to 1 (perfect discrimination) [41, 42].

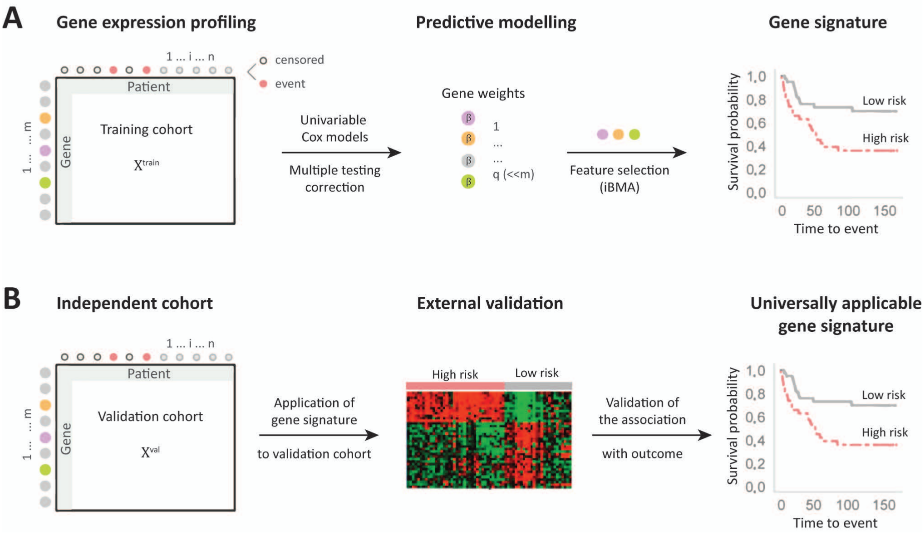

Figure 3. Identification of novel prognostic gene signatures. **A**, Selection of covariates (genes) based on Cox models and iBMA resulted in a gene signature stratifying patients into different risk groups. **B**, Validation of the gene signature in an independent validation cohort ensures universal applicability. Adapted from Zhao *et al*. (2012), and Reis-Filho and Pusztai (2011).

The time-dependent area under the receiver operating characteristic (ROC) curve function (AUC(t)) depicts the model's ability to distinguish between patients who experience the event from those who remain event-free [41]. The advantage of AUC(t) functions lies in the sequential description of accuracy over time as opposed to the C-index, which gives a global overview [41]. Predictive models should be validated to ensure that the model works for new sets of patients that were not included in the training cohort used to develop the model [43]. Evaluating model performance in external cohorts can identify overfitted models as well as other deficiencies in model development, such as small sample size or incorrect handling of missing values [44]. External validation is the first step towards the establishment of a model in clinical practice [45].

1.5 Tumour clonality

Cancer can be viewed from an evolutionary perspective as a genetically and epigenetically heterogeneous population of individual cells reflecting both the development of cancer and the challenges in curing it [46, 47]. Clonal tumour cell populations are defined as a set of cells that share similar genomic alterations arising from a common ancestor [47]. Tumour cells can gain the ability to invade other tissues and organs to generate new tumours (metastatic recurrence). Currently, there is no consensus on determining whether multiple tumours in the same patient are different entities that developed independently or a recurrence of the primary lesion (clonal relatedness; **Figure 4**). In clinical practice, the assessment of clonal relatedness is presently based on the concordance of histological tumour characteristics, such as histological subtype or hormone receptor status. Clonal evolution of tumours (also termed clonal relatedness or tumour clonality) describes the generation of genetically diverse cell populations through genomic instability resulting in distinct molecular features.

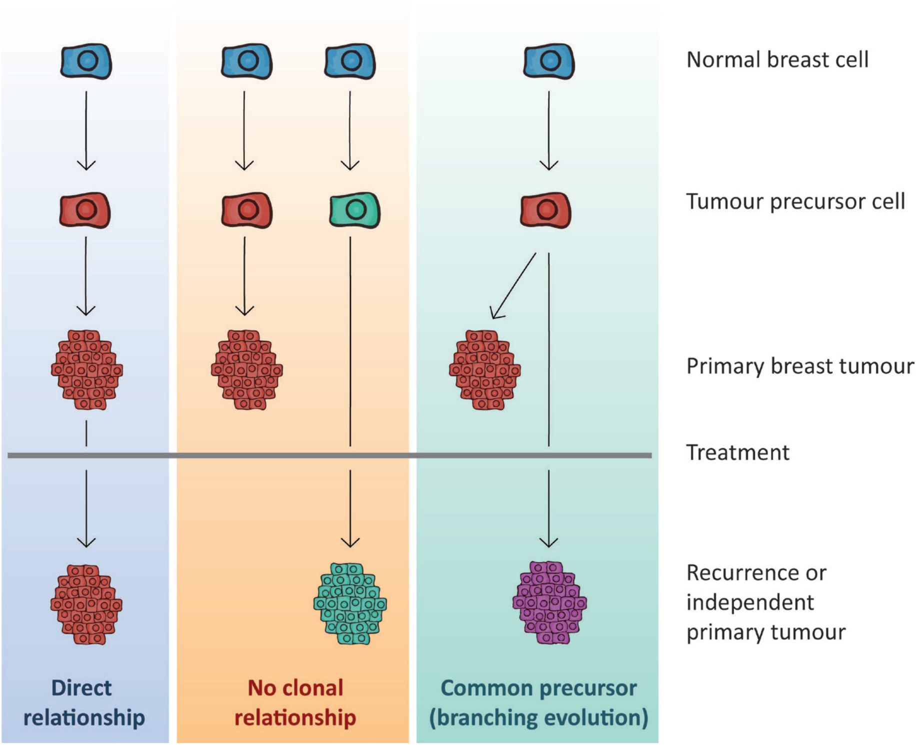

Figure 4. Clonal relatedness of multiple tumours in the same patient. In clonal tumours, both tumours are directly related and aggressive treatment is required as the recurrent tumour probably contains resistant cells. If two tumours do not share clinicopathological and molecular features, the tumours emerged independently (no clonal relationship). Indirect clonal relatedness means that two tumours share some features due to their common precursor but also show differences (branching evolution). Adapted from www.unibas.ch/en/Research/Uni-Nova/Uni-Nova-128/Uni-Nova-128-New-treatment-concepts-for-recurrent-lymphoma.html.

These features form in response to selective pressures of the tumour microenvironment and neutral changes over time (subclonal drift) and eventually lead to genetically distinct subpopulations [46, 48, 49]. Consequently, two tumours that are derived from the same tumour precursor cell will share certain features, i.e. CNAs, genetic variants, DNA methylation and gene expression patterns, in addition to nonmatching features that were acquired over time [46, 50]. Determining the degree of similarity between molecular features shared by both the primary tumour and the recurrence permits classification of tumours as independent or clonally related [51].

Genetic similarities in certain tumour features might nevertheless be due to genetic predisposition and shared environmental factors instead of indicating metastatic spread or recurrence. Furthermore, some specific chromosomal aberrations are characteristic for certain cancer types and represent non-random recurrent chromosomal aberrations [52]. Therefore, tumours that developed independently might also share common chromosomal aberrations. To assess tumour clonality, tumour-specific genetic aberrations need to be separated from the background of recurrent aberrations frequently identified in the specific cancer type [53]. Hence, clonal tumours are expected to share a higher degree of tumour-specific aberrations than can be explained through cancer-specific recurrent aberrations or randomness [53]. The discrimination between clonal and independent tumours is highly important, as an independent primary tumour has a more favourable prognosis than a recurrence [54, 55]. Thus, classification of a tumour as clonal or independent can affect the suitability of local or systemic therapy [54, 55].

1.6 Genomic instability in cancer

The evolution of a normal cell into a cancer cell requires multiple mutations, which are rare events given the low mutation rate in normal cells [56, 57]. Models for tumour evolution suggest that mutations are acquired gradually over time, eventually leading to more malignant stages of cancer [58]. However, the number of mutations commonly detected in tumours would be too high to occur within a human life span [56, 57]. One theory to describe this discrepancy is the mutator phenotype hypothesis, which proposes an elevated genome-wide acquisition of genomic aberrations as an early step in carcinogenesis [57, 59, 60].

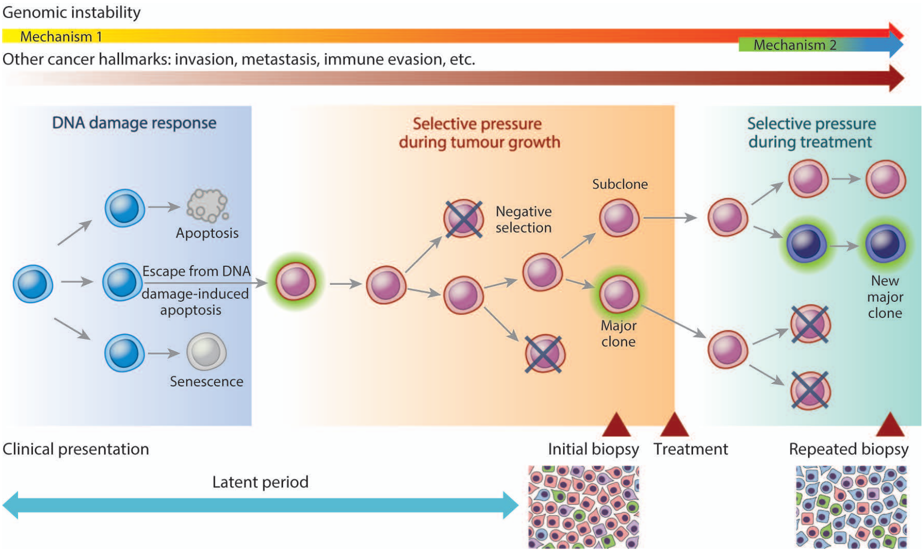

Figure 5. Mechanisms of genomic instability and consequences on tumour evolution. Cancer progression is an evolutionary process driven by somatic alterations. DNA damage results in the activation of DNA damage response pathways in the early stages of carcinogenesis, where cells with significant DNA damage undergo apoptosis or senescence. Cells that escape the DNA damage response by acquisition of genetic alterations can avoid apoptosis and accumulate further genetic alterations. Genomic instability-driven branched evolution will select for clones with higher proliferation rates, invasiveness or metastatic potential. Anticancer treatments will redirect the selective pressure, resulting in clonal repopulation. Adapted from Lee *et al.*, 2016.

These genomic aberrations can range from increased changes in DNA nucleotide sequence to large structural changes of chromosome fragments and whole genome duplication. Elevated numbers of repeats in repetitive DNA sequences are referred to as microsatellite instability [61], while chromosomal instability describes an elevated rate in the number of chromosomal aberrations (gain or loss of whole chromosomes, aneuploidy) and/or structural abnormalities (e.g. translocations, deletions, inversions, or duplications) [62-64]. Chromosomal instability is proposed to begin after at least 15-20% of the molecular time (from the last state of normal mammary development until diagnosis) has elapsed and is thus not considered the earliest source of mutations in breast cancer evolution but an on-going process in later stages [65].

Genomic instability is a hallmark of cancer and most solid cancers show evidence of genome instability with a varying degree of instability within and between cancer types [62, 63]. Proposed mechanisms for genomic instability include defects in DNA repair pathways, replication stress induced by the activation of oncogenes, inactivation of tumour suppressor genes, reactive oxygen species, chromothripsis, and breakage-fusion-bridge cycles induced by telomere dysfunction or defective mitosis [63, 66]. Genomic instability promotes inter- and intratumour heterogeneity and thereby enables the adaptation of cancer cells to environmental stress potentially leading to a more aggressive clinical behaviour and resistance to cancer therapies (**Figure 5**) [63, 67]. Genomic instability can be acquired during carcinogenesis or through germline mutations in genes responsible for genome integrity leading to predisposition to cancer [68].

1.7 Chromothripsis

Carcinogenesis has generally been perceived as a multistep process, in which cells accumulate somatic mutations over a long period of time [4]. Chromothripsis (from “chromo” (for “chromosome”) and “thripsis” (Greek for “shattering into pieces”)) presents a new paradigm of oncogenic transformation via a single catastrophic event defined by the shattering of one or more chromosomes [69, 70]. Subsequently, the DNA fragments are randomly reassembled forming a derivative chromosome [69, 70]. Typical chromothripsis patterns show at least 10-50 shifts in CNAs on a single chromosome oscillating between two or three copy number states [70]. Despite low prevalence in cancer (2-3%), chromothripsis occurs in almost all cancer types and is associated with poor patient survival [69, 70].

The exact mechanisms leading to chromothripsis remain unclear but most hypotheses assume that chromothripsis acts on condensed chromosomes explaining the highly localized shattering [69, 70]. Potential mechanisms for the induction of chromothripsis include: ionizing radiation damaging condensed chromosomes [69]; telomere attrition leading to chromosome end-to-end fusions followed by massive DNA breakage [69]; incomplete apoptosis followed by cell survival [71]; premature chromosome compaction, i.e. condensation of chromosomes before completing DNA replication [72]; DNA shattering within a micronucleus followed by reassembly into a single chromatid [70, 73]. Contrasting theories propose that chromothripsis events are not necessarily restricted to localized shattering; instead, events involving greater damage might be lethal and therefore not detected [74]. Accordingly, local shattering might manifest the upper limit of what a cell can tolerate without facing lethal consequences [74].

1.8 Radiation as a carcinogen

Radiation is the emission of energy in the form of atomic particles or waves. Atomic particles include α- and β-radiation emitted from radioactive atoms, whereas waves include electromagnetic radiation, such as radio waves, microwaves, visible light, ultraviolet light, X-rays, and γ-radiation. Radiation with sufficiently high energy can lead to radiation-matter interactions, such as breaking chemical bonds, excitation of electrons, as well as ionization of atoms and molecules by removing electrons. Ionizing radiation includes both, atomic particles and electromagnetic waves, where γ-rays, X-rays, and the high-energy spectrum of ultraviolet light are considered ionizing. Approximately 300 million particles with ionizing properties pass through each person at sea level per hour [75]. Half of this normal background radiation comes from cosmic radiation; the other half from the decay of radioactive elements within the earth [75]. The amount of energy absorbed by living tissue is measured in gray (Gy), which is defined as one joule of energy absorbed per kilogram of mass.

In tissues, ionizing radiation can cause DNA double-strand breaks (DSBs) and DNA hypomethylation among other types of cellular damage [76, 77]. DSBs can be induced directly by ionization of the sugar-phosphate backbone, or indirectly through the production of free radicals or reactive oxygen species (ROS) that can damage DNA [78, 79]. These genotoxic and carcinogenic properties of ionizing radiation can lead to chromosomal rearrangements and potentially cause severe long-term effects, such as genomic instability [77, 80]. An unstable phenotype can persist through deficiencies in DNA repair and other disturbances in cellular homeostasis including oxidative stress [81-84]. These effects are not limited to directly irradiated cells since genetic alterations, e.g. mutations, micronuclei formation, and chromosomal rearrangements, have been detected in non-irradiated cells surrounding irradiated cells (known as non-targeted or bystander effects) [81, 84, 85].

Radiation-induced genomic instability is described as an increased rate of genomic alterations initiated by ionizing radiation that persists after the exposure and ultimately promotes carcinogenesis [78, 84-87]. In addition, low doses (up to 0.1 Gy) have been shown to cause radiation-induced genomic instability [84]. The effects of ionizing radiation on humans vary by age and sex, with children and women being more radiosensitive [88, 89]. Especially, the mammary gland is known to be highly sensitive to radiation [90-93], particularly at a young age [88].

1.9 The Swedish haemangioma cohort

Infantile haemangiomas are benign vascular tumours that develop during the first weeks of life and resolve without intervention in most cases [94]. The Swedish haemangioma cohort consists of 17,200 women that were treated with ionizing radiation for haemangiomas during 1920-1965 [95, 96]. The infants were treated with encapsulated radium-226 needles that filtered α- and β-particles and allowed only γ-rays and subsequent radiation-matter interactions to penetrate the tissue. The mean and median absorbed dose to the breast were 0.18 and 0.04 Gy, respectively [95, 96]. In total, 877 cases of breast cancer (5% of the cohort) were reported, representing an excess relative risk (ERR) of 0.48 Gy^{-1} of developing breast cancer at 50 years of age, along with an excess absolute risk (EAR) of 10.4 $(10^4\ \mathrm{PYR\ Gy})^{-1}$ [95]. The ERR is the increase in risk over the background risk, while the EAR is the excess risk expressed as the difference between total risk and background risk.

Previous studies on the Swedish haemangioma cohort showed that a two-stage clonal expansion (TSCE) model incorporating radiation-induced genomic instability at an early stage of carcinogenesis significantly improved the description of the radiation risk [95, 97, 98]. However, this model was built on epidemiological data and lacked the biological evidence of dose-dependent genomic instability.

2 AIMS

The main objectives of this doctoral thesis are:

Paper I: To identify and validate a gene signature predicting breast cancer-specific survival.

Paper II: To classify multiple invasive breast tumours from the same patient as clonally related or independent based on different statistical methods and molecular data.

Paper III: To screen breast carcinomas from patients exposed to ionizing radiation in early childhood for genomic instability and investigate a potential association with absorbed dose.

3 PATIENTS AND METHODS

3.1 Patients and tumour specimens

3.1.1 Paper I and II

The breast cancer patients included in **Papers I** and **II** were diagnosed in Western Sweden between 1988 and 1999. After surgery, fresh-frozen tumour samples were stored in the tumour bank at the Sahlgrenska University Hospital Oncology Lab (Gothenburg, Sweden). Clinicopathological information were obtained from Regional Cancer Centre West (Gothenburg, Sweden), Sympathy and Melior databases (Sahlgrenska University Hospital). A subset of the tumours were stratified into different molecular breast cancer subtypes (basal-like, luminal A, luminal B, and HER2/ER-) as described elsewhere [99, 100]. Luminal B was further stratified according to the HER2 status determined by array comparative genomic hybridization (aCGH; $\log_2$ ratio $\geq$+0.5 for HER2+; and $\log_2$ ratio <+0.5 for HER2-) [23]. Routine haematoxylin and eosin stained slides from formalin-fixed paraffin-embedded (FFPE) samples were evaluated by a board certified breast pathologist. Representative imprints from each tumour specimen were stained with May-Grünwald Giemsa (Chemicon) and evaluated for neoplastic cells. Tumour specimens with at least 70% neoplastic cell content were included in downstream analyses.

In **Paper I**, 136 primary invasive breast carcinomas were selected from previously analysed patient cohorts, mainly consisting of luminal B tumours [100-102]. These 136 tumours formed the training cohort to identify the 18-marker panel. For 79 breast tumours of the training cohort complete information on established clinicopathological features could be obtained. External validation was performed using gene expression data from three publicly available datasets consisting of 1,085 breast cancer samples (**Table 2**).

In **Paper II**, 74 invasive breast carcinomas corresponding to 37 patients with two breast tumours each were selected for aCGH analysis. The patients were stratified into four groups based on the anatomic location of the breast tumours (ipsilateral or bilateral) and time interval between the diagnoses (synchronous or metachronous). Metachronous disease was defined as a time interval greater than six months between the two diagnoses. None of the patients were diagnosed with distant metastasis at the time of diagnosis of either the first or second tumours. For ipsilateral breast tumours, only samples from opposite quadrants without nipple involvement were selected.

Table 2. Overview of datasets used in **Paper I**.

Dataset	n (Subcohort)	Platform	Survival	Clinical characteristics	Ref.
Training cohort	136 (79)	HumanHT-12 Gene Expression BeadChip	DSS; OS	Age, number of positive axillary lymph nodes, histological grade, tumour size, ER, PR and HER2 status	[100-102]
GSE1456	159 (128)	Affymetrix Human Genome U133	DSS; OS; RFS	Molecular subtype, histological grade	[103]
GSE4922 (Uppsala cohort)	249 (237)	Affymetrix Human Genome U133	DFS	Age, ER status, tumour size, axillary lymph node status, and histological grade	[104]
TCGA Breast Invasive Carcinoma dataset	900 (720)	mRNA-seq	OS	Number of positive axillary lymph nodes, tumour size, age, ER and PR status	[105]

3.1.2 Paper III

The Swedish haemangioma cohort comprises 17,200 female patients that were treated for infantile haemangioma with radium-226 between 1920 and 1965 [95, 96]. A total of 877 breast cancer cases were reported by December 2009, estimating an excess relative risk at the age of 50 years of 0.48 Gy^{-1} and an excess absolute risk of 10.4 $(10^4 \text{ PYR Gy})^{-1}$ [95]. Forty-six FFPE samples for primary breast carcinomas were selected for DNA extraction representing high- and low-dose cases in the cohort. In this study, patients exposed to absorbed doses to the breast <100 mGy were defined as the “low dose” and ≥100 mGy as “high dose”. Tissue microarrays (TMAs) were used to determine the immunohistochemistry (IHC) subtype (luminal A, luminal B/HER2-negative, luminal B/HER2-amplified, HER2 positive and triple-negative) with ER, PR, HER2, and Ki-67 immunostaining [23].

3.2 Microarrays and sequencing

3.2.1 Gene expression microarray

Total RNA samples from 136 (**Paper I**) and 14 (**Paper II**) tumour specimens were processed at the Swegene Center for Integrative Biology (SCIBLU, Lund University, Sweden) using Illumina HumanHT-12 BeadChips (Illumina, CA, USA). The expression microarrays contained about 49,000 probes representing more than 25,400 RefSeq (Build 36.2, Release 22) and Unigene (Build 199) annotated genes. Illumina HumanHT-12 gene expression profiles were evaluated as described previously [100]. In brief, raw signal intensities were preprocessed and quantile normalized using the BioArray Software Environment (BASE) [106]. Further data processing was performed in Nexus Expression 2.0 (BioDiscovery) using $\log_2$-transformed, normalized expression values and a variance filter.

3.2.2 Array comparative genomic hybridization (aCGH)

In **Paper II**, whole-genome tiling arrays with 38,043 BAC reporters (UCSC May 2004 hg17: NCBI Build 35) were manufactured as previously described [107] at SCIBLU, Sweden. The clone set contained the 32K BAC clone library (BacPac Resources), the 3.4K FISH Mapped Clones Version 1.3 (BacPac Resources), clones located in telomeric regions [108], and clones covering microdeletion syndromes [109]. Male genomic DNA was used as a reference for the aCGH data. Data preprocessing and pin-based Lowess normalization were performed using BASE [106].

Segmentation into regions of gains and losses was performed using the Rank Segmentation algorithm with Nexus Copy Number Professional 7.5 software (BioDiscovery; settings: 5.0E-5 significance threshold, 1000 kb maximum contiguous probe spacing, minimum of 5 probes per segment). Minimal common regions of copy number imbalances were identified when observed in at least 25% of the tumour samples with a CNV overlap <99%. Segmented data for the segment analysis were generated using the R-package “GLAD” [110]. Log_2 ratio thresholds for low-level gain and heterozygous loss were set at ±0.3. The R-package “Clonality” [111] was used to define the LR2 (likelihood ratio with individual comparisons) and LR2 p-value and required copy number data procession with the R-package “DNAcopy” [112].

3.2.3 DNA methylation analysis

In **Paper II**, 16 samples were randomly selected to represent each clinical group (BM: bilateral-metachronous; BS: bilateral-synchronous; IM: ipsilateral-metachronous; IS: ipsilateral-synchronous) with four samples corresponding to two patients per group. Purified genomic DNA was processed at the SNP&SEQ technology platform, Uppsala, Sweden, using Illumina Infinium MethylationEPIC BeadChips (MethylationEPIC_v-1-0; mapped to UCSC Feb 2009 hg19: GRCh37). Raw data (IDAT files) were processed with the R-package “RnBeads” [113]. The probes were normalized using BMIQ (beta mixture quantile dilation) [114]. Beta values were obtained through “RnBeads”, while intensity values were extracted using the R-package “ChAMP” to generate segmented copy number data for the segment analysis [115, 116]. The R-package “conumee” was used to extract unsegmented information of CNAs on the probe level [117]. The unsegmented CNAs were used for the Similarity Index (SI), the distance measure and the clustering analysis.

3.2.4 Genome-wide SNP genotyping analysis

In **Paper II**, genome-wide SNP genotyping analysis for six tumours (three tumour pairs) was processed with Illumina Infinium HumanOmni2.5-8 v1.3 BeadChips at SCIBLU, Sweden. B-allele frequencies (BAF) and logR ratios (LRR) were calculated using the Illumina GenomeStudio Genotyping Module software (V2011.1) and hg19 build 37 reference assembly of the human genome.

3.2.5 Whole transcriptome RNA sequencing (RNA-seq)

In **Paper II**, total RNA of six tumours (three tumour pairs) were processed at SciLifeLab, Sweden. Illumina TruSeq strand-specific RNA libraries (Ribosomal depletion using RiboZero human) containing 125 bp paired-end reads were obtained for each sample on a HiSeq2000 sequencer (Illumina). Data processing was performed as described previously [118]. In brief, data was processed using FastQC (0.11.5) for quality control of raw RNA-seq reads, TrimGalore (0.3.3) to trim and filter RNA-seq reads, and STAR (2.5.1b) for alignment to the hg19 build 37 reference assembly of the human genome yielding approximately 40-50 million aligned reads per sample. HtSeq (0.6.1) [119] and Cufflinks (2.2.1) [120] were applied to calculate counts and fragments per kilobase of transcript per million mapped reads (FPKM), respectively. Quality control statistics for mapped reads were obtained using RSeQC (2.3.6).

Fusion transcripts were identified with FusionCatcher (0.99.5a) [121] using criteria to remove false positive candidate fusion events, followed by classification of "driver" fusion events (Bayesian probability scores <0.5) with oncogenic potential using Oncofuse (1.1.1) [122]. Genetic variants were identified and annotated using the Genome Analysis Toolkit (GATK 3.5.0) variant calling pipeline [123] and ANNOVAR (2016.05.11). Common genetic variants found in the human population were removed with ANNOVAR using the dbSNP (hg19_snp138) and 1000 Genomes Project databases (1000g2015aug) with a minor allele frequency (MAF) threshold of 0.01.

3.2.6 OncoScan CNV Plus Assay

In **Paper III**, 36 breast cancer samples from the Swedish haemangioma cohort were processed at the Array and Analysis Facility (Uppsala, Sweden) using the OncoScan CNV Plus Assay (Affymetrix), which identifies CNAs, loss of heterozygosity and somatic mutations. Array fluorescence intensity data (CEL files) were combined using the Chromosome Analysis Suite (ChAS; version: 3.3.0.139) to produce OSCHP files. The MAPD (Median of the Absolute Values of all Pairwise Differences) is a QC metric generated by ChAS and works as a global measure for variation in the microarray probes that is ideally below 0.3 but ranged between 0.287-0.398 for our samples [124]. The ndSNPQC (SNP Quality Control of Normal Diploid Markers) measures how well genotype alleles are resolved in the microarray data and ranged between 15.313-46.021 (ideally ndSNPQC ≥26) [124]. These shortcomings in quality were accounted for by the R-package "TAPS" (Tumour Aberration Prediction Suite; v.2.0) that generated quality metrics based on allele-specific copy number data [125]. "TAPS" detects chromosomal aberrations with higher sensitivity even for tumours with a low proportion of tumour cells by integrating allelic data [125]. Five samples were excluded based on the "TAPS" visualization of chromosomal aberrations, while the remaining 31 samples were analysed using R (v.3.5.1) [126] and Nexus Express Software for OncoScan 3.1 (BioDiscovery; Build version: 9289). In the Nexus Express analyses, combined regions of DNA gains and losses were stratified by high- and low-dose groups (p-value cut-off: 0.05; differential threshold: 25%; minimum segment size to detect copy number gains and losses, allelic imbalances, and LOH: 500 kb; minimum probes per segment: 20). Combined regions with >90% overlap with CNVs were removed.

3.3 Bioinformatics and statistical analysis

Statistical analyses were performed in the open-source programming environment R [126] using two-sided tests and a 0.05 p-value cut-off, unless stated otherwise.

3.3.1 Paper I

3.3.1.1 Multivariable predictive modelling

First, univariable Cox proportional hazards models were fitted for each probe on the gene expression microarray using the R-package “survival” (v2.40-1) [127]. A total of 9,159 transcripts in the training cohort (n = 136) were identified as significantly associated with DSS. Second, adjusting for multiple testing using Bonferroni correction reduced the number of significant transcripts to 186. Third, iterative BMA (R-packages “BMA” (v3.18.7) [128] and “iterativeBMAsurv” (v1.32.0) [38]) was used to further reduce the number of transcripts to an 18-marker panel.

3.3.1.2 Survival analysis and predictive power

Hazard ratios were obtained by fitting univariable and multivariable Cox proportional hazard models for each cohort using the R-package “survival” (v2.40-1) [127]. Patients were stratified into high- and low-risk groups according to the linear predictor η (eta) which represents the product of the covariate vector x (gene expression) and the parameter vector β (Cox coefficient). Patients with $\eta > 0$ were classified as high-risk patients and those with $\eta < 0$ as low-risk patients. If η equals 0, the patient cannot be classified in the high- or low-risk groups, because a HR of 1 means the covariate has no effect on the model. Kaplan-Meier plots were generated using the R-package “survminer” (v0.2.2) [129].

AUC(t) statistics and C-indices were calculated using the R-package “risksetROC” (v1.0.4) [130] to assess the predictive power of the 18-marker model, the established marker model and the combined model (18 markers and established markers) based on the linear predictor as a marker [41, 42]. Since complete clinical information is needed for the combined models, only patients with complete information were used for the training and validation cohorts.

3.3.1.3 Oncotype Dx analysis

The commercially available Oncotype Dx is a qRT-PCR-based (quantitative real-time reverse-transcriptase polymerase chain reaction) 21-gene signature that includes 16 cancer-related genes and 5 reference genes for normalization [28]. A multivariable model was fitted based on gene expression microarray data of the training cohort using the 16 cancer-related genes. The C-indices and AUC(t) functions were obtained as described above. Bootstrapping was performed with 1,000 iterations comparing the 18-marker panel to the Oncotype Dx–based 16-gene signature. The model was applied to the whole cohort and the ER-positive subcohort as Oncotype Dx is only clinically validated for ER-positive breast carcinomas.

3.3.1.4 Pathway analysis

The interaction between the 18 candidate genes was analysed with the Ingenuity Pathway Analysis (IPA) software (Ingenuity Systems) using Fisher's exact test (significance threshold at P <0.05). The diseases and bio functions tool was used to detect diseases and disorders that are associated with the 18 genes as well as molecular and cellular functions that overlap with the 18 genes.

3.3.2 Paper II

3.3.2.1 Similarity Index (SI)

The underlying assumption of the SI is that two tumours that share a higher degree of patient-specific CNAs than two randomly paired tumours are considered clonally related [53]. A permutation-based approach can ensure that the shared CNAs between two tumours outweigh the recurrent chromosomal aberrations commonly found in breast cancer [53]. By calculating the similarities for artificial pairings of tumours from different patients, we could approximate the reference distribution of similarities due to recurrent aberrations or randomness, which allows us to reject the null hypothesis if the SI exceeds its 95th percentile [53]. Normalized DNA copy number data were discretized into loss ($\log_2$ ratio <-0.3), normal, and gain ($\log_2$ ratio >0.3). Unique (N_U), shared (N_S), and opposite (N_O) changes were calculated for each combination of tumours to obtain the SI ranging between 0 (completely different) and 1 (identical genomic profiles):

$$SI = \frac{N_S}{N_S + N_U + N_O}$$

For SI application on gene expression microarray data, the SI remained unchanged discretising normalized $\log_2$ ratios using a 1.5 fold change cut-off (underexpressed ($\log_2$ ratio <-0.58); neutral; overexpressed ($\log_2$ ratio >0.58)).

For SI application on DNA methylation array data, the SI was modified (SI_{met}) because the SI for copy number data is based on measuring the amount of alterations from the biologically neutral state (two copies per allele). For DNA methylation however, neither methylated nor unmethylated can be defined as the neutral state of a cytosine. In the SI_{met}, beta values were discretized according to thresholds defined by Du *et al.* [131] into methylated (beta value >0.8), unmethylated (beta value <0.2), and hemimethylated (beta value range: 0.2-0.8). The SI_{met} counts the number of probes with shared methylation states between two tumours and divides it by the total number of probes, which provides the percentage of shared methylation states.

3.3.2.2 Hierarchical clustering

Unsupervised hierarchical clustering was performed using single linkage with Euclidean distance as proposed by Ostrovnaya and colleagues [132]. Clonality was defined as two tumours from the same patient clustering together in the terminal branch of the dendrogram.

3.3.2.3 Distance measure

Distance matrices of the Euclidean distances between different tumour samples were computed using the basic “stats” R-package [126]. The reference distribution of distances was approximated by calculating the distance measure for all possible combinations of tumour pairs. Tumour pairs that are more similar exhibit a shorter distance. Statistical significance for clonality was defined as the distance of a true tumour pair below the fifth percentile of the reference distribution.

3.3.2.4 Shared segment analysis

Two clonally related tumours are expected to share some altered DNA segments. A shared segment was defined as an overlap of the exact loci in both ends of the segment with the same altered direction (increase or decrease in copy number). Clonality was defined as the number of shared segments above the 95th percentile of the reference distribution.

3.3.2.5 Mutation and fusion transcript analysis

Mutational changes that were identical in both tumours were counted for all possible combinations of tumour pairs in genomic and exonic RNA-seq data as well as in a panel of 254 breast cancer mutation spots [133]. Clonality was defined as the number of shared mutations above the 95th percentile of the reference distribution. The R-package "Clonality" [111] was applied to profiles of somatic mutations with loci-specific mutation probabilities obtained from the TCGA breast cancer dataset [105]. Furthermore, fusion transcripts of all tumours were compared and transcripts with identical 5' and 3' fusion partner breakpoints were counted with clonality being defined as the number of shared fusion transcripts above the 95th percentile of the reference distribution.

3.3.2.6 Cohen's kappa

Cohen's kappa was applied to measure the chance-corrected agreement of two observations [134]. Cohen's kappa indices of agreement were calculated using the R-package "rel" [135] to detect the highest agreement between the different methods assessing clonality.

3.3.3 Paper III

3.3.3.1 Processing of DNA copy number data

Weighted $\log_2$ ratios and B allele frequencies generated by ChAS were processed by the R-package "copynumber" (v.1.22.0) to obtain allele-specific segmentation [136]. The number of segments was defined as the number of allele-specific autosomal segments generated by "copynumber". The FGA (fraction of the genome altered) was defined as the number of autosomal DNA gain or loss ($\log_2$ ratio cut-off: ±0.15) divided by the total number of autosomal probes [137]. Allele-specific segments were used for chromothripsis-like pattern (CTLP) detection using the web-based CTLPScanner (http://cgma.scu.edu.cn/CTLPScanner/; Genome assembly: GRCh37/hg19; CN status change times: ≥10; Log_{10} of likelihood ratio ≥8; Minimum segment size (Kb): 50; Signal distance between adjacent segments: 0.3; Genomic gains and losses: ±0.15) [138]. The R-package "ASCAT" (allele-specific copy number analysis of tumours; v.2.5) was used to assess tumour purity, ploidy and allele-specific copy number profiles [139].

3.3.3.2 G2I (Genomic instability index)

The G2I algorithm applies a two-parameter index based on the global level of genomic alteration (TXP: altered probes / total probes) and the number of altered genomic regions (NB: local score statistics for altered genomic regions) [140]. G2I uses differential survival data to define the cut-off that separates the samples into stable (G2I-1 and G2I-2) and unstable (G2I-3) tumour genomes [140]. The G2I algorithm was performed on copy number data (discretized with ±0.15 $\log_2$ ratio cut-offs) using the R scripts provided by Bonnet and colleagues [140].

3.3.3.3 Complex arm-wise aberration index (CAAI)

The CAAI score is a validated prognostic marker in breast cancer that captures focal DNA alterations (i.e. narrow peaks of high copy number gains), which represent the degree of local distortion [141, 142]. Samples were classified as CAAI positive if the CAAI score exceeded the threshold 0.5 in at least one chromosome arm of the sample [141, 142]. Since the CAAI algorithm classified all samples in the haemangioma cohort as unstable, we applied three CAAI-related metrics to represent tendencies of genomic instability (average score, number of unstable arms, and maximum CAAI score).

3.3.3.4 GII (Genomic instability index)

The GII is defined as the FGA relative to the baseline ploidy of the sample [143]. A threshold of GII = 0.2 was used to distinguish genomically stable from unstable tumours, as previously defined [144, 145].

3.3.3.5 Survival analysis

The R-package “survival” (v.2.43-3) was used to fit Cox models and C-indices were calculated using the R-package “risksetROC” (v.1.0.4). The R-package “survminer” (v.0.4.3) was used to generate Kaplan-Meier plots that were based on (i) the G2I classification and (ii) the linear predictor η of the interaction model combining the effects of the G2I classification and the absorbed dose (G2I* absorbed dose; equivalent to: G2I + absorbed dose + G2I:absorbed dose). Patients with $\eta > 0$ were classified as poor-prognosis patients and $\eta < 0$ as good-prognosis patients. The survival data complied with the proportional hazards assumption for fitting Cox models.

4 RESULTS AND DISCUSSION

4.1 Paper I

4.1.1 Identification of the 18-marker panel

In the training cohort, 9,159 transcripts were significantly associated with DSS in univariable Cox models. Adjusting for multiple testing reduced the number of transcripts to 186, which was further narrowed down to 18 genes using iterative BMA. Nine of the 18 genes (*ACAA1*, *BORCS6*, *CCNA2*, *CDCA5*, *FAM91A1*, *KIAA0494*, *MTURN*, *NEIL3*, and *TRIP13*) displayed a HR <1 and were thus negatively associated with breast cancer-specific mortality (**Figure 6**). The remaining nine genes (*ADGRG6*, *CDKN2A*, *HJURP*, *HSPA14*, *LRRCC1*, *PRR11*, *SKA2*, *SNX8*, and *STAM*) had an unfavourable effect on survival, i.e. positive association with breast cancer-specific mortality. *CDKN2A* was the only gene that was not significant (*P* = 0.059) in the multivariable model but was kept in the model due to its strong contribution to the predictive power. The strength of the model is that the marker selection was unbiased and presented a good trade-off between parsimony and predictive capacity. Pathway analysis showed that the molecular and cellular functions of the 18 markers included cell cycle, cellular assembly and organization, and DNA replication, recombination and repair. Hence, the unbiased statistical approach of marker selection identified biologically meaningful markers without including previous biological knowledge.

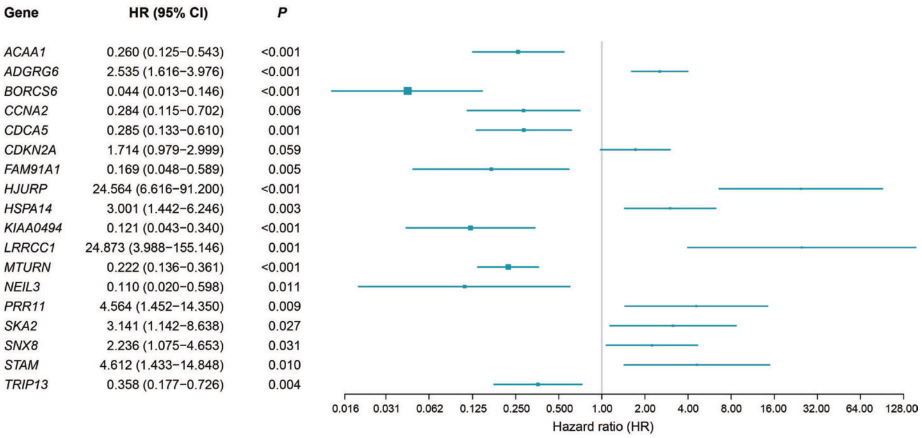

Figure 6. Forest plot depicting the hazard ratios (HRs) of the multivariable 18-marker panel in the training cohort. HRs above one indicate that a gene is positively associated with breast cancer-specific mortality. The box size is based on precision, where a bigger box size represents a more precise confidence interval (95% CI). The x-axis is plotted on logarithmic scale.

4.1.2 Survival prognosis based on the 18-marker panel

Multivariable Cox models based on the linear predictor stratified all training and validation cohorts significantly into low- and high-risk groups (**Figure 7**). In the training cohort, 15 (79-patient subcohort; HR **=** 0.069; **Figure 7A**) and 12 times (136-patient complete cohort; HR = 0.089; **Figure 7B**) as many patients died at any time in the high-risk group compared to the low-risk group. In the validation cohorts, the HR ranged from 0.417 for the TCGA dataset (**Figure 7D**) to 0.182 for RFS in the GSE1456 cohort (**Figure 7G**) representing 2-5 times as many deceased patients in the high-risk group compared to the low-risk group. All corresponding log-rank tests for the training and validation cohorts showed significant differences between the high- and low-risk groups confirming the universal applicability of the 18-marker panel.

4.1.3 High predictive power of combined model

To improve outcome prediction, multivariable models were fitted based on a) the 18 markers, b) established clinicopathological markers (patient age at diagnosis, histological grade, number of positive axillary lymph nodes, pathological tumour size, ER, PR and HER2 status), and c) a combined model of the 18 markers and established markers (**Table 3**). In the training cohort, the 18-marker model showed a C-index of 0.913 for DSS and 0.896 for OS. Combining the 18-marker panel with the established markers further improved the C-indices to 0.930 for DSS and 0.929 for OS. In the validation cohorts, the combined model for DSS showed the highest C-index (GSE1456; C-index: 0.829).

In the training cohort (n = 79), the AUC(t) function (**Figure 8A**) confirmed the accuracy of the 18-marker model over time and showed a higher predictive power than established clinicopathological markers. The combined model had the highest predictive power, which receded with time but remained above 0.8. The high C-index and robust trend of the AUC(t) function, even after 8 years of follow-up, confirmed the stability of the model. The AUC(t) functions for the GSE1456 validation cohort (**Figure 8B**) were stable over time. The 18-marker panel showed a superior performance in comparison to the established markers, while the combined model had the highest predictive power. Taken together, the 18-marker panel in connection with clinical parameters proved to be an independent predictive model for clinical outcome in breast cancer.

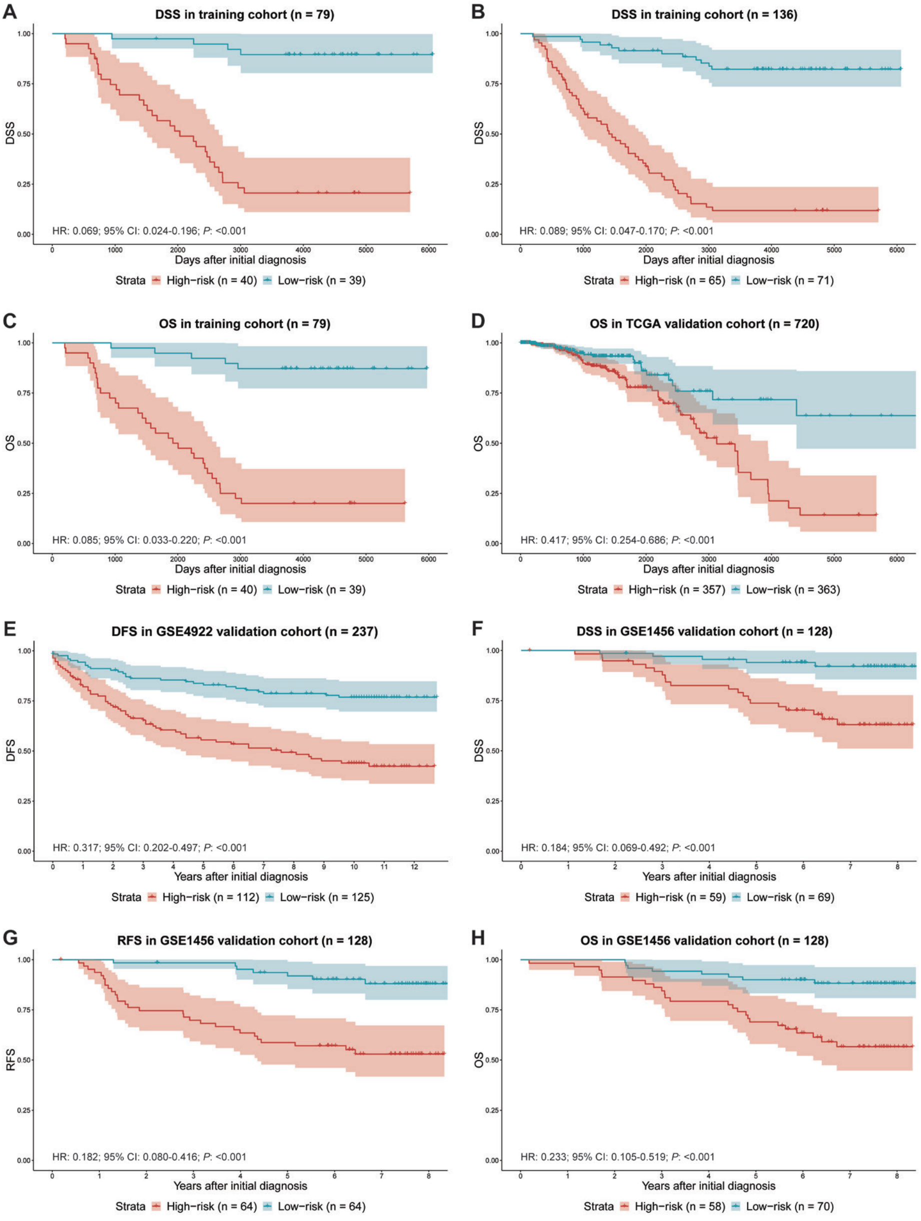

Figure 7. Kaplan-Meier analysis of the 18-marker panel in training and validation cohorts. The x-axes depict time after initial diagnosis and y-axes depict survival. **A** and **B**, Estimators of the probability of DSS in the subset training cohort (n = 79) and complete training cohort (n = 136). **C**, Estimators of the probability of OS in the subset training cohort (n = 79). **D**, Estimators of the probability of OS in the TCGA cohort (n = 720). **E**, Estimators of the probability of DFS in the GSE4922 cohort (n = 237). **F-H**, Estimators of the probability of DSS, RFS, and OS in the GSE1456 cohort (n = 128).

Table 3. C-indices of the multivariable models (18 markers, established markers, and combined model).

Cohort	18 markers	Established markers	Combined model
Training			
DSS	0.913	0.767	0.930
OS	0.896	0.778	0.929
Validation			
TCGA			
OS	0.665	0.631	0.699
GSE4922			
DFS	0.686	0.625	0.719
GSE1456			
DSS	0.803	0.707	0.829
RFS	0.754	0.678	0.769
OS	0.748	0.655	0.767

4.1.4 Comparison of the 18-marker panel to Oncotype Dx

Oncotype Dx is a clinically validated 21-gene assay for patients with ER-positive tumours based on qRT-PCR. To compare the 18-marker panel to the Oncotype Dx assay, we fitted a multivariable model based on a 16-gene subset (excluding normalization genes) using gene expression microarray data. The 18-marker panel showed a significantly higher predictive power than the Oncotype Dx-based gene signature when applied to the complete cohort (**Figure 8C**). Since Oncotype Dx is only validated for ER-positive tumours, AUC(t) functions were also generated for the ER-positive subcohort (**Figure 8D**), where the 18-marker panel showed a clearly higher predictive power. These results highlight the novelty and potential clinical benefit of the 18-marker signature, which not only exceeds the predictive power of the established clinical markers, but also the predictive power of the clinically validated Oncotype Dx signature.

4.1.5 Limitations of the study

A major limitation of the study is the lack of complete clinical information for the established markers in the training cohort, reducing the cohort to a 79-patient subcohort for the multivariable modelling.

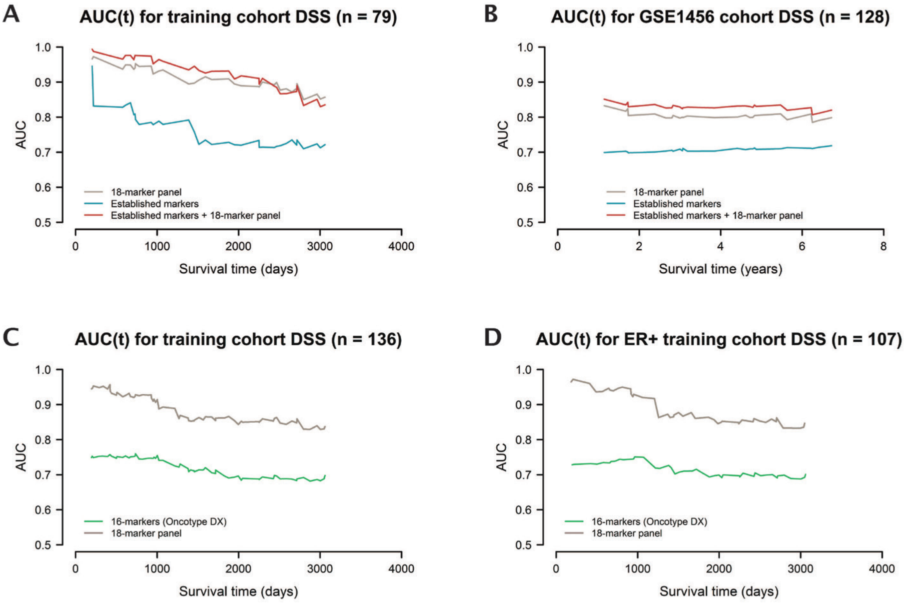

Figure 8. AUC(t) functions of multivariable models. The lines represent the time-dependent area under the ROC curve (AUC(t)) for the 18-marker panel (grey), the established markers (blue), the combined model (red), and the Oncotype Dx-based 16-marker model (green). **A**, Estimated performance of the training cohort for DSS (n = 79). Established clinical variables contain patient age at diagnosis, histologic grade, number of positive axillary lymph nodes, pathologic tumour size, ER, PR, and HER2 status. **B**, Estimated performance of GSE1456 validation cohort for DSS (n = 128). Established clinical variables contain histologic grade and subtype. **C**, Estimated performance of the 18-marker panel in comparison to the Oncotype Dx-based 16-marker panel in the complete training cohort (n = 136) for DSS. **D**, Estimated performance of the 18-marker panel in comparison to the Oncotype Dx-based 16-marker panel in the ER-positive training cohort (n = 107) cohort for DSS.

Furthermore, the training and microarray-based validation cohorts originated from Swedish Cancer Registry studies, thereby overrepresenting the Swedish population in this study. Another limitation is the comparison of probes from different microarray platforms (Illumina Human HT-12 Whole-Genome Expression BeadChip and Affymetrix Human Genome U133 Set). In most cases, the probes on the two microarray platforms did not map to the same parts of the transcript sequence, while the TCGA mRNA-seq dataset represented a different type of RNA-based experiment. Nevertheless, true biological effects should be detectable irrespective of the platform or type of experiment used to analyse gene expression.

4.2 Paper II

4.2.1 Histopathological discordances in tumour pairs

Discordant clinical factors were detected in 32% of tumour pairs (12/37). Changes in histological subtype were most prevalent (35% of tumour pairs; 6/17), while molecular subtype differed in 25% of tumour pairs (2/8), ER status in 11% (4/35 patients), and HER2 status in 8% (3/37 patients). The discordant changes were equally distributed between the different clinical groups (BM: bilateral-metachronous; BS: bilateral-synchronous; IM: ipsilateral-metachronous; IS: ipsilateral-synchronous) and showed no statistical significance when stratified by group.

4.2.2 Differential DNA copy number imbalances

Differential DNA copy number imbalances were identified in recurrent regions of DNA copy number gain (blue) and loss (red) comprising at least 25% of the tumours in the patient cohort (**Figure 9**). Very few differences in DNA copy number frequencies were found between synchronous and metachronous tumours with 59 significantly different genomic regions (**Figure 9A**).

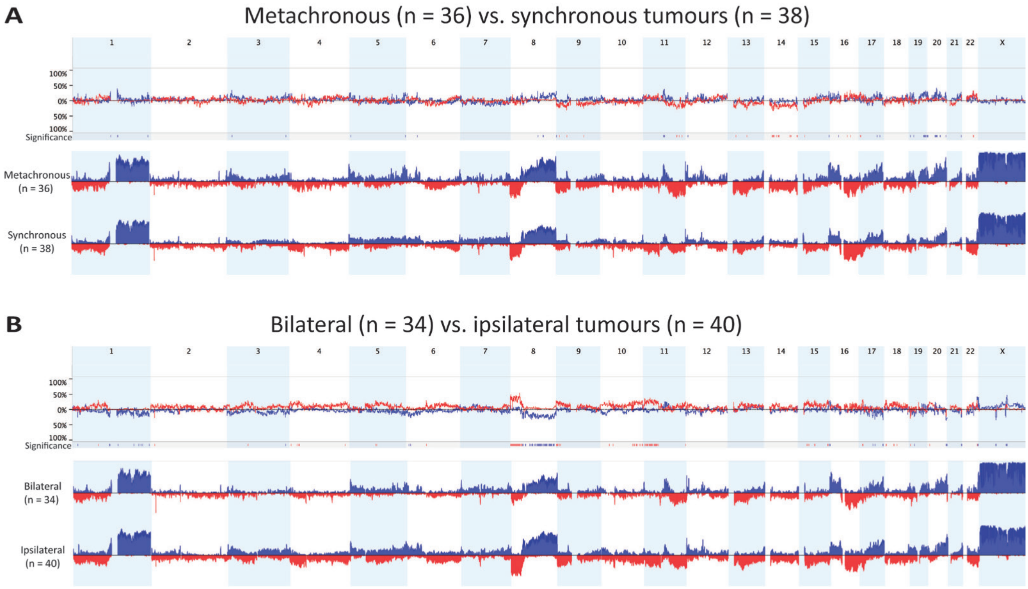

Figure 9. Genome-wide frequency plots of DNA copy number gains (blue) and losses (red) stratified by the time interval between the tumours (**A**, metachronous vs. synchronous) and the laterality (**B**, bilateral vs. ipsilateral).

Stratification by laterality resulted in 134 statistically significant regions of DNA copy number imbalances with conspicuous losses on 8p and 11p in the ipsilateral subgroup (**Figure 9B**). These results suggested that the anatomic location had a greater influence on genomic diversity in the cohort than the time interval between the tumours.

4.2.3 DNA copy number as a tool for clonal relatedness

The SI applied to the aCGH data classified 46% of tumour pairs (17/37) as clonally related, which contradicted with current classification methods for clonality using histopathology. The SI aims to assess the overlap between CNAs in two tumours (**Figure 10**). Several molecular techniques, such as CGH (comparative genomic hybridization) [146, 147], aCGH [148, 149], as well as whole exome and whole genome sequencing (WES and WGS, respectively) [150-152], have been used to assess tumour clonality along with various analytical tools [50, 51, 53, 146, 148, 153].

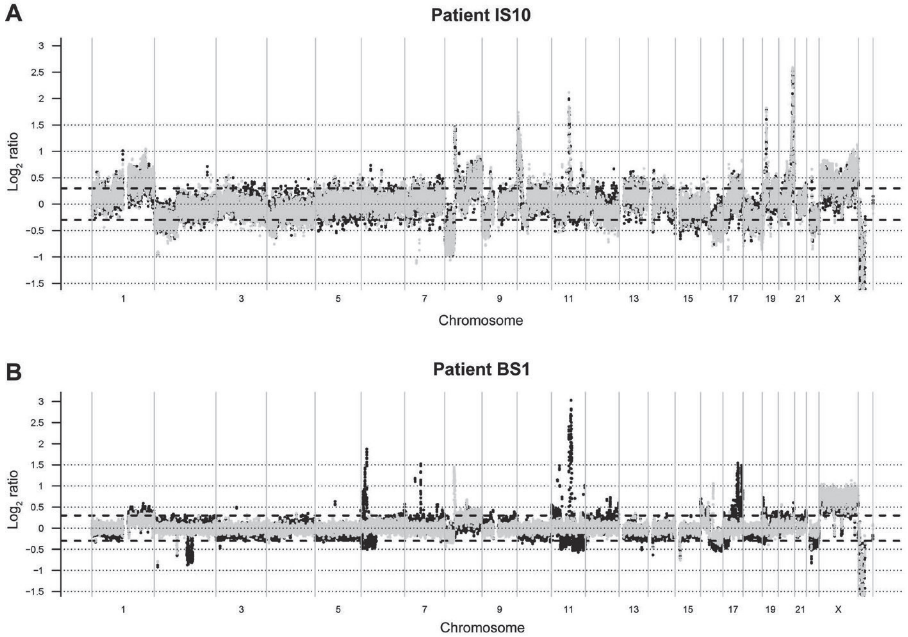

Figure 10. Overlay of DNA copy number profiles of two tumours from patient IS10 (**A**) and patient BS1 (**B**). CNAs of one tumour were plotted in black and CNAs of the other tumour in grey to show similarities and differences between the two tumours from the same patient. The SI classified the tumours of patient IS10 as clonally related based on the CNAs, while the tumours of patient BS1 were classified as independent primary tumours.

Currently, there is no consensus on which type of data and analysis method provide the most stable definition of clonality. In contralateral tumour pairs, Alkner *et al.* demonstrated clonal relatedness in 10% (1/10) of cases [152], which was lower than the clonal relatedness of bilateral tumours according to the SI in our cohort (29%, 5/17 tumour pairs). In a study on bilateral-metachronous tumour pairs, Klevebring *et al.* found 12% (3/25) of cases to be clonally related [151], which was also lower than in our study (22%, 2/9 tumour pairs). Direct comparisons of the rate of clonality between studies might vary due to differences in the study set-up, methods and statistics. However, in a cohort comprising ipsilateral-synchronous tumour pairs, Desmedt *et al.* defined 67% (24/36) of tumour pairs as clonal [154], which is comparable to the clonality rate of 64% (7/11 tumour pairs) in our study.

4.2.4 DNA methylation as a tool for clonal relatedness

In Kruskal's non-metric multidimensional scaling plot (MDS; **Figure 11**), beta values of synchronous samples showed a greater dissimilarity from each other as well as from other tumours of the cohort, while the metachronous samples formed a distinct cluster.

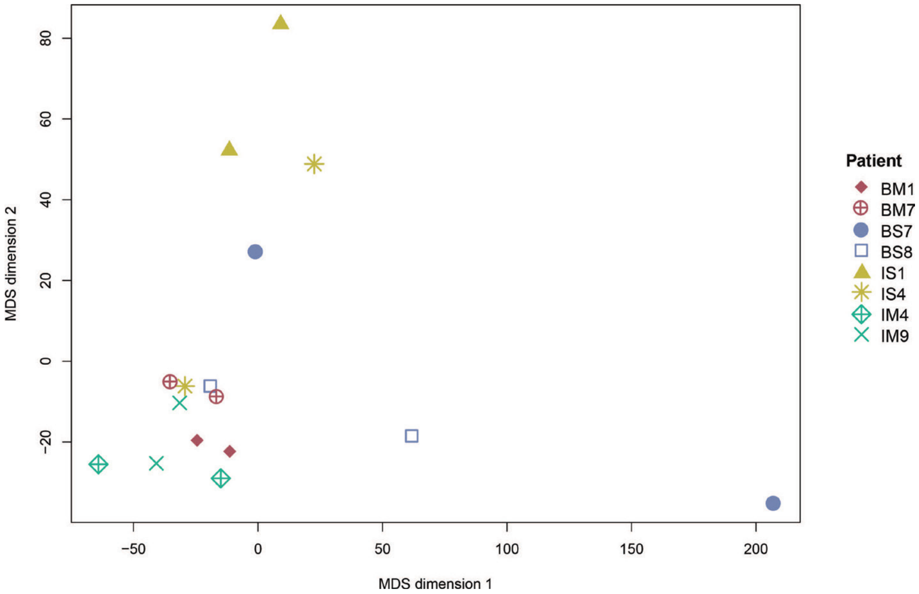

Figure 11. Kruskal's non-metric multidimensional scaling (MDS) plot of beta values from the DNA methylation cohort (n = 16). The MDS plot visualized similarities between the individual samples based on the Euclidean distance matrix.

DNA methylation array-derived intensity values offered another method to generate copy number data but presented a more liberal type of data for clonality classification than aCGH-derived copy number data. Particularly in the clustering analysis, the intensity data more frequently classified tumour pairs as similar in comparison with other types of molecular data (**Figure 12**). Concordance in clonality assessment between the copy number data generated from aCGH and DNA methylation array-generated intensity data was observed in only 50% of the cohort (BM7, BS7, BS8, and IS4). An overlap in the classification of clonality between DNA methylation beta values and aCGH data was found in 63% of the tumour pairs (BM7, BS7, IM4, IM9, and IS3), which is lower than in other studies [155, 156]. The small cohort size and the dynamic nature of DNA methylation limited the conclusions that could be drawn regarding the feasibility of using DNA methylation as a tool to assess tumour clonality.

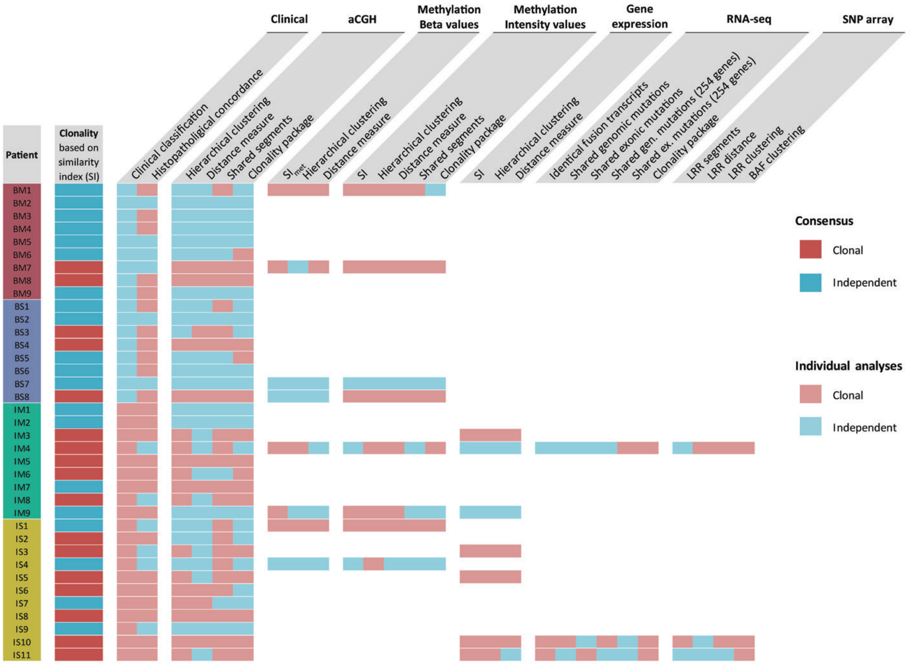

Figure 12. Classification of clonality based on statistical methods and type of data. Red boxes indicate that the analysis defined the tumour pair as clonal and blue boxes indicate independence of the tumours. BAF: B allele frequency; LRR: log R ratio; SI: Similarity Index; SI$_{met}$: modified SI for methylation data.

4.2.5 Gene expression as a tool for clonal relatedness

The gene expression subcohort consisted of seven patients with ipsilateral tumours. All synchronous cases (4/4) were identified as clonal while 2/3 metachronous cases were classified as independent tumours (**Figure 12**). Regardless of the statistical method used, consistent results were shown for clonality classification and were in line with the aCGH results (except for patient IM4). Ipsilateral synchronous tumour pairs were identified as clonal based on gene expression patterns indicating that gene expression is similar for tumour cells arising in the same breast at the same time. As gene expression is a highly dynamic process, it might depict the adjacent tumour microenvironment more than the underlying genetic clonal relatedness.

4.2.6 Agreement between the methods

Cohen's kappa indices were calculated to identify the agreement of clonality estimates between the different statistical methods. Hierarchical clustering and the SI showed the highest agreement for aCGH data (0.659 and 0.630, respectively). Hierarchical clustering is an unsupervised classification tool assuming that the number of clusters and their members are unknown, which is not the case in the assessment of clonality. Therefore, the SI is preferred over hierarchical clustering as it is easier to interpret and specifically designed for the comparison of two tumours from the same patient. The SI identified 46% (17/37) of the tumour pairs as clonal (**Figure 12**) and often showed opposite tendencies compared to the histopathological concordances. Additionally, no significant association between the SI and the clinical classification was found (Wilcoxon rank sum test: $P_{\text{Laterality}}$ = 0.247; $P_{\text{Synchronicity}}$ = 0.095; analysis of variance (ANOVA): $P_{\text{Clinical groups}}$ = 0.229), highlighting the alarming reality that there is very little connection between current clinical guidelines and the biology underlying tumour clonality.

The assessment of clonality based on the SI and hierarchical clustering showed similar tendencies in the majority of patients. The distance measure gave comparable results but seemed to be a more conservative measure since fewer tumour pairs were classified as clonal. The shared segment analysis with the aCGH data clearly favoured the clonality hypothesis with defining 21/37 tumour pairs as clonal.

In most cases, the type of molecular data presented similar tendencies regardless of the method applied. This leads to the underlying question of which biological process provides the most reliable evidence for clonality.

DNA methylation and gene expression are more dynamic than DNA mutations and CNAs. Hence, similar DNA methylation and gene expression patterns might represent responses to similar environmental factors. Tumour evolution of subclones is hypothesized to lead to a mixed pattern of shared and independent CNAs and DNA mutations [157], which advocates the use of DNA-based data to assess clonality. On the other hand, tumours developing in the same genetic background and environmental setting have been hypothesized to accumulate more frequently similar patterns of DNA mutations and CNAs, despite emerging as independent tumours [152, 158].

4.2.7 Limitations of the study

A major limitation of the study is the small cohort size, which limited the conclusions that could be drawn. Particularly in permutation-based approaches, some subcohorts were too small to perform meaningful statistics. A further disadvantage of the permutation-based approach was that it did not show a clear separation between the reference distribution and the clonal tumour pairs from the same patient. Hence, some artificial tumour pairs from different patients also showed statistical significance. Possible explanations could be intratumour heterogeneity, advanced accumulation of changes unique to the subclone, repeated occurrence of CNAs frequently identified in breast cancer [159-161], or technical artefacts increasing background noise [157]. Intratumour heterogeneity complicates clonality analyses due to biological differences in different parts of a tumour and subclone evolution, which presents a general disadvantage of bulk analyses. Single-cell DNA sequencing could help to avoid obstacles connected to intratumour heterogeneity and contamination with normal cells. In aCGH, contamination with normal cells can diminish the intensity of detected CNAs and small cell populations might not be detected. However, by using only samples that showed a tumour cell content of at least 70%, we ruled out that a lack of clonal relatedness could be due to a lack of tumour cells.

4.3 Paper III

4.3.1 Dose-dependent differences in genomic instability

CNAs were found in all analysed tumours and affected an average of 17.23 chromosomes per tumour, while ranging between 7 and 23 affected chromosomes per tumour. Breast carcinomas in the low- and high-dose groups showed significantly different patterns of CNAs across the genome (**Figure 13**). The high-dose group encompassed statistically significant regions of copy number gains on chromosomes 2q, 4, 17, 21q, and 22q, as well as copy number losses on 6q in comparison with the low-dose group.

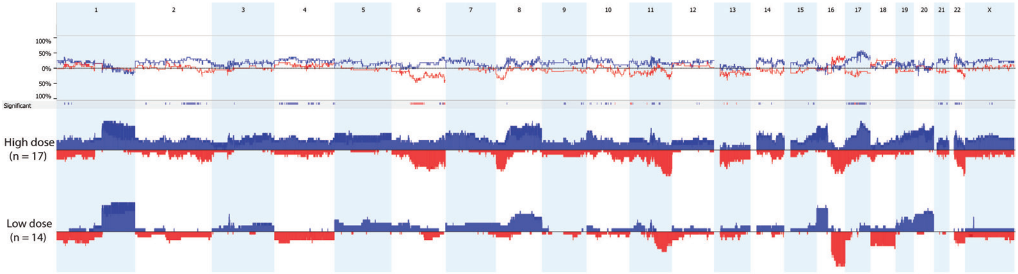

Figure 13. Genome-wide frequency plots of DNA copy number gains (blue) and losses (red) stratified by the high-dose (n = 17) and low-dose groups (n = 14).

In addition, the high-dose group showed a significantly higher number of total CNAs (**Figure 14**; *P* = 0.003) consistent with a higher fraction of the genome altered (FGA; *P* = 0.044). The TXP values (G2I-derived statistic; *P* = 0.048), the numbers of two or more copy gains (*P* = 0.019) and one copy losses (*P* = 0.035) were significantly increased in the high-dose group. Furthermore, a significantly higher percentage of the genome was changed in the high-dose group as compared to the low-dose group (Nexus-derived statistic; *P* = 0.019). These results suggest that a higher absorbed dose in the infant led to more complex genomic alterations in the breast tumour genome as an adult.

The increased complexity of genomic alterations was manifested in an increase in the number of CNAs and a higher FGA and thereby a more unstable tumour genome. These findings demonstrated that dose-dependent biological changes can persist in a patient up to 80 years after irradiation and are consistent with previous studies on A-bomb survivors reporting increased numbers of CNAs as a hallmark of genomic instability [162].

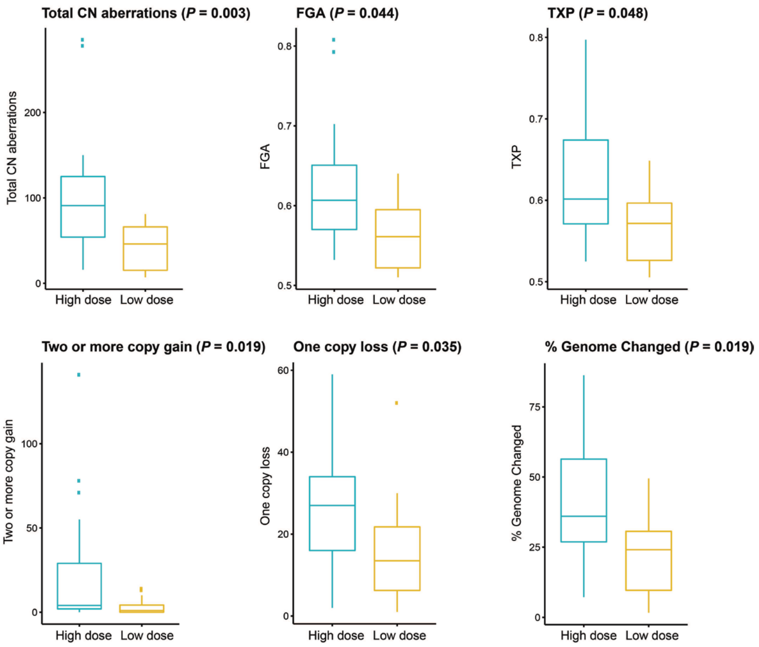

Figure 14. Boxplots of molecular tumour features stratified by the high-dose (n = 17) and low-dose groups (n = 14) with p-values calculated using Mann-Whitney U test.

Other factors such as individual radiation sensitivity, exposure to other genotoxic stressors, or lifestyle choices in the subsequent decades might confound the dose-dependent genomic instability to some extent [89]. However, Spearman’s rho (ρ) identified significant positive correlations between the absorbed dose and all three CAAI measures (average, number of unstable arms, and maximum score) along with the number of chromosomal segments, one copy losses, two or more copy gains, and the percentage of the genome changed. A strong correlation was found between the total number of CNAs and the absorbed dose, further corroborating that a higher absorbed dose is associated with a more unstable genome.

4.3.2 Increased occurrence of chromothripsis-like patterns (CTLP regions)

Chromothripsis has been described in almost all cancer types with a prevalence of approximately 2-3% [69] and about 0-21% in breast carcinomas [163]. We detected a total of 12 CTLP regions in 29% of the patient samples (9/31) presenting an increase in frequency. In most patients only one chromosome was affected, except for patient T13 (low-dose group) with three CTLP regions and patient T108 (high-dose group) with two regions. One-third of the detected CTLP regions were on chromosome 11, which is a previously identified CTLP hotspot [138]. Contrarily, the occurrence of CTLPs was independent of the absorbed dose groups for the 31 patients included in this study (*P* = 0.456). A possible explanation could be the small cohort size. Furthermore, several mechanisms are hypothesized to cause chromothripsis, such as telomere attrition, and DSB generation by other exogenous agents or oncogene-induced replicative stress [164]. Therefore, chromothripsis might occur in some breast tumours of this cohort regardless of prior irradiation exposure.

The exact mechanisms driving chromothripsis are still unknown. One theory proposes that chromothripsis is triggered through ionizing radiation during mitosis leading to DSBs in a narrow region of a chromosome or several chromosomes in close proximity [74, 75]. Subsequently, the resulting chromosome fragments could re-assemble in the consecutive G1 phase [75, 164]. CTLPs can be induced artificially by ionizing radiation using proton microbeam irradiation [165]. The mechanism inducing chromothripsis is hypothesized to vary depending on the type of radiation exposure [165]. Accordingly, when the entire cell is irradiated (as assumed in the radium-226 treatment), chromothripsis is thought to be induced via micronuclei formation. The increased occurrence of CTLPs in the Swedish haemangioma cohort could indicate that chromothripsis and genomic instability might be implemented through defects in similar pathways or mechanisms, such as micronuclei formation [165]. However, it remains unclear whether chromothripsis was a direct effect of radiation exposure or an indirect effect where radiation-activated factors lead to chromothripsis in subsequent decades.

4.3.3 Interaction between absorbed dose and genomic instability

The G2I algorithm stratified the cohort with statistical significance in the univariable Cox model but had low predictive power (**Figure 15A**; *P* = 0.006; C-index: 0.656; AIC: 38.299). Inversely, the univariable Cox model fitted using the absorbed dose showed no statistical significance but had high predictive power (*P* = 0.1; C-index: 0.800; AIC: 34.978). Multivariable models were fitted to investigate putative additive effects (G2I + absorbed dose) and possible interactions between the two observations (**Figure 15B**; G2I * absorbed dose; equivalent to: G2I + absorbed dose + G2I:absorbed dose).

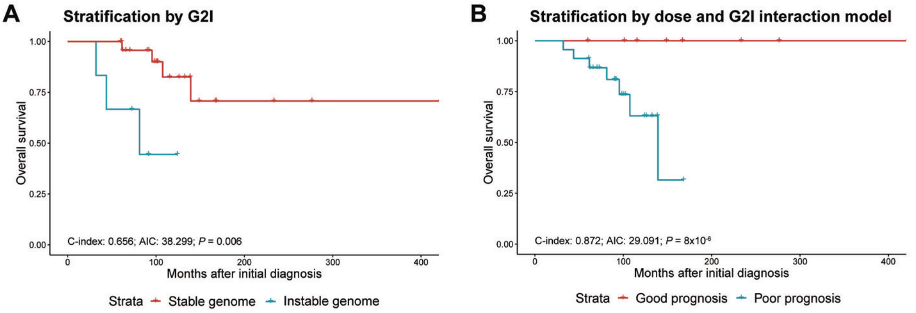

Figure 15. Kaplan-Meier analysis stratified by (**A**) a univariable Cox model using G2I, and (**B**) the linear predictor of a multivariable Cox model comprising G2I, the absorbed dose and the interaction variable (G2I:absorbed dose).

The interaction model of G2I and the absorbed dose exhibited the highest predictive power and the lowest AIC. Furthermore, the interaction model stratified all deceased patients into the poor prognosis group using the linear predictor η. As previously reported, a mechanistic model incorporating the effects of radiation-induced genomic instability at an early stage of carcinogenesis significantly improved the description of the radiation risk in the Swedish haemangioma cohort [95, 97, 98]. Taken together, the interaction model supported the hypothesis that radiation-induced genomic instability forms the missing link to the dose-dependent risk of developing breast cancer in the Swedish haemangioma cohort.

4.3.4 Limitations of the study

This study had several limitations. First, the small sample size limited the statistical power of the study. Second, the low DNA quality of FFPE samples limited the range of assays that could be applied. Third, the OncoScan array has a relatively low resolution ranging between 50-300 kb. Hence, the default settings of the CTLPScanner were increased to a minimum segment size of 50 kb and lowered to copy number status change times ≥10, which is common practice for array data [163]. Fourth, absorbed dose is a macrodosimetric measure that is not considered optimal for the analysis and interpretation of molecular events [166]. The availability of microdosimetric calculations would have provided a more accurate basis for associations between genomic instability and relative biological effectiveness of the radiation microenvironment in breast tissue. Finally, it should be noted that from a radiobiological perspective, absorbed doses around 1 Gy are considered “moderate” and that doses above several Gray are categorized as “high”.

5 CONCLUSIONS AND OUTLOOK

The 18-marker panel proved to be a robust classification model with high predictive power that effectively stratified patients into low- and high-risk prognosis groups. The predictive power was stable over time and gave the best prediction in combination with established markers for DSS. Use of the 18-marker panel in conjunction with clinical parameters can help to personalize treatment resulting in more aggressive therapy for high-risk patients (reduce under-treatment) and less aggressive therapy for low-risk patients (reduce over-treatment). Further research is needed to validate the biological impact of the 18 markers on the protein level. Understanding the interplay between these markers can facilitate the development of new therapies for high-risk breast cancer patients.

The SI was identified as the most accurate tool to assess clonal relatedness in breast tumour pairs by comparing the degree of similarity of an individual tumour pair to a reference distribution. In the majority of the analyses, the type of molecular data used had a stronger impact on the assessment of clonality than the analytical method used. In metachronous cancer, tumour clonality indicates insufficient treatment of the first tumour, hence, the patient could benefit from a change in treatment regimen. A more accurate classification of clonal relatedness may mitigate treatment failure and relapse by integrating tumour-associated molecular features and clinical parameters. Future research needs to define guidelines with exact thresholds to standardize clonality testing in a routine diagnostic setting.

In the Swedish haemangioma cohort, we found biological indications for radiation-induced genomic instability persisting after exposure and ultimately promoting carcinogenesis. Tumours from patients with higher absorbed dose showed increased levels of genomic instability demonstrating the long-term consequences of irradiation in humans. The highly predictive Cox regression model incorporating the interaction between absorbed dose and genomic instability gave further evidence for radiation contributing to carcinogenesis through genomic instability. However, the molecular mechanisms accounting for persistence of genomic instability and its manifestation in breast carcinomas remain unclear. Future research needs to investigate to which degree micronucleus formation, ROS, and radiation-induced changes in DNA methylation affect genome stability. This work gives a biological basis for improved risk assessment to minimize carcinogenesis as a secondary disease after radiation therapy.

ACKNOWLEDGEMENTS

I would like to express my gratitude to everyone who supported me during these years, and especially:

Khalil Helou, my main supervisor, for the opportunity to join your group and for introducing me to the exciting field of breast cancer. Thank you for giving me the freedom to work on the topics that I enjoyed most and for giving me the time to discover my own strengths. Knowing I can always come to you and discuss ideas has been invaluable. I want to thank you for all your support, encouragement, scientific and non-scientific guidance during this time.

Toshima Parris, my co-supervisor, for your endless help in each and every aspect of this PhD. I am incredibly grateful for all the knowledge you shared with me, all your support and encouragement, and especially for never getting tired of correcting my manuscripts.

Per Karlsson and **Anikó Kovács**, my co-supervisors, for sharing your limitless knowledge and constructive feedback. Thank you for all your help and for believing in me.

Eva Forssell-Aronsson, my co-supervisor, for your involvement in my projects.

Szilárd Nemes, my unofficial co-supervisor, for everything I have learned from you, for your patience, and for believing in me. None of this work would have been possible without you.

Britta Langen, my second unofficial co-supervisor, for your constructive feedback, advice, and support. I am grateful for your friendship and for every time you made me laugh even in the toughest moments.

I would also like to thank the members of the group: **Hanna**, for your funny remarks in the most unexpected moments; **Elisabeth**, for being the loving and caring person that you are; and of course **May** and **Ulla**, the “lab ladies”, for the nice time and technical support.

I don’t want to imagine what these years would have been like without my friends, **Ágota** and **Junchi**. I want to thank you for your friendship and your company in both, the good and the tough moments. I am very grateful for all

your encouragement throughout this time and, of course, for the enormous amounts of chocolate in the fluffy jar ☺

A very big thank you goes to my amazing **lunch group** including **Agnieszka**, **Gautam**, **Dorota**, and **Dominika** for all the laughter and occasionally incredibly irritating discussions we had ☺

I also would like to thank the people frequently met in the legendary Sahlgrenska Cancer Center elevator (which clearly demonstrates the peak of Swedish engineering capabilities): **Karoline**, **Stefan**, **Rebecca**, and **Parmida**, for your friendship since the early Master days; **Hana**, for your energetic and genuine personality, never let Sweden break you!; **Mia**, for all your funny remarks and incredible humour; and **Alexandra**, for your help in the administrative maze of PhD studies and for being the straightforward person that you are.

None of this would have been possible without the endless support from my parents, **Tina** and **Ecki**. Vielen Dank für all eure Unterstützung und dass ihr immer für mich da seid. Ich bin froh, euch zu haben! I would like to thank everyone in **my family** for all their support and encouragement, especially: **Oma Anne**, dafür dass du immer an mich glaubst; **Christopher**, for your friendship, integrity and amazing taste in movies; **Mani** and **Hubert**, for always believing in me and for actually reading my papers ☺

Last but not least, I would also like to thank my amazing friends: **Sai**, thank you so much for your support during this time and, of course, for the endless amounts of chocolate ☺; **Lena F.**, for being the positive and thoughtful friend that you are; **Sissi**, for being friends since childhood and for always keeping in touch despite the distance; my **Bachelor and Master girls**: **Lena**, **Katharina**, **Vanessa**, **Melanie**, **Vera**, and **Verena**, for all the amazing moments we had and for keeping in touch throughout the years.

REFERENCES

1. World Health Organization, **Cancer.** Access date: 01/02/2019. Available from: https://www.who.int/en/news-room/fact-sheets/detail/cancer.
2. World Health Organization, **Cancer fact sheets. Source: Globocan 2018.** International Agency for Research on Cancer. Access date: 01/02/2019. Available from: http://gco.iarc.fr/today/fact-sheets-cancers.
3. Vogelstein, B., et al., **Cancer genome landscapes**. *Science (New York, N.Y.)*, 2013. 339(6127): p. 1546-1558.
4. Stratton, M.R., P.J. Campbell, and P.A. Futreal, **The cancer genome**. *Nature*, 2009. 458(7239): p. 719-24.
5. Tomasetti, C., L. Li, and B. Vogelstein, **Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention**. *Science (New York, N.Y.)*, 2017. 355(6331): p. 1330-1334.
6. American Cancer Society, **Family Cancer Syndromes.** Access date: 01/02/2019. Available from: https://www.cancer.org/cancer/cancer-causes/genetics/family-cancer-syndromes.html.
7. Ellsworth, D.L., et al., **Single-cell sequencing and tumorigenesis: improved understanding of tumor evolution and metastasis**. *Clinical and translational medicine*, 2017. 6(1): p. 15-15.
8. Ramsay, D.T., et al., **Anatomy of the lactating human breast redefined with ultrasound imaging**. *Journal of anatomy*, 2005. 206(6): p. 525-534.
9. Gudjonsson, T., et al., **Myoepithelial cells: their origin and function in breast morphogenesis and neoplasia**. *Journal of mammary gland biology and neoplasia*, 2005. 10(3): p. 261-272.
10. Pandya, S. and R.G. Moore, **Breast development and anatomy**. *Clin Obstet Gynecol*, 2011. 54(1): p. 91-5.
11. National Cancer Institute, **PDQ Breast Cancer Treatment.** Access date: 01/02/2019. Available from: https://www.cancer.gov/types/breast/hp/breast-treatment-pdq#_627_toc.
12. World Health Organization, **World Cancer Report 2014**. 2014, International Agency for Research on Cancer.
13. Brierley, J., M.K. Gospodarowicz, and C. Wittekind, **TNM classification of malignant tumours**. Eighth edition. ed. 2017, Chichester, West Sussex, UK; Hoboken, NJ: John Wiley & Sons, Inc.
14. Dahlman-Wright, K., et al., **International Union of Pharmacology. LXIV. Estrogen receptors**. *Pharmacol Rev*, 2006. 58(4): p. 773-81.
15. Cianfrocca, M. and L.J. Goldstein, **Prognostic and predictive factors in early-stage breast cancer**. *Oncologist*, 2004. 9(6): p. 606-16.
16. Slamon, D., et al., **Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene**. *Science*, 1987. 235(4785): p. 177-182.
17. Varga, Z., et al., **Assessment of HER2 status in breast cancer: overall positivity rate and accuracy by fluorescence in situ hybridization and immunohistochemistry in a**

single institution over 12 years: a quality control study. *BMC Cancer*, 2013. 13(1): p. 615.

18. Yerushalmi, R., et al., **Ki67 in breast cancer: prognostic and predictive potential**. *Lancet Oncol*, 2010. 11(2): p. 174-83.

19. Harris, L.N., et al., **Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline**. *Journal of Clinical Oncology*, 2016. 34(10): p. 1134-1150.

20. Perou, C.M., et al., **Molecular portraits of human breast tumours**. *Nature*, 2000. 406(6797): p. 747-52.

21. Sorlie, T., et al., **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *Proc Natl Acad Sci U S A*, 2001. 98(19): p. 10869-74.

22. Calza, S., et al., **Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients**. *Breast Cancer Res*, 2006. 8(4): p. R34.

23. Goldhirsch, A., et al., **Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011**. *Ann Oncol*, 2011. 22(8): p. 1736-47.

24. Curigliano, G., et al., **De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017**. *Ann Oncol*, 2019.

25. Vijver, M., et al., **A gene-expression signature as a predictor of survival in breast cancer**. *N Engl J Med*, 2002. 347.

26. van 't Veer, L.J., et al., **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature*, 2002. 415(6871): p. 530-6.

27. Cardoso, F., et al., **70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer**. *New England Journal of Medicine*, 2016. 375(8): p. 717-729.

28. Paik, S., et al., **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer**. *N Engl J Med*, 2004. 351(27): p. 2817-26.

29. Brufsky, A.M., **Predictive and prognostic value of the 21-gene recurrence score in hormone receptor-positive, node-positive breast cancer**. *American journal of clinical oncology*, 2014. 37(4): p. 404-410.

30. Reis-Filho, J. and L. Pusztai, **Gene expression profiling in breast cancer: classification, prognostication, and prediction**. *The Lancet*, 2011. 378(9805): p. 1812-1823.

31. Wirapati, P., et al., **Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures**. *Breast Cancer Research*, 2008. 10(4): p. R65.

32. Early Breast Cancer Trialists' Collaborative Group, **Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials**. *Lancet*, 2005. 365(9472): p. 1687-717.

33. Liu, J., et al., **Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data**. *BMC Med Genomics*, 2008. 1: p. 39.

34. Clark, T.G., et al., **Survival analysis part I: basic concepts and first analyses**. *Br J Cancer*, 2003. 89(2): p. 232-8.

35. Kleinbaum, D.G. and M. Klein, **Survival Analysis: A Self-Learning Text**. 2006: Springer New York.

36. Cox, D., **Regression Models and Life Tables**. *Journal of the Royal Statistical Society, Series B*, 1972. 34.

37. Bradburn, M.J., et al., **Survival analysis part II: multivariate data analysis--an introduction to concepts and methods**. *Br J Cancer*, 2003. 89(3): p. 431-6.

38. Annest, A., et al., **Iterative Bayesian Model Averaging: a method for the application of survival analysis to high-dimensional microarray data**. *BMC Bioinformatics*, 2009. 10(1): p. 72.

39. Yeung, K.Y., R.E. Bumgarner, and A.E. Raftery, **Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data**. *Bioinformatics*, 2005. 21(10): p. 2394-2402.

40. Volinsky, C., et al., **Bayesian Model Averaging in Proprtional Hazard Models: Assessing the Risk of a Stroke**. *Applied Statistics*, 1997. 46.

41. Saha-Chaudhuri, P. and P.J. Heagerty, **Non-parametric estimation of a time-dependent predictive accuracy curve**. *Biostatistics (Oxford, England)*, 2013. 14(1): p. 42-59.

42. Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, **Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors**. *Stat Med*, 1996. 15(4): p. 361-87.

43. Rahman, M.S., et al., **Review and evaluation of performance measures for survival prediction models in external validation settings**. *BMC Medical Research Methodology*, 2017. 17: p. 60.

44. Collins, G.S., et al., **External validation of multivariable prediction models: a systematic review of methodological conduct and reporting**. *BMC Med Res Methodol*, 2014. 14: p. 40.

45. Royston, P. and D.G. Altman, **External validation of a Cox prognostic model: principles and methods**. *BMC Med Res Methodol*, 2013. 13: p. 33.

46. Nowell, P.C., **The clonal evolution of tumor cell populations**. *Science*, 1976. 194(4260): p. 23-8.

47. Merlo, L.M.F., et al., **Cancer as an evolutionary and ecological process**. *Nature Reviews Cancer*, 2006. 6: p. 924.

48. Aparicio, S. and C. Caldas, **The implications of clonal genome evolution for cancer medicine**. *N Engl J Med*, 2013. 368(9): p. 842-51.

49. Burrell, R.A., et al., **The causes and consequences of genetic heterogeneity in cancer evolution**. *Nature*, 2013. 501(7467): p. 338-45.

50. Begg, C.B., K.H. Eng, and A.J. Hummer, **Statistical tests for clonality**. *Biometrics*, 2007. 63(2): p. 522-30.

51. Ostrovnaya, I., V.E. Seshan, and C.B. Begg, **Using somatic mutation data to test tumors for clonal relatedness**. *Ann Appl Stat.*, 2015. 9.

52. Mertens, F., et al., **Chromosomal Imbalance Maps of Malignant Solid Tumors: A Cytogenetic Survey of 3185 Neoplasms**. *Cancer Research*, 1997. 57(13): p. 2765-2780.

53. Nemes, S., et al., **A diagnostic algorithm to identify paired tumors with clonal origin**. *Genes Chromosomes Cancer*, 2013. 52(11): p. 1007-16.

54. Ostrovnaya, I., et al., **A metastasis or a second independent cancer? Evaluating the clonal origin of tumors using array copy number data**. *Stat Med*, 2010. 29(15): p. 1608-21.

55. Smith, T.E., et al., **True recurrence vs. new primary ipsilateral breast tumor relapse: an analysis of clinical and pathologic differences and their implications in natural history, prognoses, and therapeutic management**. *Int J Radiat Oncol Biol Phys*, 2000. 48(5): p. 1281-9.

56. Hanahan, D. and R.A. Weinberg, **The hallmarks of cancer**. *Cell*, 2000. 100(1): p. 57-70.

57. Loeb, L.A., **Human Cancers Express a Mutator Phenotype: Hypothesis, Origin, and Consequences**. *Cancer Res*, 2016. 76(8): p. 2057-9.

58. Fearon, E.R. and B. Vogelstein, **A genetic model for colorectal tumorigenesis**. *Cell*, 1990. 61(5): p. 759-67.

59. Loeb, L.A., C.F. Springgate, and N. Battula, **Errors in DNA Replication as a Basis of Malignant Changes**. *Cancer Research*, 1974. 34(9): p. 2311.

60. Davis, A., R. Gao, and N. Navin, **Tumor evolution: Linear, branching, neutral or punctuated?** *Biochim Biophys Acta Rev Cancer*, 2017. 1867(2): p. 151-161.

61. Ionov, Y., et al., **Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis**. *Nature*, 1993. 363(6429): p. 558-61.

62. Lengauer, C., K.W. Kinzler, and B. Vogelstein, **Genetic instabilities in human cancers**. *Nature*, 1998. 396: p. 643.

63. Negrini, S., V.G. Gorgoulis, and T.D. Halazonetis, **Genomic instability — an evolving hallmark of cancer**. *Nature Reviews Molecular Cell Biology*, 2010. 11: p. 220.

64. Geigl, J.B., et al., **Defining 'chromosomal instability'**. *Trends Genet*, 2008. 24(2): p. 64-9.

65. Nik-Zainal, S., et al., **The life history of 21 breast cancers**. *Cell*, 2012. 149(5): p. 994-1007.

66. Lee, J.K., et al., **Mechanisms and Consequences of Cancer Genome Instability: Lessons from Genome Sequencing Studies**. *Annu Rev Pathol*, 2016. 11: p. 283-312.

67. Kalimutho, M., et al., **Patterns of Genomic Instability in Breast Cancer**. *Trends Pharmacol Sci*, 2019.

68. Fearon, E.R., **Human cancer syndromes: clues to the origin and nature of cancer**. *Science*, 1997. 278(5340): p. 1043-50.

69. Stephens, P.J., et al., **Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development**. *Cell*, 2011. 144(1): p. 27-40.

70. Korbel, J.O. and P.J. Campbell, **Criteria for inference of chromothripsis in cancer genomes**. *Cell*, 2013. 152(6): p. 1226-36.

71. Tubio, J.M. and X. Estivill, **Cancer: When catastrophe strikes a cell**. *Nature*, 2011. 470(7335): p. 476-7.

72. Johnson, R.T. and P.N. Rao, **Mammalian Cell Fusion : Induction of Premature Chromosome Condensation in Interphase Nuclei**. *Nature*, 1970. 226(5247): p. 717-722.

73. Crasta, K., et al., **DNA breaks and chromosome pulverization from errors in mitosis**. *Nature*, 2012. 482(7383): p. 53-8.

74. Maher, C.A. and R.K. Wilson, **Chromothripsis and human disease: piecing together the shattering process**. *Cell*, 2012. 148(1-2): p. 29-32.

75. Lieber, M.R., **The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway**. *Annual review of biochemistry*, 2010. 79: p. 181-211.

76. Kim, J.-G., et al., **Epigenetics meets radiation biology as a new approach in cancer treatment**. *International journal of molecular sciences*, 2013. 14(7): p. 15059-15073.

77. Muller, H.J., **Artifical Transmutation of the Gene**. *Science*, 1927. 66(1699): p. 84-7.

78. Boss, M.-K., R. Bristow, and M.W. Dewhirst, **Linking the history of radiation biology to the hallmarks of cancer**. *Radiation research*, 2014. 181(6): p. 561-577.

79. Desouky, O., N. Ding, and G. Zhou, **Targeted and non-targeted effects of ionizing radiation**. *Journal of Radiation Research and Applied Sciences*, 2015. 8(2): p. 247-254.

80. Morgan, W.F., **Radiation-induced genomic instability**. *Health Phys*, 2011. 100(3): p. 280-1.

81. Morgan, W.F., **Non-targeted and delayed effects of exposure to ionizing radiation: I. Radiation-induced genomic instability and bystander effects in vitro**. *Radiat Res*, 2003. 159(5): p. 567-80.

82. Hall, E.J. and A.J. Giaccia, **Radiobiology for the Radiologist**. 2006: Lippincott William & Wilkins.

83. Nagasawa, H. and J.B. Little, **Unexpected sensitivity to the induction of mutations by very low doses of alpha-particle radiation: evidence for a bystander effect**. *Radiat Res*, 1999. 152(5): p. 552-7.

84. Little, J.B., **Genomic instability and bystander effects: a historical perspective**. *Oncogene*, 2003. 22: p. 6978.

85. Goldberg, Z. and B.E. Lehnert, **Radiation-induced effects in unirradiated cells: a review and implications in cancer**. *Int J Oncol*, 2002. 21(2): p. 337-49.

86. Pampfer, S. and C. Streffer, **Increased Chromosome Aberration Levels in Cells from Mouse Fetuses after Zygote X-irradiation**. *International Journal of Radiation Biology*, 1989. 55(1): p. 85-92.

87. Morgan, W.F., et al., **Genomic instability induced by ionizing radiation**. *Radiat Res*, 1996. 146(3): p. 247-58.

88. Ronckers, C.M., C.A. Erdmann, and C.E. Land, **Radiation and breast cancer: a review of current evidence**. *Breast Cancer Research*, 2004. 7(1): p. 21.

89. Holmberg, E., et al., **Excess breast cancer risk and the role of parity, age at first childbirth and exposure to radiation in infancy**. *British journal of cancer*, 2001. 85(3): p. 362-366.

90. Boice, J.D., Jr., et al., **Frequent chest X-ray fluoroscopy and breast cancer incidence among tuberculosis patients in Massachusetts**. *Radiat Res*, 1991. 125(2): p. 214-22.

91. Mattsson, A., et al., **Radiation-induced breast cancer: long-term follow-up of radiation therapy for benign breast disease**. *J Natl Cancer Inst*, 1993. 85(20): p. 1679-85.

92. Miller, A.B., et al., **Mortality from breast cancer after irradiation during fluoroscopic examinations in patients being treated for tuberculosis**. *N Engl J Med*, 1989. 321(19): p. 1285-9.

93. Land, C.E., et al., **Incidence of female breast cancer among atomic bomb survivors, Hiroshima and Nagasaki, 1950-1990**. *Radiat Res*, 2003. 160(6): p. 707-17.

94. Darrow, D.H., et al., **Diagnosis and Management of Infantile Hemangioma**. *Pediatrics*, 2015. 136(4): p. e1060-104.

95. Eidemuller, M., et al., **Breast cancer risk and possible mechanisms of radiation-induced genomic instability in the Swedish hemangioma cohort after reanalyzed dosimetry**. *Mutat Res*, 2015. 775: p. 1-9.

96. Lundell, M., et al., **Breast cancer risk after radiotherapy in infancy: a pooled analysis of two Swedish cohorts of 17,202 infants**. *Radiat Res*, 1999. 151(5): p. 626-32.

97. Eidemuller, M., et al., **Breast cancer risk after radiation treatment at infancy: potential consequences of radiation-induced genomic instability**. *Radiat Prot Dosimetry*, 2011. 143(2-4): p. 375-9.

98. Eidemuller, M., et al., **Breast cancer risk among Swedish hemangioma patients and possible consequences of radiation-induced genomic instability**. *Mutat Res*, 2009. 669(1-2): p. 48-55.

99. Hu, H., et al. **Comparative Study of Classification Methods for Microarray Data Analysis**. in *Proceedings of the Fifth Australasian Conference on Data Mining and Analystics*. 2006. Sydney, Australia: Australian Computer Society.

100. Parris, T.Z., et al., **Clinical implications of gene dosage and gene expression patterns in diploid breast carcinoma**. *Clin Cancer Res*, 2010. 16(15): p. 3860-74.

101. Mollerstrom, E., et al., **High-resolution genomic profiling to predict 10-year overall survival in node-negative breast cancer**. *Cancer Genet Cytogenet*, 2010. 198(2): p. 79-89.

102. Parris, T.Z., et al., **Frequent MYC coamplification and DNA hypomethylation of multiple genes on 8q in 8p11-p12-amplified breast carcinomas**. *Oncogenesis*, 2014. 3: p. e95.

103. Pawitan, Y., et al., **Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts**. *Breast Cancer Research*, 2005. 7(6): p. R953-R964.

104. Ivshina, A.V., et al., **Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer**. *Cancer Res*, 2006. 66(21): p. 10292-301.

105. TCGA, **The Cancer Genome Atlas (TCGA)**. https://cancergenome.nih.gov/.

106. Saal, L.H., et al., **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data**. *Genome Biol*, 2002. 3(8).

107. Jonsson, G., et al., **High-resolution genomic profiles of breast cancer cell lines assessed by tiling BAC array comparative genomic hybridization**. *Genes Chromosomes Cancer*, 2007. 46(6): p. 543-58.

108. Knight, S.J., et al., **An optimized set of human telomere clones for studying telomere integrity and architecture**. *Am J Hum Genet*, 2000. 67(2): p. 320-32.

109. Vissers, L.E., et al., **Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities**. *Am J Hum Genet*, 2003. 73(6): p. 1261-70.

110. Hupe, P., et al., **Analysis of array CGH data: from signal ratio to gain and loss of DNA regions**. *Bioinformatics*, 2004. 20(18): p. 3413-22.

111. Ostrovnaya, I., et al., **Clonality: an R package for testing clonal relatedness of two tumors from the same patient based on their genomic profiles**. *Bioinformatics.*, 2011. 27.

112. Seshan and Olshen, **DNAcopy: A Package for Analyzing DNA Copy Data**. 2010.

113. Assenov, Y., et al., **Comprehensive analysis of DNA methylation data with RnBeads**. *Nat Meth*, 2014. 11(11): p. 1138-1140.

114. Teschendorff, A.E., et al., **A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data**. *Bioinformatics*, 2013. 29(2): p. 189-96.

115. Feber, A., et al., **Using high-density DNA methylation arrays to profile copy number alterations**. *Genome Biology*, 2014. 15(2): p. R30.

116. Morris, T.J., et al., **ChAMP: 450k Chip Analysis Methylation Pipeline**. *Bioinformatics*, 2014. 30(3): p. 428-430.

117. Hovestadt V and Zapatka M, **conumee: Enhanced copy-number variation analysis using Illumina DNA methylation arrays**. 2017.

118. Parris, T.Z., et al., **Genome-wide multi-omics profiling of the 8p11-p12 amplicon in breast carcinoma**. *Oncotarget*, 2018. 9(35): p. 24140-24154.

119. Anders, S., P.T. Pyl, and W. Huber, **HTSeq--a Python framework to work with high-throughput sequencing data**. *Bioinformatics*, 2015. 31(2): p. 166-9.

120. Trapnell, C., et al., **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nat Protoc*, 2012. 7(3): p. 562-78.

121. Nicorici, D., et al., **FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data**. *bioRxiv*, 2014.

122. Shugay, M., et al., **Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions**. *Bioinformatics*, 2013. 29(20): p. 2539-46.

123. McKenna, A., et al., **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res*, 2010. 20(9): p. 1297-303.

124. Life Technologies, **Chromosome Analysis Suite 3.2 (ChAS 3.2).** Access date: 01/02/2019. Available from: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/ChAS_Manual.pdf.

125. Rasmussen, M., et al., **Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity**. *Genome Biol*, 2011. 12(10): p. R108.

126. R Core Team, **R: A Language and Environment for Statistical Computing**. 2018. p. R Foundation for Statistical Computing.

127. Therneau, T., **survival-package: A Package for Survival Analysis in S**. 2015.

128. Adrian Raftery, J.H., Chris Volinsky, Ian Painter and Ka Yee Yeung., **BMA: Bayesian Model Averaging**. 2017.

129. A. Kassambara and M. Kosinski, **survminer: Drawing Survival Curves using 'ggplot2'**. 2017.

130. Heagerty, P.J. and P. Saha-Chaudhuri, **risksetROC: Riskset ROC curve estimation from censored survival data**. 2012.

131. Du, P., et al., **Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis**. *BMC Bioinformatics*, 2010. 11: p. 587-587.

132. Ostrovnaya, I. and C.B. Begg, **Testing Clonal Relatedness of Tumors Using Array Comparative Genomic Hybridization: A Statistical Challenge**. *Clinical Cancer Research*, 2010. 16(5): p. 1358.

133. Begg, C.B., et al., **Contralateral breast cancers: Independent cancers or metastases?** *Int J Cancer*, 2017.

134. Cohen, J., **A Coefficient of Agreement for Nominal Scales**. *Educational and Psychological Measurement*, 1960. 20(1): p. 37-46.

135. Riccardo Lo Martire, **rel: Reliability Coefficients**. 2017.

136. Nilsen, G., et al., **Copynumber: Efficient algorithms for single- and multi-track copy number segmentation**. *BMC Genomics*, 2012. 13: p. 591.

137. Wood, H.M., et al., **The genomic road to invasion-examining the similarities and differences in the genomes of associated oral pre-cancer and cancer samples**. *Genome Med*, 2017. 9(1): p. 53.

138. Cai, H., et al., **Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens**. *BMC Genomics*, 2014. 15: p. 82.

139. Van Loo, P., et al., **Allele-specific copy number analysis of tumors**. *Proc Natl Acad Sci U S A*, 2010. 107(39): p. 16910-5.

140. Bonnet, F., et al., **An array CGH based genomic instability index (G2I) is predictive of clinical outcome in breast cancer and reveals a subset of tumors without lymph node involvement but with poor prognosis**. *BMC Med Genomics*, 2012. 5: p. 54.

141. Vollan, H.K.M., et al., **A tumor DNA complex aberration index is an independent predictor of survival in breast and ovarian cancer**. *Molecular oncology*, 2015. 9(1): p. 115-127.

142. Russnes, H.G., et al., **Genomic architecture characterizes tumor progression paths and fate in breast cancer patients**. *Science translational medicine*, 2010. 2(38): p. 38ra47-38ra47.

143. Chin, S.F., et al., **High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer**. *Genome biology*, 2007. 8(10): p. R215-R215.

144. Burrell, R.A., et al., **Replication stress links structural and numerical cancer chromosomal instability**. *Nature*, 2013. 494(7438): p. 492-496.

145. Lee, A.J.X., et al., **Chromosomal instability confers intrinsic multidrug resistance**. *Cancer research*, 2011. 71(5): p. 1858-1870.

146. Waldman, F.M., et al., **Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences**. *J Natl Cancer Inst*, 2000. 92(4): p. 313-20.

147. Park, S.C., et al., **Genetic changes in bilateral breast cancer by comparative genomic hybridisation**. *Clin Exp Med*, 2007. 7(1): p. 1-5.

148. Bollet, M.A., et al., **High-resolution mapping of DNA breakpoints to define true recurrences among ipsilateral breast cancers**. *J Natl Cancer Inst*, 2008. 100(1): p. 48-58.

149. Brommesson, S., et al., **Tiling array-CGH for the assessment of genomic similarities among synchronous unilateral and bilateral invasive breast cancer tumor pairs**. *BMC Clin Pathol*, 2008. 8: p. 6.

150. Castellarin, M., et al., **Clonal evolution of high-grade serous ovarian carcinoma from primary to recurrent disease**. *J Pathol*, 2013. 229(4): p. 515-24.

151. Klevebring, D., et al., **Exome sequencing of contralateral breast cancer identifies metastatic disease**. *Breast Cancer Research and Treatment*, 2015. 151(2): p. 319-324.

152. Alkner, S., et al., **Contralateral breast cancer can represent a metastatic spread of the first primary tumor: determination of clonal relationship between contralateral breast cancers using next-generation whole genome sequencing**. *Breast Cancer Res*, 2015. 17: p. 102.

153. Ostrovnaya, I., V.E. Seshan, and C.B. Begg, **Comparison of properties of tests for assessing tumor clonality**. *Biometrics*, 2008. 64(4): p. 1018-22.

154. Desmedt, C., et al., **Uncovering the genomic heterogeneity of multifocal breast cancer**. *J Pathol*, 2015. 236(4): p. 457-66.

155. Moarii, M., et al., **Epigenomic alterations in breast carcinoma from primary tumor to locoregional recurrences**. *PLoS One*, 2014. 9(8): p. e103986.

156. Huang, K.T., et al., **Assessment of DNA methylation profiling and copy number variation as indications of clonal relationship in ipsilateral and contralateral breast cancers to distinguish recurrent breast cancer from a second primary tumour**. *BMC Cancer*, 2015. 15: p. 669.

157. Andrade, V.P., et al., **Clonal relatedness between lobular carcinoma in situ and synchronous malignant lesions**. *Breast Cancer Res.*, 2012. 14.

158. Wistuba, II, et al., **Two identical triplet sisters carrying a germline BRCA1 gene mutation acquire very similar breast cancer somatic mutations at multiple other sites throughout the genome**. *Genes Chromosomes Cancer*, 2000. 28(4): p. 359-69.

159. Hicks, J., et al., **Novel patterns of genome rearrangement and their association with survival in breast cancer**. *Genome Res*, 2006. 16(12): p. 1465-79.

160. Fridlyand, J., et al., **Breast tumor copy number aberration phenotypes and genomic instability**. *BMC Cancer*, 2006. 6(1): p. 96.

161. Haverty, P.M., et al., **High-resolution genomic and expression analyses of copy number alterations in breast tumors**. *Genes Chromosomes Cancer*, 2008. 47(6): p. 530-42.

162. Oikawa, M., et al., **Significance of genomic instability in breast cancer in atomic bomb survivors: analysis of microarray-comparative genomic hybridization**. *Radiation oncology (London, England)*, 2011. 6: p. 168-168.

163. Luijten, M.N.H., J.X.T. Lee, and K.C. Crasta, **Mutational game changer: Chromothripsis and its emerging relevance to cancer**. *Mutat Res*, 2018. 777: p. 29-51.

164. Forment, J.V., A. Kaidi, and S.P. Jackson, **Chromothripsis and cancer: causes and consequences of chromosome shattering**. *Nat Rev Cancer*, 2012. 12(10): p. 663-70.

165. Morishita, M., et al., **Chromothripsis-like chromosomal rearrangements induced by ionizing radiation using proton microbeam irradiation system**. *Oncotarget*, 2016. 7(9): p. 10182-92.

166. Li, W.B., W. Hofmann, and W. Friedland, **Microdosimetry and nanodosimetry for internal emitters**. *Radiation Measurements*, 2018: p. Pages 29-42.

SAMMANFATTNING PÅ SVENSKA

Bröstcancer är den vanligaste cancertypen hos kvinnor med mer än 2 miljoner nya fall och cirka 627 000 bröstcancerrelaterade dödsfall i världen år 2018. Flera gensignaturer har visat sig kunna förutsäga prognosen för bröstcancerutfall och i vissa fall också kunna predicera effekter av olika behandlingar. Nuvarande behandlingsriktlinjer fokuserar ännu främst på etablerade patient- och tumörspecifika egenskaper och i mindre utsträckning på gensignaturer. Vi har identifierat en gensignatur baserad på uttrycket av 18 gener som gör att vi verkar kunna förutsäga bröstcancerspecifik överlevnad mer exakt än med den kliniskt validerade Oncotype Dx-signaturen.

Trots ökad överlevnad i bröstcancer under senare tid så får cirka 6-23% av patienterna återfall inom fem år. Detta talar för behandlingssvikt vid den initiala behandlingen. Det är mycket viktigt att skilja mellan återfall som beror på klonal evolution och de som beror på nya primära tumörer. För närvarande finns det ingen tydlig riktlinje för att avgöra om det är återfall med klonal evolution eller en ny primär bröstcancer. Vi jämförde olika statistiska metoder och datatyper och identifierade likhetsindexet (SI) som det mest pålitliga verktyget för att klassificera tumörklonalt återfall.

Bröstkörteln är känd för att vara mycket känslig för joniserande strålning, särskilt hos unga. Under åren 1920-1965 behandlades spädbarn med joniserande strålning för hudhemangiom och uppföljning har visat en ökad risk att utveckla bröstcancer senare i livet. Joniserande strålning har visat sig inducera genomisk instabilitet. Därför analyserade vi brösttumörerna från patienter i den svenska hemangiomkohorten för genomisk instabilitet. Patienter med högre absorberade doser i bröstet uppvisade ökad genomisk instabilitet jämfört med patienter som fick lägre doser. Strålningsinducerad genomisk instabilitet kan vara förklaringen till att joniserande strålning i ung ålder kan leda till ökad risk för bröstcancer under de följande decennierna.

Sammanfattningsvis belyser detta arbete hur molekylärbiologi och statistiska modeller kan öka möjligheterna till riskbedömningar i tillägg till redan etablerade kliniska och patologiska faktorer. Detta gör att riskbedömningarna kan förbättras och med hjälp av det kan strategierna för bröstcancerbehandling anpassas till det bättre.