

From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

Molecular epidemiology studies on risk factors for breast cancer and disease aggressiveness

Emilio Ugalde Morales



**Karolinska
Institutet**

Stockholm 2020

All previously published papers were reproduced with permission from the publisher.
If not otherwise stated, illustrations are by author.

Cover painting by the author, 2020.

Published by Karolinska Institutet.

Printed by Universitetservice US-AB

© Emilio Ugalde Morales, 2020

ISBN 978-91-7831-947-3

Molecular epidemiology studies on risk factors for breast cancer and disease aggressiveness
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Emilio Ugalde Morales

Principal Supervisor:

Professor Kamila Czene
Karolinska Institutet
Department of Medical Epidemiology and
Biostatistics

Co-supervisors:

Dr. Jingmei Li
Genome Institute of Singapore
Laboratory of Women's Health and Genetics

Dr. Felix Grassmann
University of Aberdeen
Institute of Medical Sciences

Professor Keith Humphreys
Karolinska Institutet
Department of Medical Epidemiology and
Biostatistics

Professor Per Hall
Karolinska Institutet
Department of Medical Epidemiology and
Biostatistics

Opponent:

Associate Professor Lao Saal
Lund University
Department of Clinical Sciences, Lund
Division of Oncology

Examination Board:

Associate Professor Carsten Daub
Karolinska Institutet
Department of Biosciences and Nutrition

Associate Professor Sofia Carlsson
Karolinska Institutet
Institute of Environmental Medicine

Professor Charlotta Dabrosin
Linköping University
Department of Biomedical and Clinical Sciences
Division of Surgery, Orthopedics and Oncology

To all affected by and involved in *cancer*.

ABSTRACT

Breast cancer is a heterogeneous disease. Aggressive subtypes are characterized by faster growth rates, increased capability to invade and metastasize, leading to poorer clinical outcomes. In this thesis, we use a molecular epidemiology approach to investigate the association between risk factors and aggressive breast cancer defined by tumor characteristics, intrinsic subtypes, mode of detection, and survival. Using a variety of methods, we analyzed data from well-characterized breast cancer cohorts in Sweden, genome-wide association studies, and gene expression profiling of tumors.

In Paper I, we found that breast cancer genetic load, defined by rare deleterious variants in 31 breast cancer genes, and unlike common variants, is positively associated with unfavorable tumor characteristics, patient survival, and mode of detection.

In Paper II, we observed that women with low breast cancer risk defined by the Tyrer-Cuzick risk score were more likely to develop aggressive tumors. We computed a low-risk gene expression profile that was consistently associated with worse prognosis. In addition, our analysis showed that increased proliferation rather than estrogen status underlie this association.

In Paper III, we examined gene expression profiles in a subset of aggressive breast cancer tumors, known as interval cancers. By taking mammographic density and intrinsic PAM50 subtypes into account, we found an interval cancer gene expression profile to be associated with immune subtypes in breast cancer, particularly those involving interferon response.

In Paper IV, we show that breast cancer has a shared immune-related genetic component with celiac disease, an autoimmune disorder. In consistency with previous epidemiological findings, we found that a higher genetic load for celiac disease was associated with lower breast cancer risk.

Overall, this thesis aims to provide scientific evidence towards a better understanding of the factors underlying the development of aggressive breast cancers that could shed light on the design of better preventative strategies aimed at lowering disease mortality.

LIST OF SCIENTIFIC PAPERS

- I. Jingmei Li, **Emilio Ugalde-Morales**, Wei Xiong Wen, Brennan Decker, Mikael Eriksson, Astrid Torstensson, Helene Nordahl Christensen, Alison M. Dunning, Jamie Allen, Craig Luccarini, Karen A. Pooley, Jacques Simard, Leila Dorling, Douglas F. Easton, Soo Hwang Teo, Per Hall, Kamila Czene
Differential burden of rare and common variants on tumor characteristics, survival, and mode of detection in breast cancer.
Cancer Research. 2018 Nov 1;78(21):6329-6338.

- II. **Emilio Ugalde-Morales**, Felix Grassmann, Keith Humphreys, Jingmei Li, Mikael Eriksson, Nicholas P. Tobin, Åke Borg, Johan Vallon-Christersson, Per Hall, Kamila Czene
Association between breast cancer risk and disease aggressiveness: characterizing underlying gene expression patterns.
International Journal of Cancer. (In press)

- III. **Emilio Ugalde-Morales**, Felix Grassmann, Keith Humphreys, Jingmei Li, Nicholas P. Tobin, Jonas Bergh, Åke Borg, Linda Sofie Lindström, Johan Vallon-Christersson, Per Hall, Kamila Czene
Interval breast cancer gene expression profile is associated with immune subtypes.
(Manuscript)

- IV. **Emilio Ugalde-Morales**, Jingmei Li, Keith Humphreys, Jonas F. Ludvigsson, Haomin Yang, Per Hall, Kamila Czene
Common shared genetic variation behind decreased risk of breast cancer in celiac disease.
Scientific Reports. 2017; 7: 5942.

TABLE OF CONTENTS

1	Background.....	7
1.1	Breast cancer aggressiveness	8
1.1.1	Tumor characteristics	8
1.1.2	Intrinsic subtypes.....	9
1.1.3	Interval cancers.....	10
1.2	Breast cancer risk factors	11
1.2.1	Non-genetic risk	11
1.2.2	Genetic risk.....	12
1.3	Molecular aspects of breast cancer biology.....	16
1.3.1	Carcinogenesis	16
1.3.2	Cancer hallmarks.....	16
1.3.3	Heterogeneity and evolution.....	18
1.3.4	Immunogenicity	18
2	Aims.....	19
3	MATERIALS AND METHODS	21
3.1	Underlying study populations	21
3.1.1	LIBRO-1	21
3.1.2	KARMA	21
3.1.3	BCAC	21
3.1.4	Ethical approvals	22
3.2	Data material.....	22
3.2.1	Tumor characteristics, treatment, and survival	22
3.2.2	Risk factors	23
3.2.3	Interval cancer and mammographic density.....	23
3.2.4	Genetic data	23
3.2.5	Gene expression data: tumor RNA sequencing	25
3.3	Summary variables	26
3.3.1	Protein-truncating variants (Paper I)	26
3.3.2	Polygenic risk score (Paper I and IV).....	26
3.3.3	Tyrer-Cuzick risk score (Paper II).....	26
3.3.4	Gene expression profiles (Paper II and III)	27
3.3.5	Molecular subtypes (Paper II and III).....	27
3.3.6	Immune subtypes (Paper III)	27
3.4	Study designs	28
3.4.1	Case-only study (Paper I-III)	28
3.4.2	Case-control study (Paper IV)	28
3.4.3	Genetic correlation and overlap (Paper IV)	28
3.5	Statistical methods.....	29
3.5.1	Logistic regression	29
3.5.2	Multinomial logistic regression	29
3.5.3	Cox Proportional-Hazards regression.....	29

3.5.4	Gene expression analysis	30
3.5.5	Gene set enrichment analysis (GSEA)	30
3.5.6	Cross-trait LD Score (LDSC) regression	31
3.5.7	SNP Effect Concordance Analysis (SECA).....	31
3.5.8	Confounding.....	32
4	MAIN RESULTS AND INTERPRETATIONS.....	33
5	Concluding remarks	41
6	Future perspectives.....	43
	Acknowledgements	45
	References.....	47

LIST OF ABBREVIATIONS

95% CI	95% confidence intervals
BCAC	Breast Cancer Association Consortium
LDSC	Cross-trait LD Score
ER	Estrogen receptor
GSEA	Gene Set Enrichment Analysis
GWAS	Genome Wide Association Analysis
HER2	Human epidermal growth factor receptor 2
IHC	Immunohistochemical
KARMA	KARolinska MAMmography Project for Risk Prediction of Breast Cancer
LD	Linkage disequilibrium
LIBRO-1	Linné-Bröst 1 study
Ki-67	marker of proliferation Ki-67
MSigDB	Molecular Signature Database
NGS	Next Generation Sequencing
OR	Odds ratio
PAM50	PAM50, breast cancer molecular subtypes
PD	Percent mammographic density
Per 1-SD	per-one standard deviation
PRS	Polygenic risk score
PR	Progesterone Receptor
PTVs	Protein-truncating variants
RR	Relative risk
SECA	SNP Effect Concordance Analysis
TCGA	The Cancer Genome Atlas
TILs	Tumor infiltrating lymphocytes
TC	Tyrer-Cuzick

1 Background

Breast cancer is the most commonly occurring malignancy among women, and its incidence is increasing.[1] In 2018, over two million new cases and more than 600,000 deaths were estimated worldwide.[2] Global differences in incidence and mortality are largely explained by age and country-level income.[3] The increasing trend in incidence is mainly attributed to reproductive and lifestyle patterns such as older age at first birth, decrease in childbearing and breastfeeding, lower physical activity, and obesity.[4, 5] Breast cancer survival has improved over the last decades as the result of the development of adjuvant and targeted therapies,[6] and introduction of mammographic screening,[7] primarily in more developed countries.[8] Nevertheless, strong differences in survival and other clinical outcomes are observed between groups of patients,[9] particularly in women diagnosed with triple-negative breast cancer for whom optimal therapies are lacking.[10]

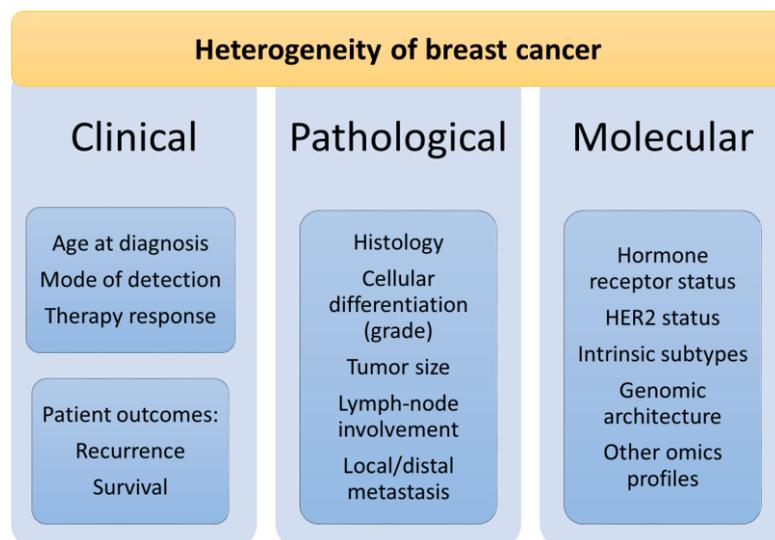


Figure 1. Breast cancer heterogeneity. Clinical, pathological, and molecular features depict the heterogeneous nature of breast cancers.

From both a biological and a clinical perspective, breast cancer is considered to be a heterogeneous disease that can be characterized by a number of clinical, pathological, and molecular features (**Figure 1**). Generally, breast carcinomas are divided into ductal or lobular, according to their localization, and as in situ or invasive. The most common type of breast carcinomas are invasive ductal carcinomas (IDC) that account for more than 50% of invasive cases, followed by invasive lobular carcinoma (ILC), observed in 5% to 15% of patients.[11, 12] The main clinico-pathological features used to describe invasive breast carcinomas are tumor grade (undifferentiated vs well-differentiated), tumor stage (larger tumor size and number of lymph node involved), and cellular receptor status (negative vs positive). Based on these features, breast cancers can be categorized according to the WHO recommendations.[13,

14] The current state of knowledge supports that breast cancers are a mixture of multiple and diverse dynamic entities, from which a phenotype emerges based on the interplay between intrinsic tumor characteristics and host factors, imposing great challenges for diagnosis and treatment.[15]

Because of the observed heterogeneity in clinical outcomes, tumors more likely to display aggressive features and to have poor prognosis require special attention. Therefore, patient stratification is important for adequate clinical management, i.e. treatment decision and patient care.[16, 17] Also, an accurate tumor classification can allow epidemiological and functional studies to unravel mechanisms of carcinogenesis and disease progression,[18] which can prove useful for design of early intervention studies towards prevention. In this thesis, we present results in the field of molecular epidemiology of invasive breast cancer, and discuss how our findings contribute towards better a understanding of the relationship between risk factors and disease aggressiveness.

1.1 BREAST CANCER AGGRESSIVENESS

Aggressive subtypes of breast cancer can be described as tumors with higher capacity for proliferation, invasiveness, and metastasis, leading to poorer prognosis and ultimately, higher mortality rates. An imperative task in breast cancer research is to identify factors associated with poorer outcomes, in order to effectively reduce breast cancer burden. The following subsections describe subsets of invasive breast cancers defined according to clinico-pathological information (tumor characteristics), gene expression profiling (classification into intrinsic subtypes), and the mode of detection in the context of mammographic screening in Sweden (interval cancers as compared to screen-detected).

1.1.1 Tumor characteristics

During the recent decades, large efforts have been made to identify markers that can facilitate to predict prognosis (clinical outcomes such as recurrence and death) and therapy response. Breast cancer tumor characteristics are the most broadly studied prognosticators and represent a relevant measure of aggressive disease.[16] Stage, tumor size, lymph-node involvement, and histological grade, are accepted as prognostic markers, where tumors of larger than 20 mm, node-involvement, and distant metastasis (stage IV tumors) exhibit poorer prognosis.[19] Based on molecular targets such as hormone nuclear receptors of estrogen (ER) and progesterone (PR), and the human epidermal growth factor receptor 2 (HER2), which are commonly measured by immunohistochemistry lab techniques, tumors can be classified into hormone positive (HR+) when expression of either ER or PR is detected,[20] as HER2-positive, or as triple-negative when lacking expression of either marker. HR+ status has been associated with lower risks of mortality independently of demographic and other tumor characteristics.[21, 22] In contrary, 15 to 25 % of breast cancers overexpressing HER2, a transmembrane tyrosine kinase receptor involve in cell growth,[23] are associated with poorer survival,[24] whereas 10% to 20% of cases classified as triple-negative tumors are associated with worse prognosis.[25]

1.1.2 Intrinsic subtypes

The development of high-throughput molecular technologies has allowed for the characterization of biological samples at very high resolution. In 2000, Perou and colleagues introduced the concept of ‘molecular intrinsic’ breast cancer subtypes based on the idea that the observed phenotypic diversity could be described by distinct gene expression patterns.[26] By comparing expression profiles from 22 paired sample specimens, the authors could identify an ‘intrinsic’ subset comprised of 496 genes for which variation across samples was larger than the variation within pairs of samples from the same tumors. Using hierarchical clustering, samples could be separated into two main groups distinctive on their ER receptor status, but with a considerable amount of residual variation within each group, indicating the existence of additional breast cancer subtypes.

Following this principle, classification tools with clinical relevance have been developed. A so called PAM50 classifier[27] was found to predict breast cancer subtypes based on a fifty gene expression signature into five previously reported breast cancer subtypes: luminal A, luminal B, HER2-enriched, basal-like, and normal-like. Analysis of more than 500 tumors concluded that the multiple levels of biological variation could be captured by four main PAM50 subtypes and explain a fair amount phenotypic heterogeneity.[28] The PAM50 subtypes have also been shown to be robust in spite of intra-tumor heterogeneity,[29] in line with the original proposal by Perou and colleagues. More importantly, PAM50 subtypes were clinically validated by predicting significant differences in patient survival independently of clinical predictors (i.e. tumor characteristics), with HER2- and basal-like the tumors with poorer outcome.[30] In addition, PAM50 subtypes were shown to provide more clinically relevant information than histopathological parameters and with the potential to improve treatment strategies. **Table 1** shows a summary of the relation between the main intrinsic subtypes and the most common immunohistochemical (IHC) markers and clinical outcome.

Table 1. Overall description of breast cancer intrinsic subtypes by IHC markers and clinical outcome.

Intrinsic subtype	Main IHC markers	Clinical outcome (survival)
Luminal A	ER+, PR+, HER2-, low proliferation	Good
Luminal B	ER+, PR+, HER2+/-, high proliferation	Intermediate-poor
HER2-enriched	ER-, PR-, HER2+	Poor
Basal-like	ER-, PR-, HER2-, basal marker+	Poor

1.1.3 Interval cancers

Interval cancer is defined as breast cancers diagnosed after a negative mammographic screening and before the next programmed screening.[31] Upon blinded re-review of mammogram images, interval cancers can be classified into ‘true’ cases with no signs of malignant lesions, or ‘missing’ cases where re-examination reveals abnormal signs.[32] Based on epidemiological surveillance in mammographic screening programs, ‘true’ interval cancers have been observed to occur in 14.7% of cases at annual screening intervals, 17-30% at biennial, and 32-38% in triennial programs, and about 20-25% of cases were missed at screening.[33]

It has been proposed that ‘true’ interval cancers correspond to fast growing tumors and therefore are enriched in aggressive breast cancer subtypes.[34-36] When compared with screen-detected tumors, interval cancers are more likely to have larger size, lymph node involvement, higher grade, to be triple-negative or HER2-positive, hormone receptor-negative, and are associated with poorer survival.[32, 33] Molecular characterization of tumors using sequencing technologies showed that interval cancers are associated with molecular intrinsic subtypes of poor survival independently of mammographic density.[37] In that same study, interval cancers were found to be enriched in luminal B and basal-like tumors. Moreover, association with higher mutational load in *TP53*, *PPP1R3A*, and *KMT2B* cancer-related genes as well as differences in somatic copy number aberrations were found, suggesting that key biological features drive aggressiveness of interval cancers.

Because mammographic dense tissue affects screening specificity, i.e. increases false-negative cases (a phenomenon referred to as ‘masking’), interval cancers in women with low mammographic density are more likely to be enriched in ‘true’ interval cancers (**Figure 2**). Previous studies in our group found pronounced differences when comparing invasive tumors with low mammographic density ($\leq 20\%$) on interval cancers versus screen-detected regarding lymph node involvement, ER-negative status, HER2-positive, progesterone receptor-negative and triple-negative.[38]

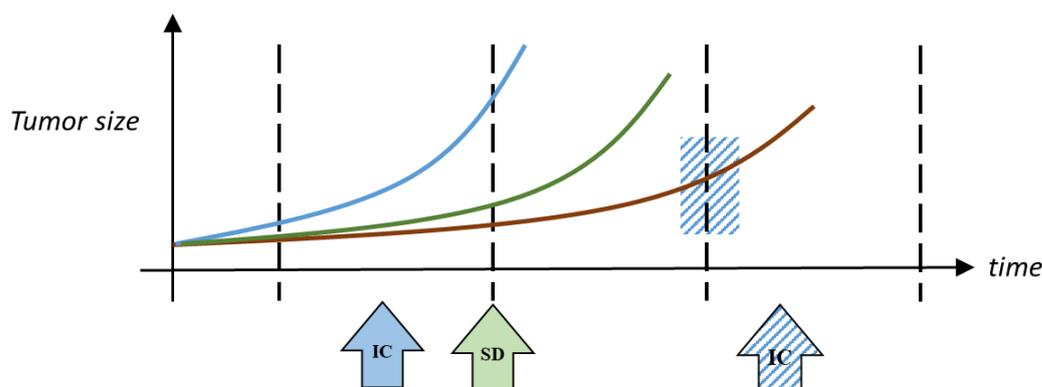


Figure 2. Graphical description of breast cancer mode of detection. Arrows indicate a breast cancer diagnosis within a mammographic screening setting, and shaded area represents high mammographic density. Tumors can be missed at screening due to a “masking effect” in high dense breasts. IC, interval cancer. SD, screen-detected breast cancer.

1.2 BREAST CANCER RISK FACTORS

Breast cancer is a complex disease involving hereditary (genetic) and environmental (non-genetic) risk factors. An essential task in cancer epidemiology is to identify and quantify the contribution of these risk factors on the disease development.

1.2.1 Non-genetic risk

Established risk factors for breast cancer are related to age, estrogen exposure, reproductive history, and mammographic density. The relationship between the main non-genetic exposures and breast cancer risk is shown in **Table 2**. Mammographic density has been discovered to be a strong and independent risk factor for breast cancer[39], and seem to be independent of molecular subtypes.[40, 41] Family history of breast cancer is an important risk factor reflecting the complex interaction between genetic and environmental factors involved in breast cancer etiology. It is defined as having first-degree (e.g. mother, sister, or daughter) or second-degree relatives that have been diagnosed with breast cancer. Family history is associated with an intermediate to high risk independently of mammographic density.[42] The risk is about doubled in first-degree than second-degree family history, and the risk is higher risk when both mother and sister have been affected.[43]

Table 2. Main breast cancer risk by non-genetic risk factors.

Risk factor	Exposure	Effect	Aggressive subtype
Sex	Female	↑↑↑↑	
Age	Older (> 40 or >60 years old)	↑↑↑	
Family history	Yes vs No	↑↑↑	<i>BRCA</i> , basal-like
Mammographic density	High vs Low	↑↑↑	
Benign breast disease	Yes vs No	↑-↑↑↑	
Age at menarche	At age < 12 years old	↑	
Age at menopause	At age > 55 years old	↑	
Parity	Yes vs No	↓	<i>BRCA</i> , basal-like
Age at first birth	Older age (> 35 years old)	↑	
Breast feeding	No vs yes, (e.g. < 1 year)	↑↑	Basal-like
Postmenopausal obesity	Body mass index > 30 Kg/m ²	↑	
Oral contraceptive use	Ever vs Never	↑	
HRT	Ever vs Never	↑↑	No (Luminal A)
Oophorectomy	Yes vs No	↓↓↓	

Arrows represent the direction and strength of association: ↑↑↑↑ very strong, ↑↑↑ strong, ↑↑ intermediate, ↑ low.

Hormonal exposure, mainly of estrogen, has a pivotal role in the risk to develop breast cancer.[44] Other conventional risk factors that most presumably act by modifying hormone exposure are: female sex and older age with the largest risk estimates, followed by low to

intermediate relative risks associated with age at natural menopause, age at menarche, age at first birth, breast feeding, parity, postmenopausal obesity, oophorectomy, and exogenous estrogen exposures such as oral-contraceptive use and hormone/estrogen-replacement therapy.[43]

As breast cancer is a heterogeneous disease, efforts are made to identify specific risk factors, particularly on aggressive subtypes.[45] Full-term pregnancy and breastfeeding were the most important protective factors in hereditary (*BRCA* carriers) breast cancer.[46] A study from our group assessing the heterogeneity of different risk factors including reproductive and genetic factors, also found breastfeeding to be protective factor mainly for basal-like tumors, which were enriched in *BRCA* mutations when compared with luminal A tumors; ever use of hormone replacement therapy was differentially associated with increased risk of luminal A tumors.[47] Regarding mammographic density, the risk for breast cancer does not seem to differ by ER or HER2 status.[40]

1.2.2 Genetic risk

A genetic (inherited) component in breast cancer is well established in the etiology of breast cancer.[48] Estimation of heritability based on twin studies found 25 to 31 percent to be explained by genetic factors.[49-51] Currently, high-risk women are primarily identified on the basis of family history and mutation screening of the *BRCA1*[52] and *BRCA2*[53] genes located on chromosome 17 and 13, respectively, which convey a lifetime risk between 50 to 85% and account for approximately 15% of familial breast cancer.[54]

A large effort to investigate the genetic component of breast cancer has taken place since the discovery of the *BRCA1/2* genes.[55] Genetic risk variants can be classified into high, moderate, and low risk, based on their penetrance expressed in relative risks (RR) as: 1) high risk if $RR > 4$, 2) moderate risk if RR between 2 to 4, and 3) low risk if $RR < 1.5$. [56] Base on their allele frequency, genetic variants are classified into common ($>1\%$) or rare ($<1\%$). Put together, it has been observed that highly pathogenic variants are rare and that susceptibility variants with lower risk are more frequent (**Figure 3**). Moderate-to-high risk variants have been mainly identified in familial breast cancer cases through genetic linkage studies followed by positional cloning, and candidate gene-panel sequencing of unrelated individuals which search for protein-coding deleterious variants found at a frequency $<1\%$ in general population.[57] On the other hand, common variants have been identified through Genome Wide Association Analysis (GWAS). While rare variants tend to be of higher penetrance, common variants are often of low-penetrance.

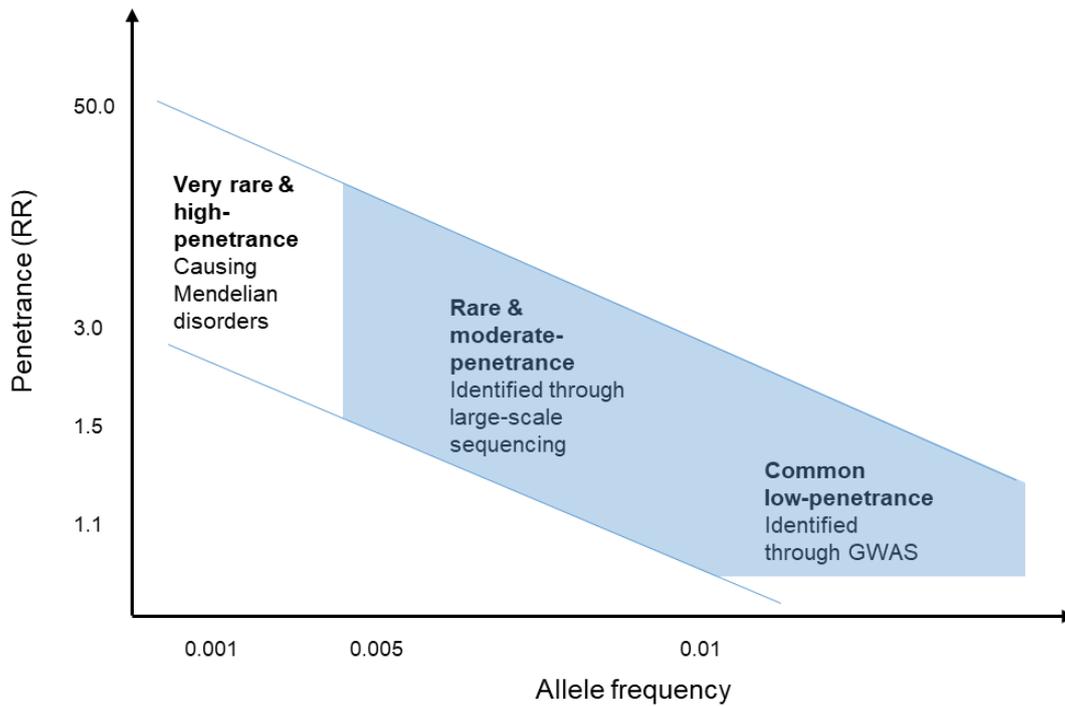


Figure 3. Penetrance and frequency of breast cancer genetic variants. Variants within the shadowed area, thought to explain breast cancer heritability. Variants below the lower boundary (rare and low-penetrance) are hard to be detected and of little utility, while variants above the upper boundary (common and high-penetrance) are subjected to strong negative-selection, thus difficult to be observed. Inspired by Manolio, T.A., et al.[58] RR, relative risks.

Common variants

Initiated with the discovery of common predisposition variants by the first breast cancer GWAS in 2007,[59] low-risk variants have been identified in large cohorts under the hypothesis of a ‘common-disease common-variants’.[57] A series of successful studies with continuously increasing number of participants, allow for pooled analysis through international collaborations.[60] The largest GWAS in European population up to date analyzed 118,474 breast cancer cases and 96,201 controls, as well as *BRCA1* mutations carriers, 9,414 affected and 9,494 unaffected.[61] The study included participants from 82 studies from the Breast Cancer Association Consortium (BCAC), and from 60 studies from the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA). The authors reported 32 new loci were identified in addition to the 178 loci reported in previous GWAS from the BCAC.[62, 63] The 210 variants were found to explain 54.2, 37.6 and 26.9% of the genome-wide chip heritability for luminal-A-like, triple-negative, and *BRCA1* carriers, respectively, and about 18% of the familiar risk for invasive breast cancer.[61] Analysis on the genetic correlation between breast cancer subtypes showed that luminal A-like breast cancer is less correlated with triple-negative and *BRCA1* subtypes (0.46 and 0.39, respectively), while highest correlation was observed for the *BRCA1* subtype with triple-negative and HER2-enriched-like subtypes (0.84 and 0.80, respectively) (**Figure 4**).

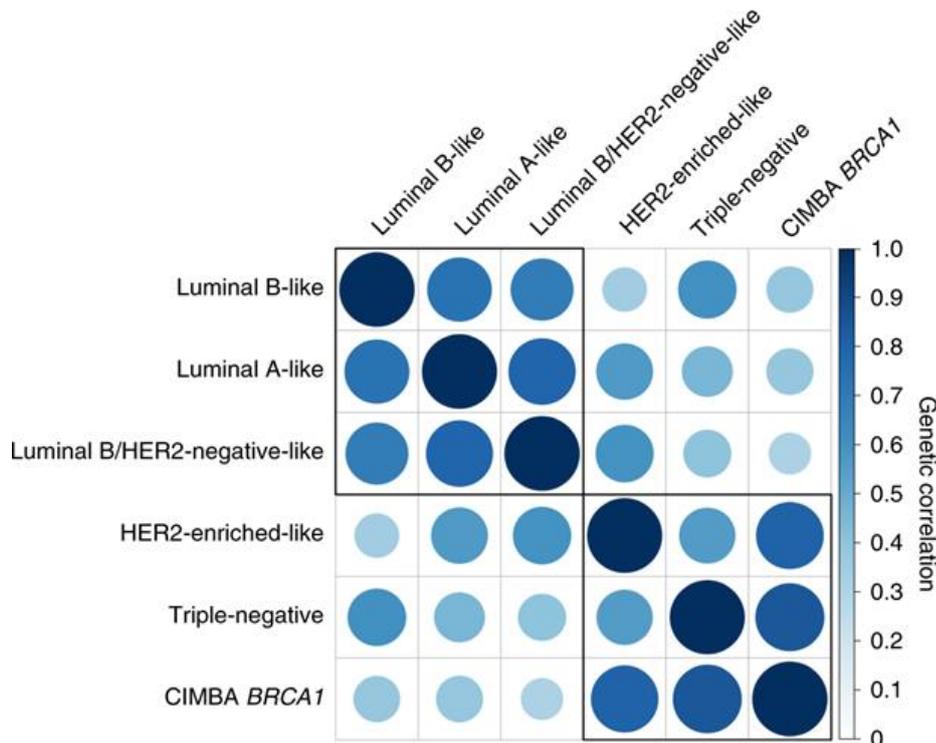


Figure 4. Genetic correlation between breast cancer intrinsic-like subtypes, estimated through LDSC regression. Reprinted by permission from Springer Nature. Zhang et al 2020.[61]

When combined into polygenic risk scores (PRSs),[64] common variants are able to explained a larger proportion of phenotypic variation and have proven useful for risk stratification.[65, 66] Recently, a PRS including 330 single nucleotide polymorphisms (SNPs) was found to confer 83% to 65% higher risk for luminal-A-like and triple-negative subtypes, respectively.[61] These results are similar to a previous PRS based on 313 SNPs that was associated with 61% higher risk of overall breast cancer, 4.37-fold risk of ER-positive, and 2.78-fold risk of ER-negative breast cancer on women in the highest centile as compared with women in the middle PRS quintile.[67]

Rare variants

It has been proposed that the “missing heritability” observed in GWAS studies could be potentially explained by rare variants with moderate-to-high risk effects that require the systematic characterization of a large number of samples.[68] The advent of Next Generation Sequencing (NGS) technologies, that allow for reading entire coding regions of DNA (exon-sequencing), or whole genomes (whole-genome sequencing), has enable the discovery of novel breast cancer germline pathogenic variants which could not have been identified through family studies.[69-72] Multigene (sequencing) panels are proving to be useful in identifying

breast cancer susceptibility genetic variants involved in DNA repair (similar to the *BRCA* genes), cell-cycle control or mitotic signal transduction pathways.[73]. Carriers of high- and moderate-risk germline mutations in genes such as *BRCA1*, *BRCA2*, *CHEK2* and *PALB2*, have been found to be predisposed to specific subtypes of breast cancer.[74-77] In particular, mutated *BRCA1* gene is highly enriched for basal-like tumors,[78, 79] and *ATM* and *CHEK2* have been observed to be associated with higher risk of ER-negative disease.[80] In addition to the *BRCA* genes, germline mutations in *PALB2*, *RAD51D*, and *BARD1* have been found to be associated with triple-negative breast cancer.[81] Combined, mutations in high-to-moderate risk genes account for 9% to 14% of triple-negative cases and were associated with more aggressive phenotypes, as found through gene-panel sequencing of patients unselected for family history of breast cancer.[82, 83]

While sequencing studies allow for identification of rare deleterious variants, GWAS variants lie most presumably on gene regulatory elements.[57] Interestingly, sequencing of exon-intron boundaries of 56 genes identified through GWAS studies, only found weak evidence of rare deleterious variants being associated with breast cancer risk in non-*BRCA* families.[84] A similar conclusion was drawn from a large study sequencing 38 genes neighboring 38 leading GWAS SNPs.[85] However, this does not discard the possibility of finding regulatory variants conferring high risk effects on GWAS studies, and does not mean that all variants in protein-coding regions are of high penetrance.

Immune-related genetic factors

Immune and inflammatory responses play a key role in the different stages of cancer disease.[86] It is possible that immune-related genetic variants affect breast cancer susceptibility by influencing immunosurveillance mechanisms and could potentially provide prognostic information, particularly on tumor subtypes with higher immunogenicity. For instance, studies on candidate genetic predisposition variants have found immune-related genes to be associated with ER-negative breast cancer.[87, 88]

Large-scale genotyping initiatives have been undertaken in order to characterize the genetic architecture underlying the phenotypic variation of immune traits[89, 90] and the susceptibility to develop autoimmune diseases in which an altered immune response is exhibited.[91] These data sources represent an opportunity to explore the role of immune-related factors on breast cancer in order to identify common etiological and prognostic factors. Autoimmune diseases are known to be associated with breast cancer based on epidemiological data.[92] Celiac disease in particular, a gastrointestinal immune-mediate disease triggered by gluten intake, has been associated with reduced risk of breast cancer.[93, 94]. Base on this idea, overall trends or correlations between traits based on genomic variant information can be used to guide the search for shared etiological factors, also referred as pleiotropy.[95] For that, different methodologies have been proposed to exploit the ‘hidden’ information that can be capture from GWAS studies.[96-98]

1.3 MOLECULAR ASPECTS OF BREAST CANCER BIOLOGY

This section aims to describe key concepts in cancer biology that provide a theoretical framework for the understanding of the observed molecular features in cancer, and with particular remarks on breast cancer. In this context, mutations are thought as the main drivers of cancer cells (carcinogenesis), where biological features emerge (cancer hallmarks) in a complex and dynamic fashion (heterogeneity and evolution) (**Figure 5**). A separate section on immunogenicity, that in a sense provides a link between the former concepts, is also briefly described.

1.3.1 Carcinogenesis

Carcinogenesis, also known as tumorigenesis or oncogenesis, refers to the processes by which genomic alterations, acquired and/or inherited, lead to the formation of cancer cells.[99] Mutated genes driving this process are broadly classified into two categories according to their biological function: tumor-suppressor genes, which act as “guardians of the genome”, and oncogenes or proto-oncogenes. In breast cancer, a number of oncogenes (e.g. *ErbB2*, *PI3KCA*, *MYC*, and *CCND1*) and tumor-suppressor genes (*BRCA1*, *BRCA2*, *PTEN*, *CHK2*, *NBS1*, *RAD50*, *PALB2*, *BRIP*) have been identified.[100, 101]

The prevailing theory of cancer proposes that for the formation of cancer cells, a stepwise acquisition of cancer-favoring mutations is required for the clonal evolution of cancer cells.[102-104] The nature of these mutations can be inheritance, DNA-damaging environmental factors, and consequence of errors in DNA replication.[105] Causal mutations are referred as driver mutations, to distinguish them from passenger or neutral mutations.[106, 107] In a recent pan-cancer analysis 299 driver genes were identified, of which about 10% were found in more than half of the cancer types, while more than 50% of genes were unique to one subtype.[108] Still, the role of passenger mutations is debated,[109] as they may provide evolutionary advantages to intermediate cancer cell phenotypes.[110] Generally, mutational signatures provide valuable information on cancer etiology, prognosis, and potential therapeutic targets.[111] For instance, mutational load across cancer genomes was used to identify diagnostic and prognostic gene expression signals,[112] and to predict positive response to immunotherapy in different cancer types.[113]

1.3.2 Cancer hallmarks

By acquisition of key biological properties, called ‘cancer hallmarks’,[114, 115] abnormal cells are able to become tumorigenic and invasive. Cancer hallmarks include: sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming of energy metabolism, and evading immune destruction, all of which are underlined by two enabling characteristics: genome instability, and tumor-promoting inflammation. Tumors can also be conceptualized as tissues composed of multiple cell types that interact with the tumor microenvironment through signaling processes.

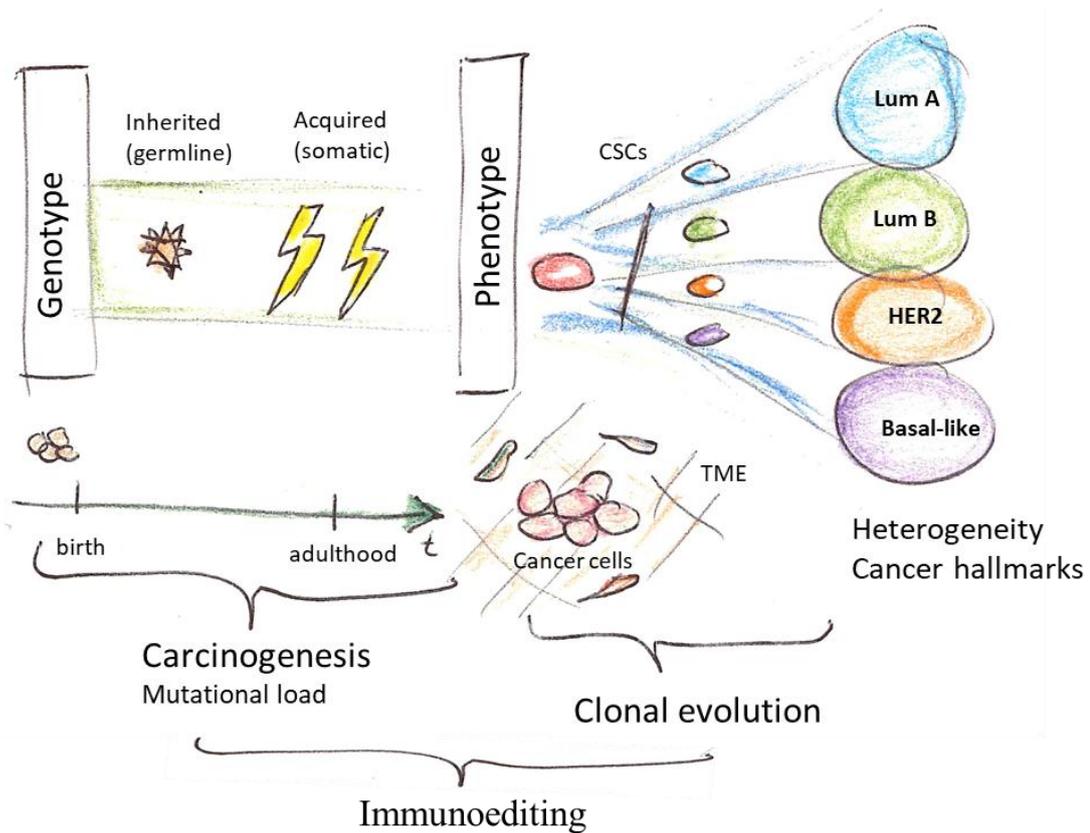


Figure 5. Molecular aspects of breast cancer biology. Mammary cells portray a unique mutational landscape (genotype); somatic mutations can trigger transformation into cancer cells, which evolve into a distinct breast cancer subtype through clonal evolution; Immunoediting events occur along this process. CSCs, cancer-stem cells; TME, tumor microenvironment.

In the specific case of breast cancer, these principles can be used to describe breast cancer heterogeneity in a more coherent way by assigning tumor subtypes onto cancer hallmarks.[116] For instance, ‘sustaining proliferative signaling’ is proposed to be the principal mechanism driving tumors with hormonal (e.g. ER | PR positive, luminal A subtype) and growth receptor positivity (tumor with HER2+). Over-expressed proliferation markers such as TOP2A, Ki-67, and cell cycle genes can further differentiate [ER+ | PR+, HER2-] tumors into more aggressive subtypes. ‘Activating invasion and metastasis’ hallmark is more characteristic in tumors with poorer prognosis such as triple negative [ER-, PR- and HER2-] and basal like subtype, where basal markers such as cytokeratins are linked with tumorigenesis and metastasis. Properties related to the same hallmark, such as epithelial to mesenchymal transition processes and cell stemness, are enriched in triple negative tumors. Other breast cancer subtype that can be distinguished in relation with ‘Evading immune destruction’ cancer hallmark, is the so called interferon-rich tumors (accounting for ~10% of cases), a subset of triple-negative cancers with intermediate survival outcome.

1.3.3 Heterogeneity and evolution

Biological features of cancer, such as intratumoral heterogeneity in breast carcinomas,[117] can be described from a developmental perspective.[118, 119] Under this paradigm, cancer cell populations (clones and sub-clones) arise from carcinogenic and tumorigenic events in an evolution-like process, which can explain fundamental differences between breast cancer subtypes at the single-cell level.[120] For instance, sub-clonal mutations identified in the *AKT*, *FGFR*, *PIK3CA*, and *TP53* genes through multi-region NGS sequencing of breast cancer primary tumors, were found to explain aggressive phenotypes such as increased proliferation, chemoresistance, invasiveness, and metastasis.[121] Another theory to explain tumor heterogeneity forwarded in 1977, is the existence of different cells-of-origin,[122] referred as cancer stem cells (CSCs). Other models integrating clonal evolution and CSCs, to explain intratumoral heterogeneity, have also been proposed.[123]

1.3.4 Immunogenicity

Clonal selection is influenced by the host immune system in a process called immunoediting, which can be differentiated into three phases: elimination, equilibrium, and scape.[124] Elimination refers to the processes by which the innate effector cells and adaptive immune system lead newly formed cancer cell towards apoptosis. The equilibrium phase is characterized by the acquisition of molecular modifications which allow malignant cells to avoid immune system detection and elimination. In the scape phase, tumor cells are capable of inducing an immunosuppressive microenvironment allowing further proliferation.[125]

Immunogenicity of the tumor reflects the extent of immune response involvement and can be described by the presence of tumor infiltrating lymphocytes (TIL). For breast cancer in particular, TILs have been observed predominantly on triple-negative tumors, i.e. more than 50% lymphocytic infiltrate, which correlates with better prognosis from each 10 percent increase in TIL; other breast cancer subtypes exhibit lower TIL levels and could benefit from TIL enhancing therapies.[126] Immunogenicity is determined by the ability of immune cells to recognize tumor-specific epitopes defining the antigenicity of a tumor. Higher mutational load is positively correlated with formation of tumor neoantigens, making them more immunogenic, and is usually higher on ER-negative tumors. Immunogenicity can also be enhanced by specific mutational signatures affecting DNA repair mechanism, as it is for *BRCA* genes linked to basal-like tumors. Tumors with lower proliferation and genomic instability such as luminal subtype may result in lower antigenicity. However, heterogeneity in ER-positive tumors could be influenced by high genomic driver diversity favoring immune-scape mechanisms.[124]

2 Aims

The overall aim of this thesis was to generate new knowledge on breast cancer epidemiology by leveraging of molecular data, with a particular focus on disease aggressiveness. The specific aims, corresponding to the constituent scientific papers included in this thesis, are the following:

- I. To evaluate the contribution of rare and common germline genetic variants on disease aggressiveness.
- II. To characterize tumor gene expression patterns behind the association between breast cancer risk, defined by the 5-year Tyrer-Cuzick risk score, and disease aggressiveness.
- III. To investigate underlying biological features of interval breast cancers independent of the PAM50 intrinsic subtypes.
- IV. To assess the genetic correlation and overlap between breast cancer and celiac disease.

3 MATERIALS AND METHODS

In this thesis, we included women who participated in two breast cancer studies in Sweden designed to investigate risk factors, tumor characteristics, and clinical outcomes, among other aspects of breast cancer. After ethical approval and participant informed consent, detail level information was collected from questionnaires, medical records, and from linkage to high quality registers. Additionally, data from external studies in Swedish and European populations was used in some of the analysis. The following sections describe the study populations, data material and summary variables, study designs, and statistical methods used across the four papers.

3.1 UNDERLYING STUDY POPULATIONS

In all our studies, we included women recruited under the Linné-Bröst 1 (LIBRO-1) study. In papers II, III, and IV, we included women who participated in the KARolinska MAMmography Project for Risk Prediction of Breast Cancer (KARMA) study. In paper I and IV, some analyses were based on data obtained from The Breast Cancer Association Consortium (BCAC). Gene expression validation datasets obtained from two independent breast cancer cohorts, The Cancer Genome Atlas (TCGA), and the MERCK, are also briefly described.

3.1.1 LIBRO-1

The LIBRO-1 study has been described in previous publications.[38, 127] Briefly, it consists of women diagnosed with invasive breast cancer between January of 2001 and December of 2008 in the Stockholm/Scotland regions of Sweden, who were alive in 2009. In total, more than 9,000 women were identified through the Regional Cancer register and invited, of which 5,715 accepted to participate in the study.

3.1.2 KARMA

KARMA is a large and well characterized prospective breast cancer cohort.[128] It is derived from population-based mammographic-screening or clinical radiology examinations conducted at five participating hospitals from the Stockholm and Skåne regions of Sweden (Stockholm South General Hospital, Helsingborg Hospital, Skåne University Hospital, Lund Hospital, and Landskrona Hospital). Between January 2011 and March 2013, more than 210,000 women were invited to participate. In total, more than 70,000 women with or without breast cancer diagnosis were included in study, of which approximately 3,000 have been diagnosed with invasive breast cancer.

3.1.3 BCAC

BCAC is the largest international initiative to characterize the genetic susceptibility of breast cancer (<http://bcac.ccge.medschl.cam.ac.uk/>). In our studies, we used GWAS summary results reported for overall, ER-positive, and ER-negative breast cancer risk, based on European population, here referred as *GWAS summary statistics*, which are further discussed in section 3.2.4.3 Briefly, in Paper IV we used summary results published in 2015,[129] in which over

85,000 women of European ancestry (45,290 cases) were genotyped under the Collaborative Oncological Gene-Environment Study (COGS), using an Illumina iSelect SNP Array that covered 211,155 SNPs, the iCOGS.[130] In Paper I, we used GWAS summary results published in 2017,[62] where over 100,000 women of European ancestry (61,282 cases) were genotyped using the iCOGS, or the OncoArray, an Illumina SNP array targeting more than 500,000 genomic variants.[131]

3.1.4 Ethical approvals

All women participating in the LIBRO-1 and KARMA studies gave written informed consent to extract data from medical records and national registers, provided information on risk factors, and donated a blood sample for genetic analysis. The studies were approved by the Regional Ethical Review Board at Karolinska Institutet (LIBRO-1, DNR: 2009/254-31/4, amendments 2011/2010-32, and 2012/465-32; KARMA, DNR: 2010/958-31/1, amendment 2013/2090-32), and were conducted in accordance with the Declaration of Helsinki. In brief, all personal data was pseudonymised by the Swedish National Board of Health and Welfare (in Swedish, Socialstyrelsen) and analyzed in secure local servers at the Department of Medical Epidemiology and Biostatistics, following data management guidelines.

3.2 DATA MATERIAL

Data material consisted of the main outcomes and exposures are described in this section. The main outcomes included tumor characteristics (all papers), breast cancer specific survival (Papers I and II), and interval cancers (Paper III). The main exposure variables were based on breast cancer risk factors (Paper II), genetic (Papers I and IV), and gene expression data (Papers II and III).

3.2.1 Tumor characteristics, treatment, and survival

Data on molecular markers was retrieved from medical and pathology records. ER and PR percentage staining was determined using radioimmunoassay or IHC techniques and dichotomized into positive (if $\geq 10\%$) or negative status, otherwise. HER2 status was dichotomized as negative status if protein expression from IHC/immunocytochemistry was 0 or 1+, or higher, and no gene amplification by FISH, and assigned positive status if gene amplification by FISH was observed. Proliferation marker Ki67 was measured in hotspot regions following routine guidelines and was reported as low if percent staining $< 20\%$, or as high otherwise. Information on prior breast cancer diagnoses, lymph node involvement (dichotomized into positive or negative), tumor size diameter measured in millimeters, and tumor grade recorded using the Nottingham Histologic Grade system, as well as treatment regimen (adjuvant chemotherapy, endocrine therapy, and radiotherapy) was obtained through the Swedish National Cancer Register (INCA)[132] and the Stockholm-Gotland Regional Breast Cancer Quality Register[133] through the Swedish personal identity numbers.[134] Information on date and breast cancer-specific cause of death (code “C50*”) was obtained from the Swedish Cause of Death Register.[135]

3.2.2 Risk factors

Information on various aspects of women's health was extracted from questionnaire data provided by each participant in the LIBRO-1 and KARMA study, at time of entry. Data on reproductive history, family history of breast cancer, exposure to exogenous estrogen (i.e. use of oral contraceptives, and hormone-replacement therapy) was obtained. This information was used to estimate absolute breast cancer risks (Tyrrer-Cuzick score) in Paper III, as described below in section 3.3.3.

3.2.3 Interval cancer and mammographic density

In this thesis, we assessed mammographic screening history (i.e. dates at mammographic screening visits) to define breast cancer mode of detection (i.e. screen-detected or interval cancer). Screening information regarding data and outcome from each visit was obtained from the population-based mammography screening database[136] at the Stockholm-Gotland Regional Cancer Center. Since 1989, in Stockholm, all women aged 50 to 69 years have been invited to screening at 24-month intervals, and 2005, women aged 40 to 49 years have also been invited to screening at 18-month intervals. Breast cancer diagnosis from women regularly attending mammographic screening, were classified by mode of detection into interval cancers, or screen-detected breast cancer. Interval cancers were defined as breast cancer diagnosis occurring after a negative screening mammogram and before the next programmed screen. Screen-detected tumors were defined as breast cancer diagnosis occurring after a positive screening mammogram. Breast mammographic density, expressed in percentage (PD), was measured from mammograms of healthy breasts prior breast cancer diagnosis using a machine-learning algorithm, STRATUS,[137] developed by our group.

3.2.4 Genetic data

Three types of germline genetic data were analyzed in this thesis: 1) sequencing data, 2) raw genotype data, and 3) GWAS summary statistics. The first type consisted of targeted sequencing of 31 breast cancer related genes to measure carriership of rare deleterious variants (section 3.3.1). The second and third type inform about common genetic variants across the genome that can be summarized into PRSs (section 3.3.2) or used to estimate genetic correlation (section 3.4.3).

3.2.4.1 Sequencing data (Paper I)

Germline DNA sequencing was performed in LIBRO-1 patients using a custom-made gene panel. The gene panel was design to target exome and intro/exon boundary regions of 31 genes breast cancer predisposition genes included in commercial gene panels.[80] DNA variant calling was obtained for 5,099 (99.55%) of 5,122 patients successfully sequenced at the Centre for Cancer Genetic Epidemiology, at University of Cambridge (see eMethods in Data Supplement 2, Paper I). In brief, the panel consisted of an amplicon-based (targeted) custom panel, applied to a NGS platform. The in total 1,350 amplicons and primer sequences used to target the intro/exon boundaries of the 31 genes are shown in Paper I, Table S1 in Data

Supplement 2. Library preparation (enrichment) was performed using the Fluidigm Access Array 48.48 system. Libraries were sequenced on a single lane of an Illumina platform (Hi-Seq200) yielding 100-base paired-end reads. Bioinformatics processing of raw data consisted of reads alignment to a human reference genome (hg19) using the Burrows-Wheeler Aligner (BWA).[138] Variant calling was performed using the Genome Analysis Toolkit (GATK)[139] UnifiedGenotyper pipeline (see Figure S1, in Data Supplement 2, Paper I). A number of quality control and hard filtering criteria were applied to obtain high confidence variant calls for SNP, and INDELS, separately, as shown in Paper I, Supplementary Table 3. Rare variants were defined at <2% frequency. For annotation of variants (i.e. classification of variants into nonsense, frameshift, splicing, missense, etc.), the ANNOVAR[140] software was used.

3.2.4.2 Raw genotype data (Paper I and IV)

In this thesis, we used individual-level genotype data from women who participated in the LIBRO-1 or KARMA study, and that were genotyped as part of the iCOGS initiative.[129] In Paper I, a case-only study, more than 5,000 women diagnosed with breast cancer who participated in the LIBRO-1 study were included. These women were also included in paper IV, in addition to 5,433 women without cancer diagnosis (controls) who participated in the KARMA study. Women subjected to genotyping had donated blood samples at study entry. From these samples, germline DNA was extracted and genotyped using the iCOGS, a custom Illumina iSelect SNP array.[130] Because the array only targets approximately 200,000 independent SNPs, a standard strategy is to impute genotypes at genome-wide coverage based on the principle of linkage disequilibrium (LD).[141] In our studies, imputation was performed using the IMPUTE version 2 software,[142] and a genome reference panel of densely genotyped individuals from the 1000 Genome Project, which contains information for over 88 million common variants.[143]

3.2.4.3 GWAS summary statistics (Paper I and IV)

We used data on common genetic variation of breast cancer and celiac disease measured in large GWAS studies, referred as GWAS summary statistics. Advantages of this type of data is its de-personalized nature, making easier to become publically available and to be used for research purposes.

Breast cancer

Breast cancer GWAS summary statistics were obtained from data published as part of the BCAC consortium, based on the iCOGS and the OncoArray genotyping arrays (see section **3.1.3**).

Celiac disease

GWAS summary statistics for celiac disease (133,352 SNPs) were downloaded from the ImmunoBase (<https://www.immunobase.org/>), a web based resource focused on the genetics

and genomics of immunologically related human diseases. Celiac disease data has been reported in a GWAS study by Trynka and colleagues[144] on 12,041 celiac disease cases and 12,228 controls of European ancestry using the Illumina Infinium High-Density array (ImmunoChip), designed to target 195,806 SNPs located at immune-related genome regions.[145]

3.2.5 Gene expression data: tumor RNA sequencing

3.2.5.1 LIBRO-1/KARMA dataset

Genome-wide expression data was measured for a subset of LIBRO-1 and KARMA patients. This subset of tumors were sequenced under two sequencing initiatives: the ClinSeq,[146] and the SCAN-B.[147] In brief, total RNA was extracted from tumor samples using ribosomal depletion (RiboZero; Illumina, US) in the ClinSeq study, and a poly-A enrichment dUTP library protocol in the SCAN-B. High quality RNA (RIN > 7) was assured. RNA sequencing was performed using the Illumina HiSeq technology. At minimum, more than 5 million paired-end RNA fragments (reads) were obtained on each samples, and less than 60% duplication, meaning that a unique read is mapped twice.

Quantification of gene expression levels was performed using a fast-alignment algorithm (quasi-mapping based mode), Salmon version 0.9.1.[148] For that, an index reference transcriptome (genome assembly version GRCh38) was built using the `--type quasi -k 19` flag. Then, transcript-level estimates were extracted using the `tximportData` R package, and aggregated into gene-level expression values using the `tximport` R package.[149] Because the SCAN-B library protocol was design to target mRNAs, we filtered approximately 19,000 gene-coding mRNAs in both datasets.

3.2.5.2 External datasets

Two external gene expression datasets were used as validation sets. A publically available breast cancer cohort from the TCGA database[152] was used in Papers II and III, and a nested breast cancer cohort of the MERCK study, which has been previously described [127, 150, 151] was used in Paper III. In brief, the TCGA dataset consisted of 975 primary invasive breast cancer tumors, from women of age 26 to 90 years. Pre-processed RNA-sequencing data, in form of transcript count computed with HTseq software,[153] was available for retrieval trough the GDC Data Transfer Tool. Data was downloaded on November 7th, 2018, together with patient clinical information. The MECK study comprised of 621 patients diagnosed with invasive and metastatic breast cancer, from which we identified 111 interval cancer and 109 screen-detected breast cancers. In these samples, gene expression was measured using an Affymetrix microarray assay.

3.3 SUMMARY VARIABLES

Variables that combine information from accumulated knowledge have the potential to improve our ability to evaluate the relationship between multiple factors involved in complex traits such as breast cancer. Particularly important are predictions on aggressive subtypes. In the following subsections summary variables used in this thesis are described. These include breast cancer risk scores (genetic and non-genetic), as well as gene expression patterns related to the disease aggressiveness (gene expression profiles and breast cancer subtypes).

3.3.1 Protein-truncating variants (Paper I)

Protein-truncating variants (PTVs) in 31 breast cancer genes sequenced on a gene panel (section 3.2.4.1), were used to summarize breast cancer genetic load by rare deleterious variants. PTVs were defined as variants disrupting gene function by introducing a stop codon (nonsense mutation), by frameshift insertion/deletions, or through splice site mutations. For *BRCA1/2* genes, PTVs annotations were refined based on previously confirmed nonsense and frameshift mutations,[154] as well as missense pathogenic variants in *BRCA1/2* genes confirmed by the ENIGMA international expert panel (<http://brcaexchange.org/>). The Maftools software[155] was used for summary, analysis, and visualization of the annotated variants. PTV carriership was defined as having at least one PTV in any of the 31 genes included in the gene panel, and was analyzed as a binary exposure.

3.3.2 Polygenic risk score (Paper I and IV)

A PRS is a tool to summarize the small effects of multiple loci associated with a polygenic disease,[156] and has potential clinical applications.[157] PRSs are calculated as a weighted combination of the number of risk alleles an individual has, where the weights/effect size estimates are typically obtained from an independent study population.[67] In Paper I, we included 162 GWAS significant ($P\text{-value} < 5 \times 10^{-08}$) SNPs associated with breast cancer. In Paper IV, we computed PRSs based on 199, 276, 1284, and 3803 SNPs associated with celiac disease, selected under four P-value thresholds (5×10^{-08} , 1×10^{-05} , 1×10^{-02} , 5×10^{-02}), respectively.

We computed each PRS under the following multiplicative (log-additive) model:

$$\text{PRS} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

where β_k ($k = 1, \dots, n$) is the per-allele log-odds ratio, x_k is the number of risk alleles (e.g. 0, 1, or 2) for SNP_{*k*}, and n is the number of SNPs included in the PRS. The model assumes no genetic interactions, an assumption which has been shown to be reasonable.[158, 159]

3.3.3 Tyrer-Cuzick risk score (Paper II)

In Paper II, we measured the risk of women to develop breast cancer expressed as 5-year absolute risk. The risk was calculated using the Tyrer-Cuzick (TC) model,[160] a breast cancer risk assessment tool from the International Breast Cancer Intervention Study (IBIS). We entered information about established risk factors of breast cancer into the model. These

included factors related to an increased lifetime endogenous estrogen exposure (i.e. early age at menarche, late age at first birth, late age at menopause), as well as exogenous exposures (i.e. use of oral contraceptives, and have undergone hormone-replacement therapy), to previous benign breast disease including hyperplasia, atypical hyperplasia, and lobular cancer in situ, in addition to height, weight, family history of breast and ovarian cancer, Ashkenazy descent, and *BRCA* mutation status.

In brief, the TC model estimates the probability of a woman to develop breast cancer within 5 years, 10 years, or during her lifetime. The risk is age-specific, which is calculated by considering the incidence rate of breast cancer at five year interval observed in the background population (e.g. UK or Sweden). The model multiplies the age-specific incidence by the probability of carrying a hypothetical predisposition gene accounting for all unknown predisposition genes explaining the familial aggregation. This probability is derived from *BRCA* frequencies in the population using a Bayes theorem. We computed the scores using the IBIS tool version 7 (<http://www.ems-trials.org/riskevaluator/>).

3.3.4 Gene expression profiles (Paper II and III)

In Paper II and III, we profiled tumors based on gene expression patterns associated with breast cancer risk (Paper II) or with interval breast cancer (Paper III). The profiles included genes showing strongest association with the exposure of interest, and were computed as the following:

$$\text{Profile} = W_1g_1 + W_2g_2 + \dots + W_n g_n$$

where W_k ($k = 1, \dots, n$) is the gene weight obtained from the discovery gene expression analyses (explained in section 3.5.4), g_k are the log₂-scaled and normalized gene expression levels for gene k , and n is the number of genes included in the profile.

3.3.5 Molecular subtypes (Paper II and III)

In Paper I, surrogate molecular subtypes were assigned to samples using a machine learning algorithm fed with data on immunohistochemistry marker status (ER, PR, HER2, and ki67) as previously implemented by our group.[47]

In paper II and III, breast cancer intrinsic molecular subtypes, also known as PAM50 subtypes, were inferred from gene expression data using a research-based classifier, the Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype (AIMS).[161] This machine learning algorithm was developed to assign subtypes based on a set of patient-level gene expression rules, which are not affected by differences in array/sequencing platforms and cohort composition, both of which the PAM50 classifier is sensitive to.

3.3.6 Immune subtypes (Paper III)

Tumors were classified into distinct immune subtypes based on gene expression profiles following methodology published by Amara and colleagues.[162] The underlying principle is

that it is possible to extract immune-related signals from “bulk” RNA sequencing data. In that paper, the authors assessed 57 published immune expression signatures that could be assigned into one of four co-expression modules: core-serum response (CSR), T-cells and/or B-cells (T/B-cell), interferon (IFN), and transforming growth factor beta (TGFB). Using hierarchical cluster analysis,[163] samples could be classified based on five immune subtypes representing coherent immune-related expression patterns: Immune Low, CSR-High, IFN/CSR High, T/B-Cell/IFN High, and TGFB High.

3.4 STUDY DESIGNS

3.4.1 Case-only study (Paper I-III)

In Papers I to III, we used a case-only design in order to study patterns of breast cancer aggressiveness. In this way, we could assess whether differences in adverse outcomes (more aggressive vs. less aggressive) defined by a number of prognostic factors (e.g. ER-negative tumors), could be explained by a variable of interest (exposure). In this thesis, we evaluated exposures such as different types of breast cancer genetic load (Paper I), level of non-genetic breast cancer risk (Paper II), and unfavorable mode of detection (Paper III).

3.4.2 Case-control study (Paper IV)

Case-control studies are meant to evaluate the association between a variable of interest (exposure) and the disease status (cases or disease-free controls). In Paper IV, we used this study design to estimate breast cancer risk by the amount of genetic predisposition to celiac disease. The main difference with case-only studies, is that the reference (control) group is comprised of non-affected individuals. This approach offers a cost-effect strategy to quantify whether an exposure variable is likely to reduce or increase the probability of developing a disease. A higher level of scientific evidence stems from prospective cohort studies and from randomized-control trials, however, these study designs are costly and often not viable.

3.4.3 Genetic correlation and overlap (Paper IV)

In Paper IV, we performed analysis on the shared genetic component between breast cancer and celiac disease, using two type of methods: 1) based on raw genotype data we computed a celiac disease PRS (see section 3.3.2), which was then used to estimate risk to develop breast cancer, and 2) estimation of genetic correlation and overlap based on analysis of GWAS summary statistics (section 3.2.4.3) using methods described in sections 3.5.6 and 3.5.7. In brief, this type of study aim to assess the extent of genetic variation that is shared between polygenic traits (e.g. complex diseases). Ideally, genetic variation on each trait should be measured using set of samples independent from the discovery studies in order to avoid over-estimation. The existence of a widespread shared genetic variation across complex traits is based on the concept of pleiotropy, meaning that genetic variants can be involved in multiple disease pathways.[164] In essence, genetic overlap is similar to the concept of pleiotropy, while in genetic correlation analysis the direction of the association is taken into account.[98] Future

studies might explore the role of cell-type or tissue specific effects on the genetic predisposition shared between different diseases.[165]

3.5 STATISTICAL METHODS

All statistical analyses were performed in the open-access statistical software R. Binary outcomes were analyzed using logistic regressions, and categorical outcomes using multinomial logistic regressions. Survival analyses were performed using Cox regressions. Methods to analyze gene expression differences and their effect at the level of biological processes, are also included. Finally, methods to assess genetic correlation are described, as well as modeling of potential confounding.

3.5.1 Logistic regression

Logistic regression was used to make inference on binary variables (outcomes) such as ER-negative vs ER-positive status or interval cancer vs screen-detected tumor. Effect estimates for the associations of explanatory variables (exposures) such as high vs low PRS, or carriership of rare variants, with the main outcomes, were expressed as odds ratios (OR), and 95% confidence intervals, were calculated. With logistic regression, the probability of an event Y (binary outcome), conditional on an explanatory variable, X , is modelled as

$$P(Y = 1|X) = P(Y) = \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}}$$

where β_0 and β_1 are two parameters inferred from the data, e is the exponential, and X is a covariate (but can easily be extended to a set of covariates). The log-odds is defined as $\log[P(Y)/1 - P(Y)]$, which is known as the logit transformation. It can be shown that unbiased estimates of ORs can be obtained from case-control data by fitting a logistic regression model and by taking the exponent of the estimate of β_1 . Under this model, β_1 represents the log-odds ratio, and is defined as $\log[\text{odds of } Y | X = 1 / \text{odds of } Y | X = 0]$. Logistic regression models were fitted as special cases of generalized linear models using the R *function* `glm`.

3.5.2 Multinomial logistic regression

Multinomial logistic regression was used to estimate ORs, with 95% confidence intervals, when assessing association with categorical outcome variables (reference category vs exposure categories), using the *nnet* R package. The multinomial regression model can be viewed in terms of separate logistic regression models for each exposure category, each against the reference group.

3.5.3 Cox Proportional-Hazards regression

Cox Proportional Hazards (PH) regression was used to estimate hazard-ratios (HRs), with 95% confidence intervals. Under this method, a time-to-event (e.g. from death due to breast cancer) is modeled as a function of explanatory variables, and individuals are considered to be at risk

from the time to entry to the end of follow up, and are censored (no longer considered as risk) if the event has occurred, the participant has left the study, or the study has ended. The Cox PH regression model specifies the hazard of an event at time t as a function of a baseline hazard and the independent explanatory variable(s), such that

$$\lambda(t|X_i) = \lambda_0(t) \exp(X_i \cdot \beta)$$

where $\lambda_0(t)$ is the baseline risk of the event (i.e. with covariates set to zero) per unit change in the underlying time scale, X_i is the vector of covariates for i individuals, and β is the vector of coefficients explaining the hazards. In our analysis, time since diagnosis, in years, was used as the underlying time scale.

3.5.4 Gene expression analysis

Gene expression analysis were performed in Papers II and III to quantify gene-level differences across samples explained by the exposure of interest. Methods under the generalized linear model framework that allowed for including co-variables were developed to analyze gene expression levels in the form of intensities yielding from Microarray platforms.[166] With the advent of RNA-sequencing technologies, which instead produces count data (e.g. number of RNA fragments mapped into a gene-coding region), new types of statistical model have been required. In our studies, we used the methodology implemented in the *edgeR* package,[167, 168] which is based on the negative binomial distribution and quasi-likelihood tests. Data preprocessing included normalization for differences on library composition (e.g. amount the RNA fragments) using the trimmed mean of M-value method.[169]

3.5.5 Gene set enrichment analysis (GSEA)

Gene set enrichment analysis (GSEA) methods assess the information coherence at gene level (association with a given trait) when grouped into gene sets defined based on prior knowledge (gene annotations). In this thesis, we used a well curated annotation source, the Molecular Signature Database (MSigDB) comprised of fifty biological hallmark gene sets (not to be confused with the hallmarks of cancer!).[170] Generally, gene sets are said to be enriched if there is significant statistical evidence that its constituent genes are differentially expressed in a consistent manner. Since a number of GSEA methodologies have been developed based on different statistical assumptions and hypothesis testing,[171, 172] one approach is to look for consistent results produced by different methods.

In Papers II and III, we used a comprehensive workflow analysis implemented in the *Piano* R package.[173] The input data is described on each paper. Gene set-level statistics were computed using six different GSEA methods: Wilcoxon rank-sum test, tail strength, mean, median, sum, reporter features, and Stouffer's method. To summarize findings, the Piano workflow generates a consensus score to rank gene set based on their consistency for association across the different GSEA methods. In addition, gene set enrichment is distinguished by the direction of gene expression changes (e.g. positive or negative) into: non-directional class (only gene-level P-value information is considered), mixed-directional class

(gene set can have subset of genes associated in opposite directions, one of which dominates), and distinct-directional class, which indicates a clear trend for association in either direction (genes associated in opposite direction will cancel each other out). Statistical significance was assessed based on the null distribution computed by permutation of gene labels. To control for multiple testing, we allowed the false discovery rate (FDR) to be lower than 5%. Significantly enriched gene sets were reported as having a median adjusted P-value lower than 0.05.

3.5.6 Cross-trait LD Score (LDSC) regression

In Paper IV, we used LDSC regression to estimate genetic correlation.[98] This method is based on modeling GWAS summary statistics based on LD (linkage disequilibrium).[98, 174] LD is the non-random association of alleles at different loci (the combinations of alleles at different loci on the same chromosome are called haplotypes). If two SNPs were independent and associated randomly, for alleles A_1 and A_2 at locus A, with respective frequencies $p_i (i = 1,2)$, and alleles B_1 and B_2 at locus B, with frequencies $q_k (k = 1,2)$, the expected haplotype probabilities would be defined by $p_i \times q_k$. [141] LDSC regression is based on modeling the genetic covariance as the relationship between SNP effect estimates for two traits (i.e. the product of z scores, $z_{1j}z_{2j}$ for SNP j) explained by the amount of information, LD score, SNPs carry. In such a case, the LDSC regression assumes that an SNP in high LD with other SNPs summarizes the effects those SNPs. The LDSC regression is estimated in such a way that the expected value of $z_{1j}z_{2j}$ is regressed on the LD scores, under the equation

$$E[z_{1j}z_{2j} | \ell_j] = \frac{\sqrt{N_1 N_2} \rho_g}{M} \ell_j + \frac{\rho N_s}{\sqrt{N_1 N_2}}$$

where N_i is the study sample size, ρ_g is the genetic covariance, ℓ_j is the LD score, N_s is the total number of individual, ρ is the phenotypic correlation among N_s overlapping samples, and M is the number alleles in the reference panel with minor allele frequency between 5% and 50%. Genetic correlation is then calculated as

$$r_g := \rho_g / \sqrt{h_1^2 h_2^2}$$

where the genetic covariance ρ_g is normalized by the SNP heritabilities h_i^2 from study i .

3.5.7 SNP Effect Concordance Analysis (SECA)

In addition to the LDSC regression used in Paper IV, we measured genetic correlation and overlap using the SECA methodology,[96] which is based on a different statistical approach. While in the LDSC regression genetic correlation is modelled as a function of the LD scores, the SECA method is based on a pre-selection of independent SNPs (e.g. in low LD with each other) so that genetic correlation is not inflated by SNPs in high LD. SNPs are filtered using a two-step LD pruning procedure. Of notice, genetic correlation is referred as genetic concordance under this method.

In order to assess the genetic shared component between two traits, effect estimates (P-values) from each trait (P_i) are plotted against each other into a grid. The grid consists of 144 squares defined by combinations of 12 x 12 equally increasing P-value thresholds: $\{P_1, P_2\} = \{0.01, 0.05, 0.1, 0.2, 0.3, \dots, 1.0\}$, where the strongest evidence of genetic overlap would be expected to occur at the lowest P-value combination $\{P_1 < 0.01, P_2 < 0.01\}$. Genetic overlap is defined as the excess in overlapping SNPs (observed > expected) across the grid assessed through binomial tests. To avoid overestimation due to powered GWAS yielding low P-values, the expected frequency of overlapping SNPs is defined as the observed frequency for one of the traits. In paper IV, we used the observed proportion of celiac disease SNPs as the expected value. Analogous to genetic correlation, SECA test assesses for consistency in the direction of effects between overlapping SNPs through Fisher's tests across the P-value grid. An odds ratio > 1 indicates genetic concordance (e.g. positive correlation), < 1 indicates genetic discordance (e.g. negative correlation), and an odds ratio equal to 1 means no evidence of genetic correlation. Significance testing was performed by generating empirical null distribution through random permutation of SNP effect estimates. Following the SECA approach, the subset of overlapping SNPs with strongest evidence of genetic correlation can be identified.

3.5.8 Confounding

Because of the cross-sectional nature of case-only and case-control studies, potential confounding effects are of major concern. Confounding variables are factors correlated with both the exposure and the outcome, and that lead to spurious associations or masked effects (biased estimates) when not taken into account. Modeling of covariates and stratified analysis are two common approaches to deal with potential confounding. We used both approaches in this thesis, while crude effects were obtained from unadjusted analyses. The main confounding variables were: chronological age, age at breast cancer diagnosis, PAM50 subtypes, mammographic density, tumor characteristics, and treatment. In genetic correlation analysis, the effect of phenotypic covariates is assumed to be minimal, and other sources of confounding such as genetic correlation by sample overlap or relatedness, and LD structure, are considered under each methodology.

4 MAIN RESULTS AND INTERPRETATIONS

Unlike common genetic variation, carriership of rare deleterious variants in breast cancer predisposition genes was associated with more aggressive tumors and poorer survival.

In Paper I, we analyzed the contribution of germline genetic variants towards disease aggressiveness in a case-only study. We found that common genetic variants tend to predispose to tumors of more favorable clinicopathology, whereas rare deleterious variants were associated with more aggressive disease defined by tumor characteristics and the PAM50 subtypes (**Figure 6**). In particular, stronger differences were observed for tumor grade, luminal B, and basal-like subtypes. Of note, we did not observe differences in common genetic load (i.e. for overall, ER-positive, and ER-negative PRS) by rare variant carriership status (see paper I, Figure S3). In addition, rare deleterious variants in any of the 31 breast cancer genes, PTV carriership, was associated with interval cancers in women with low mammographic density, as well as with worse survival independently of treatment and tumor characteristics (see paper I, table 3 and 4). The strongest association with worse survival was observed for women below age 50 and carriers of non-*BRCA1/2* rare variants, whereas no association was observed for common variants. Likewise, *BRCA1/2* rare variants seemed to drive the association with ER-negative and basal-like subtypes in younger women (OR: 1.75; 95% CI, 1.11 to 2.75, and OR: 5.24; 95% CI, 2.35 to 11.66, respectively), and this is consistent with previous knowledge about the enrichment of *BRCA1* variants in basal-like tumors. Interestingly, non-*BRCA1/2* rare variants remained significantly associated with poorly-differentiated and luminal-B tumors in older women (OR: 1.65; 95% CI, 1.10 to 2.48, and OR: 2.21; 95% CI, 1.36 to 3.59, respectively). Together, our analysis indicates that carriership of rare deleterious variants in any of the 31 predisposition genes predispose to more aggressive disease, independently of age group.

Discussion

In our study, we leveraged large genotyping data and previous GWAS findings in order to compute genetic risk scores, as well as on targeted exome sequencing to inquire into rare deleterious mutations. Our findings suggest that rare and common variants act as distinct risk entities. This is consistent with the hypothesis that rare predisposition mutations with moderate effects (e.g. heterozygous mutations in genes for which biallelic mutations are known to be causal of genetic syndromes) are more likely to explain the missing heritability observed in GWAS studies.[175] However, there are some methodological challenges in interrogating the contribution of rare variants using NGS technologies, such as low power.[176] To overcome that, we used an approach by aggregating mutations. Thus, we assumed that predisposition to breast cancers of an aggressive phenotype could occur through disruption of any of the genes tested.

In that context, the association of rare variants with unfavorable disease outcomes suggests a direct link to the etiopathology of breast cancer. Because the analyzed genes are involved in important processes of genome maintenance such as DNA-repairing mechanisms,[177] deleterious mutations in any of these genes could increase the probability to develop more aggressive tumors by acquisition of further mutations. Interestingly, a recent whole-exome sequencing study of 54 non-BRCA familial breast cancer index patients found that 44% of them carried one or maximum 3 rare deleterious variants.[178] The authors reported that mutations in DNA repairing genes conferred a two-fold increased risk as compared to 120 matched controls, and novel variants in genes not known to be related to cancer were found, highlighting the usefulness of sequencing approaches.

Regarding common variants, their utility to predict risk to develop either breast cancer subtype, will require the incorporation of subtype-specific risk variants identified through ongoing efforts, particularly of triple-negative breast cancer,[63] and as shown in the latest comprehensive GWAS published in May, 2020.[61] As noticed in our analysis, genetic load from common variants weighted by ER-negative disease did not show association with unfavorable prognosticators, and although correctly predicted higher risk for ER-negative disease, had a small effect (OR per 1-SD: 1.10, 95% CI, 1.01 to 1.19). In the contrary, analysis based on variants weighed by ER-positive disease showed similar associations as weighting by the overall risk, supporting the idea that discovery based on overall breast cancer is biased towards the most frequent ER-positive disease.

Lower breast cancer risk defined by the Tyrer-Cuzick score was associated with more aggressive disease. Gene expression analysis highlighted the involvement of proliferative processes.

In Paper II, we found that breast cancer risk defined by the TC score was inversely associated with basal-like and HER2-enriched surrogate subtypes (as compared with luminal A surrogate subtype), and with ki-67 proliferative marker. In order to better understand the association between lower TC and disease aggressiveness, we characterized underlying molecular differences. Using transcriptomic data, we correlated gene expression to TC score and summarized it into a low-risk TC-expression profile (TC-Gx), based on the top 37 genes showing strongest correlation with TC. The low-risk TC-Gx was able to discriminate tumors and indicated an overlap with more aggressive subtypes, particularly with basal-like tumors (**Figure 7**). Regression analyses showed that the low-risk TC-Gx was associated with the more aggressive PAM50 subtypes such as basal-like (validation dataset, OR per 1-SD: 13.20, 95% CI, 7.10 to 24.57) and HER2-enriched (validation dataset, OR per 1-SD: 4.79, 95% CI, 2.95 to 7.79) as compared with luminal A tumors, and with higher breast cancer-specific mortality (\geq mean vs $<$ mean low-risk TC-Gx, HR: 2.29, CI, 1.21 to 4.35). The association with higher mortality was partially explained by the PAM50 subtypes and by higher levels of ki-67 gene expression, but not by ER expression. Interestingly, we found that low-TC gene expression was significantly enriched in proliferative and oncogenic signaling processes.



Figure 6. Differential association of rare and common genetic variants with breast cancer aggressiveness. Rare variants are represented by carriership of at least one PTV in any of the 31 panel genes (violet) or excluding *BRCA1/2* variants (blue). Common variants are depicted by overall (green), ER-positive (mustard), and ER-negative (orange) PRSs.

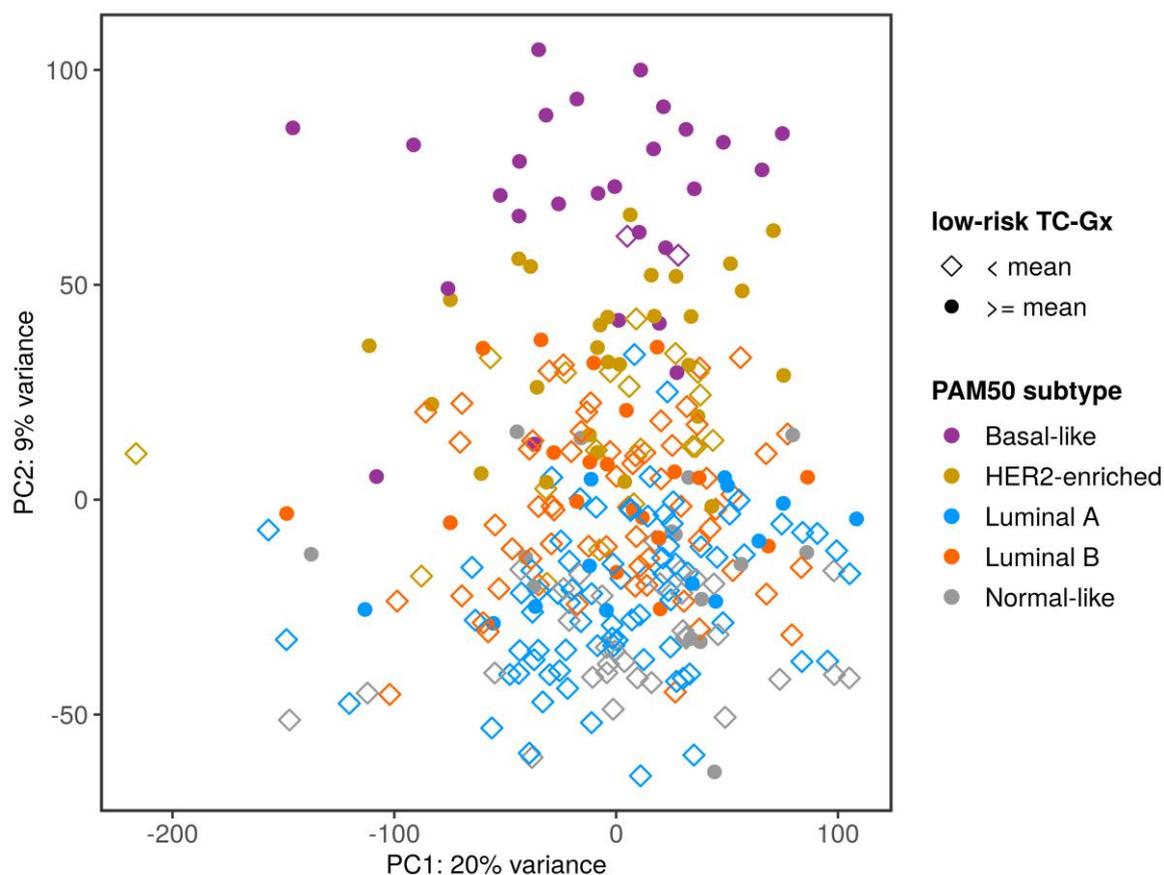


Figure 7. Relationship between low-risk TC-Gx profile and breast cancer intrinsic subtypes. Principal Component Analysis (PCA) plot showing similarity between 296 samples in the discovery set based on transcriptomic data (whole-genome expression levels). Samples are labeled according to the low-risk TC-Gx profile: open square if decreased ($<$ mean distribution) or solid dot if increased (\geq mean distribution). In addition, tumor samples are colored to show their relationship with intrinsic subtypes.

Discussion

In our study, we used a molecular epidemiology approach to investigate gene expression features behind the association of low TC score with the breast cancer subtypes of unfavorable prognosis, and this is consistent with previous work from our group.[38] In addition, our low-risk TC-Gx was found to be associated with worse survival independently of ER status, suggesting the existence of underlying risk factors beyond the involvement of estrogen exposure. Therefore, risk modeling incorporating factors associated with basal-like and HER2-enriched disease, and in particular risk factors favoring higher proliferation, could have an important contribution to improve risk assessment tools. Interestingly, the association between low-risk TC-Gx and survival was only weakened after adjusting by the PAM50 subtypes and high proliferation status, indicating the existence of poor prognosis tumors within the Luminal A subtype.

Findings from our enrichment analysis highlighted oncogenic and signaling pathways as potential mechanisms underlying the increased aggressiveness associated with low TC. Interestingly, these included MYC and E2F oncogenic targets, as well as mTORC1 and WNT beta catenin pathways, which have been associated with aggressive breast cancer subtypes. For instance, MYC overexpression has been associated with poorer outcomes in basal-like, Luminal A breast cancer with lymph-node involvement, and HER2-positive tumors,[179], whereas E2F transcription factors, mTORC1 and WNT beta catenin, have been found to be important in triple-negative breast cancer.[180-182]

The TC model, similar to other risk assessment tools (such as the Gail model, and genetic scores), can be used to stratify women into distinct risk groups. This stratification can be used to tailor preventative strategies. For instance, women with a risk higher than the average risk observed for their age-group, could be considered for personalized screening,[183] so that early detection is achieved more successfully in the entire screening population. We argue that in order to effectively reduce disease mortality, risk assessment tools should be able to identify women at increased risk of developing breast cancer aggressive subtypes.

Compared to screen-detected tumors, interval cancers in women with low-dense breasts exhibited gene expression patterns associated with interferon immune subtypes, independently of PAM50 subtypes.

In Paper III, we characterized gene expression for interval cancer in women with low-dense breasts as compared with screen-detected tumors, in order to identify underlying biological features independently of the PAM50 subtypes. Through enrichment analysis using the MSigDB database, a curated collection of hallmark gene sets representing well-defined biological processes, we found that altered gene expression in interval cancers was mainly related to immune response. We then profiled tumors based on genes found to be strongly associated with interval cancer by computing the IC-Gx profile. The IC-Gx was found to be associated with breast cancer subtypes, particularly with subtypes involving a high interferon signal, and this was replicated in an independent cohort from the TCGA database (**Figure 8**).

Discussion

It is not well understood why some tumors, referred as interval cancers, are less likely to be detected through regular mammographic screening. Because interval cancers tend to have more adverse tumor characteristics,[184] and molecular subtypes[185] one hypothesis is that fast growing tumors commonly of ER-negative, basal-like and HER2-enriched subtype, are able to reach symptomatic detectability in a short time span and therefore are more likely to become interval cancer. Another complementary hypothesis is that high mammographic density, which is associated with interval cancers,[186, 187] reduces mammographic screening sensitivity, also known as masking effect.[33] Nevertheless, in addition to conventional tumor

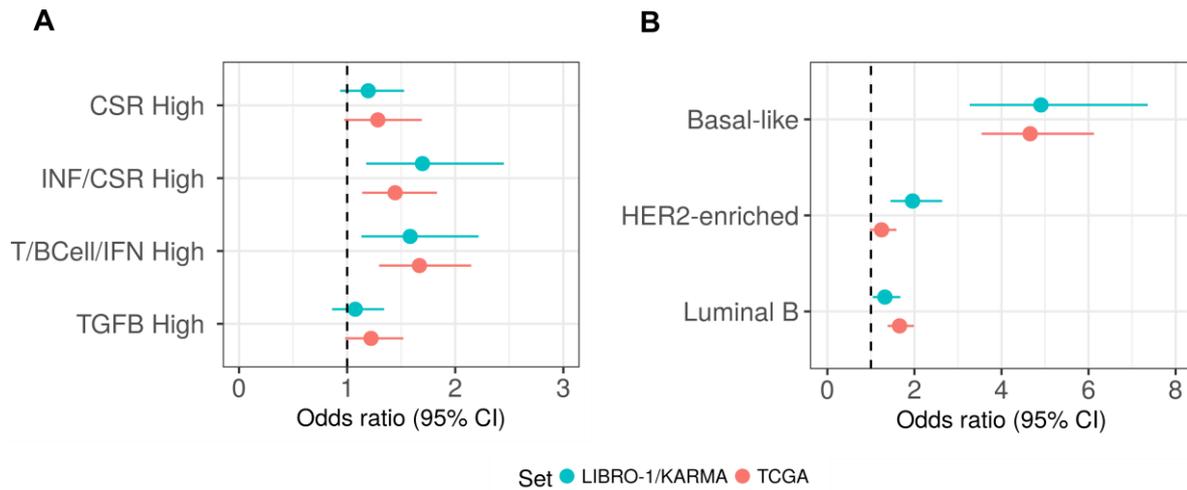


Figure 8. Association between the IC-Gx profile with breast cancer subtypes. **A)** Association with immune subtypes as compared with the *Immune Low* subtype, and adjusted for PAM50 subtypes. CSR, core-serum response; T/B-cell, T-cells and/or B-cells; IFN, interferon; TGFB, transforming growth factor beta. **B)** Association with main intrinsic subtypes as compared with the Luminal A subtype. In both figures, estimates were obtained from multinomial logistic regressions in the discovery (LIBRO-1/KARMA, n=672) and external validation set (TCGA, n=975). Odds ratio and 95% confidence intervals are shown per one-standard deviation in the IC-Gx profile.

characteristics, other factors are needed to be discovered in order to explain the poorer outcomes observed for interval cancer.[188] Previous work from our group have shown that aggressive interval cancers are over represented in women with mammographically low dense breasts.[38] Also, molecular characterization of interval cancers lead to the conclusion that most features were explained by the PAM50 subtypes.[37]

In order to improve early detection of breast cancer, better biological understanding of interval cancers is needed.[189] By dissecting gene and molecular process likely to underlie interval cancers, our study contributes toward this goal. In particular, our study highlights the involvement of the interferon immune response as a potential target. These findings are further supported by preliminary data from our group showing that germline genetic variants associated with interval cancers are significantly enriched in a network of IC-Gx genes (*CXCL7/8*, *CXCR1/2*) interacting with type I and II interferon genes. To our knowledge, this is the first study to characterize gene expression patterns on interval cancers in women with low dense breast, and to bring forward the potential role of interferon genes into this subset of aggressive breast cancers.

Shared common genetic variation between breast cancer and celiac disease was found to be inversely correlated, consistent with previous epidemiological findings on the reduced risk of breast cancer in celiac disease patients.

In Paper IV, we measured the extent of the shared genetic variation between breast cancer and celiac disease, guided by previous epidemiological findings on the observed reduced risk of breast cancer in celiac disease patients. We found a statistically significant inverse genetic correlation between the two diseases using two different analytical approaches (overall breast cancer and celiac disease, LDSC, $r = -0.17$, s.e. 0.05; SECA, OR: 0.60, 95% CI: 0.44 to 0.82). In a third analysis, we performed a case-control analysis to estimate breast cancer risk by celiac disease genetic load, namely, celiac disease polygenic risk score (celiac-PRSs). We found that a higher genetic load for celiac disease was associated with 6% to 13% decrease risk of breast cancer when comparing highest versus lowest quartile distribution (quartile 4 vs. quartile 1 of celiac-PRS based on 3,803 associated with celiac disease at nominal P-value, OR: 0.83, 95% CI 0.75 to 0.93). Associations by ER status showed similar results between overall breast cancer and ER-positive disease, whereas no association was observed for ER-negative breast cancer, nor for other tumor characteristics. Further assessment of the genetic overlap between the two diseases showed that top SNPs were significantly overrepresented in pre-defined gene sets such as: induction of apoptosis and programmed cell death, MAPK and other protein-protein interaction subnetworks, as well as gene sets related to immune phenotypes. A prioritization analysis highlighted fifteen top SNPs as the most relevant loci for the genetic overlap between the two diseases (**Figure 9**).

Discussion

The immune system has an important role in breast cancer, both in the etiology and the progression of the disease.[126, 190] As a strategy to investigate the immune-related genetic component in the etiology of breast cancer, we exploited potential pleiotropic effects with celiac disease by leveraging of the largest GWAS summary statistics from breast cancer and celiac disease available at the time. Our findings were consistent with the observed reduced risk of breast cancer in celiac disease patients, which has been reported to be 10% to 15% lower, in Nordic populations.[92-94, 191] In addition, we pinpointed genetic loci and molecular pathways as most likely underlying a shared etiology between the two diseases. Our findings forward the hypothesis that an increased genetic susceptibility to celiac disease could be protective against breast cancer pathogenesis by regulation of key immune processes directing cancer cells towards apoptosis, in which immunosurveillance processes can prevent mammary cancer cells to proliferate.[192] Likewise, an unfavorable immune-related genetic load could be involved in the breast cancer etiopathology by predisposing mammary cancer cells to evade the immune system. It is possible that an increased propensity for the formation of cancer cells (such as mutations in genome-stabilizing genes), together with an altered immunogenic microenvironment, could lead towards an equilibrium phase of cancer cells with the host immune system. In such a phase, cancer cells would adapt and acquire further mutations favoring tumor development.[125]

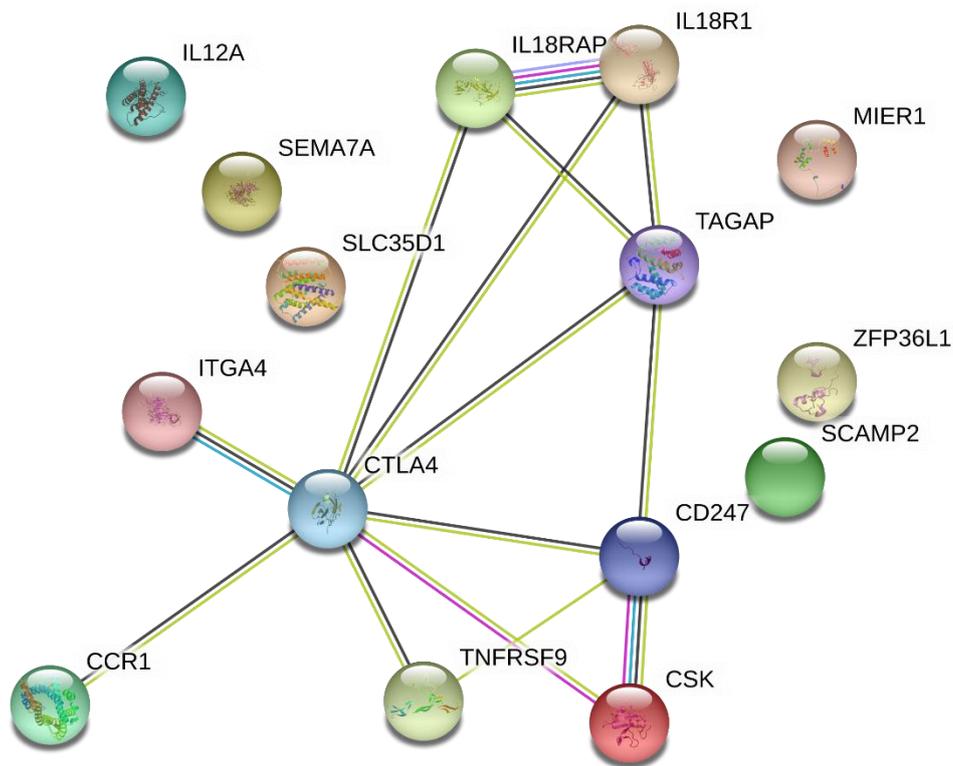


Figure 9. Network of the 15 immune-related genes most likely to underlie the genetic overlap between breast cancer and celiac disease. Genes were deemed as significantly relevant (P -value <0.05) in a prioritization analysis from a list of 52 top-overlapping SNPs between breast cancer (BC) and celiac disease (CD) ($P_{BC}\leq 0.05$, $P_{CD}<1\times 10^{-05}$), and are described in Paper IV, Table 3. Gene network was generated based on known gene-gene interactions using STRINGv11.[193]

5 Concluding remarks

- i. The study of breast cancer genetic germline variants is useful to improve our clinical and biological understanding of the disease, particularly when linked to detailed phenotypic data.
- ii. We observed that unlike common variants, rare deleterious variants in breast cancer predisposition genes were associated with more aggressive disease. This supports the hypothesis that rare variants with moderate penetrance in key pathways have an important contribution in the disease progression.
- iii. The observed association between lower risk of breast cancer with more aggressive disease, is likely due to the lack of accuracy to predict highly proliferative and invasive tumors by means of established risk factors (non-genetic) as modelled in the Tyrer-Cuzick score.
- iv. Interval breast cancer, a subset of tumors not detected at the time of regular mammographic screening visits, was found to display unique gene expression patterns correlated with interferon-immune response that could be involved in their aggressive phenotype.
- v. The potential role of an inherited immunogenic environment in the etiology of breast cancer was highlighted by our findings on the shared genetic component with celiac disease, an autoimmune disease.
- vi. Findings in this thesis could inform further efforts toward the identification of women at high risk to develop aggressive breast cancer.

6 Future perspectives

After the conclusion of the four studies presented in this thesis, some questions come to light:

Would rare deleterious variants be associated with breast cancer (subtype-specific) risk in a case-control study, or are these type of variants mainly contributing to the disease prognosis? Would sequencing of pairs of germline and somatic tissue show that protein-truncating variants become homozygous driver mutations, or that rather act through other mechanisms?

What is the optimal genotyping strategy in breast cancer epidemiological studies needed to identify significant genetic variants associated with aggressive subtypes? Would well-powered, subtype-specific GWAS on basal-like, HER2-enriched, and highly proliferative tumors, be sufficient? Or are whole-exome and/or whole-genome sequencing approaches necessary to identify genes and regulatory regions implicated in the development of aggressive subtypes of breast cancer?

Either way, large studies using surrogate subtype classifiers based on IHC markers, could be of great importance, while more refined analysis to understand additional biological aspects will require omics data such as gene expression. In that way, the discovery of novel risk factors specific to more aggressive disease subtypes and their incorporation into risk assessments tools will improve our ability to identify women at increased risk.

Another future direction is whether existence of additional aggressive subsets of breast cancer require our attention, for instance therapy-resistant Luminal A tumors. Or, is that other aggressive features, such as unfavorable immune response and mutational load, act independently of phenotype from the PAM50 classification and are more important in this context?

Our findings suggest that fine mapping of the genetic variation in immune-related genomic regions will be also useful to understand the role of the immune system in susceptibility to develop breast cancer. In addition, future genomic and functional studies are required to better understand the role of immune phenotypes in the disease progression of aggressive subsets such as interval cancers. For instance, are there specific interferon genes also associated with BC prognosis? If so, what are the mechanisms favoring tumor progression that could be targeted?

Finally, because higher mortality is observed in non-European populations from often less socioeconomically developed countries, worldwide reduction of disease mortality will require the transfer of knowledge and improvements of their healthcare systems.

Acknowledgements

First of all, I would like to thank my supervisors. My main supervisor, **Kamila Czene**, thank you for having me as one of your doctoral students, for the hard work and dedication, for the constructive criticism, and for all the support. To **Jingmei Li**, thank you for being so helpful and patient helping me improve. Thanks for welcoming me at MEB and for together with Kamila, choose me to join the group. That phone interview has definitely been one of the most life-changing calls I've had. Thanks to **Keith Humphreys** for been so friendly and helpful with all the statistical details, I have learned a lot from that. To **Felix Grassmann**, thanks for your support and amazing contribution on the last manuscripts. To **Per Hall**, thank you for providing the wider perspective and for making this possible together with Kamila.

To friends and colleagues from Kamila's and Pelle's group: **Fredrik Strand, Johanna Holm, Haomin Yang, Mikael Eriksson, Erwei Zeng, Xinhe Mao, Natalie Holowko, Wei He, Pui San Tan, Zhadi Azam, Marike Gabrielson, and Ami Rönnberg**. Special thanks to **Fredrik** for welcoming me into the group, for the nice talk, lunches, after works, and hangouts. Thanks for being so generous and supportive. To **Johanna**, thanks for being so friendly and cool person. Your thesis is definitely a great legacy. To **Haomin**, thanks also for being always helpful and so friendly. To **Mikael**, thanks for being so helpful, the data quality is just great. To **Ami**, thanks for being so kind and supportive.

To my close friends at MEB: **Andreas Yangmo, Carolyn Cesta, Isabell Brikell, Ida Karlsson, Dylan Williams, Jet Termorshuizen, Qian Yang, Laura Ghirardi, Elisabeth Dahlqwist, Isabella Ekheden**. Thanks for welcoming me into Sweden, for all the support during the hard times, and even more, for making me feel at home. Thanks for the good moments and cool adventures. Special thanks to **Andreas**, thanks so much for the friendship and generosity, for organizing all sort of cool stuff, for being such a great host and remind me of the Mexican culture. To **Carolyn**, my great gratitude for all your kindness, help, and support, for finding me places to live and meals to eat, and for saving my thesis from funny hazards. To **Ida**, thanks for being an amazing friend, it would not have been the same without that shared love for Weissbier and nice talks. To **Dylan**, for being such a great pal, person, and researcher. To **Elisabeth**, thanks for been such a great friend, for the nice talks, hard-core exercising and hangouts.

To everyone at MEB, thank you all for making this place a great one to be part of. To **Qing Shen**, with whom shared co-chair of the PhD group. To **Jiayao Lei**, and **Malin Ericsson**, thanks for helping with the application process. To **Frida Lundberg, Mark Taylor, Mina Rosenqvist**, was great to join you for fika. To **Andreas Karlsson** for the nice talks at lunch and dinner. **Robert Karlsson**, thanks for the help and cool mood. To the amazing Harry Potter cast: **Kat Bokenberger, Elisabeth Dahlqwist, Hannah Bower, Shuyang Yao, Gabriel Isheden, and Zhen Ning**. Thanks for all the great fun organizing Christmas MEB's dinner. To the dear, friendly and amazing persons with whom I have a Latin connection at MEB: **Laura Ghirardi, Marco Trevisan, Marica Leone, Elisa Longinetti**, thanks for your friendliness and

great mood. Thanks as well to **Tong Gong, Tyra Lagerberg, Ash Thompson, Jingru Yu, Tingting Huang, Fei Yang, Cecilia Radkiewicz, Bronwyn Brew, Anna Johansson, Rikard Strandberg, Henrik Olsson, Wenjiang Deng, Daniela Mariosa, Anna Plym, Camilla Sjörs**, and all the friendly people at MEB.

Special thanks to **Shadi, Wenjiang**, and **Rikard** for reading the thesis and helping me prepare for the defence. To my mentor, **Sarah Bergen**, thanks for being supportive and for the nice talks.

To the friendly research staff and professors at MEB, **Patrick Magnusson, Paul Dickman, Mark Clements, Kristina Johnell, Fredrik Wiklund, Mark Divers, Mattias Rantalainen**, and many more. To MEB administrative staff, thanks for always helping out and making things work. Particularly to **Gunilla Sonnebring** and **Marie Jansson**. To the directors of PhD studies and educational administrators at MEB: **Paul Lichtenstein** and **Amelie Plymoth, Gunilla Nilsson Ross**, and **Alessandra Nanni**. Special thanks to **Alessandra** for assisting and helping throughout this process.

Thanks to all the people who had been involved in the studies and generation of data used in this thesis. To my **co-authors** and **collaborators**, thank you very much for your work and contribution to this thesis. Thanks to the Swedish Bioinformatics Advisory Program, in particular to **Sebastian DiLorenzo** and **Björn Nystedt**.

I would like to thank as well to people from my masters at Wageningen University, The Netherlands. To my thesis supervisors **Guido Hooiveld** and **Leo Lahti**. To my dearest and beloved friends **Ale Hernández, Vicente Sedano, Noora Ottman** and **Armando Garcia**. Thanks to people from my BSc program at Universidad Iberoamericana, Leon, Mexico. In particular to **Luis Adolfo Torres**, for being my mentor and friend, and for inspiring me to pursue a scientific carrier. To my thesis supervisor **Teresa Tusié-Luna**, friends at the National Institute of Nutrition, Mexico City. Special thanks to **Gabriela Fonseca** for your friendship and encouragement. Thanks to **Karla Sánchez-Lara** at the Cancer Center in MedicaSur Hospital, Mexico City.

To my *beloved ones*, thanks so much for your being there for me in the hard times and for cheer me in the good ones. Thanks for believe in me and making me feel loved. To my dear family: **Paco, Lucy, Adrian, Natan, e Tania**, and the beautiful little branches, thanks for all the effort and sacrifices you have made, for letting me be, for worrying and taken care of me, for walking with me along the way regardless of the physical distance. You are my roots and part of my identity. *Aún en la distancia, su esencia siempre ha estado y estará presente en mi. Los amo.*

And last but not least, thanks to **all the women** that have participated in the studies included in this work.

REFERENCES

1. Forouzanfar, M.H., et al., Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. *Lancet*, 2011. 378(9801): p. 1461-1484.
2. Bray, F., et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, 2018. 68(6): p. 394-424.
3. Bellanger, M., et al., Are Global Breast Cancer Incidence and Mortality Patterns Related to Country-Specific Economic Development and Prevention Strategies? *J Glob Oncol*, 2018. 4: p. 1-16.
4. Momenimovahed, Z. and H. Salehiniya, Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer (Dove Med Press)*, 2019. 11: p. 151-164.
5. Parkin, D.M. and L.M. Fernandez, Use of statistics to assess the global burden of breast cancer. *Breast J*, 2006. 12 Suppl 1: p. S70-80.
6. Early Breast Cancer Trialists' Collaborative, G., et al., Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet*, 2012. 379(9814): p. 432-44.
7. Berry, D.A., et al., Effect of screening and adjuvant therapy on mortality from breast cancer. *N Engl J Med*, 2005. 353(17): p. 1784-92.
8. Youlten, D.R., et al., The descriptive epidemiology of female breast cancer: an international comparison of screening, incidence, survival and mortality. *Cancer Epidemiol*, 2012. 36(3): p. 237-48.
9. Delgado-Ramos, G.M., et al., Real-world evaluation of effectiveness and tolerance of chemotherapy for early-stage breast cancer in older women. *Breast Cancer Res Treat*, 2020. 182(2): p. 247-258.
10. Furlanetto, J. and S. Loibl, Optimal Systemic Treatment for Early Triple-Negative Breast Cancer. *Breast Care (Basel)*, 2020. 15(3): p. 217-226.
11. Ehemann, C.R., et al., The changing incidence of in situ and invasive ductal and lobular breast carcinomas: United States, 1999-2004. *Cancer Epidemiol Biomarkers Prev*, 2009. 18(6): p. 1763-9.
12. Makki, J., Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. *Clin Med Insights Pathol*, 2015. 8: p. 23-31.
13. Sinn, H.P. and H. Kreipe, A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition. *Breast Care*, 2013. 8(2): p. 149-154.
14. Hoon Tan, P., et al., The 2019 WHO classification of tumours of the breast. *Histopathology*, 2020.
15. Yeo, S.K. and J.L. Guan, Breast Cancer: Multiple Subtypes within a Tumor? *Trends in Cancer*, 2017. 3(11): p. 753-760.
16. Waks, A.G. and E.P. Winer, Breast Cancer Treatment: A Review. *JAMA*, 2019. 321(3): p. 288-300.
17. Moo, T.A., et al., Overview of Breast Cancer Therapy. *PET Clin*, 2018. 13(3): p. 339-354.
18. Russnes, H.G., et al., Breast Cancer Molecular Stratification From Intrinsic Subtypes to Integrative Clusters. *American Journal of Pathology*, 2017. 187(10): p. 2152-2162.
19. Chavez-MacGregor, M., et al., Incorporating Tumor Characteristics to the American Joint Committee on Cancer Breast Cancer Staging System. *Oncologist*, 2017. 22(11): p. 1292-1300.
20. Hammond, M.E., et al., American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J Clin Oncol*, 2010. 28(16): p. 2784-95.
21. Grann, V.R., et al., Hormone receptor status and survival in a population-based cohort of patients with breast carcinoma. *Cancer*, 2005. 103(11): p. 2241-51.

22. Dunnwald, L.K., M.A. Rossing, and C.I. Li, Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. *Breast Cancer Research*, 2007. 9(1).
23. Yarden, Y. and M.X. Sliwkowski, Untangling the ErbB signalling network. *Nat Rev Mol Cell Biol*, 2001. 2(2): p. 127-37.
24. Piccart-Gebhart, M.J., et al., Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *N Engl J Med*, 2005. 353(16): p. 1659-72.
25. Boyle, P., Triple-negative breast cancer: epidemiological considerations and recommendations. *Ann Oncol*, 2012. 23 Suppl 6: p. vi7-12.
26. Perou, C.M., et al., Molecular portraits of human breast tumours. *Nature*, 2000. 406(6797): p. 747-52.
27. Parker, J.S., et al., Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, 2009. 27(8): p. 1160-1167.
28. Koboldt, D.C., et al., Comprehensive molecular portraits of human breast tumours. *Nature*, 2012. 490(7418): p. 61-70.
29. Karthik, G.M., et al., Intra-tumor heterogeneity in breast cancer has limited impact on transcriptomic-based molecular profiling. *Bmc Cancer*, 2017. 17.
30. Prat, A., et al., Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast*, 2015. 24: p. S26-S35.
31. McCarthy, A.M., et al., Breast Cancer With a Poor Prognosis Diagnosed After Screening Mammography With Negative Results. *JAMA Oncol*, 2018. 4(7): p. 998-1001.
32. Rayson, D., et al., Comparison of Clinical-Pathologic Characteristics and Outcomes of True Interval and Screen-Detected Invasive Breast Cancer Among Participants of a Canadian Breast Screening Program: A Nested Case-Control Study. *Clinical Breast Cancer*, 2011. 11(1): p. 27-32.
33. Houssami, N. and K. Hunter, The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *NPJ Breast Cancer*, 2017. 3: p. 12.
34. Porter, P.L., et al., Breast tumor characteristics as predictors of mammographic detection: Comparison of interval- and screen-detected cancers. *Journal of the National Cancer Institute*, 1999. 91(23): p. 2020-2028.
35. Gilliland, F.D., et al., Biologic characteristics of interval and screen-detected breast cancers. *Journal of the National Cancer Institute*, 2000. 92(9): p. 743-749.
36. Meshkat, B., et al., A comparison of clinical-pathological characteristics between symptomatic and interval breast cancer. *Breast*, 2015. 24(3): p. 278-82.
37. Li, J.M., et al., Molecular Differences between Screen-Detected and Interval Breast Cancers Are Largely Explained by PAM50 Subtypes. *Clinical Cancer Research*, 2017. 23(10): p. 2584-2592.
38. Holm, J., et al., Risk factors and tumor characteristics of interval cancers by mammographic density. *J Clin Oncol*, 2015. 33(9): p. 1030-7.
39. Eriksson, L., et al., Mammographic density and molecular subtypes of breast cancer. *Br J Cancer*, 2012. 107(1): p. 18-23.
40. Antoni, S., et al., Is mammographic density differentially associated with breast cancer according to receptor status? A meta-analysis. *Breast Cancer Research and Treatment*, 2013. 137(2): p. 337-347.
41. Velasquez Garcia, H.A., et al., Mammographic density parameters and breast cancer tumor characteristics among postmenopausal women. *Breast Cancer (Dove Med Press)*, 2019. 11: p. 261-271.
42. Ahern, T.P., et al., Family History of Breast Cancer, Breast Density, and Breast Cancer Risk in a US Breast Cancer Screening Population. *Cancer Epidemiology Biomarkers & Prevention*, 2017. 26(6): p. 938-944.

43. Clamp, A., S. Danson, and M. Clemons, Hormonal risk factors for breast cancer: identification, chemoprevention, and other intervention strategies. *Lancet Oncology*, 2002. 3(10): p. 611-619.
44. Rojas, K. and A. Stuckey, Breast Cancer Epidemiology and Risk Factors. *Clin Obstet Gynecol*, 2016. 59(4): p. 651-672.
45. Barnard, M.E., C.E. Boeke, and R.M. Tamimi, Established breast cancer risk factors and risk of intrinsic tumor subtypes. *Biochimica Et Biophysica Acta-Reviews on Cancer*, 2015. 1856(1): p. 73-85.
46. Toss, A., et al., The impact of reproductive life on breast cancer risk in women with family history or BRCA mutation. *Oncotarget*, 2017. 8(6): p. 9144-9154.
47. Holm, J., et al., Assessment of Breast Cancer Risk Factors Reveals Subtype Heterogeneity. *Cancer Res*, 2017. 77(13): p. 3708-3717.
48. Teugels, E. and S. De Brakeleer, An alternative model for (breast) cancer predisposition. *NPJ Breast Cancer*, 2017. 3: p. 13.
49. Mucci, L.A., et al., Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *Jama-Journal of the American Medical Association*, 2016. 315(1): p. 68-76.
50. Czene, K., P. Lichtenstein, and K. Hemminki, Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish family-cancer database. *International Journal of Cancer*, 2002. 99(2): p. 260-266.
51. Moller, S., et al., The Heritability of Breast Cancer among Women in the Nordic Twin Study of Cancer. *Cancer Epidemiol Biomarkers Prev*, 2016. 25(1): p. 145-50.
52. Miki, Y., et al., A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, 1994. 266(5182): p. 66-71.
53. Wooster, R., et al., Identification of the breast cancer susceptibility gene BRCA2. *Nature*, 1995. 378(6559): p. 789-92.
54. Shiovitz, S. and L.A. Korde, Genetics of breast cancer: a topic in evolution. *Annals of Oncology*, 2015. 26(7): p. 1291-1299.
55. Mavaddat, N., et al., Genetic susceptibility to breast cancer. *Mol Oncol*, 2010. 4(3): p. 174-91.
56. Easton, D.F., et al., Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. *New England Journal of Medicine*, 2015. 372(23): p. 2243-2257.
57. Chandler, M.R., E.P. Bilgili, and N.D. Merner, A Review of Whole-Exome Sequencing Efforts Toward Hereditary Breast Cancer Susceptibility Gene Discovery. *Human Mutation*, 2016. 37(9): p. 835-846.
58. Manolio, T.A., et al., Finding the missing heritability of complex diseases. *Nature*, 2009. 461(7265): p. 747-53.
59. Easton, D.F., et al., Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 2007. 447(7148): p. 1087-93.
60. Lilyquist, J., et al., Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. *Cancer Epidemiol Biomarkers Prev*, 2018. 27(4): p. 380-394.
61. Zhang, H., et al., Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet*, 2020. 52(6): p. 572-581.
62. Michailidou, K., et al., Association analysis identifies 65 new breast cancer risk loci. *Nature*, 2017. 551(7678): p. 92-+.
63. Milne, R.L., et al., Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nature Genetics*, 2017. 49(12): p. 1767-1778.
64. Mavaddat, N., et al., Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants. *Jnci-Journal of the National Cancer Institute*, 2015. 107(5).
65. Gail, M.H., Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *J Natl Cancer Inst*, 2008. 100(14): p. 1037-41.

66. Dite, G.S., et al., Using SNP genotypes to improve the discrimination of a simple breast cancer risk prediction model. *Breast Cancer Res Treat*, 2013. 139(3): p. 887-96.
67. Mavaddat, N., et al., Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet*, 2019. 104(1): p. 21-34.
68. Zuk, O., et al., Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A*, 2014. 111(4): p. E455-64.
69. Dong, L., et al., Detection of novel germline mutations in six breast cancer predisposition genes by targeted next-generation sequencing. *Hum Mutat*, 2018. 39(10): p. 1442-1455.
70. Kraemer, D., et al., Prevalence of genetic susceptibility for breast and ovarian cancer in a non-cancer related study population: secondary germline findings from a Swiss single centre cohort. *Swiss Med Wkly*, 2019. 149: p. w20092.
71. Lu, H.M., et al., Association of Breast and Ovarian Cancers With Predisposition Genes Identified by Large-Scale Sequencing. *JAMA Oncol*, 2019. 5(1): p. 51-57.
72. Patel, A.P., et al., Association of Rare Pathogenic DNA Variants for Familial Hypercholesterolemia, Hereditary Breast and Ovarian Cancer Syndrome, and Lynch Syndrome With Disease Risk in Adults According to Family History. *JAMA Netw Open*, 2020. 3(4): p. e203959.
73. Colas, C., et al., "Decoding hereditary breast cancer" benefits and questions from multigene panel testing. *Breast*, 2019. 45: p. 29-35.
74. Teo, Z.L., et al., Tumour morphology predicts PALB2 germline mutation status. *British Journal of Cancer*, 2013. 109(1): p. 154-163.
75. Roy, R., J. Chun, and S.N. Powell, BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nature Reviews Cancer*, 2012. 12(1): p. 68-78.
76. Honrado, E., et al., Pathology and gene expression of hereditary breast tumors associated with BRCA1, BRCA2 and CHEK2 gene mutations. *Oncogene*, 2006. 25(43): p. 5837-5845.
77. Heikkinen, T., et al., The Breast Cancer Susceptibility Mutation PALB2 1592delT Is Associated with an Aggressive Tumor Phenotype. *Clinical Cancer Research*, 2009. 15(9): p. 3214-3222.
78. Prat, A., et al., Molecular features of the basal-like breast cancer subtype based on BRCA1 mutation status. *Breast Cancer Res Treat*, 2014. 147(1): p. 185-91.
79. Turner, N.C. and J.S. Reis-Filho, Basal-like breast cancer and the BRCA1 phenotype. *Oncogene*, 2006. 25(43): p. 5846-53.
80. Decker, B., et al., Rare, protein-truncating variants in ATM, CHEK2 and PALB2, but not XRCC2, are associated with increased breast cancer risks. *J Med Genet*, 2017. 54(11): p. 732-741.
81. Shimelis, H., et al., Triple-Negative Breast Cancer Risk Genes Identified by Multigene Hereditary Cancer Panel Testing. *J Natl Cancer Inst*, 2018. 110(8): p. 855-862.
82. Buys, S.S., et al., A study of over 35,000 women with breast cancer tested with a 25-gene panel of hereditary cancer genes. *Cancer*, 2017. 123(10): p. 1721-1730.
83. Couch, F.J., et al., Inherited mutations in 17 breast cancer susceptibility genes among a large triple-negative breast cancer cohort unselected for family history of breast cancer. *J Clin Oncol*, 2015. 33(4): p. 304-11.
84. Li, N., et al., Evaluating the breast cancer predisposition role of rare variants in genes associated with low-penetrance breast cancer risk SNPs. *Breast Cancer Res*, 2018. 20(1): p. 3.
85. Decker, B., et al., Targeted Resequencing of the Coding Sequence of 38 Genes Near Breast Cancer GWAS Loci in a Large Case-Control Study. *Cancer Epidemiol Biomarkers Prev*, 2019. 28(4): p. 822-825.
86. Grivennikov, S.I., F.R. Greten, and M. Karin, Immunity, inflammation, and cancer. *Cell*, 2010. 140(6): p. 883-99.

87. Li, J., et al., 2q36.3 is associated with prognosis for oestrogen receptor-negative breast cancer patients treated with chemotherapy. *Nat Commun*, 2014. 5: p. 4051.
88. Lei, J., et al., Assessment of variation in immunosuppressive pathway genes reveals TGFBR2 to be associated with prognosis of estrogen receptor-negative breast cancer after chemotherapy. *Breast Cancer Res*, 2015. 17: p. 18.
89. Roederer, M., et al., The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis. *Cell*, 2015. 161(2): p. 387-403.
90. Orru, V., et al., Genetic variants regulating immune cell levels in health and disease. *Cell*, 2013. 155(1): p. 242-56.
91. Parkes, M., et al., Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet*, 2013. 14(9): p. 661-73.
92. Hemminki, K., et al., Effect of autoimmune diseases on risk and survival in female cancers. *Gynecol Oncol*, 2012. 127(1): p. 180-5.
93. Ludvigsson, J.F., et al., Reduced risk of breast, endometrial and ovarian cancer in women with celiac disease. *Int J Cancer*, 2012. 131(3): p. E244-50.
94. Askling, J., et al., Cancer incidence in a population-based cohort of individuals hospitalized with celiac disease or dermatitis herpetiformis. *Gastroenterology*, 2002. 123(5): p. 1428-35.
95. Gratten, J. and P.M. Visscher, Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med*, 2016. 8(1): p. 78.
96. Nyholt, D.R., SECA: SNP effect concordance analysis using genome-wide association summary results. *Bioinformatics*, 2014. 30(14): p. 2086-8.
97. Lee, S.H., et al., Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 2012. 28(19): p. 2540-2542.
98. Bulik-Sullivan, B., et al., An atlas of genetic correlations across human diseases and traits. *Nat Genet*, 2015.
99. Nik-Zainal, S., From genome integrity to cancer. *Genome Med*, 2019. 11(1): p. 4.
100. Pranavathiyani, G., et al., Integrated transcriptome interactome study of oncogenes and tumor suppressor genes in breast cancer. *Genes Dis*, 2019. 6(1): p. 78-87.
101. Lee, E.Y. and W.J. Muller, Oncogenes and tumor suppressor genes. *Cold Spring Harb Perspect Biol*, 2010. 2(10): p. a003236.
102. Greaves, M. and C.C. Maley, Clonal evolution in cancer. *Nature*, 2012. 481(7381): p. 306-13.
103. Paduch, R., Theories of cancer origin. *Eur J Cancer Prev*, 2015. 24(1): p. 57-67.
104. Vogelstein, B., et al., Cancer genome landscapes. *Science*, 2013. 339(6127): p. 1546-58.
105. Tomasetti, C., L. Li, and B. Vogelstein, Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 2017. 355(6331): p. 1330-1334.
106. Bozic, I., et al., Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci U S A*, 2010. 107(43): p. 18545-50.
107. Haber, D.A. and J. Settleman, Cancer: drivers and passengers. *Nature*, 2007. 446(7132): p. 145-6.
108. Bailey, M.H., et al., Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, 2018. 174(4): p. 1034-1035.
109. Gatenby, R.A., J.J. Cunningham, and J.S. Brown, Evolutionary triage governs fitness in driver and passenger mutations and suggests targeting never mutations. *Nat Commun*, 2014. 5: p. 5499.
110. Wodarz, D., A.C. Newell, and N.L. Komarova, Passenger mutations can accelerate tumour suppressor gene inactivation in cancer evolution. *J R Soc Interface*, 2018. 15(143).

111. Helleday, T., S. Eshtad, and S. Nik-Zainal, Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet*, 2014. 15(9): p. 585-98.
112. Liu, H., et al., Prognostic gene expression signature revealed the involvement of mutational pathways in cancer genome. *J Cancer*, 2020. 11(15): p. 4510-4520.
113. Samstein, R.M., et al., Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet*, 2019. 51(2): p. 202-206.
114. Hanahan, D. and R.A. Weinberg, The hallmarks of cancer. *Cell*, 2000. 100(1): p. 57-70.
115. Hanahan, D. and R.A. Weinberg, Hallmarks of Cancer: The Next Generation. *Cell*, 2011. 144(5): p. 646-674.
116. Dai, X.F., et al., Cancer Hallmarks, Biomarkers and Breast Cancer Molecular Subtypes. *Journal of Cancer*, 2016. 7(10): p. 1281-1294.
117. Torres, L., et al., Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. *Breast Cancer Res Treat*, 2007. 102(2): p. 143-55.
118. Horne, S.D., S.A. Pollick, and H.H. Heng, Evolutionary mechanism unifies the hallmarks of cancer. *Int J Cancer*, 2015. 136(9): p. 2012-21.
119. Fouad, Y.A. and C. Aanei, Revisiting the hallmarks of cancer. *Am J Cancer Res*, 2017. 7(5): p. 1016-1036.
120. Wang, Y., et al., Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 2014. 512(7513): p. 155-60.
121. Yates, L.R., et al., Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med*, 2015. 21(7): p. 751-9.
122. Hamburger, A.W. and S.E. Salmon, Primary bioassay of human tumor stem cells. *Science*, 1977. 197(4302): p. 461-3.
123. Zhang, M., A.V. Lee, and J.M. Rosen, The Cellular Origin and Evolution of Breast Cancer. *Cold Spring Harb Perspect Med*, 2017. 7(3).
124. Luen, S., et al., The genomic landscape of breast cancer and its interaction with host immunity. *Breast*, 2016. 29: p. 241-250.
125. Jiang, X. and D.J. Shapiro, The immune system and inflammation in breast cancer. *Mol Cell Endocrinol*, 2014. 382(1): p. 673-82.
126. Stanton, S.E. and M.L. Disis, Clinical significance of tumor-infiltrating lymphocytes in breast cancer. *Journal for Immunotherapy of Cancer*, 2016. 4.
127. Li, J., et al., Breast cancer genetic risk profile is differentially associated with interval and screen-detected breast cancers. *Ann Oncol*, 2015. 26(3): p. 517-22.
128. Gabrielson, M., et al., Cohort profile: The Karolinska Mammography Project for Risk Prediction of Breast Cancer (KARMA). *Int J Epidemiol*, 2017.
129. Michailidou, K., et al., Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature Genetics*, 2015. 47(4): p. 373-U127.
130. Michailidou, K., et al., Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet*, 2013. 45(4): p. 353-61, 361e1-2.
131. Amos, C.I., et al., The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev*, 2017. 26(1): p. 126-135.
132. Barlow, L., et al., The completeness of the Swedish Cancer Register: a sample survey for year 1998. *Acta Oncol*, 2009. 48(1): p. 27-33.
133. Emilsson, L., et al., Review of 103 Swedish Healthcare Quality Registries. *J Intern Med*, 2015. 277(1): p. 94-136.
134. Ludvigsson, J.F., et al., The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. *Eur J Epidemiol*, 2009. 24(11): p. 659-67.

135. Johansson, L.A. and R. Westerling, Comparing Swedish hospital discharge records with death certificates: implications for mortality statistics. *Int J Epidemiol*, 2000. 29(3): p. 495-502.
136. Lind, H., et al., Breast Cancer Screening Program in Stockholm County, Sweden - Aspects of Organization and Quality Assurance. *Breast Care (Basel)*, 2010. 5(5): p. 353-357.
137. Eriksson, M., et al., A comprehensive tool for measuring mammographic density changes over time. *Breast Cancer Res Treat*, 2018. 169(2): p. 371-379.
138. Li, H., Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*, 2012. 28(14): p. 1838-44.
139. 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.
140. Wang, K., M. Li, and H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 2010. 38(16): p. e164.
141. Goode, E.L., Linkage Disequilibrium, in *Encyclopedia of Cancer*, M. Schwab, Editor. 2011, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 2043-2048.
142. Howie, B.N., P. Donnelly, and J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 2009. 5(6): p. e1000529.
143. Genomes Project, C., et al., A global reference for human genetic variation. *Nature*, 2015. 526(7571): p. 68-74.
144. Trynka, G., et al., Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet*, 2011. 43(12): p. 1193-201.
145. Cortes, A. and M.A. Brown, Promise and pitfalls of the ImmunoChip. *Arthritis Res Ther*, 2011. 13(1): p. 101.
146. Rantalainen, M., et al., Sequencing-based breast cancer diagnostics as an alternative to routine biomarkers. *Sci Rep*, 2016. 6: p. 38037.
147. Saal, L.H., et al., The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. *Genome Med*, 2015. 7(1): p. 20.
148. Patro, R., et al., Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*, 2017. 14(4): p. 417-419.
149. Sonesson, C., M.I. Love, and M.D. Robinson, Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res*, 2015. 4: p. 1521.
150. Lindstrom, L.S., et al., Gene signature model predicts metastatic onset better than standard clinical markers - Nested case-control design uniquely enables enrichment for biologically relevant features. *Cancer Research*, 2013. 73.
151. Cunha, S.I., et al., Endothelial ALK1 Is a Therapeutic Target to Block Metastatic Dissemination of Breast Cancer. *Cancer Res*, 2015. 75(12): p. 2445-56.
152. Cancer Genome Atlas, N., Comprehensive molecular portraits of human breast tumours. *Nature*, 2012. 490(7418): p. 61-70.
153. 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.
154. Borg, A., et al., Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Hum Mutat*, 2010. 31(3): p. E1200-40.
155. Mayakonda, A., et al., Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res*, 2018. 28(11): p. 1747-1756.

156. Dudbridge, F., Power and Predictive Accuracy of Polygenic Risk Scores. *Plos Genetics*, 2013. 9(3).
157. Yanes, T., et al., Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field. *Breast Cancer Res*, 2020. 22(1): p. 21.
158. Milne, R.L., et al., A large-scale assessment of two-way SNP interactions in breast cancer susceptibility using 46,450 cases and 42,461 controls from the breast cancer association consortium. *Hum Mol Genet*, 2014. 23(7): p. 1934-46.
159. Joshi, A.D., et al., Additive interactions between susceptibility single-nucleotide polymorphisms identified in genome-wide association studies and breast cancer risk factors in the Breast and Prostate Cancer Cohort Consortium. *Am J Epidemiol*, 2014. 180(10): p. 1018-27.
160. Tyrer, J., S.W. Duffy, and J. Cuzick, A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in Medicine*, 2004. 23(7): p. 1111-1130.
161. Paquet, E.R. and M.T. Hallett, Absolute assignment of breast cancer intrinsic molecular subtype. *J Natl Cancer Inst*, 2015. 107(1): p. 357.
162. Amara, D., et al., Co-expression modules identified from published immune signatures reveal five distinct immune subtypes in breast cancer. *Breast Cancer Res Treat*, 2017. 161(1): p. 41-50.
163. Wilkerson, M.D. and D.N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 2010. 26(12): p. 1572-3.
164. Solovieff, N., et al., Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet*, 2013. 14(7): p. 483-95.
165. Hekselman, I. and E. Yeger-Lotem, Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat Rev Genet*, 2020. 21(3): p. 137-150.
166. Ritchie, M.E., et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, 2015. 43(7): p. e47.
167. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 2010. 26(1): p. 139-40.
168. McCarthy, D.J., Y. Chen, and G.K. Smyth, Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*, 2012. 40(10): p. 4288-97.
169. Robinson, M.D. and A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, 2010. 11(3): p. R25.
170. Liberzon, A., et al., The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*, 2015. 1(6): p. 417-425.
171. Ackermann, M. and K. Strimmer, A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 2009. 10: p. 47.
172. Maleki, F., et al., Gene Set Analysis: Challenges, Opportunities, and Future Research. *Front Genet*, 2020. 11: p. 654.
173. Varemo, L., J. Nielsen, and I. Nookaew, Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res*, 2013. 41(8): p. 4378-91.
174. Bulik-Sullivan, B.K., et al., LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*, 2015. 47(3): p. 291-5.
175. Skol, A.D., M.M. Sasaki, and K. Onel, The genetics of breast cancer risk in the post-genome era: thoughts on study design to move past BRCA and towards clinical relevance. *Breast Cancer Res*, 2016. 18(1): p. 99.
176. Lee, S., et al., Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*, 2014. 95(1): p. 5-23.

177. Nielsen, F.C., T. van Overeem Hansen, and C.S. Sorensen, Hereditary breast and ovarian cancer: new genes in confined pathways. *Nat Rev Cancer*, 2016. 16(9): p. 599-612.
178. Shahi, R.B., et al., Identification of candidate cancer predisposing variants by performing whole-exome sequencing on index patients from BRCA1 and BRCA2-negative breast cancer families. *BMC Cancer*, 2019. 19(1): p. 313.
179. Green, A.R., et al., MYC functions are specific in biological subtypes of breast cancer and confers resistance to endocrine therapy in luminal tumours. *Br J Cancer*, 2016. 114(8): p. 917-28.
180. Li, Y., et al., Expression patterns of E2F transcription factors and their potential prognostic roles in breast cancer. *Oncol Lett*, 2018. 15(6): p. 9216-9230.
181. Costa, R.L.B., H.S. Han, and W.J. Gradishar, Targeting the PI3K/AKT/mTOR pathway in triple-negative breast cancer: a review. *Breast Cancer Res Treat*, 2018. 169(3): p. 397-406.
182. Gangrade, A., et al., Preferential Inhibition of Wnt/beta-Catenin Signaling by Novel Benzimidazole Compounds in Triple-Negative Breast Cancer. *Int J Mol Sci*, 2018. 19(5).
183. Shieh, Y., et al., Breast Cancer Screening in the Precision Medicine Era: Risk-Based Screening in a Population-Based Trial. *J Natl Cancer Inst*, 2017. 109(5).
184. Kirsh, V.A., et al., Tumor characteristics associated with mammographic detection of breast cancer in the Ontario breast screening program. *J Natl Cancer Inst*, 2011. 103(12): p. 942-50.
185. Cabioglu, N., et al., Poor Biological Factors and Prognosis of Interval Breast Cancers: Long-Term Results of Bahcesehir (Istanbul) Breast Cancer Screening Project in Turkey. *JCO Glob Oncol*, 2020. 6: p. 1103-1113.
186. Krishnan, K., et al., Mammographic density and risk of breast cancer by mode of detection and tumor size: a case-control study. *Breast Cancer Res*, 2016. 18(1): p. 63.
187. Nguyen, T.L., et al., Interval breast cancer risk associations with breast density, family history and breast tissue aging. *Int J Cancer*, 2020. 147(2): p. 375-382.
188. Chuang, S.L., et al., Using tumor phenotype, histological tumor distribution, and mammographic appearance to explain the survival differences between screen-detected and clinically detected breast cancers. *APMIS*, 2014. 122(8): p. 699-707.
189. Shieh, Y., E. Ziv, and K. Kerlikowske, Interval breast cancers - insights into a complex phenotype. *Nat Rev Clin Oncol*, 2020. 17(3): p. 138-139.
190. Edechi, C.A., et al., Regulation of Immunity in Breast Cancer. *Cancers (Basel)*, 2019. 11(8).
191. Viljamaa, M., et al., Malignancies and mortality in patients with coeliac disease and dermatitis herpetiformis: 30-year population-based study. *Dig Liver Dis*, 2006. 38(6): p. 374-80.
192. Dunn, G.P., et al., Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol*, 2002. 3(11): p. 991-8.
193. Szklarczyk, D., et al., STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*, 2019. 47(D1): p. D607-D613.