THE UNIVERSITY OF CHICAGO

THE INFLUENCE OF GERMLINE GENETIC VARIATION ON EARLY ONSET BREAST CANCER INCIDENCE AND MORTALITY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PUBLIC HEALTH SCIENCES

BY
MOLLY SCANNELL BRYAN

CHICAGO, ILLINOIS
DECEMBER 2016

Copyright © 2016 by Molly Scannell Bryan

All rights reserved

To Mom, Dad, Kyle, and Penelope. I know the infinite because of your love and support.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

For their guidance, input, and edits, I thank my thesis committee, Habibul Ahsan, Maria Argos, Lin Tong, and Dan Nicolae.

For their tecnical help with the TCGA data, I thank Dezheng Huo and Zhenyu Zhang.

For the incredible support and encouragement from the Department of Public Health Sciences, particularly: Liane Kurina, Diane Lauderdale, Brandon Pierce, Tamara Konetzka, James Dignam, Robert Gibbons, and Jameca Lozano-Lott.

For their help with the Chicago-based data, I thank Lin Tong, Stephanie Melkonian, Brandon Pierce, Farzana Jasmine, Muhammad Kibriya, and Chenan Zhang.

For her support and encouragement, I thank my friend and classmate, Chenan Zhang.

For pre-BSD colleague and friends: Jen Niedziela, As Meninas, Donna Monahan, David Bittle, and Pamela Hayward.

For raising me with love and curiosity, I thank my parents David and Mary Scannell.

And my husband Kyle and daughter Penelope, who have both supported my education in all ways for as long as they have known me.

ABSTRACT

In the United States, breast cancer is the most frequently diagnosed non-skin cancer in women, and one in five women who are diagnosed develop breast cancer before age 50. Germline genetic variation is a known risk factor for breast cancer risk, and a suspected risk factor for breast cancer mortality, but previous investigations have not comprehensively identified all of the genetic variation that is expected to be associated with breast cancer. One possible explanation for this gap in knowledge is the only relatively recent ability to investigate the effect of rare germline genetic variation, which up until recently has been too expensive and technically challenging to measure in the a large number of participants that are necessary for genetic epidemiologic studies, and the methodological challenges of identifying rare variants.

This thesis uses three complementary methods (single marker regression analysis, SKAT-O gene-based tests, and candidate gene) to identify individual risk loci and three additional complementary methods (Kriging whole genome prediction, polygenic risk scores, and whole genome heritability estimates) to predict breast cancer risk and breast cancer mortality using a population of women who were diagnosed with breast cancer before the age of 50. Suggestively associated risk loci were examined for evidence of replication using an independent sample.

For breast cancer risk, the identification analyses find three genes in which variation is associated with risk of breast cancer: FGFR2 (discovery $p = 2.18 \cdot 10^{-5}$; replication $p < 10^{-30}$), NEK10 (discovery $p = 1.20 \cdot 10^{-3}$; replication $p < 10^{-30}$), and MKL1 (discovery $p = 2.62 \cdot 10^{-4}$; replication $p < 10^{-30}$). Previous studies had identified loci near each of these genes as being associated with breast cancer risk, but conditional analyses indicate that the associations in the MKL1 and NEK10 genes are driven by risk loci distinct from those previously reported, and are driven by risk loci that would not have been identified

using a single variant regression. The genetic data alone is able to predict breast cancer risk with an AUC of 0.618 (95% CI 0.610-0.629). When the influence of a limited set of non-genetic predictors is also incorporated, the combined model is able to predict breast cancer risk with an AUC of 0.655 (95% CI: 0.649-0.660). This combined model is a significant improvement over models that include only the genetic information or only the non-genetic risk factors.

In contrast to the analyses of the genetic determinants of breast cancer development, this analysis does not find any compelling evidence that breast cancer mortality is strongly driven by germline genetics that could be measured by our study.

The identified genes all represent possible pharmacological targets for cancer chemoprevention. The prediction model for breast cancer risk improves upon existing methods of prediction, and is strong enough to be useful at the population level. From a clinical perspective, the model still has low levels of discrimination, but may be strong enough to be used in very specific scenarios, such as interpretation of ambiguous screening results, or to help individuals to understand their personal risk when considering other medical treatments that may increase the risk of breast cancer such as hormone replacement therapy or hormone-assisted reproductive therapy. In the context of breast cancer prognosis, these investigations support other lines of evidence that suggest that for many women who are diagnosed with breast cancer, germline genetic variation does not strongly influence the risk of mortality.

CHAPTER 1

INTRODUCTION

1.1 Background

Breast cancer is the most common cancer in women, and one in eight American women will develop breast cancer over her lifetime.[1] Almost twenty five percent of women diagnosed with breast cancer eventually die of the disease,[2] and fear of recurrence and mortality lowers the quality of life for women who are diagnosed.[3–6] Breast cancer is a heterogeneous disease that is caused by and progresses due to a complex array of risk factors. While any individual woman's cancer develops due to the unique set of exposures that she accrues over a lifetime, these individual exposures give rise to patterns of risk. This thesis investigates in-depth the risk factor of germline genetic variation, with a focus on germline genetic variants that are rare and located within gene regions. The patterns of this risk factor have been in not comprehensively described in breast cancer risk and prognosis, and a better characterization of this risk factor will improve knowledge of biological mechanisms of breast cancer, identify possible targets for therapeutic intervention, and translate into more precise estimators of risk. Each analytic component of this thesis is motivated by one of two complementary goals: to identify genetic risk factors for early onset breast cancer, and to predict the overall risk women have from the disease. The results of this thesis develop a cohesive narrative that identifies loci that are associated with breast cancer risk and prognosis, describes the underlying characteristics of the genetic determinants of breast cancer development and progression, and comprehensively quantifies the genetic contribution to a given woman's risk of breast cancer.

Results from previous studies and biological plausibility indicate that germline genetic variation can influence the risk breast cancer development. The exact mechanism of how

genetics influences risk is not fully understood, but genetic variation may predispose a person to genomic instability, provide a fertile cellular environment for a tumor, or impede immune response to proto-oncogenic cells.[7]

The association between germline genetic variation and prognosis is less well established than the association with risk, but previous research has implicated particular risk variants, and the relationship is biologically plausible. Germline genetic variation may affect a patient's ability to metabolize a drug, which in turn can affect survival by altering the amount of available active metabolites of pharmaceutical treatments, or increasing the probability of treatment-limiting adverse events.[8–12] Similarly, germline genetics may be responsible for a cellular environment that favors metastases, or otherwise aggressive tumors, and may alter cellular functions that are crucial to tumor proliferation such as angiogenesis, growth signaling, telomere length, inflammation, immune response, DNA repair, apoptosis, and cell cycle control.[13–22]

Age has a complex relationship with the risk and prognosis of breast cancer. While the causal nature of the relationship is not completely understood, women who are diagnosed before the age of 50 (one in five of those diagnosed[2]) have worse outcomes than those who are diagnosed later in life.[23–28] Some non-genetic risk factors, such as reproductive history and obesity, change the direction of their effect in women who are diagnosed early when compared to their effect in women who are diagnosed later.[29] While several germline genetic variants have been implicated in the risk of the late onset disease, their effect on the development of the early onset disease has not been well characterized. Better understanding of this relationship between age breast cancer etiology can help to both understand the underlying biological mechanisms of breast cancer, and also help to develop more precise risk scores for women.

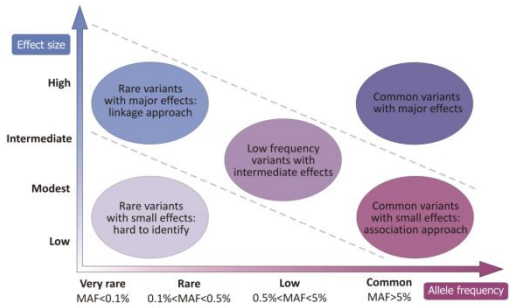
1.2 Identification of Risk Variants in Breast Cancer

Given the many possible pathways by which germline genetic variation may influence breast cancer risk and prognosis, analyses that identify individual risk loci may illuminate the cellular pathways and implicate cellular processes that are involved in oncogenesis or metastases. This can improve understanding of the underlying biological mechanisms that are integral to breast cancer development. Beyond this biological insight, genes that are associated with breast cancer may suggest targets for future pharmaceutical interventions for chemoprevention or treatment of cancer.

Genetic data possess several distinct characteristics that must be accounted for when attempting to identify germline genetic variation as a risk factor in any disease, and it is necessary to use study designs and statistical methods that account for these characteristics. One characteristic of genetic data that distinguishes it from other epidemiologic risk factors is its high dimensionality. Each study participant has three billion possible variants that may be associated with disease, in addition to other genetic abnormalities such as copy number variations and insertions and deletions. While many nucleotides are constant, the number of loci that do vary is still much larger than can be handled by many statistical methods. Methods have been developed that reduce this high dimensionality that incorporate prior information to limit the search for associations. Investigations that scan genome-wide for evidence of risk loci must balance the use of prior information while also remaining agnostic enough to allow the data to implicate novel loci.

A framework that can be used to approach this balance considers two separate characteristics of genetic variants that truly are risk factors: the relationship between its rarity in the population and the magnitude of its effect on the trait, and its predicted functionality. These characteristics of the causal variants have suggest the appropriate study design

Figure 1.1. Possible Distributions of the Strength of Variant Associations with their Frequencies



and statistical method that is required to identify them, but these characteristics are often unknown for most diseases, including breast cancer.

1.2.1 Risk as a Function of Variant Rarity and Effect Size

A framework to consider the relationship between the rarity of the variant and the magnitude of its effect on the trait was outlined by Manolio et al.[30] and McCarthy et al.[31] (Figure 1.1, from Zemunik and Boraska[32]). For polygenic diseases that are affected by multiple risk variants, causal variants have been discovered at multiple places along this effect size-rarity distribution. Since no statistical method is optimal to detect risk variants for all combinations of effect size and rarity, it is common that multiple complementary statistical methods will be required to fully characterize the genetic variants that drive polygenic disease.

In breast cancer risk, prior research has established that risk variants are located at least two quadrants of this spectrum-rare variants of large effect, and common variants of modest effect. Rare variants of large effect, such as mutations in BRCA1, BRCA2, and TP53,[33] were largely discovered by linkage approaches that studied affected families. These implicated genes are responsible for cellular processes such as DNA repair and cell cycle control, and their identification as risk loci has confirmed the importance of these

processes in oncogenesis. The risk variants are rare in the general population, and therefore do not dominate population-level risk estimates, but for the people that carry those variants, they confer a large risk of breast cancer.

Common variants that have a modest to low effect have also been implicated as risk factors for both breast cancer risk and prognosis. For variants such as these, a person can carry any single risk variant without having a risk that is dramatically increased, but, since these variants are common, collectively they can contribute to a large risk burden. Single marker regression association studies in genome-wide association study (GWAS) frameworks successfully identified 128 variants throughout the genome that increase a woman's risk of breast cancer,[34] and a smaller number of variants have been suggestively identified as possibly associated with breast cancer mortality.

However, despite these successes, there still remains missing heritability in breast cancer risk. Despite studies and meta-analyses of 50,000 participants or more, variants that have been identified only contribute about half of the total expected risk due to genetics that is expected from family studies.[35] This suggests that the variants responsible for this missing heritability may be characterized other combinations of effect size and rarity.

This thesis investigates variants that are rare in the general population, and confer intermediate-to-modest risk of breast cancer (the center of Figure 1.1). This class of variants requires a different statistical approach to identify them. They cannot be interrogated by single marker regression analyses of GWASs, either because they are not present at a high enough frequency to be observed in studies of realistic sample sizes, or, if they are seen, they are too so rare for a logistic regression to produce well defined odds ratios for the effect size of that variant. Their modest effect size also makes them difficult to identify in linkage analyses. Current statistical approaches to identify the effects of this class of variation require additional, sometimes restrictive, assumptions. These assumptions are necessary in order to interrogate rare variations, but do place certain limitations on the

interpretations of the analyses, and the appropriateness of these assumptions needs to be evaluated in the context of each disease of interest.[36]

In this case of breast cancer, one plausible assumption is to posit that variants within the same gene act collectively to alter risk. This behavior has already been observed in the BRCA1 and BRCA2 genes where variants at multiple loci all are capable of inactivating the gene product, damaging the DNA repair capacity of the cell, and increasing breast cancer risk.[36] If variants do act in this collective manner, these genes can be identified by implementing a family of tests known as gene-based tests. Gene-based tests shift the hypothesis from the variant-level to the gene-level.

Gene-based tests do have some limitations, most prominently that variants that are outside gene regions (roughly 98% of the genome) cannot be interrogated. However, the tests have benefits as well. Gene-based tests can incorporate rare variation, unlike single marker regression analyses, and the method can be applied to study participants that are selected from the general population using a standard epidemiologic case/control study design. The direct functional relevance of gene products can make results of gene-based tests easier to interpret than the result of single marker regression tests. While gene-based tests will not be able to identify all causal loci, they will be able to well-interrogate gene regions, which are highly likely to harbor at least some of the variation that is associated with disease. These benefits justify their use in circumstances where prior biological understanding suggests that low frequency variants of modest effect do affect the risk of disease. Evidence for this includes diseases (such as breast cancer) where multiple well-powered single marker regression analyses have only identified causal loci, but their collective effect still falls sort of the estimated heritability in the trait.

A wide variety of gene-based tests have been proposed, and their appropriateness depends on the sparsity of causal variants, and their distribution throughout the genome. In circumstances where the assumptions of the given gene-based test match the genetic archi-

tecture of the disease under study, gene-based tests are an effective method to implicate a gene in disease. The optimal sequence kernel association test (SKAT-O[37]) has emerged as a strong gene-based test which is robust to some deviations from its underlying assumptions. SKAT-O has not been used in the context of any breast cancer phenotype to test for variation within a gene that can collectively act to increase risk.

1.2.2 Risk as a Function of Variant Functionality

In addition to considering the frequency/effect size distribution of the causal variants, it can also be helpful to consider the predicted functionality of possible causal variants. The predicted functionality of a variant can help to suggest variants that are more likely to be causal. Implicitly, gene-based tests incorporate variant functionality by restricting to variants that can be grouped to a single gene, but more nuanced classifications are also possible. An extreme way to incorporate functionality is to restrict the analysis to variants that are predicted to cause a change in amino acid translation. If variants that confer risk of breast cancer are mostly variants that cause changes in amino acid translation, then future studies would be able to focus on just those variants. By only assaying them, the multiple testing burden would be reduced and fewer truly causal variants would be identified as not associated. However, there is strong evidence that in the case of breast cancer, disease causing variants act through additional mechanisms of action besides changes in amino acid translation. Many of the variants that have been identified through single marker regression tests are exonic variants that do not cause changes in protein coding (although they may tag a protein-coding variant by way of linkage disequilibrium), or are within a gene region that are not in the exons (such as intronic variants), or are intergenic.[34]

For these reasons, analyses that only focus on variants that are predicted to alter amino acid translation are expected to miss many truly causal variants. A more agnostic approach

would be to up-weight variants that were likely to be causal, while still keeping those that have less strong prior evidence of association. The SKAT-O test can incorporate variant-level weights. In most weighting scenarios, incorporating even incorrect weights will not introduce bias or reduce power,[38] and incorporating weights that do reflect the true association of a variant with disease can substantially increase power.[37]

However, the optimal method to translate past information on predicted functionality into weights is still a matter of study. Several studies that have used gene-based tests have weighted variants based only on their rareness. Other studies do not attempt to weight at all, and restrict their analyses to variants that are either rare or predicted to cause protein changes. However, given the late onset of breast cancer, the variants that are associated with the disease would have a smaller-than-expected effect on fitness, and therefore may not be as rare as would be expected from evolutionary models. A method that incorporates a more nuanced understanding of predicted pathogenicity would be preferable. However, several annotations and pathogenicity scores have been developed, and it is not clear which annotation is best able to highlight variants that are likely to be involved in disease. Single-dimensional annotations classify variants based on any of several features, such as predicted functionality, evidence of evolutionary constraint, previous association with disease, and evidence of regulatory function. Translating these concepts into a single weight that incorporates each of the dimensions has not been widely done. The Combined Annotation Dependent Depletion (CADD) score is an overall deleterious score that incorporates each of these single-dimensional annotations by estimating the extent each is able to predict whether a variant has reach fixation in the general population.[39] Weighting by an overall deleteriousness score, as created by the developers of CADD would allow gene-based tests to include all variants near gene regions, and would reflect the multi-dimensional characteristics that define the relationship between germline genetic variation and disease. This kind of deleterious score allows for an explicit incorporation of evolutionary and other

constraints on the test, which is expected to be necessary for gene-based tests to perform optimally.[40]

1.3 Prediction of Breast Cancer Risk and Prognosis

The above discussion focuses on the ability to identify particular loci or genes that are associated with disease, with that will highlight a particular cellular mechanism as being associated with disease. A complementary question involves the ability to predict an individual woman's risk of breast cancer using genetic data. The goal of prediction is less to identify the causal risk factors, but rather to infer their collective effect on risk and prognosis to produce an individual quantification of risk.

Both breast cancer risk and mortality have several known non-genetic risk factors that are reproducibly associated with disease. For both outcomes, the predictive power of these models is modest. These models can be used to predict the risk of a population, but their low discrimination makes them less relevant for individual clinical risk decisions.[41] Prediction models that incorporate germline genetic variation can allow the whole genome to be used to collectively infer the total burden of germline genetic variation on breast cancer risk and prognosis, and will lead to a better prediction of those who are at high risk of developing the disease or dying from it.

There are several methods that have been proposed to incorporate germline genetic information into a prediction model. Genetic relatedness matrix restricted maximum likelihood-based (GREML) prediction models, including Kriging,[42] allow for genetic variation throughout the genome to contribute to prediction. GREML models do not require that the causal variants already be identified in order to contribute to the model. Breast cancer risk has already been determined to be a polygenic disease, in that multiple variants contribute to any woman's individual risk. Moreover, the missing heritability in breast cancer indicates

that some of the causal variants are not yet identified. These features make the Kriging method of prediction a strong choice for breast cancer risk.

Kriging can be implemented in a way that allows for different sets of variants to be grouped together. The form of the variants' collective association with disease can differ between these groups. These groups can be selected to reflect the different annotated functionalities of the possible risk variants throughout the genome. While not as comprehensive as weighting (as the variants can only be divided into a relatively small number of groups), this method of whole genome prediction does incorporate prior information about the expected predicted functionality of a disease, and has been successful in incorporating germline genetic variation into prediction of other traits.

1.4 Analysis

With the preceding as background, this thesis investigates the genetic determinants of breast cancer risk and prognosis. The analyses focus on two complementary lines of questioning. First, to identify genes that contain variants (including rare variants of modest effect) that collectively contribute to risk and prognosis, and second, to predict overall breast cancer risk using genome-wide measures of variation.

The primary data that are available to investigate these questions come from ten ongoing studies designed to assess the risk factors associated with early onset breast cancer. Participants in these studies are women of European descent who were 51 years or younger at the time of their diagnosis (for cases) or enrollment (for controls) and not known to carry pathogenic mutations in the genes BRCA1 or BRCA2. DNA was available for 4914 participants (3,876 cases and 1,038 controls) through blood draws. Each of these participants was genotyped on an Illumina exome-chip genotyping array that measured 238,524 variants. The chip was designed to more closely interrogate often-rare variants in gene regions,

with particular emphasis on nonsynonymous variants. A subset of 3357 participants additionally was genotyped on an Illumina genome-wide genotyping array, which interrogated 3,310,158 variants after imputation.

SKAT-O is implemented to identify genes associated with risk (Chapter 2) and prognosis (Chapter 3), and CADD weights are applied to the variants for each analysis. Any genes that are identified are subject to conditional analyses which determine whether the associations are driven by (1) variants that are common enough or of strong enough effect size to be identified through a GWAS framework in the same participants, or (2) variants that are already known to be associated with breast cancer phenotypes through previously published work.

Any genes that are identified are also examined for evidence of replication in an independent data set, which also gives evidence on whether the implicated genes are also involved in the genetic architecture for women who are diagnosed later. The replication data for the risk analyses are summary statistics from a meta-analysis of a single marker logistic regression case control studies of breast cancer risk. This meta-analysis combined data from 15,863 breast cancer cases and 41,461 controls and interrogated 2,608,508 variants after imputation. The prognosis replication sample is derived from the participants of The Cancer Genome Atlas (TCGA) study (data generated by the TCGA Research Network: http://cancergenome.nih.gov/). The women included were matched on race, and ultimately 711 cases and 6,087,804 genotyped or imputed variants are used for prognosis replication.

In addition to identifying genes associated with overall prognosis, Chapter 3 investigates whether there is evidence that variation in genes is associated with known prognostic indicators that can be discerned at the time of diagnosis: estrogen receptor status, progesterone receptor status, HER2 status, grade, and stage. Any genes that these analyses identify may be responsible for the development of particular subtypes of cancer that tend to be more aggressive. A susceptibility to these histopathologically distinct tumors may suggest

personalized targets for chemoprevention.. Additionally, Chapter 3 examines whether genetic variation in any of the CYP family of genes is associated with mortality, and if this association differs for women with estrogen receptor positive tumors. It has been a matter of recent controversy whether mutations in the CYP family of genes are associated with poorer outcomes.[43,44] These genes encode enzymes that metabolize tamoxifen, an effective treatment for women with estrogen receptor positive tumors, but it is unclear whether an altered ability to metabolize tamoxifen translates into higher mortality, and the analyses of Chapter 3 look for evidence of this association.

The whole genome prediction model (Chapter 4) is carried out using the Kriging method, and separates the variants into categories based on their rareness and predicted functionality. Chapter 4 culminates in by presenting an optimal prediction model that combined genetic data with non-genetic predictors. The Kriging method also investigates whether the ability to predict breast cancer risk and prognosis is driven by variants that have already been reported, or whether additional risk variants remain to be identified. This analysis further probes the conclusions from family studies that suggest that there remains undiscovered risk variants for both breast cancer risk and prognosis.

1.5 Gaps in Knowledge and Implications of Results

The investigations of this thesis provide new biological insight into the genetic determinants of breast cancer risk and prognosis. The effect of germline genetic variation on breast cancer risk has been the focus of many studies, but gaps in knowledge remain about both individual risk loci that are involved and the overall genetic influence on risk. Investigations into the genetic determinants of breast cancer mortality have been less widely reported, and this thesis presents multiple complementary investigations into this relationship.

The gene-based analyses of Chapters 2 and 3 are designed to identify genes that are associated with breast cancer risk and prognosis. This can implicate unappreciated cellular processes that affect breast cancer development and progression. Given the young age of breast cancer onset of the participants in this study, the gene-based analyses will also provide evidence on whether genetic influences of the better-studied late onset disease are also risk factors for women who will be diagnosed early. Similarly, these investigations suggest whether the genes that are responsible for increased risk of developing breast cancer are also involved in poorer prognosis in women who have already been diagnosed. While the implications for risk assessment are large for each of these questions, their answers remain unsettled.

There have been a limited number of prior studies that use the GWAS framework of single marker regression association tests to investigate mortality in breast cancer cases at the genome-wide level, and those that have been done were largely underpowered, making it difficult to draw firm conclusions from this previous work. These previous studies found few suggestive associations, most of which were not significant in the original study's replication sample, and none of which have been replicated in subsequent studies. This dearth of prior research on an important health topic may be due to the long amount of time needed to collect mortality data prospectively when compared to the relatively short amount of time GWAS-style studies have been around. It may also reflect a publication bias, where studies that do not find any association are not easily shared publicly.

Given the low effect size of most of the identified associations with breast cancer, most women's individual level of risk is not well defined by her genotype at one risk loci. For that reason, risk models that incorporate all genetic variation are needed to quantify a woman's overall risk of disease. The prediction models of Chapter 4 will improve upon previous work that does not currently incorporate germline genetic variation. Currently, models that do not incorporate genetics can predict breast cancer risk with an area under the receiver

operating characteristic curve (AUC) between 0.6 and 0.7,[45,46] and there have not been any published models that are specific to early onset breast cancer. Similarly, models that predict prognosis without genetic information can predict mortality with an AUC of around 0.7.[47] The Kriging method of whole genome prediction may improve upon this predictive power.

The prediction presented in Chapter 4 also informs the extent to which different classes of predicted variant functionality are likely to contain the undiscovered variants that are associated with breast cancer risk and prognosis. This knowledge can suggest the appropriate methods that can be used to identify the individual risk variants in future studies. While the primary goal of prediction models is not to identify the specific variants that are associated with disease, the Kriging prediction method can be used to characterize the unidentified causal variants. In particular, the results of Chapter 4 suggest whether the variation that drives breast cancer is common or rare, and whether that variation has a particular type of predicted functionality. This knowledge helps to resolve long-running questions about the relative importance of different portions of the genome in the genetic architecture of cancer.

In addition to biological insights, these investigations represent the first application of many statistical tools to the question of breast cancer. Whole genome prediction has not been applied to either breast cancer risk or prognosis, and gene-based tests have only been incorporated twice. These investigations represent only the third study of breast cancer risk that has interrogated rare variation directly. In 2013 Haiman et al.[48] and more recently (September 2016) Haddad et al.[49] and Zhou et al. used exome arrays, but each restricted their analysis to rare putative functional variants rather than weighting, and Haiman used a burden style test rather than a SKAT-O test. A third study[50] that investigated rare variation using gene-based tests in a genome-wide setting in breast cancer also used a burden test, and additionally did not directly interrogate the rare variants in gene regions, but rather

inferred their existence by imputation. Their reported results are therefore limited by the assumptions of the models they used. These study designs leave open questions about rare variation and the effect of age and ancestry which can be answered by the results of this thesis. While the data available for this thesis is of modest sample size, the success or lack thereof of the analytic techniques employed in this thesis can inform whether additional, better powered investigations using gene-based analyses or whole genome prediction are likely to be fruitful.

In no previously published work that applies gene-based tests genome-wide for any disease have the CADD weights been used (one candidate gene study used the CADD score multiplied by minor allele frequency to investigate cardiovascular disease[51]). The results of the analyses that weight with CADD weights suggest whether this method is an appropriate method to increase power by incorporating prior information.

Neither whole genome prediction models nor linear mixed model whole genome heritability estimates, which are based on the same conceptual methodology, have been published for either breast cancer risk or prognosis. The estimations from Chapter 4 therefore put their results into a larger context of heritability estimates. The current understanding of heritability for breast cancer has been estimated from family studies, which may be biased by shared environment.

The goals forwarded by this work, identification and predication, are complementary. The results of both lines of inquiry may result in more efficient, optimized medical care, and could confer many clinical benefits. Identified genes could provide valuable insight into the genetic etiology of breast cancer risk and prognosis. Additionally genes may be identified that can not affect population-level risk but are still important for a given person, and could identify possible targets for pharmaceutical intervention.

Whole-genome prediction can quantify risk for a person without identifying the distinct variant that drives that predictive power and suggest classes of variants that may be

more likely to hold causal variation. When combined with non-genetic information, the prediction model will both more accurately predict population-level risk and also improve upon current risk estimates to move towards prediction models that are clinically actionable. A strong risk model would identify low-risk women who could be screened less often, which would allow them to devote less energy searching for symptoms of a disease they are unlikely to develop. Given the high prevalence of breast cancer, even a modest increase in the total ability to predict risk could potentially impact the interpretation of ambiguous screening results for many women,[52] and could reduce both over-treatment and unidentified tumors. Additionally, a risk model could provide additional information for women who are considering other medical interventions that may increase their breast cancer risk, such as menopausal hormone therapy or hormonal assisted reproductive therapies.[53]

In the context of breast cancer mortality, a stronger prediction model could suggest more aggressive monitoring and treatment for high-risk subgroups of patients, and could also help to identify women who could pursue less aggressive treatments. These classifications would reduce the morbidity associated with exposure to chemotherapies,[54,55] while at the same time identifying those at high risk of mortality who may want to be treated more aggressively.

CHAPTER 2

THE EFFECT OF GERMLINE GENETIC VARIATION IN GENE REGIONS ON THE RISK OF EARLY ONSET BREAST CANCER

2.1 Background

Breast cancer is the most frequently diagnosed cancer in American women,[1] and one in five women who are diagnosed develop breast cancer before age 50.[2] Genetic variation has been identified as a risk factor for breast cancer. It has been hypothesized that germline variants interact with somatic mutations within the tumor during tumorigenesis, and the same somatic mutation may develop into a cancer cell in one women but not another due to germline variation.[7] While the exact mechanism of how genetics influences breast cancer risk is not fully understood, it is plausible that genetic variation may predispose a woman to breast cancer through a predisposition to genomic instability, by providing a fertile cellular environment for a tumor, or by impeding immune response to proto-oncogenic cells.[7] Several non-genetic risk factors such as parity and the use of synthetic hormones confer an increased risk of breast cancer early in life, but offer a protective benefit against the disease later in life. Although the implications for risk assessment are large, it is still unclear the extents to which genetic influences of the better-studied late onset disease are also risk factors for women who will be diagnosed early.

Investigating genetics as a risk factor can be a challenge due to the large number of potential disease-causing variations, and the unknown pathway by which each variation may contribute to disease risk. The most appropriate statistical method to investigate risk will depend on how many variants ultimately are associated with disease,[56] their sparsity throughout the genome,[57] the form of the relationship between the variation and disease risk,[58] their rareness, and the strength of their effect on disease.[30] Each of these charac-

teristics may differ from variant to variant and between diseases, and are often unknown. Single marker regression analysis examine the variants as independent predictors of disease, and is a method employed by genome-wide association studies (GWASs). Single marker regressions have been successful in identifying causal variants for diseases where the causal variants are common enough to include in a regression framework (this threshold will vary depending on the sample size, but variants with a minor allele frequency greater than $\left(\frac{1}{2n}\right)^{\frac{1}{2}}$ are typically included[59]), and variants that have a strong enough association with the disease in question. For context, the study sizes of single studies that have investigated breast cancer phenotypes with single marker regression techniques between 1000 and 5000 cases per study, and the median effect size of genome-wide significant results is an odds ratio of 1.1.[34,60]

However, single marker regressions have limitations. They cannot identify risk loci where variants are too rare or whose effect is too weak, and they are also limited by concerns about type I error rate. Single marker regression analyses conduct a large amount of tests, which requires employing a strict significance thresholds in order to exclude false positives. In many cases, these thresholds can exclude many truly causal variants.[61] There is also evidence that in breast cancer risk, common variants are tagging the effect of rare variants that are not always directly assayed.[62]

Much of the recent research into the genetic determinants of breast cancer has incorporated information from common variation assayed on genome-wide arrays, and the variants that can be reliably imputed from them.[50,63–67] These studies suggest that some of the genes associated with risk of late-onset disease also influence the early onset disease.[50,68–70]However, there still remains "missing heritability" in breast cancer,[35,71–73] where genes that have been identified by research only contribute about half of the total expected risk due to genetics.

Variants in protein-coding regions of the genome are expected to harbor some of the undiscovered variation that is associated with risk of breast cancer. While analyses that focus on gene regions exclude a large percentage of the genome the central role of genes in transcription and ultimately amino acid translation makes variants that reside in genes represent biologically plausible candidates for association with disease,[74] which justifies the use of methods that can well-interrogate these regions, even if other complementary methods will then be required to examine the rest of the genome.

A class of suitable tests has been developed for examining variation within a single region, called set-based tests. Set-based tests shift the hypothesis from whether an individual variant is associated with disease to whether a collection of variants is associated with disease. The sets are often taken to be variants within gene boundaries to give the set an immediate biological interpretation. In the context of genes, if a gene-defined set-based analysis identifies a gene as associated with disease, this suggests that any variation within that gene or gene proxy region collectively contributes to disease. Set-based tests allow for variants that are too rare to test individually to contribute evidence for risk, and also common variants whose effects are too modest to detect using standard single marker regression approaches.

Many set-based tests have been developed that can be implemented as gene-based tests,[37,75–77] with two of the most commonly used being burden tests and the sequence kernel association test (SKAT). Burden tests[76] sum the number of risk alleles within a gene, and then estimate the combined effect of that number of risk variants on the disease. SKAT[37] estimates the effect of each variant within a gene-based on a linear mixed effect model and tests for non-zero variation explained by genetic factors via a variance component approach. The burden test is more powerful than the SKAT test if all of the variants in a gene increase risk of disease. The SKAT test is more powerful than the burden test if the variants within a gene may increase or decrease the disease risk. Since these assump-

tions about the density of risk variants and the direction of their association are typically not known,[78] an omnibus test that combines the test statistics from burden and SKAT was developed, the optimal sequence kernel association test (SKAT-O).[37] SKAT-O calculates both the burden test statistic and a SKAT test statistic for each gene, and then uses the data adaptively to weight and combine the two test statistics by a mixing factor, ρ, which ranges from zero (where the test statistic is equivalent to the SKAT test statistic) to one (where the test statistic is equivalent to the burden test statistic). A value of ρ that is small (less than 0.1) indicates that the relationship between the gene and the risk of breast cancer was better characterized by the assumptions of the SKAT test, and ρ greater than 0.5 indicates that the relationship is better characterized by the assumptions of the burden test. The distribution, effect size, and sparsity of as-yet-unidentified causal variants within gene regions that are associated with breast cancer risk is not totally established (although previously studied single marker regression results indicate that the minor allele at a risk locus can be both protective and deleterious[34]). Given this uncertainty, the omnibus test may be a more appropriate tool to identify genes harboring risk loci than either the SKAT or burden test alone, especially since in most situations, SKAT-O is more powerful than either test alone.[37]

Many studies that have implemented gene-based tests include in their analysis only variants that are either rare, or variants that independent annotation sources identify as "functional" (e.g.: nonsynonymous variants). This decision is often justified as necessary to remove noise and improve power by excluding variants that are unlikely to be associated with disease. However, SKAT-O can also incorporate prior knowledge about variants that are more likely to be associated with disease without fully excluding them by applying weights to the individual variants.[?,79] Weighting allows analyses that use SKAT-O to include those variants that may be causal but are not yet defined by characteristics that have been identified as suggestive of disease in the still-nascent understanding of molecular bi-

ology. In most genome-wide analytic scenarios, the use of weights will not increase type I or type II error rates,[38] and a weight that reflects the true disease process can significantly improve power.[37]

Currently, weighting is largely implemented by weighting according to the rareness, or minor allele frequency (MAF) of the variant. Weighting by MAF operationalizes the assumption that evolutionary constraints keep variants that strongly increase a risk of disease at low frequency in the population. However, not all variants that cause disease are kept at a low frequency.[80] For this reason, many annotations have also been developed that incorporate more broad indicators of pathogenicity beyond MAF. These functional annotations such as SIFT,[81] PolyPhen,[82] and CADD[39] operationalize the knowledge from previous research that variation at certain portions of the genome are expected to have a greater effect on disease risk. Of these, the CADD algorithm combines many single-dimensional annotations into one score of the predicted "deleteriousness" of that variant into a reproducible single score. This score can then be used to up-weight variants in the SKAT-O tests that are expected to cause disease.

In many cases summary statistics from a given study are more easily accessible due to fewer privacy restrictions. In these cases, SKAT-O cannot calculate the significance of a gene set. For this reason, other methods have been developed to calculate the significance of a gene-based on summary-level statistics.[?,83–89] Of these, one of the most straightforward test is Fisher's method.[90] This method combines the p-value of i separate variants within a gene using the formula $-2\sum ln(p_i)$. Under certain assumptions, this statistic is distributed as χ^2 with i degrees of freedom. However, in the case of correlated p-values, which is common in genetic regions with linkage disequilibrium (LD), Fisher's method and others that also do not take into account the LD, inflate the type I error rate of the genes tested. To correct for this, the VEGAS method[88] incorporates public use genetic data from HapMap

to control for correlation among p-values within a gene to infer the significance of the gene-based test statistic.

With this as background, this manuscript will investigate whether, in the context of early onset breast cancer, SKAT-O using CADD weights is able to identify genes that are associated with disease risk. The analysis will use conditional analyses to investigate whether any results are driven by common variation that would have been identified by a single marker regression analysis. This manuscript will also employ conditional analysis to determine whether any genes identified are driven by variants that are already known to be associated with breast cancer phenotypes, which will determine whether the identified gene contains novel risk loci. Simulations suggest that it is unlikely that there remain undiscovered risk variants for breast cancer with a minor allele frequency greater than 5% and a magnitude of effect that produces an odds ratio greater than 2,[91] but this has not yet been definitively empirically confirmed. The results will provide an opportunity to clarify whether rare variants of modest effect size are important in the genetic etiology of breast cancer, and whether gene-based tests are a useful tool to examine them. The investigations in this analysis will implement gene-based tests that reflect a hypothesis that collectively variation within the same gene can contribute to risk, and the choice of the specific SKAT-O test reflects the hypothesis that in some genes that hold causal variation, the distribution of causal alleles within that gene is relatively sparse, and in some causal variants, minor alleles may be protective of brat cancer. The choice of weights in this investigation reflect the hypothesis that variants that are predicted to be deleterious via the CADD algorithm have a higher probability of being causal for breast cancer risk, and if this does indeed reflect the underlying biological processes that drive breast cancer risk, the weighted analysis will have more power to detect causal genes than the unweighted one. No previous study of any trait has used the CADD scores directly as weights.

These investigations represent only the fourth study of breast cancer risk that has interrogated rare variation directly. In 2013 Haiman et al.[48] and more recently (August 2016 and September 2016) Haddad et al.[49] and Zhou et al.[92] all used exome arrays. However, each of these studies made methodological decisions that were sub optimal. Haiman et al. used a burden style test rather than a SKAT-O, and all three restricted their analysis to rare putative functional variants rather than weighting. A third study[50] that investigated rare variation using gene-based tests in a genome-wide setting in breast cancer also used a burden test, and additionally did not directly interrogate the rare variants in gene regions, but rather inferred their existence by imputation. The Zhou study controlled for things that were possibly in the causal pathway of breast cancer risk. None of the investigations found genes that were significant at the genome-wide level. Several studies that have implemented whole-exome and whole-genome sequencing are underway, but their results are not yet published. There have been several published gene-based analyses at the genome-wide level that investigate the genetic determinants of many diseases, and the vast majority has restricted the analysis to variants of a particular functionality or rareness. This analysis will instead up-weights variants that are predicted to be functional, and will include all resumed variants, and will assess whether any identified gene is driven by common variation through conditional analyses. This approach balances incorporating prior knowledge, while also allowing identification of strong associations that occur between not-yet-well-understood risk loci and disease.[93] This manuscript will discuss the appropriateness of three weighting methods in the context of the results they give.

The participants of this study are all aged 50 or younger. The analyses presented here will therefore provide evidence on whether genetic influences of the better-studied late onset disease are also risk factors for women who will be diagnosed early, or if instead that the genetic etiology of breast cancer risk differs for early onset cases.

Any genes that are identified will implicate particular gene products as being responsible at the cellular level for early onset breast cancer oncogenesis. This knowledge will improve the understanding of the underlying biology of early onset breast cancer risk, and may help to suggest possible targets for pharmaceutical chemoprevention. If genes of large effect are discovered, the new risk loci would continue to expand the ability to predict what patients are at risk for early onset breast cancer.

2.2 Methods

2.2.1 Population

The participants for these analyses were selected from ten ongoing studies designed to assess the risk factors associated with early onset breast cancer. Participants were women of European descent who were not known to carry pathogenic mutations in the genes BRCA1 or BRCA2. Details of the recruitment are found in Table 2.1. Ninety eight percent of the cases were younger than 50 years old at the time of their diagnosis (for cases) and all controls were younger than 50 at the time of enrollment. Six of the study sites (Australia, Northern California, Ontario, Philadelphia, and New York) were members of the Breast Cancer Family Registry (BCFR), whose methods have been described elsewhere.[63] Briefly, two of the BCFR centers (Northern California and Canada) recruited through population-based registries, three (Utah, Philadelphia, and New York) recruited through clinic- and community-based outreach, and one (Australia) recruited through a mix of population and clinic-based outreach. Participants were also included from four studies not included in the BCFR consortium. The German Genetic Epidemiologic Study of Breast Cancer;[64] and Long Island Breast Cancer Study Project;[65] and the Seattle study[66] were population-based case-control studies described elsewhere. The Chicago participants were identified from the Chicago Multiethnic Breast Cancer Epidemiologic Cohort, a hospital-based study of

Table 2.1: Characteristics of Studies Included in Exome-wide Analysis

Study Name	Study Location	Years Recruiting	Case Criteria	Control Criteria	Cases	Controls
Breast Cancer Family Registry	Australia	1992-2000	Living in the Melbourne and Sydney metro areas, family recruited from the Victoria and NSW cancer registries	Randomly selected from electoral rolls, matched to cases on age and city	473	118
Breast Cancer Family Registry	Northern California	1996-2003	SEER Cancer registry in the San Francisco metro area	Random digit dialing in study area, matched to cases on age and race/ethnicity	176	65
Breast Cancer Family Registry	Ontario	2001-2010	Ontario Cancer Registry	Random digit dialing in study area, matched to cases on age	582	152
Breast Cancer Family Registry	Philadelphia, Pennsylvania	1996-2000	Living in Philadelphia	N/A	333	0
Breast Cancer Family Registry	New York, New York	1996-2000	Living in New York, New Jersey, or Connecticut	N/A	551	0
Breast Cancer Family Registry	Utah	1996-2012	Living in Salt Lake City	N/A	152	0
Genetic Epidemiologic Study of Breast Cancer by Age 50	Germany	1992-1995	38 clinics in the Rhein-Neckar-Odenwald and Freiburg regions	Randomly selected from local population registries	466	437
Long Island Breast Cancer Study Project	New York	1996-1999	Nassau and Suffolk counties	Random digit dialing in study area, matched to cases on age	162	98
Seattle	Seattle, Washington	1990-1992	King, Pierce, and Snohomish counties; age less than 45 at diagnosis	Random digit dialing in study area, matched to cases on age and race	288	103
University of Chicago	Chicago, Illinois	1998-2010	Treated at the University of Chicago Cancer Center	N/A	326	0

Cases and controls are numbers included in the analysis after QC

breast cancer at the University of Chicago.[94,95] For the Chicago study, demographic factors, clinical, and pathological data, were abstracted from medical chart, epidemiologic risk factors, such as reproductive and lifestyle factors, were collected via structured questionnaire, and cancer relapse and survival were ascertained via patient medical records and linkage to the national death index.

2.2.2 Genotyping

DNA was available for 4914 participants (3,876 cases and 1,038 controls) through blood draws. The samples were whole genome amplified using the Qiagen Repli-G mini kit. 3956 (3121 cases and 835 controls) were genotyped on the Illumina HumanExome 12v1.0 chip, and 958 (755 cases and 203 controls) were genotyped on the Illumina HumanExome 12v1.1 chip. The samples were processed using 49 plates in two batches, and the process was carried out according to the manufacturer's protocol. To improve the quantity and quality of available genomic DNA, the samples were whole genome amplified using the Qiagen Repli-G mini kit,[22] and were processed using 49 plates in two batches, following the manufacturer's protocol. TeCan Evo was used for automation. Raw data was processed by Genome Studio on 2010.3 software, and the no-call threshold was set at 0.15, per Illumina's recommendation for Infinium chips. Clustering was done using the Illumina supplied cluster files. After keeping only variants that were on both chips, 238,524 variants were interrogated.

2.2.3 Analysis methods

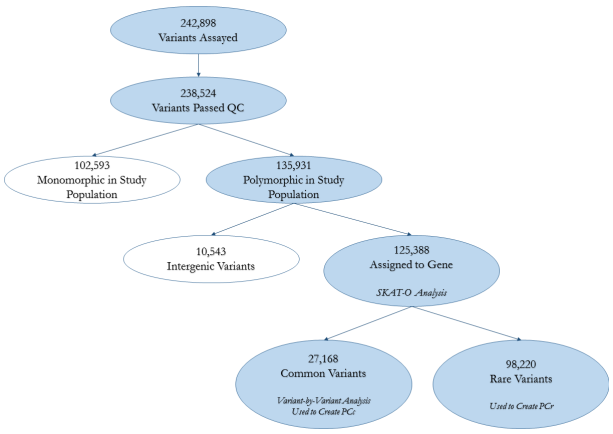
2.2.3.1 Quality Control

The quality control followed the protocol suggested by Guo et al.[96] Participants were excluded for low genotyping rate (rate < 95%; 248 excluded), male sex (nine excluded), high heterozygosity (F statistic greater than three standard deviations from the mean, or heterozygosity greater than four standard deviations from the mean; 52 excluded), one of each pair of duplicated genotypes (twenty four samples excluded; three replicates, twelve duplicates from the same center, nine recruited into both the Long Island and New York City studies), principal component outliers (three participants whose first or second principal components (constructed from common variants) were more than six standard deviations away from the mean). Additionally, due to the family-based case ascertainment of some of the studies, we also excluded 126 participants whose genotypes were highly correlated (estimated relatedness from a GCTA-created genetic relatedness matrix greater than 0.4).[97]

Variants were excluded from the analysis if they had a low call rate (rate < 95%; 4335 excluded), or if they were common variants (defined below) with Hardey-Weinberg equilibrium p-values of less than $2.5 \cdot 10^{-7}$ in controls (p = 0.05 Bonferroni corrected for 200,000 tests; 39 excluded). The final variant-level exclusions were the result of evidence that on some plates variants were unreliably assigned (a plate-by-plate single marker regression analysis found that in some cases genotype could predict plate). For these variant-plate combinations, variants were excluded for all participants on that plate if this single marker regression p-value was smaller than $2.5 \cdot 10^{-7}$. As a result of this QC step, 100 variant-plate combinations were set to missing.

After these exclusions, the analysis set contained 3479 cases, 973 controls, and 238,524 variants. Of these, 135,931 were polymorphic in the study population. Variants were assigned to genes using the ANNOVAR software,[98] and excluded if they were annotated to

Figure 2.1. Variants Used in Analysis



intergenic regions, leaving 125,388 polymorphic variants, which were annotated to 16,815 genes. Variants were classified as "common" and "rare" based on their MAF, with a threshold at MAF equal to $\left(\frac{1}{2n}\right)^{\frac{1}{2}} = 0.0106$.[59] A schematic of the variants used in this analysis is shown in Figure 2.1.

2.2.3.2 Controlling for Population Stratification

Rare variants and common variants have different correlations with ancestry, and therefore will have different potential to induce confounding in genetic association studies.[99,100] To counter this potential for inflated type I error rates, EIGENSTRAT[101,102] constructed two sets of principal components from the analysis set. One set was constructed using "common" variants assayed by the array (PC_c), and one using "rare" variants (PC_r). In a logistic regression that did not include genetic information, the first five PC_c and the first three PC_r were associated with case status. Including any other principal components did not improve the logistic model fit, as determined by a likelihood ratio test. These eight PC were included in all subsequent analyses.

2.2.3.3 Common Variation

To identify individual variants that would have been identified through single marker regression GWAS methods as being associated with risk of early onset breast cancer, common variants that could be assigned to a gene were analyzed using a logistic regression single marker regression framework with the PLINK software.[103,104] This analysis assumed an additive model of inheritance. Results were visualized using the qqman[105] and ggplot2[106] R software packages.[107] Variants whose p-values were smaller than the Bonferroni-corrected level of $1.8 \cdot 10^{-6}$ would have been considered suggestive of association with early onset breast cancer.

2.2.3.4 All Variation in Gene Regions

To examine whether variants within a gene collectively are associated with the risk of early onset breast cancer, the variants were analyzed using the SKAT-O method.[37] The analysis was conducted using the SKAT package for R, with the "SKATO" method in the function SKATBinary with efficient resampling.[108] The analysis was repeated three times: with equal weights; with heavy weights on rare variants (as suggested by the SKAT authors, weights on each variant equal to the beta function evaluated at the MAF of that variant in controls with shape parameters $\alpha = 1$ and $\beta = 25$); and with weights on each variant equal to the PHRED-like CADD score for that variant. For each of the methods, the significance threshold was determined by correcting a $p<0.05$ threshold by the effective number of tests computed, which was determined by the SKAT package. Genes whose p-values were less than this threshold using any weighting method were considered suggestively associated with early onset breast cancer.

Table 2.2: Characteristics of Studies Included GAME-ON/DRIVE

Study	Country	Case Ascertainment	Control Ascertainment	Genotyping Platform	Cases	Controls
ABCFS	Australia	Recruitment through cancer registries in Victoria and NSW	Recruitment from electoral rolls in Melbourne and Sydne matched to cases by age in 5-year categories	Illumina 610k	282	285
DFBBCS	Netherlands	BRCA1/2 mutation negative familial bilateral breast cancer patients selected from five clinical genetics centers	Rotterdam study; 55 years or older at time of inclusion.	Cases: Illumina 610k; Controls: Illumina 550k	464	3255
HEBCS	Finland	Helsinki University Central Hospital	Population controls from Finish Genome Centre (NordicDB)	Cases: Illumina 550k + 610; Controls: Illumina 370k	726	1012
BBCS	UK	UK Cancer Registries	WTCCC2: 1958 Birth Cohort + UK National Blood Service	Cases: Illumina 370k; Controls: Illumina 1.2M	1609	2663
GCHBOC	Germany	BRCA1/2 mutation negative cases from university clinics in Cologn and Munic	KORA (Cooperative Health Research in the Region Ausburg)	Cases: Affymetrix 5.0k; Controls Affymetrix 6.0k	634	477
UK2	UK	Cancer genetics clinics and oncology clinics	WTCCC2: 1958 Birth Cohort + UK National Blood Service	Cases: Illumina 370k; Controls: Illumina 1.2M	3628	2663
SASBAC	Sweden	Population-based postmenopausal women with breast cancer	Population-based controls, age-matched to cases	Cases: Illumina 317k+240k; Controls: Illumina 550k	790	756
MARIE	Germany	Sample of ductal and lobular carcinomas from the MARIE study, oversampled 2:1 for lobular	KORA (Cooperative Health Research in the Region Ausburg)	Cases: Illumina 370k; Controls: Illumina 550k	652	470
BPC3	USA, Europe, Poland	Sample of ER negative cases from eight cohort studies	Controls from eight cohort studies	Illumina 660k, Illumina 550k, Illumina 300k	2188	25519
BCFR*	USA, Europe, Canada, Australia	Population-based registries and clinic-based enrollment	Population-based controls, age-matched to cases	Illumina 610k, Cyto12	3523	2702
SardiNIA	Italy	Clinic-based Sardinian origin breast cancer patients	Sardinians with no history of cancer in first degree relatives recruited at communinty blood donation centers	Cases: Affymetrix 500k; Controls: Affymetrix 6.0	1367	1659

Cases and controls are numbers included in the analysis after QC

*The participants referred to as "BCFR" in the GAME-ON/DRIVE meta-analysis differ from the BCFR participants from the exome array analysis. The participants labeled "BCFR" in the meta-analysis are from sites in Australia, Ontario, California, Long Island, Germany, Seattle, and USC. 2323 cases (17% of the cases in the replication) and 1034 controls (2% of the controls in the replication) overlap with the participants in the exome array analysis.

2.2.4 *Replication: GAME-ON/DRIVE Summary Statistics*

Genes and variants that showed suggestive association were then investigated for evidence of significance in a population of breast cancer cases of all ages using the variant-level summary statistics provided by the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study in the Genetic Associations and Mechanisms in Oncology (GAME-ON[109]) consortium. The DRIVE study data combined information from twelve genome-wide association studies of breast cancer. Details of the recruitment are found in Table 2.2. Eight of the studies (Australia Breast Cancer Familial Study (ABCFS[110]); Rotterdam Study (DFBBCS[111]); Finland Breast Cancer Study (HEBCS[112,113]); British Breast Cancer Study (BBCS[114]); German Hereditary Breast and Ovarian Cancer Study (GCHBOC[115]); UK Breast Cancer Study 2 (UK2[116]); Singapore and Sweden Breast Cancer Study (SASBAC[113]); Mammary carcinoma Risk factor Investigation (MARIE[117]) were analyzed together, as described elsewhere,[118,119] and the other three studies (National Cancer Institute Breast and Prostate Cancer Cohort Consortium (BPC3[120,121]); Breast Cancer Family Registry subset and associated trials (BCFR[63]); and Sardinia[122]) were analyzed separately using slightly different quality control, and then combined. The twelve studies ultimately contributed 15,863 cases and 41,461 controls to the meta-analysis of 2,608,508 variants after imputation. The methods of this meta-analysis are detailed elsewhere.[50]

If any variants passed the genome-wide significance threshold ($1.8 \cdot 10^{-6}$) in the early onset participants, they would be compared to the GAME-ON/DRIVE summary statistics for replication and evidence that the same loci was causal in both the early onset and overall. If fewer than 20 variants had p-values smaller than the threshold in the early onset population, then the variants with the 20 smallest p-values were compared to the GAME-ON/DRIVE summary statistics. The variant would be identified as suggestively associated with both early onset and all-ages breast cancer if the meta-analysis p-values were less

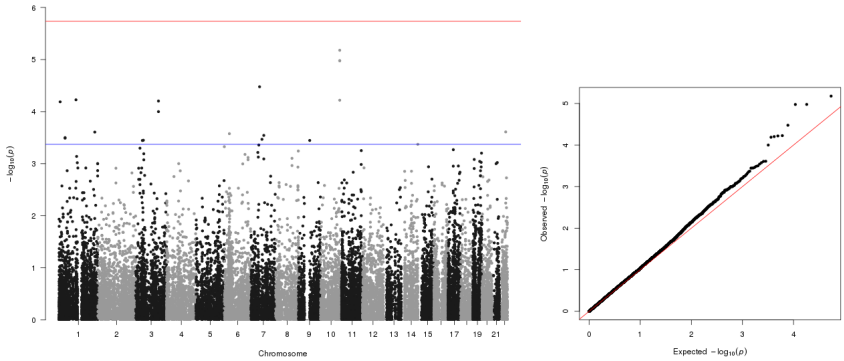
than Bonferroni-corrected threshold (determined by 0.05 divided by the number of variants tested for replication).

For the gene-based analyses, genes that any of the three SKAT-O weighting methods identified as suggestively associated with early onset breast cancer would be examined for evidence of also being involved in breast cancer diagnosed at any age. If fewer than 20 genes were suggestively associated with early onset breast cancer, then genes identified as the 20 most significant in any weighting method were examined. This was done by first annotating the GAME-ON variants into gene sets using ANNOVAR, and then combining the p-values of the GAME-ON summary statistics into a gene-based statistic using the VEGAS method. If the VEGAS method could not calculate a p-value for the summary statistics, a p-value was calculated with the Fisher method. The resulting p-values of the suggestive genes were compared to a Bonferroni-corrected threshold determined by 0.05 divided the total number of genes tested for replication. If the p-value was below this threshold, that gene was considered suggestively associated with both breast cancer that is diagnosed early and in all ages.

2.2.5 Comparison with GWAS-Identified Variants and Known Variants

In order to establish whether any suggestive findings were driven by variants that would have been identified through a single marker regression, genes were re-analyzed while controlling for any variants that were suggestively associated with early onset breast cancer or breast cancer of any age at diagnosis. Genes were selected for this analysis if their analysis in the early onset cases met one of two criteria. First, if they were suggestively associated with early onset breast cancer by having a p-value in any weighting method that was below the genome-wide p-value threshold as determined by SKAT, and second if they were suggestively associated with both the early onset disease and all-ages disease by the

Figure 2.2. Single Marker Logistic Regression Results for Common Variation Assayed on the Exome Array



The red line represents a p-value of $1.8 \cdot 10^{-6}$, and the blue line represents the p-value of the twentieth most significant SNV.

replication analysis. This conditional analysis was repeated using all three weighting methods, and implemented using the prepCondScores and skatOMeta functions of the skatMeta R package.[123] This package calculates the SKAT-O test statistic and controls for linkage disequilibrium with variants that are being conditioned upon by calculating p-values via permutation.

In order to establish whether any suggestive findings were driven by variants that were already known to be associated with a breast cancer phenotype, suggestive genes (genes defined by the three categories in the previous paragraph) were cross-referenced with the NHGRI-EBI GWAS Catalog.[34,60] If the GWAS catalog reported established associations (p-value $<5 \cdot 10^{-8}$) between a single nucleotide variant (SNV) in that gene and a breast cancer phenotype, then the gene was re-analyzed using used the skatMeta R, conditional on those variants.

Table 2.3: Most Significant Variants from Single Marker Regression Logistic Regression for Common Variation Assayed on the Exome Array

CHR	Position	Variant Name	Gene	Minor Allele	Major Allele	MAF	Annotated Function*	OR (95% CI)	p-value
10	123346116	rs2981575	FGFR2	G	A	0.450	intronic	1.269 (1.144, 1.407)	6.640E-06
10	123337335	rs2981579	FGFR2	A	G	0.460	intronic	1.258 (1.136, 1.394)	1.052E-05
10	123346190	rs1219648	FGFR2	G	A	0.447	intronic	1.262 (1.138, 1.399)	1.056E-05
7	55433884	exm622146	LANCL2	C	A	0.338	nonsynonymous SNV	1.264 (1.132, 1.412)	3.338E-05
1	109265029	exm79606	FNDC7	G	T	0.399	synonymous SNV	1.242 (1.117, 1.380)	5.937E-05
10	123352317	rs2981582	FGFR2	A	G	0.438	intronic	1.236 (1.115, 1.372)	6.041E-05
3	145938619	exm357004	PLSCR4	C	T	0.351	nonsynonymous SNV	0.803 (0.722, 0.894)	6.266E-05
1	7797503	exm10535	CAMTA1	G	C	0.133	nonsynonymous SNV	0.744 (0.644, 0.860)	6.486E-05
3	145917761	exm356986	PLSCR4	C	T	0.346	nonsynonymous SNV	0.808 (0.726, 0.900)	1.001E-04
22	40820151	rs5757949	MKL1	C	T	0.312	intronic	0.817 (0.733, 0.910)	2.453E-04
1	229772932	exm157648	URB2	G	A	0.065	nonsynonymous SNV	1.560 (1.230, 1.979)	2.474E-04
6	32975808	exm537347	HLA-DOA	A	G	0.011	nonsynonymous SNV	0.456 (0.299, 0.696)	2.653E-04
7	82453708	exm630212	PCLO	C	A	0.416	nonsynonymous SNV	1.214 (1.093, 1.347)	2.861E-04
1	39352271	exm47042	RHBDL2	T	G	0.455	nonsynonymous SNV	1.212 (1.091, 1.345)	3.161E-04
1	39340282	exm46983	GJA9	T	C	0.455	nonsynonymous SNV	1.211 (1.091, 1.344)	3.242E-04
7	70736168	exm2266425	WBSCR17	A	G	0.493	intronic	0.826 (0.743, 0.917)	3.402E-04
3	49054692	exm313521	DALRD3	T	C	0.233	synonymous SNV	0.808 (0.719, 0.908)	3.548E-04
9	71865988	exm753720	TJP2	T	C	0.082	nonsynonymous SNV	1.453 (1.184, 1.785)	3.581E-04
3	42814668	exm2239553	LOC729083	T	A	0.207	ncRNA_intronic	1.272 (1.115, 1.452)	3.587E-04
14	105225979	exm2273525	SIVA1(NM_006427:c.*157C>G,NM_021709:c.*157C>G)	G	C	0.189	UTR3	1.276 (1.114, 1.462)	4.248E-04

*From ANNOVAR
positions refer to the HG19 assembly
CHR=Chromosome
MAF=Minor Allele Frequency
OR=Odds Ratio
CI=Confidence Interval

2.3 Results

2.3.1 Common variation

After controlling for the principal components, the single marker regression analysis of the 27,168 common variants has a genomic inflation factor of 1.065. A summary of the association results is shown in Figure 2.2, and details of the twenty SNVs with the smallest p-values are shown in Table 2.3.

None of the SNVs are significant at the pre-set threshold. The most significant SNV is located at chr10:123346116 in the intron of FGFR2. The p-value of this SNV is $6.64 \cdot 10^{-6}$. Each additional allele increased the odds of breast cancer by 27% (OR: 1.268; 95% CI: 1.14-1.41). The most significant SNV that was predicted to cause a change in a translated amino acid is a nonsynonymous SNV in LANCL2, located at chr7:55433884. The p-value of this variant is $3.3 \cdot 10^{-5}$, and each additional risk allele is estimated to the odds of breast cancer by 26% (OR: 1.264; 95% CI: 1.132-1.412).

Of the twenty SNVs with the smallest p-values, twelve are interrogated in the GAME-ON/DRIVE meta-analysis, and the p-values of these variants are shown in Table 2.4. Of these, three in the gene FGFR2 are significant at the Bonferroni-corrected level of $2.5 \cdot 10^{-3}$ in the GAME-ON/DRIVE data. They are located at chr10:123337335, chr10:123346190, and chr10:123352317.

2.3.2 Gene-based tests

Figure 2.3 summarizes the coverage of the exome array, and characterizes the rarity of the variants within the gene. The median number of variants per gene was five, the median number of total minor alleles in a gene was 898, and the median number of individuals with at least one minor allele in a gene was 816.

Table 2.4: Replication of Single Marker Regression Findings in GAME-ON/DRIVE

			Exome Array			GAME-ON/DRIVE		
Location	Gene	Minor/Major Allele*	MAF	OR (95% CI)	p-value	EAF	OR (95% CI)	p-value
chr10:123346116	FGFR2	A/G	0.450	1.269 (1.144, 1.407)	6.640E-06	NA	N/A	
chr10:123337335	FGFR2	G/A	0.460	1.258 (1.136, 1.394)	1.052E-05	0.451	1.287 (1.243, 1.332)	2.961E-46
chr10:123346190	FGFR2	A/G	0.447	1.262 (1.138, 1.399)	1.056E-05	0.436	1.280 (1.325, 1.237)	7.986E-45
chr7:55433884	LANCL2	A/C	0.338	1.264 (1.132, 1.412)	3.338E-05	NA	N/A	
chr1:109265029	FNDC7	T/G	0.399	1.242 (1.117, 1.380)	5.937E-05	NA	N/A	
chr10:123352317	FGFR2	G/A	0.438	1.236 (1.115, 1.372)	6.041E-05	0.431	1.287 (1.242, 1.333)	9.383E-45
chr3:145938619	PLSCR4	T/C	0.351	0.803 (0.722, 0.894)	6.266E-05	0.337	0.986 (0.952, 1.022)	4.366E-01
chr1:7797503	CAMTA1	C/G	0.133	0.744 (0.644, 0.860)	6.486E-05	NA	N/A	
chr3:145917761	PLSCR4	T/C	0.346	0.808 (0.726, 0.900)	1.001E-04	0.333	0.990 (0.955, 1.026)	5.871E-01
chr22:40820151	MKL1	T/C	0.312	0.817 (0.733, 0.910)	2.453E-04	0.295	0.945 (0.910, 0.981)	2.985E-03
chr1:229772932	URB2	A/G	0.065	1.560 (1.230, 1.979)	2.474E-04	0.064	1.084 (1.164, 1.011)	2.407E-02
chr6:32975808	HLA-DOA	G/A	0.011	0.456 (0.299, 0.696)	2.653E-04	NA	N/A	
chr7:82453708	PCLO	A/C	0.416	1.214 (1.093, 1.347)	2.861E-04	0.433	1.040 (1.076, 1.005)	2.485E-02
chr1:39352271	RHBDL2	G/T	0.455	1.212 (1.091, 1.345)	3.161E-04	0.458	1.017 (1.053, 0.984)	3.171E-01
chr1:39340282	GJA9	C/T	0.455	1.211 (1.091, 1.344)	3.242E-04	0.459	1.017 (1.053, 0.983)	3.193E-01
chr7:70736168	WBSCR17	G/A	0.493	0.826 (0.743, 0.917)	3.402E-04	0.440	0.981 (0.949, 1.015)	2.808E-01
chr3:49054692	DALRD3	C/T	0.233	0.808 (0.719, 0.908)	3.548E-04	0.236	0.996 (1.036, 0.958)	8.353E-01
chr9:71865988	TJP2	C/T	0.082	1.453 (1.184, 1.785)	3.581E-04	NA	N/A	
chr3:42814668	LOC729083	A/T	0.207	1.272 (1.115, 1.452)	3.587E-04	NA	N/A	
chr14:105225979	SIVA1	C/G	0.189	1.276 (1.114, 1.462)	4.248E-04	NA	N/A	

*Refers to the allele frequency in the exome data analysis set
MAF: Minor allele frequency in BCFR data
EAF: Effect of allele that is minor in BCFR data in GAME-ON/DRIVE
positions refer to the HG19 assembly

Figure 2.3. Distribution of Variants Per Gene, Minor Alleles Per Gene, and Participants with Minor Alleles Per Gene

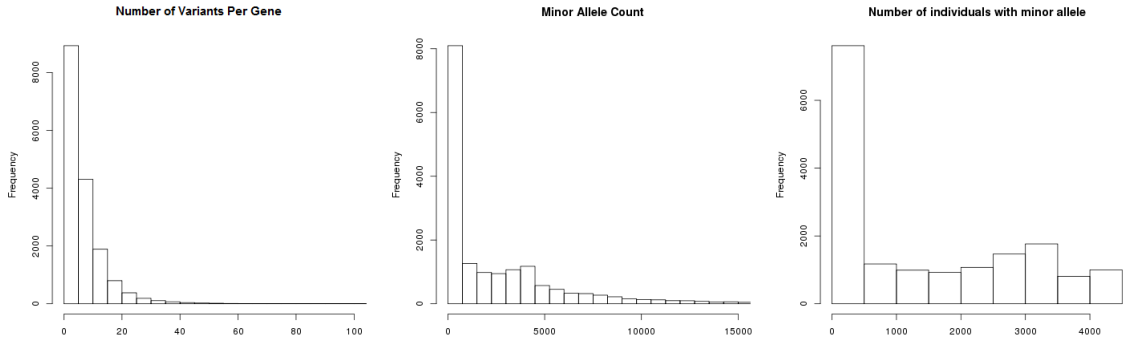


Figure 2.4. Distribution of Variant Weights for Variants Analyzed from the Exome Array

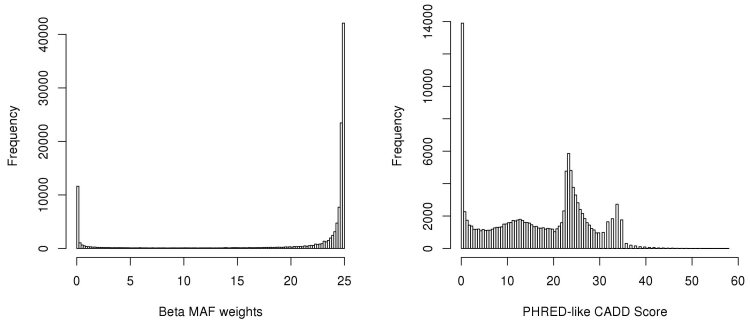
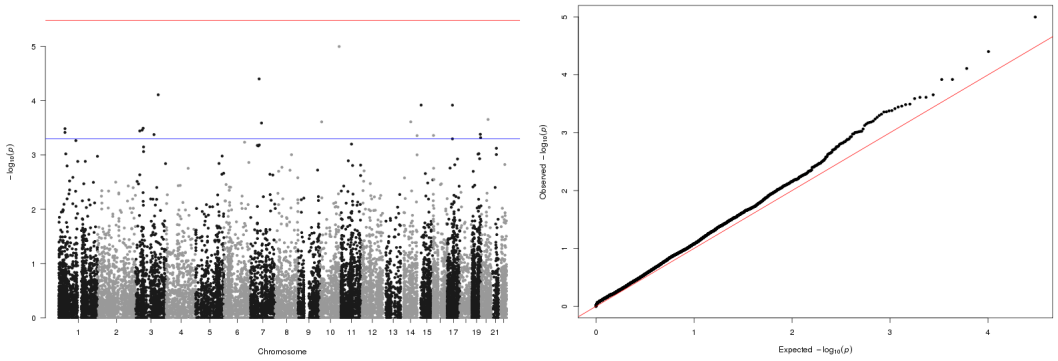


Figure 2.5. Sequence Kernel Association Test-Optimal Results for Exonic Variants Assayed on the Exome Array with Equal Weights

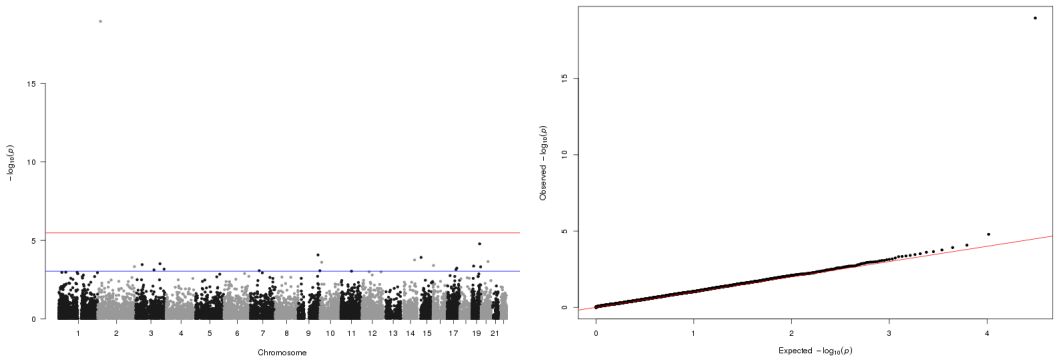


The red line represents a p-value threshold based on a Bonferroni correction of the effective number of tests, as calculated by the SKAT package; the blue line represents the p-value of the twentieth most significant SNV.

The SKAT-O gene-based tests were repeated using three weighting methods: once with equal weights; once with each variant weighted by the beta function with $\alpha = 1$, $\beta = 25$, evaluated at the minor allele frequency of that variant in controls; and once by the CADD deleteriousness score of that variant, transformed to a PHRED-like scale. The distributions of the beta weights and CADD weights for analysis set are shown in Figure 2.4. The distribution of the beta weights is skewed towards the maximum weight of 25, a result of the rareness of most of the variants in the exome array.

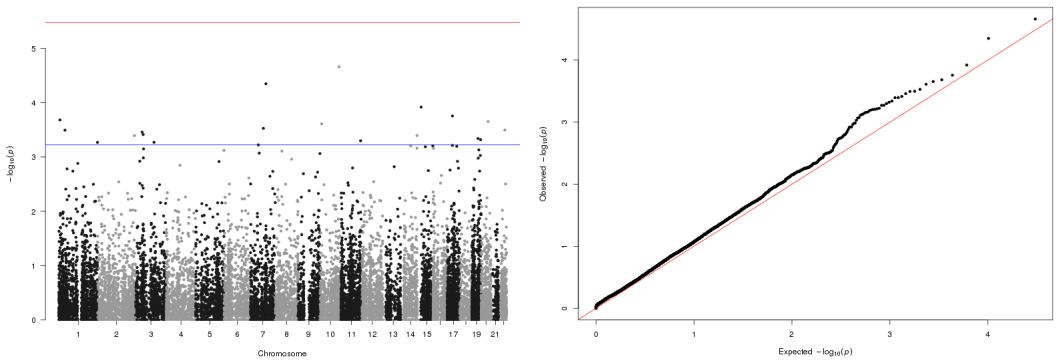
After controlling for the principal components, the genomic inflation factor of the gene-based tests produced by the SKAT-O analysis with equal weights is 1.12. A summary of the gene-based results with equal weights is shown in Figure 2.5. The genomic inflation factor is 1.09 in the SKAT-O analysis when the variant weights were equal to the beta function, and is summarized in Figure 2.6. The genomic inflation factor is 1.10 in for the SKAT-O analysis when the variant weights were equal to the variant's CADD score, and this analysis is summarized in Figure 2.7. P-values for all weighting schemes for genes that are in the 20 most significant in any weighting scheme are shown in Table 2.5.

Figure 2.6. Sequence Kernel Association Test-Optimal Results for Exonic Variants Assayed on the Exome Array with Beta Weights



The red line represents a p-value threshold based on a Bonferroni correction of the effective number of tests, as calculated by the SKAT package; the blue line represents the p-value of the twentieth most significant SNV.

Figure 2.7. Sequence Kernel Association Test-Optimal Results for Exonic Variants Assayed on the Exome Array with CADD Weights



The red line represents a p-value threshold based on a Bonferroni correction of the effective number of tests, as calculated by the SKAT package; the blue line represents the p-value of the twentieth most significant SNV.

Table 2.5: Most Significant Genes Identified by Sequence Kernel Association Test-Optimal with Three Weighting Methods

Gene	Variants in Gene	Total Count of Minor Alleles in Gene	p-value equal weights	p-value beta weights	p-value CADD weights
ABCC5	6	3486	3.84e-01	6.72e-04*	4.33e-01
ACE	26	4269	3.04e-02	5.82e-04*	6.36e-04
AP2B1	2	4262	5.05e-04*	5.64e-01	6.11e-04
CAMTA1	24	12762	1.05e-01	5.29e-01	2.08e-04*
CDON	18	16625	1.53e-03	1.00e+00	5.03e-04*
CELF2	2	11	2.45e-04*	2.45e-04*	2.45e-04*
DALRD3	4	2080	3.22e-04*	5.16e-01	3.86e-04*
DDC	7	7064	6.80e-04	3.66e-01	5.99e-04*
EML5	6	2265	8.50e-02	1.74e-04*	5.71e-02
FGFR2	8	16044	1.01e-05*	1.30e-01	2.19e-05*
GJA9	10	4456	3.85e-04*	1.00e+00	3.50e-01
HSF5	5	1947	3.69e-02	7.20e-04*	2.30e-02
HSPBP1	2	16	4.79e-04*	4.79e-04*	4.79e-04*
ILF3	4	2926	8.59e-02	4.28e-04*	3.73e-03
KPNA7	4	634	6.25e-01	3.46e-01	4.48e-05*
LANCL2	4	3047	3.97e-05*	8.24e-01	9.86e-02
MAP4K1	6	60	9.64e-04	2.00e-03	4.58e-04*
MAPKAP1	4	4051	7.55e-01	8.43e-05*	2.26e-01
MKL1	18	9844	1.50e-03	3.80e-02	3.19e-04*
MSGN1	3	7258	1.00e+00	1.10e-19*	1.00e+00
NEK10	11	10381	3.61e-04*	6.85e-01	1.20e-03
NOXA1	4	2840	3.18e-01	8.54e-04*	8.65e-04
OR11L1	12	3889	1.06e-03	1.14e-03	5.37e-04*
PCLO	57	19585	5.86e-03	2.55e-01	2.97e-04*
PLSCR4	9	8088	7.78e-05*	2.83e-01	6.21e-01
PSPH	5	33	6.54e-04	8.46e-04*	8.48e-04
PTPRCAP	5	236	6.30e-04	9.09e-04*	2.06e-01
RAB26	5	26	4.36e-04*	3.92e-04*	6.95e-04
RHBDL2	6	4135	3.28e-04*	1.00e+00	3.21e-04*
RUVBL2	3	4452	7.40e-01	1.64e-05*	1.44e-01
SH3BP4	19	2016	4.55e-01	4.64e-04*	4.04e-04*
SIVA1	4	1744	4.40e-04*	6.43e-01	4.03e-04*
SLFN14	10	12265	1.21e-04*	5.43e-01	1.75e-04*
SNURF	3	6	1.21e-04*	1.21e-04*	1.21e-04*
SYNE2	104	19289	2.45e-04*	1.33e-01	6.24e-04
UPK1B	8	388	4.21e-04*	7.52e-04*	5.36e-04*
VEPH1	23	3267	4.10e-01	3.05e-04*	1.60e-01
WBSCR17	10	9522	2.58e-04*	8.31e-01	5.89e-02
WFDC11	1	19	2.22e-04*	2.22e-04*	2.22e-04*
ZNF665	6	8054	4.15e-04*	1.71e-01	9.37e-04

If the gene was one of the top 20 most significant genes for that weighting scheme, its p-value is marked with an asterisk

Table 2.6: Annotation and Distribution of Variation within MSGN1 in Participants of Exome Array Study

Variant	Annotation	Minor Allele in Cases	Minor Allele in Controls
2:17998025, A→T	exonic, synonymous SNV	2844	795
2:17998027, C→G	exonic, nonsynonymous SNV	0	2
2:17998095, C→T	exonic, synonymous SNV	2848	800

Annotation from ANNOVAR

One gene was found to be significant at the genome-wide level when weighting by the beta function transformation of their minor allele frequency. For that weighting scheme, the p-value of MSGN1 on chromosome 2 is highly significant, with a p-value of $1.10 \cdot 10^{-19}$. The exome array assayed three polymorphic variants within this gene, whose variants are characterized in Table 2.6. Two of the MSGN1 variants are common, and the third is very rare, with a MAF of 0.000224, and was observed in two heterozygous controls. This SNV, positioned at chr2:17998027 in the HG19 assembly, is given a weight of 24.86 by the beta function. The p-values for MSGN1 for the other two methods are not significant (the p-value for both equal weights and CADD weights is 1.00).

The other two weighting methods do not identify any genes as associated with early onset breast cancer at the pre-set significance threshold.

Given the importance of the HLA regions in many disease processes, the suggestive significance of a variant within that region in the single marker regression analyses suggested that a closer investigation of variants that were near that suggestive variant may be fruitful. Figure 2.8 plots the significance of all variants within a 500 kilobase region around the HLA-DOA gene that were included in the single variant regression analysis of the BCFR participants. The p-values for the HLA-DOA gene were 0.548, 0.008, and 0.323 for the equal weights, MAF weights, and CADD weights respectively. While this region may be a promising region to explore for candidate gene studies in the future, there is no convincing

Figure 2.8. Single Marker Logistic Regression Results for Common Variation Assayed on the Exome Array near HLA-DOA

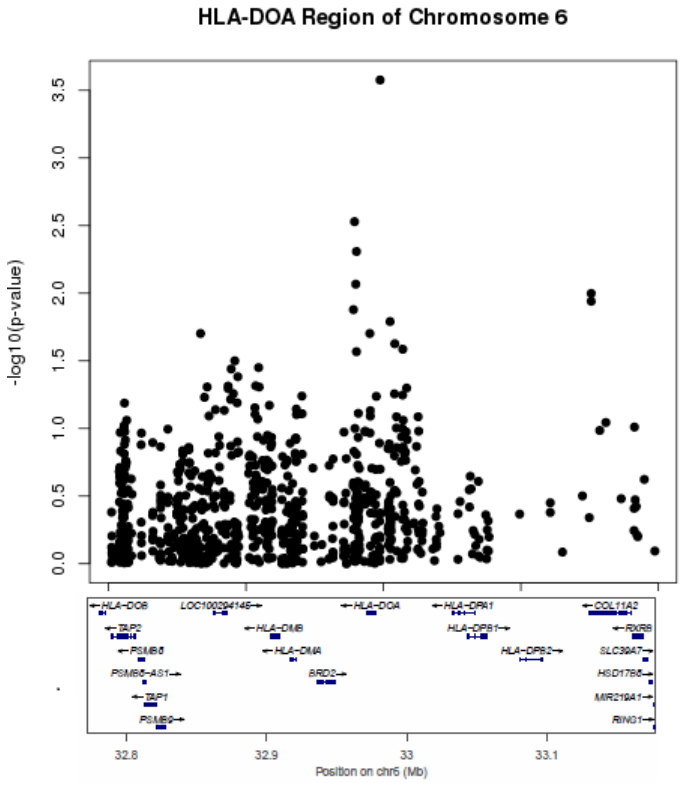
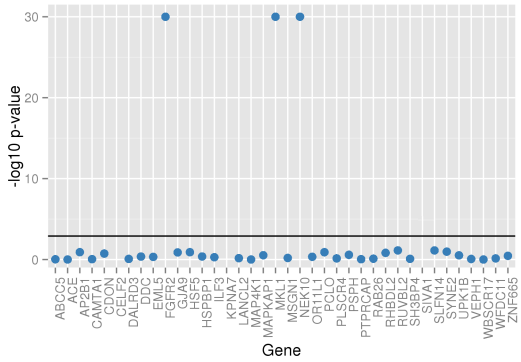


Figure 2.9. Evidence of Shared Genetic Risk for Early- and All-Ages Onset of Breast Cancer from Gene-Based-Tests in GAME-ON/DRIVE



The horizontal line represents a p-value threshold based on a Bonferroni correction of $1.25 \cdot 10^{-3}$. Genes whose p-values were smaller than 10^{-30} are presented with $p = 10^{-30}$.

evidence in this population that variation within the HLA-DOA gene is associated with breast cancer risk.

2.3.2.1 Comparison with Breast Cancer of All Ages of Onset

Forty genes were among the twenty most significant genes calculated by at least one of the three weighting methods. Of these, thirty-eight contain at least one variant interrogated by the GAME-ON/DRIVE summary statistics, and could therefore have a gene-based test constructed for replication. The two genes that are not found in GAME-ON/DRIVE are NOX1A and SNURF, CELF2, SIVA1, and KPNA7. Figure 2.9 summarizes the results of the gene-based tests based on the GAME-ON/DRIVE summary statistics of the genes. The VEGAS software is limited in its ability to calculate very small p-values, and reports a p-value of zero if 10,000 permutations continue to find smaller p-values. In Figure 2.9, these genes are represented as having a p-value of 10^{-30} for purposes of scale.

Three genes, FGFR2, NEK10 and MKL1, are significant in the GAME-ON/DRIVE results using the VEGAS method of combining p-values.

Table 2.7: Summary of ρ Mixing Parameter for Suggestive Genes

Gene	Variants in Gene	ρ Equal Weights	ρ MAF weights	ρ CADD Weights
FGFR2	8	0.0	0.0	0.5
MKL1	18	0.0	0.0	0.0
MSGN1	3	1.0	1.0	1.0
NEK10	11	1.0	0.0	1.0

2.3.2.2 Appropriateness of SKAT-O

The genes MSGN1, NEK10, FGFR2, MKL1, and SIVA1 are identified as being suggestively associated with breast cancer. In many weighting scenarios for these genes, the value of ρ, the mixing parameter that combines the SKAT with the burden test in the SKAT-O analysis is zero. In these genes (all weighting methods for MKL1; and under equal weighting and beta weighting for FGFR2 and SIVA1), the value of ρ indicates that the SKAT test was more appropriate than the burden test. In contrast, for MSGN1, ρ is estimated to be 1 for all weighting methods, indicating that the burden test was more appropriate than the SKAT test. The analysis of NEK10 requires a mixing parameter of 1 in the equal weights and CADD weights scenario, and a mixing parameter of 0 in the beta weighting scenario, as summarized in Table 2.7.

2.3.2.3 Novelty of Associations

The five genes were analyzed using conditional analyses that controlled for any SNVs that were suggestively associated with breast cancer in the single marker regression analyses of the same participants. Of these five genes, only FGFR2 contained variants that were suggestively associated with breast cancer in the single marker regression analyses, and these three are noted with a ^ in Table 2.8. After conditioning on these three variants, in none of the weighting methods does the analysis produce a p-value that was less than 0.05. This

Table 2.8: Annotation and Distribution of Variation within FGFR2 in Participants of Exome Array Study

Variant	Annotation	Minor Allele in Cases	Minor Allele in Controls
10:123239388, T→C	exonic, nonsynonymous SNV	2	0
10:123256135, A→G	exonic, nonsynonymous SNV	1	1
10:123310871, G→A	exonic, nonsynonymous SNV	22	3
10:123325158, A→G	exonic, nonsynonymous SNV	17	11
10:123337335, A→G*^	intronic	3322	809
10:123346116, G→A*	intronic	3249	786
10:123346190, G→A*^	intronic	3205	780
10:123352317, A→G*^	intronic	3162	774

Annotation from ANNOVAR
Positions refer to the HG19 assembly
*Variants identified by previous research as associated with a breast cancer phenotype
^Variants identified by the GWAS as associated with early onset breast cancer

suggests that the variants that could have been identified using single marker regression methods, or those in close LD with them, drive the bulk of the association between FGFR2 and early onset breast cancer risk in this sample. The other four genes contain associations with early onset breast cancer risk that would not have been identified using single marker regression methods.

Next, these five genes were queried in the NHGRI-EBI catalog to determine if any harbor variants that are known to be associated with a breast cancer phenotype from an earlier study. For FGFR2, the NHGRI-EBI catalog contains six variants that were significantly associated with a breast cancer phenotype in at least one study, and four were assayed on the exome array. These four variants were located at chr10:123337335, chr10:123346116, chr10:123346190, and chr10:123352317. They are noted with an asterisk in Table 2.8. After conditioning on the four previously-identified variants, in none of the weighting methods does the analysis produce a p-value that was less than 0.05: the p-value for equal weights was 0.14; the p-value for beta weights was 0.16; and the p-value for CADD weights was 0.08. This suggests that the already-identified variants, or those in close LD with them, drive the bulk of the association between FGFR2 and breast cancer risk in this sample.

Table 2.9: Annotation and Distribution of Variation within MKL1 in Participants of Exome Array Study

Variant	Annotation	Minor Allele in Cases	Minor Allele in Controls
22:40813413, A→G	exonic, nonsynonymous SNV	2	1
22:40814500, C→T	exonic, nonsynonymous SNV	2770	750
22:40814533, T→C	exonic, nonsynonymous SNV	1	0
22:40814542, A→G	exonic, nonsynonymous SNV	1	0
22:40814581, T→C	exonic, nonsynonymous SNV	0	1
22:40814749, T→C	exonic, nonsynonymous SNV	1	2
22:40814878, T→C	exonic, nonsynonymous SNV	2	1
22:40814950, T→C	exonic, nonsynonymous SNV	2	0
22:40814988, A→G	exonic, nonsynonymous SNV	2	0
22:40815256, T→C	exonic, nonsynonymous SNV	27	1
22:40815309, T→C	exonic, nonsynonymous SNV	1	0
22:40816431, A→G	exonic, nonsynonymous SNV	1	1
22:40816443, A→G	exonic, nonsynonymous SNV	1	1
22:40819589, T→G	exonic, nonsynonymous SNV	0	1
22:40820151, C→T	intronic	2127	664
22:40820273, T→C	exonic, synonymous SNV	27	1
22:40820311, T→C	exonic, nonsynonymous SNV	1	0
22:40849704, C→A	intronic	2684	810

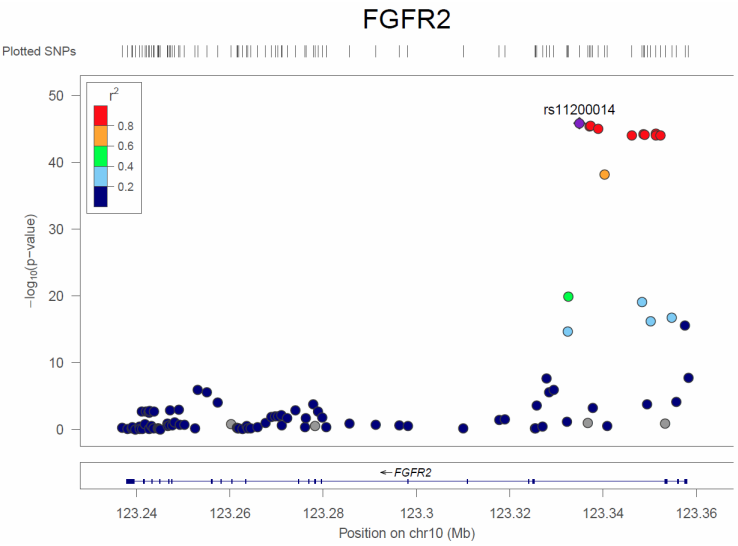
Annotation from ANNOVAR
Positions refer to the HG19 assembly

Table 2.10: Annotation and Distribution of Variation within NEK10 in Participants of Exome Array Study

Variant	Annotation	Minor Allele in Cases	Minor Allele in Controls
3:27326097, T->G	exonic, synonymous SNV	2049	492
3:27326131, C->T	exonic, nonsynonymous SNV	2	0
3:27326451, G->A	exonic, synonymous SNV	2106	507
3:27332820, G->A	exonic, nonsynonymous SNV	2053	495
3:27333024, T->C	exonic, nonsynonymous SNV	1	0
3:27338698, C->T	exonic, nonsynonymous SNV	99	22
3:27338730, A->G	exonic, synonymous SNV	1	0
3:27343261, T->C	exonic, nonsynonymous SNV	70	16
3:27349047, A->G	intronic	1893	515
3:27385817, T->C	exonic, nonsynonymous SNV	1	0
3:27387641, T->C	exonic, nonsynonymous SNV	40	14

Annotation from ANNOVAR
Positions refer to the HG19 assembly

Figure 2.10. Linkage Disequilibrium and p-values of Variants from GAME-ON/DRIVE for FGFR2



Annotation for each of the variants in these genes is found in Table 2.9 through Table2.10. SIVA1 and MSGN1 both had no known associations with a breast cancer phenotype. The NHGRI-EBI catalog reports two variants that are known to be associated with breast cancer in the MKL1 gene-rs6001930 and rs17001868. Both of these known variants are annotated to intronic regions, and neither was assayed by the exome array. In the CEU population of the 1000 Genomes, the highest r^2 measure of LD between either of these SNVs and any of the 18 SNVs in the exome array data was 0.03,[124] suggesting that the association reported at MKL1 was not driven by already-known single-variant associations. The segment of chromosome 3 that contains NEK10 is gene dense. Previous single marker regression associations have been reported both for the intron of NEK10 itself and the 3 prime UTR of SLC4A, which is immediately adjacent to NEK10. None of these variants were directly interrogated by the exome array, and none were in high LD with any of the measured variants.

Since the identification of the four suggestively associated genes relied on replication from the GAME-ON/DRIVE summary statistics which assayed different variation than the

Figure 2.11. Linkage Disequilibrium and p-values of Variants from GAME-ON/DRIVE for NEK10

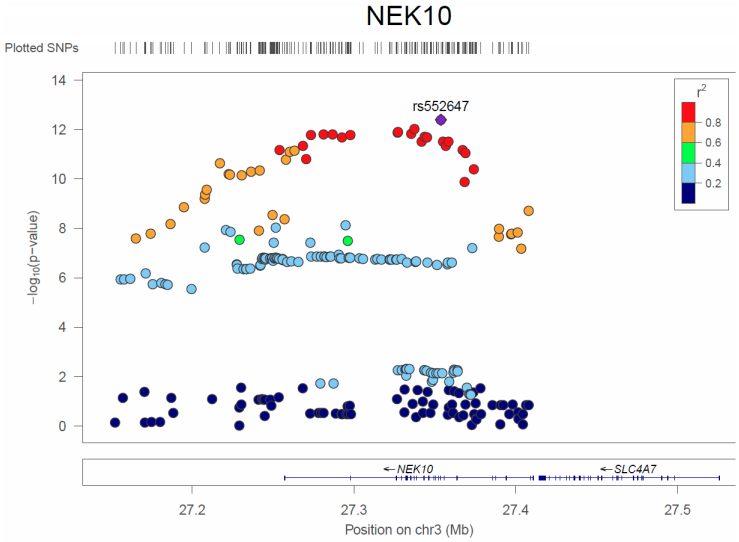
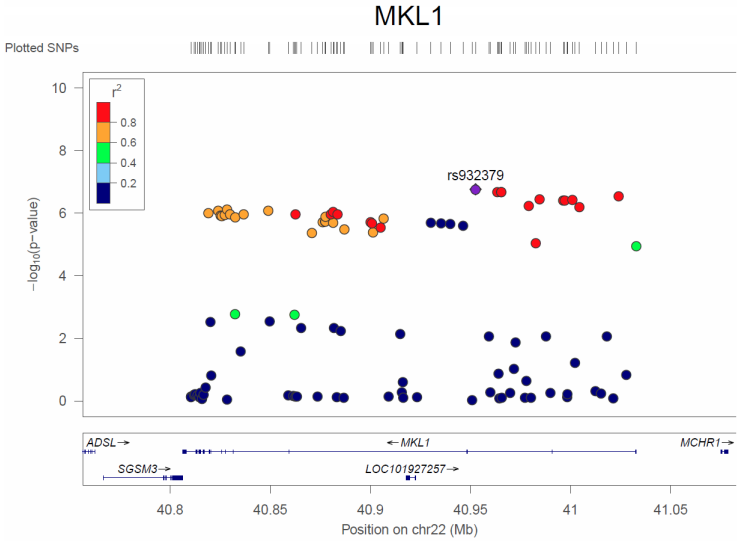


Figure 2.12. Linkage Disequilibrium and p-values of Variants from GAME-ON/DRIVE for MKL1



BCFR participants that were measured using an exome array, Figures 2.10 through **??** allow for a closer examination of the p-values and correlations between the GAME-ON/DRIVE variants that were included in the gene-based tests for each of the three genes that were suggestively associated with early onset breast cancer. In these figures, the variant with the most significant p-value in the GAME-ON/DRIVE analysis is highlighted, and the linkage disequilibrium of every other variant with that SNV is presented (calculated as an r^2).

One insight from this investigation is that some variants within NEK10 were genome-wide statistically significant in the GAME-ON/DRIVE summary statistics ($p < 5 \cdot 10^{-8}$), but had not been individually reported within the manuscripts of the original study. Therefore, the NHGRI-EBI catalog of GWAS results does not list NEK10 as associated with breast cancer, even though previous study populations did contain evidence of its association.

Other insights from the LocusZoom plots suggest which variants may be driving the gene-based results. Within the GAME-ON/DRIVE participants, the association within FGFR2 is driven by variants with very small p-values; re-running the VEGAS analysis including only the variants with a p-value less than $5 \cdot 10^{-8}$ no longer identified a statistically significant p-value. In contrast, the variants that were included within MKL1 were all not statistically significant on their own, but collectively produce a gene-based p-value that was statistically significant. The association found from the GAME-ON/DRIVE summary statistics within NEK10 appears to be driven by some variants that are statistically significant on their own, and some that did not reach the level of statistical significance, as a re-run of the VEGAS analysis including only the variants with a p-value less than $5 \cdot 10^{-8}$ still produces a p-value that was less than 10^{-30}.

2.4 Discussion

This analysis identifies three genes in which variation is associated with risk of early onset breast cancer: FGFR2 (discovery $p = 2.18 \cdot 10^{-5}$; replication $p < 10^{-30}$), NEK10 (discovery $p = 1.20 \cdot 10^{-3}$; replication $p < 10^{-30}$), MKL1 (discovery $p = 2.62 \cdot 10^{-4}$; replication $p < 10^{-30}$), and MSGN1 (discovery $p = 1.10 \cdot 10^{-19}$; replication p unavailable). Three of these genes are also suggestively associated with an overall breast cancer sample (FGFR2, MKL1, and NEK10). The association at MSGN1 appears to be an artifact that is induced when MAF is used to weight.

All three of these genes are known to be instrumental in mechanisms that are associated with cancer: FGFR2 is part of the known cancer pathway of PI3K-AKT,[125] NEK10 is involved in cell cycle control,[126] and MKL1 has been linked to oncogenic phenotypes.[127] The gene-based nature of the SKAT-O test also suggests that the product of these genes may be involved in breast cancer development, and the EMBL-EBI gene expression atlas[128] was able to verify this. All four are expressed in breast tissue, and there is evidence that three are differentially expressed in breast cancer tissue when compared to normal breast tissue: FGFR2 (over expressed), NEK10 (under expressed), and MKL1 (under expressed).

Conditional analyses suggest that with the exception of FGFR2, the associations cannot be explained by risk loci that either were already known to be associated with disease, or that would be identified in a single marker regression analysis. This validates the additional resources needed to interrogate the variants assayed with the exome array, and analyze those variants using gene-based tests. In contrast, this investigation suggests that the association found between FGFR2 and early onset breast cancer is largely driven by loci that were already known to be associated with breast cancer risk.

This analysis also suggests that weighting by predicted functionality can highlight genes that would not otherwise be identified, but that weighting by minor allele frequency

alone may be problematic. While in neither of the weighting methods is the genomic inflation factor larger than the genomic inflation factor using equal weights (suggesting that there is no systematic inflation of type I error rates for either weighting method), weighting by minor allele frequency results in the MSGN1 gene returning an association p-value of $1.10 \cdot 10^{-19}$, while it was not even nominally significant using either of the other weighting methods, or the replication data. While it is possible that the MAF weighting highlighted a true causal gene that was missed by the other weighting methods, it is more likely that this instead demonstrates an over-sensitivity of the MAF weighting method to very rare variants in genes with a small amount of variation. Since rareness is not the strongest predictor of risk, the CADD weights represent a possible preferable method to allow prior knowledge to increase the power of gene-based tests. If an allele that is rare, but has no other *a priori* expectation of being involved in disease would be given a small weight by the CADD score. The MKL1 gene, which was found to be suggestively associated was only identified as one of the top 20 most significant genes using the CADD weighting method.

This analysis also speaks to the appropriateness of using a gene-based test that incorporates both common and rare variants. The collective effect of the "common" variants in the genes MKL1 and NEK10 are of sufficient strength to highlight these genes in gene-based tests in the modest sample size of the primary data analysis, but no risk loci in those genes were suggestively associated with risk in a single marker regression.

Similarly, this analysis also supports including all non-intergenic variants a gene-based test, and not only those that are expected to cause a change in amino acid translation. Sensitivity analyses (not shown) demonstrated that a analysis that was restricted to only nonsynonymous variants would not have produced any associations, and the associations were instead driven by variants that do not alter amino acid translation. The associations at FGFR2 and MKL1, and NEK10 are all driven by variants that were annotated to the introns

of those genes (that is, conditional analyses controlling for the intronic variants produced non-significant p-values for almost all weighting methods).

Additionally, the analysis here provides possible insight into the genetic architecture of genes that are associated with early onset breast cancer risk. The ρ mixing parameter estimates indicate that SKAT is a more appropriate test than the burden test in many scenarios. While the estimates of the mixing parameter varied from test to test, in general, the SKAT test was more appropriate for genes that contained even a modest number of variants. This suggests that that few variants in these genes were causal, and that the effects of the causal variants may be both protective and deleterious. Conversely, the ρ estimate for the smaller genes suggests that they were better interrogated by the burden test, which is consistent with the assumptions of the burden test that each allele have the same magnitude and direction of risk, and are all causal. It appears that genes with a small number of variants are best analyzed using the burden test, as it is more likely that the small number of variants within them all have the same direction and magnitude of effect. In contrast, the genes that have more loci with variation are better assayed using SKAT. Similarly, the single marker regression analyses, which identified nominally significant SNVs that were both protective and deleterious, support the use of a gene-based test such as SKAT that allows for bidirectional effects. The heterogeneity of the ρestimates validate the small loss in power that is incurred by using the omnibus SKAT-O test.

This analysis suggests several next steps. Given the concerns with the MAF weighting method that lead to the identification of MSGN1, and the not-genome-wide-significant p-values of the FEFR2, MKL1, and NEK10 in the primary analysis, these associations need to be replicated in an independent analysis. The ideal study population for replication would be restricted to women who developed breast cancer early, in order to better elucidate any differences in the genetic risks between the early and late onset disease. Additional extensions of this analysis may also be able to identify genes in which rare variation is

associated with risk if they incorporated sequencing data. This would allow for a more complete understanding of the relationship between exonic variation and early onset breast cancer. Similarly, as methods emerge that allow non-exonic variants to be annotated to a given gene, set-based tests will be able to expand understanding of how intergenic regions can be associated with breast cancer risk. At the present time, there is only a rudimentary ability to annotate non-exonic variants to genes, but this is a subject of much study. As the understanding of biological pathways improves, variants will be able to be annotated to a particular gene in ways that are more sophisticated than ANNOVAR's annotation. Regulatory variants that are not spatially near the genes that they regulate could be included in the analysis. When this is combined with the increased ability to assay rare variants that will be provided by next generation sequencing technologies, if these variants are responsible for breast cancer risk, their inclusion will improve the ability of gene-based tests to identify genes responsible for breast cancer.

The participants of these studies are all of a homogeneous age (younger than 51 at diagnosis), ancestral background (European), and gender (women). As breast cancer affects people of all ages, ancestral backgrounds, and genders, additional SKAT-O analyses in populations with different characteristics will help to determine whether the genes that harbor variation that is associated with breast cancer risk differ across populations. The same analyses that were carried out in this thesis, when applied to different populations, may uncover additional insight into the genetic basis of differences in risk and mortality that are associated with these non-genetic traits.

This analysis has some limitations. Since the study population has an unequal distribution of cases and controls, rare variants are more likely to be seen in the cases, resulting in more power to detect rare deleterious variants over rare protective ones. The participants in the replication data set were selected from breast cancer patients of all ages, so it can best provide evidence of replication for genes and variants whose effects are similar for both the

early- and late-onset disease, and is only able to replicate evidence of generic drivers of the early onset disease in the scenario that the gene is also causal in the late-onset disease. The replication data set also did not explicitly interrogate rare variation, so this analysis could only fully examine variants for replication if they could be well-imputed. The composition of the replication data set also included many of the same participants as the discovery analysis (17% of the cases in the replication and 2% of the controls), and replication with a fully independent population would have been preferable. A third concern is related to the external validity of the results. Different populations have distinct rare variants that are present in disease associated genes.[129] It would need to be verified whether novel SNVs in the implicated genes have the same effect on breast cancer risk as the ones interrogated by this study.

In conclusion, this analysis continues to suggest a role for the genes NEK10, FGFR2 and MKL1 in the genetic etiology of breast cancer. The associations in the MKL1 and NEK10 genes are driven by variants that had not been reported before, there may still be additional variants to be discovered that contribute to disease risk. These three genes, if validated, represent an increase in the understanding of the underlying biology of breast cancer carcinogenesis, and implicate their gene-products as being associated with tumor development. These four genes represent possible targets for future attempts at chemoprevention of breast cancer. The analysis indicates that the SKAT-O method identifies exonic variation that would not be identified using single marker regression methods, and suggests that analyses that restrict themselves to only single marker regressions will continue to find missing heritability in early onset breast cancer risk. This validates the extra cost and statistical complexity that is needed to measure them. This analysis also incorporates prior knowledge about each variant in a novel way, and avoids discarding data in favor of weighting each variant by the CADD deleterious score, and suggests that this weighting method that can be used when investigating the genetic architecture of other diseases.

CHAPTER 3

THE EFFECT OF GERMLINE GENETIC VARIATION IN GENE REGIONS ON SURVIVAL OF EARLY ONSET BREAST CANCER

3.1 Background

One in eight American women will develop breast cancer over her lifetime.[1] While treatments and survival rates improve over time, almost 25 percent of women who are diagnosed with breast cancer will eventually die of the disease. Fear of recurrence and mortality results in a lower quality of life for women who are diagnosed.[3–6] The risk factors that contribute to mortality are still imperfectly understood even though a better understanding of the drivers of mortality could help to develop interventions that prolong life.

Known risk factors for mortality include the maturity of the tumor at the time of detection, comorbidities of the patient, and treatment decisions,[130,131] as well as molecular markers that quantify the inherent aggressiveness of the cancer.[132–135] However, even among women whose cancers are detected and treated similarly, differences in survival persist after taking tumor subtype and stage at diagnosis into account.[136,137] Women who are diagnosed with breast cancer before the age of fifty comprise one group of women who are disproportionately likely to die from breast cancer. These women account for one in five of those diagnosed with invasive breast cancer,[2] and have lower three-, five-, and ten-year survival rates than women diagnosed after age 50.[28] These younger women tend to have both more aggressive tumor subtypes,[138–141] and independent of tumor subtype, poorer prognosis.[23–27]

Germline genetic variation may be responsible for a portion of the heterogeneity in mortality outcomes. Many biologically plausible pathways exist to connect germline genetic variation and breast cancer mortality, and any given causal variant may affect risk

through any of these different pathways. Germline variation may affect a patient's ability to metabolize a drug, which in turn can affect survival by altering the amount of available active metabolites of pharmaceutical treatments,[11,15,20,142–145] or increase the probability of treatment-limiting adverse events.[8–10] Similarly, germline genetic variation may be responsible for a cellular environment that favors metastases,[7,12,146,147] and may alter cellular functions such as angiogenesis,[13,15] growth signaling,[16] telomere length,[17] inflammation,[18] immune response,[19] DNA repair,[20,148,149] apoptosis,[21,142] and cell cycle control.[14]

At least three lines of evidence have implicated germline genetic variation as a risk factor for mortality in breast cancer patients: animal studies, family studies, and identified loci from linkage and candidate gene studies. Recent *in vivo* animal studies identified several variants in possible possibly drugable pathways that prevent metastases in mice.[150,151] Family studies have compared outcomes of related and unrelated women, and found that after controlling for shared environmental influences, the related women had more similar disease trajectories.[152,153] Specific variants and copy number alterations have been associated with mortality,[21,144,148,154–165] most of which have been identified through linkage analyses or candidate gene studies.

There is evidence that germline genetic variation plays a larger role in the etiology of early onset breast cancer than the etiology of the late onset disease.[26,166,167] This suggests that a unique set of variants may be responsible for poor outcomes in younger patients, although this has not been definitively established yet.

There is suggestive evidence that germline genetic variation may play a role in mortality in women diagnosed with breast cancer by mediating the effectiveness of pharmaceutical treatments, particularly in women who are treated with adjuvant tamoxifen therapy. In the early 1980's, tamoxifen was established as effective treatment to reduce mortality for women whose tumors were estrogen receptor positive (ER+). By 2002, adjuvant treatment

of these tumors with tamoxifen or other estrogen receptor antagonists has been recommended by the American Society of Clinical Oncology, which has lead to between 70 and 90% of women with ER+ tumor using tamoxifen as part of their therapy.[140,141]

Tamoxifen is metabolized by several enzymes that are encoded by genes that contain several highly polymorphic variants. Variation in the gene CYP2D6 has most convincingly associated with differential presence of the active metabolites of tamoxifen. This has lead to the suggestion that the presence of certain CYP2D6 genotypes could be used clinically to determine an appropriate dose of tamoxifen.[145,168,169] However, to this point, researchers have not been able to definitively demonstrate that the variation in internal dose that is a result of these polymorphisms translates to different survival outcomes.[43,44] A previous genome-wide investigation suggested that haplotype analyses that included CYP2D6 polymorphisms were associated with mortality,[145] although taken alone the CYP2D6 polymorphisms were not statistically significant, resulting in a still unknown significance of the effect of these CYP2D6 polymorphisms on mortality.

In addition to estrogen receptor antagonists, the last decade has seen the release of multiple other new pharmaceutical treatments for breast cancer that do not rely on cytotoxic chemotherapies. These newer drugs have pharmacodynamics that are not completely understood, and variants that are with genes that encode enzymes that metabolize these treatments (or variants that effect the expression of these enzymes) have a strong chance to be able to influence survival. This has not yet been determined.

Many statistical methods have been developed to identify germline genetic variation that is associated with disease. The most appropriate statistical method will depend on the still-unknown characteristics of those variants: how many are associated with mortality;[56] how they are distributed throughout the genome,[57] and the what is the form and strength of the relationship between the variant and mortality.[58,153] For polygenic traits, it is likely that the causal variants are distributed such that they may be found in multiple combinations of

these aspects, and therefore multiple methods may need to be employed in order to identify them all.

One method that has been employed to instigate the relationship between germline genetic variation and mortality is single marker regression analyses. Single marker regression analyses are an appropriate tool to identify loci that are associated with disease if the causal variant is common and of at least moderate effect size, and if the single marker that is being regressed is either the causal variant or tags it well. Studies have investigated the genome-wide genetic determinants of breast cancer mortality using single marker regression are summarized in Table 3.1 and Table 3.2.

In contrast to the evidence from animal studies, family studies, and candidate gene/linkage studies, the results of these studies have largely been null, and also poorly replicated. Many of the single marker regression analyses were carried out in small sample sizes (all but one meta analysis had a sample size of less than 2000), and they were only able to follow their cohort for a short amount of time relative to the expected median survival of breast cancer patients (the median study of single marker regressions followed the cases for 6 years, and none were able to follow for longer than 7 years).

Besides sample size, limitations inherent in single marker regression may have been the cause of the largely null results. Single marker regressions cannot identify risk loci where variants are too rare or too weak, and they are also limited by concerns about type I error rate. Single marker regression analyses conduct a large amount of tests, which requires employing strict significance thresholds in order to exclude false positives. In many cases, these thresholds can exclude many truly causal variants.[61]

To move beyond single marker regression tests, a class of tests has been developed to examine variation within a single region: set-based tests. Set-based tests shift the hypothesis from whether an individual variant is associated with a trait to whether any variation within a predefined set is associated with a trait. The sets are often defined as gene

Table 3.1: Genome-wide Studies of the Association between Germline Genetic Variation and Breast Cancer Mortality

Study Title	Year	Population Description	Outcome	N	Median Follow Up Time	Events	Variants	Replication Description	Findings
A Genome-Wide Association Study of Prognosis in Breast Cancer[91]	2010	Postmenopausal women with invasive breast cancer	Breast cancer specific survival	1145	6 years	93	528,252	Top 10 genotyped in 4335 women with invasive breast cancer with 38,148 years at risk	Nothing genome-wide significant
A Genome-wide Association Study Identifies Locus at 10q22 Associated with Clinical Outcomes of Adjuvant Tamoxifen Therapy for Breast Cancer Patients in Japanese[145]	2011	Japanese patients with hormone receptor-positive, invasive breast cancer receiving adjuvant tamoxifen therapy	Recurrence-free survival	240	7 years	30	470,796	Two independent sets of 105 and 117 cases	15 SNVs in the primary analysis; rs10509373 (chr10:76397814) replicated (combined $p = 1.26 \cdot 10^{-10}$)
Novel Genetic Markers of Breast Cancer Survival Identified by a Genome-Wide Association Study[170]	2012	Shanghai-resident Chinese women	Total mortality	1950	6 years	299	613,031	Top 49 associations replicated in 4160 Shanghai women with breast cancer; Top association examined in Nurses Health Study	rs3784099 (chr14:68283210; $p = 1.44 \cdot 10^{-8}$ in discovery only)
Identification of Inherited Genetic Variations Influencing Prognosis in Early-onset Breast Cancer[171]	2013	UK women aged 40 or younger at diagnosis	Breast cancer specific survival	536	4 years	236	487,496	Top 35 associations genotyped in 1,516 independent cases from the same early-onset cohort	Nothing genome-wide significant
Genome Wide Meta-Analysis Study for Identification of Common Variation Associated with Breast Cancer Prognosis[172]	2014	UK women aged 40 or younger at diagnosis, and Finish women of all ages	Breast cancer specific survival	1341	6 years	237	475,141, imputed to 7.5 million	1523 additional participants of the POSH study	Nothing genome-wide significant

positions refer to the HG19 assembly
Studies that published both single-study results and contributed to a meta analysis will be represented twice

Table 3.2: Genome-wide Studies of the Association between Germline Genetic Variation and Breast Cancer Mortality (continued)

Study Title	Year	Population Description	Outcome	N	Median Follow Up Time	Events	Variants	Replication Description	Findings
Identification of Novel Genetic Markers of Breast Cancer Survival[173]	2015	Meta-analysis of studies in populations of European ancestry	Breast cancer specific survival	37,954	5 years	2900	200,000-700,000; imputed to 9 million	N/A	rs148760487 (chr2:162922103; $p = 1.5 \cdot 10^{-8}$) and 27 others in high LD; rs2059614 (chr11:125389528; $p = 1.3 \cdot 10^{-9}$ in ER- cases
Polymorphism at 19q13.41 Predicts Breast Cancer Survival Specifically after Endocrine Therapy[174]	2015	Meta analysis of UK women aged 40 or younger at diagnosis, and Finish women of all ages	Breast cancer specific survival	1341	7 years	547	486,478	Two independent data sets with 5011 patients	Nothing genome-wide significant
Prediction of Breast Cancer Survival Using Clinical and Genetic Markers by Tumor Subtypes[175]	2015	Incident breast cancer cases in Seoul, South Korea	Recurrence-free survival	1732	4 years	214	2,210,580 genotyped and imputed	Any SNVs identified with $p < 10^{-6}$ and MAF > .1, and any common variants in high ($r^2 > 0.4$) LD with them were genotyped in 1494 additional women from South Korea	Nothing genome-wide significant

positions refer to the HG19 assembly
Studies that published both single-study results and contributed to a meta analysis will be represented twice

boundaries, so that the results can be interpreted easily in the context of cellular biology. Gene-based tests reduce the multiple testing burden when compared to a single marker regression, and also allow for variants to contribute evidence for risk that could not be assessed using standard single marker regression approaches, such as variants that are too rare to test individually, and common variants whose effects are too modest to detect with the strict significance thresholds necessitated by single marker regression tests. In the scenario where any disruption to a gene product can increase risk of disease, gene-based tests can detect those genes, even if any single variant is too weakly associated with disease to be detected with single marker regression analyses.

In many diseases, variants in protein coding regions of the genome harbor much of the variation that is associated with disease risk. While analyses that focus on gene regions exclude a large percentage of the genome, the central role of genes in transcription and ultimately amino acid translation makes variants that reside within gene boundaries represent biologically plausible candidates for association with disease.[34,74,176,177] In addition, variation outside of gene regions can be less reliably attributed to a particular gene, and therefore are problematic to include in gene-based tests. These considerations can justify the use of methods such as gene-based tests that can well-interrogate these regions, even if other complementary methods will then be required to examine the rest of the genome.

Next generation sequencing would comprehensively interrogate all variation within gene boundaries, but these technologies are still expensive to implement at a scale needed for epidemiologic genome-wide studies. In contrast, exome-based arrays directly measure some rare variation in gene regions, and cost less than whole-exome or whole-genome sequencing. Gene-based tests can also be easily implemented in studies that have measured genetic variation with exome arrays.

Many set-based tests have been developed that can be implemented as gene-based tests.[37,75–77] SKAT-O[37] combines two of the most commonly used methodologies: bur-

den tests and the sequence kernel association test (SKAT). The burden test is more powerful than the SKAT test if all of the minor variants in a gene increase risk of disease; and the SKAT test is more powerful than the burden test if the minor variants within a gene both increase and decrease the disease risk.[57] SKAT-O calculates both the burden test statistic and a SKAT test statistic for each gene, and then uses the data adaptively to weight and combine the two test statistics by a mixing factor. In most situations, SKAT-O is more powerful than either test alone.[37] The SKAT-O methodology has been extended to be implemented as a Cox regression,[178] allowing for an estimation of the hazard associated with each additional minor allele.[179] No set-based tests have as yet been applied to investigate the genome-wide genetic determinants of breast cancer prognosis.

Many studies that implement gene-based tests further restrict the variants, and include only those that are (1) rare or (2) independent annotation sources identify as "functional" (for example: nonsynonymous variants). "Nonfunctional " or "common" variants are excluded in an attempt to remove noise that may be introduced if those variants are not associated with the trait. However, these exclusions rest on one of two strong assumptions: (1) that rare variants hold all disease causing variants or (2) that previous knowledge of genetic function will continue to predict future variants that are associated with disease. Since the era of genome-wide analyses has frequently found new discoveries of biology in variants that were thought to be "junk" DNA,[180,181] an approach that allows all variants in a set to be interrogated would be preferable.

To this end, SKAT-O can incorporate prior knowledge about variants by means of a weight on the individual variants.[77,79] In most genome-wide analytic scenarios, the use of weights will not increase type I error rates,[38] and a weight that reflects the true disease process can significantly improve power.[37]

Currently, weighting is largely implemented in a way that up-weights variants that are rare, which operationalizes the assumption that evolutionary constraints keep variants that

strongly increase risk of disease at low frequency in the population. However, not all variants that cause disease are kept at a low frequency.[80] Several annotations have been developed that more comprehensively assess the probability that a given variant may influence a trait. These functional annotations include SIFT,[81] PolyPhen,[82] and combined annotation dependent depletion (CADD) score.[39] They each operationalize the knowledge that variation at certain portions of the genome are expected to have a greater effect on disease risk. Of these, the CADD algorithm combines many single dimensional annotations into one continuous score of the predicted "deleteriousness" of each variant in the genome. While there have been few attempts to translate these annotations into weights, the CADD scaled score has a range of values that is similar to the frequency weighting that is recommended by the SKAT authors. The scaled CADD score can be directly used to up-weight variants in the SKAT-O tests that are expected to affect survival, although this has not yet been done for any trait.

With this as background, this manuscript will investigate whether variation in genes is associated with mortality in a cohort of women who have been diagnosed with early onset breast cancer using a SKAT-O methodology. Since the SKAT-O approach examines the effect of all variants collectively, gene regions may be identified that contain rare variants or common variants of weak effect that would not have been identified through a single marker regression analysis alone. This manuscript will be the first to investigate directly the influence of rare variants in gene regions on breast cancer mortality, as all previous genome-wide investigations into the genetic determinants of breast cancer mortality have incorporated information from common variation assayed on genome-wide arrays, and the variants that can be reliably imputed from them.

This manuscript will also investigate whether germline genetic variation influences tumor characteristics that are identified at the time of diagnosis and are themselves known to be associated with prognosis: the tumor's estrogen receptor (ER), progesterone receptor

(PR), and human epidermal growth factor receptor 2 (HER2) expression statuses, and the grade and stage of the tumor at diagnosis. While the participants of this study are followed for a long time compared to most previous genome-wide survival studies in breast cancer, investigating the effect of germline genetic variations on these intermediate markers of tumor aggressiveness will complement the mortality analysis, and may produce valuable insight into genetic determinants of more deadly cancers that can be detected in their early stages.

Despite the benefits of understanding how genetic variation is associated with mortality in breast cancer patients, genome-wide investigations have not yet full characterized this relationship. This manuscript will investigate to what extent this lack of consensus is due to lack of use of gene-based tests, which can identify different classes of variation than single marker regression tests.

In order to use all data available from the exome array, this analysis applies weights that incorporate prior knowledge of the expected contribution of the variant. This approach allows the analysis to include all measured variation in a gene region, which would allow for the discovery of novel associations within that gene that were not a priori considered to be likely associations.

Gene-based tests will also allow for an investigation into whether all variation in the CYP2D6 region (or any other putative pharmacogenomic gene) collectively translates into differences in survival outcomes, either in all participants, or those with ER+ tumors (in younger women with ER+ tumors from these areas, 80-90% of them are likely to have been treated with tamoxifen[140,141]). Previous haplotype analyses have implicated this region, which suggests that gene-based tests, which also combine strength across several causative variants, may be an appropriate way to investigate this region.

This manuscript will also examine previously identified loci that have been identified with mortality in our relatively larger sample. While the analyses of this manuscript inter-

rogate a different set of variants than the previous genome-wide studies of mortality. Given the low levels of replication in the mortality analysis, this will help bring more evidence to whether the results from smaller studies represent a robust finding.

This analysis will also result in a better description of the similarities and differences of the genetic determinants of breast cancer that may be a function of age, and also compare the loci that are associated with risk with those that are associated with prognosis. The participants in the primary analysis are largely under the age of 50. Since most previous studies of breast cancer mortality were comprised of participants who were relatively older than this study population, the results will compare the genetic determinants of mortality between women who are diagnosed earlier and those diagnosed later. There is increasing evidence that many of the variants that are responsible for risk do not play a large role in prognosis,[182] and this manuscript will investigate whether previously identified variants for risk are also involved in either mortality or the development of more aggressive tumors in ways that can be measured at diagnosis (as measured by ER/PR/HER2 status, stage, and grade).

If genes are identified that are associated with breast cancer survival, this would implicate that gene and allow for an improved understanding of the biological processes that influence breast cancer mortality. If the effect sizes of the genes are large, the mortality analysis could identify genetic markers that could be used in conjunction with other non-genetic risk factors to identify patients who may benefit from additional treatment, and also those who may safely be able to decide upon a less aggressive treatment with fewer side effects. If the genetic prognostic factors from this population are similar than those found in the late onset breast cancer cases, these results could support using the same genetic risk scores for mortality for women with all ages of diagnosis, and if they differ, then further work will be needed to develop a prediction model that is most appropriate for early onset

Table 3.3: Characteristics of Studies Included in Primary Analysis

Study Name	Study Location	Years Recruiting	Case Criteria	Cases
Breast Cancer Family Registry	Australia	1992-2000	Living in the Melbourne and Sydney metro areas, family recruited from the Victoria and NSW cancer registries	477
Breast Cancer Family Registry	Ontario	2001-2010	Ontario Cancer Registry	559
Breast Cancer Family Registry	Philadelphia, PA	1996-2000	Living in Philadelphia	272
Breast Cancer Family Registry	New York, NY	1996-2000	Living in New York, New Jersey, or Connecticut	393
Breast Cancer Family Registry	Utah	1996-2012	Living in Salt Lake City	100
Genetic Epidemiologic Study of Breast Cancer by Age 50	Germany	1992-1995	38 clinics in the Rhein-Neckar-Odenwald and Freiburg regions	382
Long Island Breast Cancer Study Project	New York	1996-1999	Nassau and Suffolk counties	145
Seattle	Seattle, WA	1990-1992	King, Pierce, and Snohomish counties; age less than 45 at diagnosis	288
University of Chicago	Chicago, IL	1998-2010	Treated at the University of Chicago Cancer Center	181

Participants are those included in the analysis after QC

breast cancer. These results may also identify biological pathways that are responsible for tumor aggressiveness, which might result in the discovery of drug-able targets.

In the event that no genes are clearly associated with risk of mortality, this would provide further evidence that any as-yet-undiscovered mortality loci have either small effect sizes, or are driven by variants that were not measured by the exome array. This may help to guide future studies of this topic.

3.2 Methods

3.2.1 Primary Data: Breast Cancer Family Registry and Associated Studies

The participants for the primary analyses were identified from nine ongoing studies designed to assess the risk factors associated with early onset breast cancer. Participants were women of European descent and not known to be carriers of pathogenic mutations in the

genes BRCA1 or BRCA2. Ninety eight percent of the cases were younger than 50 years old at the time of their diagnosis. Details of the recruitment are found in Table 3.3. Five of the study sites (Australia, Ontario, Philadelphia, and New York) are members of the Breast Cancer Family Registry (BCFR), whose recruiting methods are described elsewhere.[63] Briefly, two of the BCFR centers (Northern California and Canada) recruited index patients through population-based registries, three (Utah, Philadelphia, and New York) recruited through clinic- and community-based outreach, and one (Australia) recruited through a mix of population- and clinic-based outreach. Participants were also included from four studies not included in the BCFR consortium. Three of these, the German Genetic Epidemiologic Study of Breast Cancer;[64] and Long Island Breast Cancer Study Project;[65] and the Seattle Study,[66] are population-based case control studies whose recruiting methods are described elsewhere. The Chicago participants were enrolled from the Chicago Multi-ethnic Breast Cancer Epidemiologic Cohort, a hospital-based study of breast cancer at the University of Chicago.[94,95] The Chicago participants were identified through a clinic-based recruitment. Their demographic, clinical, and pathological data were gathered from medical chart, epidemiologic risk factors were collected via structured questionnaire, and mortality outcomes were ascertained via medical records and linkages with the national death index.

3.2.2 Genotyping

Peripheral blood and mortality information was available for 3232 cases. The samples were whole genome amplified using the Qiagen Repli-G mini kit. The Illumina HumanExome 12v1.0 chip was used on 2527 cases, and the Illumina HumanExome 12v1.1 chip was used on 480 cases. The samples were processed using 49 plates in two batches, and the process was carried out according to the manufacturer's protocol. To improve the quantity and quality of available genomic DNA, the samples were whole genome amplified using

the Qiagen Repli-G mini kit,[22] and were processed using 49 plates in two batches, following the manufacturer's protocol. TeCan Evo was used for automation. Raw data was processed by Genome Studio on 2010.3 software, and the no-call threshold was set at 0.15, per Illumina's recommendation for Infinium chips. Clustering was done using the Illumina supplied cluster files. After keeping only variants that were on both chips, 238,524 variants were interrogated.

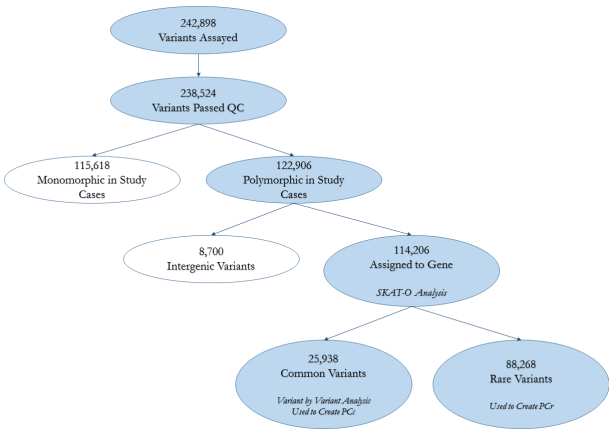
3.2.3 Primary Analysis

3.2.3.1 Quality Control

The quality control followed the protocol suggested by Guo et al.[96] Participants were excluded for low genotyping rate (rate < 95%; 166 excluded), male sex (one excluded), high heterozygosity (F statistic greater than three standard deviations from the mean, or heterozygosity greater than four standard deviations from the mean; 20 excluded), duplicated genotypes (one of each pair of six duplicates excluded), and principal component outliers (one participant whose first component was more than six standard deviations away from the mean). The recruitment process for the BCFR studies targeted individuals who were related to the index case, so an additional 16 participants who were likely related were excluded (those with estimated relatedness from a GCTA-created genetic relatedness matrix greater than 0.4).[97]

A schematic of the variants used in this analysis is shown in Figure 3.1. Variants were excluded from the analysis if they had a low call rate (rate < 95%; 4335 excluded). Variants were excluded from the analysis if they had a low call rate (rate < 95%; 4335 excluded), or if they were common variants (defined below) with Hardey-Weinberg equilibrium p-values of less than $2.5 \cdot 10^{-7}$ in controls (p = 0.05 Bonferroni corrected for 200,000 tests; 39 excluded). The final variant-level exclusions were the result of evidence that on some plates

Figure 3.1. Variants Used in Primary Analysis



variants were unreliably assigned (a plate-by-plate single marker regression analysis found that in some cases genotype could predict plate). For these variant-plate combinations, variants were excluded for all participants on that plate if this GWAS p-value was smaller than $2.5 \cdot 10^{-7}$. As a result of this step, seventy variant-plate combinations were set to missing.

After these exclusions, the analysis set contained 2954 cases and 238,524 variants. Of these, 122,906 were polymorphic in the study population. Variants were assigned to genes using the ANNOVAR software,[98] and variants were excluded if they were annotated to intergenic regions. This included in the analysis variants that were annotated as exonic (overlapping a coding region; n=105,888 variants), splicing (within two base pairs of a splicing junction; n=811), non-coding exonic RNA (n=163), non-coding intronic RNA (n=871), 5' and 3' untranslated regions (n=186 and n=474, respectively), introns (n=5456), and variants within 1 kilobase of a transcription start or transcription end site (n=187 and n=170, respectively). These 114,206 polymorphic variants from the exome array were annotated to 16,317 genes. Variants were classified as "common" and "rare" based on their minor allele frequency (MAF), with a threshold at MAF equal to $\left(\frac{1}{2n}\right)^{\frac{1}{2}} = 0.0130$.[59]

Table 3.4: Characteristics of Participants in Primary Analysis

Number of Participants	2954
Mean Age at Diagnosis (sd)	41 (6.2)
Median Years of Follow Up (IQR)	15 (9.5-17)
Number of Deaths Observed	728
Estrogen Receptor Status	
Positive	1066
Negative	719
Missing	1169
Progesterone Receptor Status	
Positive	1015
Negative	754
Missing	1185
HER2 Status	
Positive	280
Negative	378
Missing	2296
Tumor Grade	
High Tumor Grade	865
Low Tumor Grade	925
Missing	1164
Tumor Stage	
High Tumor Stage	351
Low Tumor Stage	1289
Missing	1314

Characteristics of the 2954 primary data participants that were included after quality control are found in Table 3.4. Cases were followed up for a median of 15.2 years (interquartile range: 9.5-17.0 years), and 728 deaths were observed.

3.2.3.2 Population Stratification

Rare variants and common variants have different correlations with ancestry, and therefore have different potential to induce confounding in genetic association studies.[99,100] While the study enrollment was limited to women of the same race (self-identified non-Hispanic white women) the study included women from multiple centers in four different countries, with an uneven case/control mix from each study. To counter this potential for spurious associations between variation and prognosis, EIGENSTRAT[101,102] constructed two sets

of principal components from the analysis set. One set was constructed using "common" variants assayed by the array (PC_c), and one using "rare" variants (PC_r).

In a Cox regression that did not include genetic information, the first three PC_c and the second PC_r were associated with mortality status. Including any other principal components did not improve the model fit, as determined by a likelihood ratio test. These four principal components were included in all subsequent mortality analyses. Similar analyses with non-genetic information were done to determine the optimal number of PC_c and PC_r to include in each of the five tumor characteristic logistic analyses.

3.2.3.3 All Variation in Gene Regions

To examine whether variants within a gene collectively were associated with mortality, the variants were aggregated into their annotated genes and analyzed using SKAT-O.[37] Each variant was weighted by the CADD scaled score of the minor allele. In addition to controlling for principal components, the analysis also controlled for center. The analysis was conducted in a Cox regression semi parametric framework to estimate the hazard ratio associated with each additional minor allele using the skatMeta R package.[123] The score statistics for the individual variants were calculated using the likelihood ratio test.[178] Genes whose p-values were smaller than the Bonferroni-corrected level of $3.06 \cdot 10^{-6}$ would be considered associated with early onset breast cancer mortality. Results were visualized using the qqman[105] and ggplot2[106] R software packages.[107]

The analysis was repeated twice: on all cases (N = 2954), and on all cases with ER+ tumors (n=1067), to investigate whether variation germline genetics may be particularly influential by way of genes that influence the metabolism of drugs that target the estrogen receptor growth signaling pathway. While treatment information was not available for these participants, and the date of initial diagnosis was not always available, it is likely that most

were treated with tamoxifen, which has been a proven beneficial adjuvant therapy since the 1980's,[183] and recommended by the American Society of Clinical Oncology as adjuvant therapy for women with ER+ breast cancer tumors since 2002.[184]

To examine whether variants within a gene collectively were associated with tumor characteristics that are known to be predictors of mortality, the variants were analyzed using the SKAT-O method in a logistic regression framework. Five tumor characteristics were assessed: ER status (ER status was non-missing for n=1785 cases), PR status (n=1769 cases), HER2 status (n=658 cases), whether the tumor grade was three or higher (n=1790 cases), and whether the tumor stage was three or higher (n=1640 cases). The analysis was conducted using the SKAT package for R, with the "SKATO" method in the function SKATBinary with efficient resampling.[108] The analysis weighted each variant by the CADD score for the minor allele, and controlled for principal components as described above. For each of the methods, the significance threshold was determined by correcting a $p<0.05$ threshold by the effective number of tests computed, which was determined by the SKAT package. Genes whose p-values were less than this threshold would be considered associated with that tumor characteristic.

3.2.4 Replication Data and Comparison with Breast Cancer of All Ages of Onset: TCGA

Data from participants of The Cancer Genome Atlas (TCGA) breast cancer study (data generated by the TCGA Research Network: http://cancergenome.nih.gov/) were utilized to replicate any suggestive findings (defined below) from the primary analysis, and to compare results between early onset breast cancer cases and cases of all ages of onset. Clinical and single nucleotide variant (SNV) data for all available breast cancer cases were downloaded from the TCGA data portal in June 2015. The germline SNV data were measured using

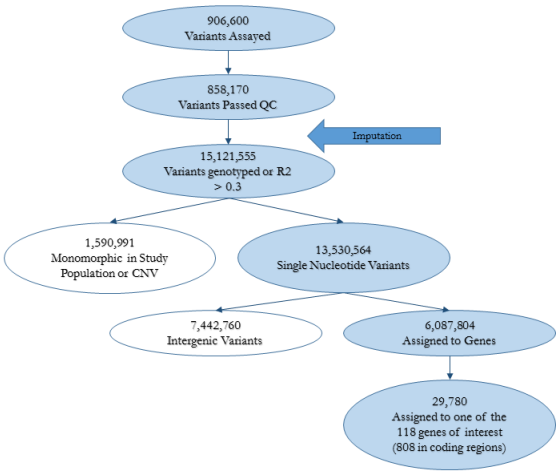
the Affymetrix Genome-Wide Human SNP 6.0 array, and the intensities were converted to genotype calls using the Broad Birdsuite.[185] To be comparable with the primary analysis, the analysis was restricted to female cases of European ancestry with mortality information available, excluding any participants or samples annotated "DNU" (Do Not Use) (900,380 variants from 768 participants).

These samples were then subjected to the same quality control steps outlined above, resulting in the following exclusions: two participants were excluded for high levels of missingness; thirty nine were excluded for high heterozygosity; sixteen for outlying principal components, and none were highly related or duplicates. 42,150 variants were excluded due to their low call rate, and 60 were excluded for failing Hardey-Weinberg equilibrium. After these quality control procedures, 711 cases and 858,170 variants were brought forward for imputation.

A schematic of the variants used at each stage of the analysis is found in Figure 3.2. Imputation was implemented by the Michigan imputation server,[186] employing ShapeIt[187] to pre-phase the variants and minimac3[188] to impute variants that were not measured. In order to best impute rare variants,[189,190] the entire 1000 Genomes phase 3 release[132] was used for a reference panel. Since the data from the TCGA participants was used to replicate suggestive associations, it was decided to use a liberal threshold for imputation quality that could still exclude low-quality variants, so variants with an imputation r^2greater than 0.3 were kept (15,121,555 variants).[191,192] These genotyped and imputed variants were then annotated using ANNOVAR, and only polymorphic variants that could be annotated to a gene were considered for analysis (6,087,804 variants). EIGENSTRAT was used to create ten principal components out of common (MAF > 0.0265) variants. Characteristics of the participants used for the replication analysis are found in Table 3.5.

To examine whether variants within genes that had been identified by the primary analysis were collectively associated with mortality in the TCGA population, the TCGA variants

Figure 3.2. Variants Used in Replication Analysis



were aggregated into genes in the same manner described above, and those genes identified as suggestively associated in the primary analysis were analyzed in the TCGA population using a Cox regression and the SKAT-O method, controlling for the minimum necessary principal components as described above. All 711 TCGA cases were included in the mortality analysis, and the 538 ER+ cases were included in the ER+ only mortality analysis. The tumor characteristic analysis included all TCGA participants that had non-missing clinical data for that tumor characteristic: 670 cases for the analysis of ER status, 667 cases for PR status, 492 for HER2 status, and 699 cases for tumors with high stage (grade was not available in the protected access TCGA clinical data).If fewer than 20 genes were associated with the trait in the primary analysis, the top 20 genes associated with each trait in the primary analysis were then investigated for evidence of association with that trait in the TCGA population. Genes with a p-value in the TCGA analysis that was less than the Bonferroni corrected level threshold 0.0025 ($p = \frac{0.05}{20}$) would be considered suggestively associated with the trait.

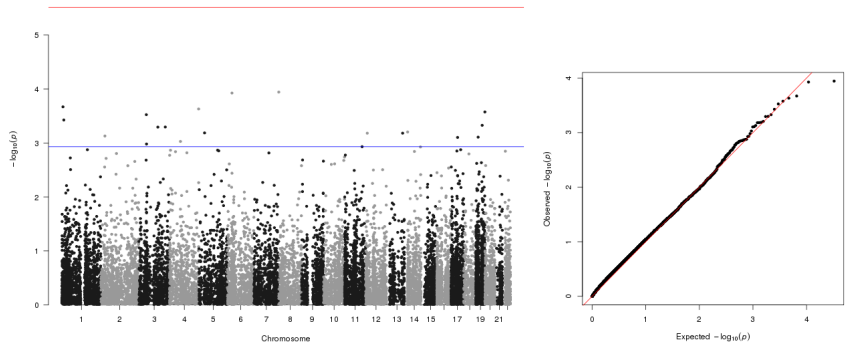
Table 3.5: Characteristics of Participants in Replication Analysis

Number of Participants	711
Mean Age at Diagnosis (sd)	59 (13)
Median Years of Follow Up (IQR)	1.2 (0.4-3.4)
Number of Deaths Observed	73
Estrogen Receptor Status	
Positive	538
Negative	132
Missing	41
Progesterone Receptor Status	
Positive	469
Negative	198
Missing	44
HER2 Status	
Positive	106
Negative	386
Missing	219
Tumor Stage	
High Tumor Stage	183
Low Tumor Stage	516
Missing	12

3.2.5 Comparison with Loci Identified through Single Marker Regression

It is unclear whether genes that would be identified by the SKAT-O approach contain loci that would also be identified using a single marker regression approach. To investigate this in the context of the data provided by the primary study participants, the exonic genetic variation measured on the primary sample was additionally analyzed in a single marker regression framework. Common variants that could be assigned to a gene (MAF >0.0130, 25,938 common variants) were analyzed using a Cox regression with an additive model of inheritance controlling for principal components and center using the GenABEL[193,194] package for the R software. Variants whose p-values were smaller than the Bonferroni corrected level of $1.93 \cdot 10^{-6}$ would be considered associated with early onset breast cancer mortality. Similarly, the common variants were assessed for their association with tumor characteristic phenotypes using a logistic regression framework and the PLINK software.[103,104] All analyses controlled for principal components as described above.

Figure 3.3. SKAT-O Cox regression Mortality Results for All Cases



The red line represents a p-value of $3.06 \cdot 10^{-6}$, and the blue line represents the p-value of the twentieth most significant gene.

3.2.6 Comparison with Previously Identified Risk Loci

Previous genome-wide studies that examined the effect of germline genetics on mortality have had low levels of replication. To investigate whether loci in gene regions that had been identified by previous research show evidence of association with mortality or tumor characteristics in the gene-based analysis, the significance of genes near loci with established associations to breast cancer phenotypes were highlighted in the analysis of the primary data.

Genes were considered to have established associations if a variant was listed the NHGRI-EBI GWAS catalog[60] with a p-value $< 5 \cdot 10^{-8}$ and mapped to a non-intergenic region. The NHGRI-EBI catalog contained 348 entries of a breast cancer phenotype (excluding alopecia as a response to chemotherapy, and excluding telomere length). Of these, 174 variants met the p-value requirement, and 113 were annotated to 65 unique genes.

Table 3.6: Comparison of Genes Suggestively Associated with Mortality in Primary and Replication Analyses

	Exome Chip			TCGA		
Gene	Minor Allele Count	Variants in Gene	p-value	Minor Allele Count	Variants in Gene	p-value
AKT2	13	5	4.70e-04	43,789	184	6.90e-01
ASXL2	77	13	7.43e-04	170,121	496	6.11e-01
BIVM	33	3	6.58e-04	68,029	193	4.57e-01
CDHR4	406	7	1.05e-03	9331	38	1.89e-01
CFAP97	3006	7	2.34e-04	13,1011	251	4.46e-01
CLCN6	14	3	3.74e-04	61,332	249	5.21e-02
COL7A1	644	39	2.99e-04	5738	73	8.80e-01
HLA-A	121	1	1.19e-04	164,007	308	5.22e-01
HSPBAP1	231	5	5.08e-04	85,394	309	6.98e-01
MPP2	47	4	7.88e-04	49,636	159	6.36e-01
OR4K14	208	3	6.25e-04	9093	17	3.46e-01
PIH1D2	330	5	1.17e-03	3426	36	1.21e-01
PRLR	496	12	6.50e-04	172,470	871	1.51e-02
SKIL	595	5	5.05e-04	38,076	178	7.79e-01
STYK1	4	4	6.58e-04	146,794	304	1.70e-01
TAS1R1	582	25	2.14e-04	18,748	78	5.13e-01
UGT2A3	33	3	9.36e-04	36,350	141	7.20e-01
ZNF134	651	11	2.66e-04	5266	44	5.13e-02
ZNF333	1152	12	7.79e-04	104,105	243	1.30e-01
ZNF596	501	8	1.14e-04	25,040	81	7.56e-01

3.3 Results

3.3.1 Association between Variation in Gene Regions and Mortality

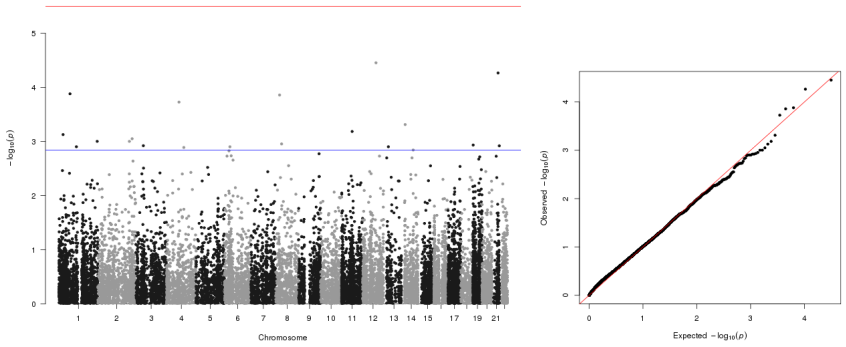
A summary of the gene-based analysis of mortality in breast cancer cases is shown in Figure 3.3. There is no evidence of systemic inflation of p-values, since the estimate of genomic inflation is $\lambda = 1.04$. No genes are associated with mortality with a p-value of $3.06 \cdot 10^{-6}$ or smaller. To determine if the results were sensitive to the weighting method, the analysis was repeated using weights that were a beta transformation of the MAF (as suggested by the SKAT authors), and equal weights. These two additional weighting methods produce substantively similar null results.

Analysis of the genetic data provided by the participants of the TCGA study does not provide any evidence that any of the genes suggestively identified in the primary data set have an association with mortality in the TCGA population. Table 3.6 displays the p-values of the top 20 most significant genes from this analysis, and contrasts this with the p-values of those genes in the TCGA population. None of the genes that are considered suggestive in the primary participants (labeled "Exome Chip") are significant at the Bonferroni corrected level in the TCGA analysis. One gene, PRLR is significant at the less stringent threshold of $p < 0.05$.

The subset of primary analysis patients with known ER+ tumors was included in a second Cox analysis that assessed the association between genetic variation and mortality. The results of this analysis when using CADD weights on each variant are summarized in Figure 3.4. Genomic inflation is $\lambda = 0.995$. No genes are associated with mortality with a p-value of $3.06 \cdot 10^{-6}$ or smaller. Weighting by MAF and weighting using equal weights produce substantively similar null results.

Table 3.6 displays the p-values of the top 20 most significant genes from the analysis in the primary data, and contrasts this with the p-values of those genes in the analysis

Figure 3.4. SKAT-O Cox regression Mortality Results for ER+ Cases

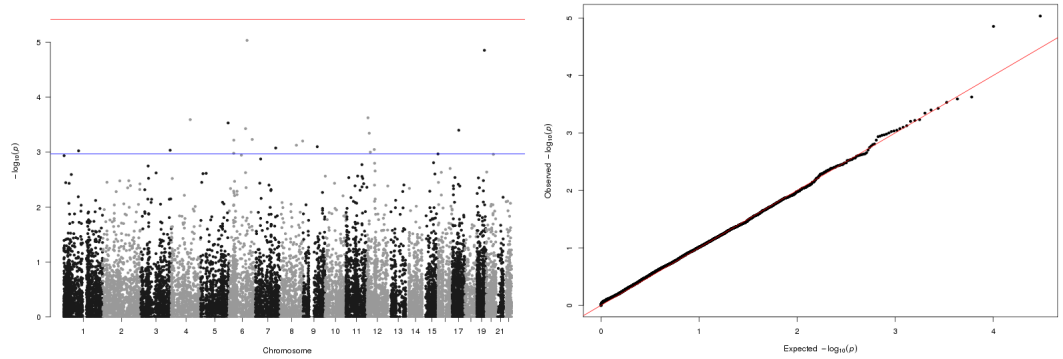


The red line represents a p-value of $3.06 \cdot 10^{-6}$, and the blue line represents the p-value of the twentieth most significant gene.

Table 3.7: Comparison of Genes Suggestively Associated with Mortality in ER+ Cases in Primary and Replication Analyses

	Exome Chip			TCGA		
Gene	Minor Allele Count	Variants in Gene	p-value	Minor Allele Count	Variants in Gene	p-value
ADGRA2	59	13	1.11e-03	12,396	101	6.02e-01
ALX1	33	3	3.52e-05	5423	82	9.35e-01
ANXA3	16	6	1.88e-04	27,042	237	6.87e-01
CFLAR	239	1	9.95e-04	17,399	124	3.19e-01
CHIA	1616	13	1.25e-03	56,892	292	4.13e-01
COL6A1	1127	14	1.20e-03	23,491	162	3.57e-02
CTH	323	3	1.32e-04	13,178	139	6.05e-01
DLST	30	2	1.44e-03	14,386	106	3.48e-01
ERG	942	2	5.43e-05	186,443	1562	9.65e-02
GAR1	0	1	1.29e-03	3587	57	5.56e-01
HNRNPU	9	3	9.93e-04	1776	23	7.31e-01
JSRP1	84	4	1.16e-03	1061	6	5.23e-01
MEDAG	147	2	1.25e-03	334	8	7.42e-01
NEFM	105	10	1.39e-04	987	17	8.92e-01
PNPLA1	1246	10	1.25e-03	35,219	351	9.75e-01
PTPRCAP	18	3	6.53e-04	1272	8	2.81e-01
SEPN1	465	7	7.45e-04	11,724	66	2.56e-01
SHISA5	1	1	1.20e-03	5935	109	1.76e-01
THTPA	93	1	4.87e-04	299	6	3.73e-01
WNT10A	34	5	8.90e-04	1989	23	1.59e-01

Figure 3.5. SKAT-O Logistic Regression Results for ER Status



The red line represents a p-value threshold based on a Bonferroni correction of the effective number of tests, as calculated by the SKAT package; the blue line represents the p-value of the twentieth most significant gene

of TCGA participants. None of the genes that are considered suggestive in the primary analysis are significant at the Bonferroni corrected level in the TCGA analysis. One gene, COL6A1, is significant at the less stringent nominal p-value threshold of $p < 0.05$. The gene CYP2D6, which encodes the enzyme which metabolizes tamoxifen into its active form, is not associated with mortality in participants with ER+ breast cancers (p=0.880 in the primary analysis). Other genes in the CYP family produced similar non-significant associations.

3.3.2 Association between Variation in Gene Regions and Tumor Subtype

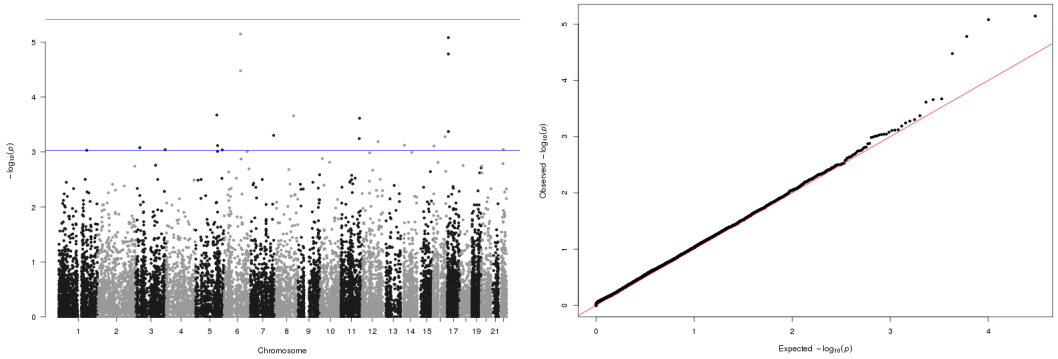
A summary of the gene-based analysis of tumor characteristics in the BCFR cases is shown in Figures 3.5 to 3.9. Estimated genomic inflation is low for each analysis: $\lambda = 1.013$ for ER status; $\lambda = 1.043$ for PR status; $\lambda = 1.017$ for HER2 status; $\lambda = 0.998$ for high tumor grade; and $\lambda = 1.066$ for high tumor stage.

None of the genes reach the genome-wide threshold for significance in the primary analysis for any of the tumor characteristics. Two neighboring genes on chromosome 17 that are located 90 kilobases downstream from the tumor suppressor gene TP53 are among

Table 3.8: Comparison of Genes Suggestively Associated with ER Status in Primary and Replication Analyses

	Exome Chip			TCGA		
Gene	Minor Allele Count	Variants in Gene	p-value	Minor Allele Count	Variants in Gene	p-value
ARHGAP29	23	9	9.47e-04	21631	216	2.25e-01
ARL10	14	2	2.94e-04	5050	33	1.66e-01
C12orf60	1666	5	4.53e-04	12083	80	2.37e-01
C9orf47	1661	7	7.96e-04	4675	33	5.66e-01
IL1RAP	976	5	9.24e-04	95561	761	1.42e-01
KATNA1	10	4	5.88e-04	42989	200	5.19e-01
KCNJ8	9	2	1.00e-03	397	14	5.37e-01
KLF10	13	4	7.49e-04	3667	38	1.00e+00
LRRK1	2511	23	1.08e-03	141047	954	8.03e-01
LY6G5B	1936	5	6.06e-04	1607	25	2.97e-01
MEIS3	17	4	1.40e-05	4214	36	7.19e-03
P3H3	2884	15	2.37e-04	12076	61	1.07e-01
POU5F1	5215	6	1.05e-03	17916	95	6.13e-01
PRDM5	2531	5	2.56e-04	205478	1257	8.89e-01
QRSL1	507	6	3.74e-04	37784	242	5.62e-02
SLC25A39	1676	2	3.99e-04	4748	30	1.00e+00
SLC38A4	210	3	8.98e-04	39639	356	5.01e-01
TMEM209	33	6	8.42e-04	24913	171	2.75e-01
TSPYL1	583	4	9.25e-06	4728	33	5.29e-01

Figure 3.6. SKAT-O Logistic Regression Results for PR Status

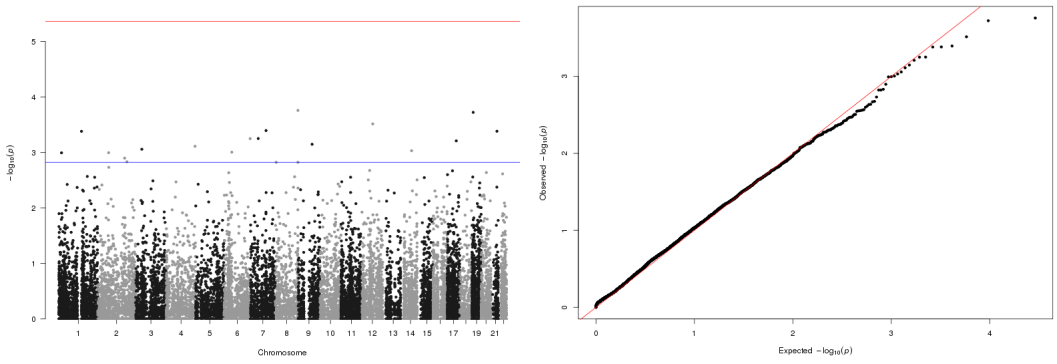


The red line represents a p-value threshold based on a Bonferroni correction of the effective number of tests, as calculated by the SKAT package; the blue line represents the p-value of the twentieth most significant gene

Table 3.9: Comparison of Genes Suggestively Associated with PR Status in Primary and Replication Analyses

	Exome Chip			TCGA		
Gene	Minor Allele Count	Variants in Gene	p-value	Minor Allele Count	Variants in Gene	p-value
AIM1	2169	16	3.34e-05	67054	411	1.00e+00
BCL9L	68	8	2.44e-04	2868	41	6.70e-01
C12orf42	1835	5	6.50e-04	111518	989	8.73e-01
CD68	913	5	8.33e-06	2203	15	8.79e-01
IL1RAP	964	5	9.08e-04	95092	761	7.60e-01
LIF	43	2	9.03e-04	4224	30	3.82e-01
MON1B	72	4	5.30e-04	15354	80	4.41e-01
MPDU1	572	4	1.65e-05	1635	14	1.00e+00
MYH6	1643	13	7.55e-04	13928	135	4.45e-01
NPM1	1554	1	9.19e-04	22640	86	4.36e-01
PCDHA4	1750	6	7.66e-04	4486	22	3.49e-02
PDIA4	118	13	4.99e-04	14658	139	6.83e-01
QRSL1	500	6	7.17e-06	37658	242	3.68e-01
RNF214	1259	2	5.71e-04	38468	274	3.76e-01
SHBG	403	4	4.25e-04	9594	68	8.89e-01
SLC4A7	2604	9	8.35e-04	81498	538	3.72e-01
TDRD5	2514	8	9.34e-04	88788	487	1.65e-01
TGFBI	46	16	2.13e-04	32096	183	6.85e-01
UBN1	141	11	7.77e-04	32105	185	1.68e-02
UTP23	540	2	2.20e-04	3557	43	2.69e-01

Figure 3.7. SKAT-O Logistic Regression Results for HER2 Status

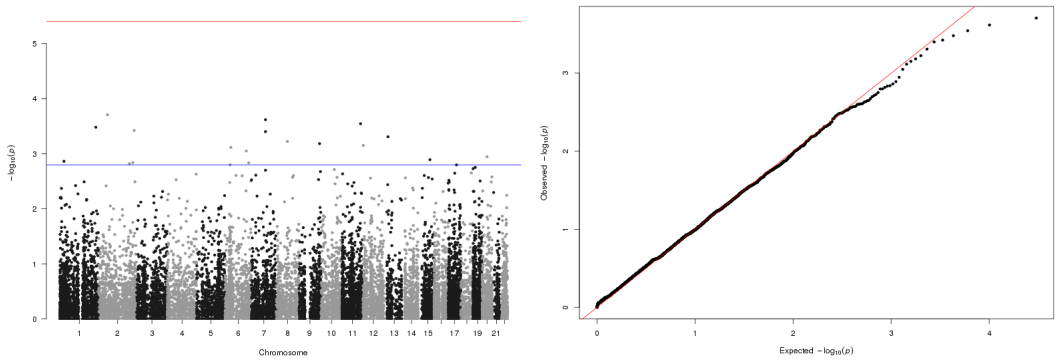


The red line represents a p-value threshold based on a Bonferroni correction of the effective number of tests, as calculated by the SKAT package; the blue line represents the p-value of the twentieth most significant gene

Table 3.10: Comparison of Genes Suggestively Associated with HER2 Status in Primary and Replication Analyses

	Exome Chip			TCGA		
Gene	Minor Allele Count	Variants in Gene	p-value	Minor Allele Count	Variants in Gene	p-value
ANGPT2	14	3	1.50e-03	80784	593	9.01e-01
CERS4	720	6	1.89e-04	32624	202	1.22e-01
CPM	23	2	3.06e-04	54047	636	6.27e-01
CROCC	199	13	1.01e-03	15275	152	1.42e-01
CTSL	8	2	7.12e-04	2255	31	8.91e-01
DACT2	515	6	5.63e-04	22474	203	4.29e-01
FAM171B	10	3	1.47e-03	23841	190	3.39e-01
GALNT16	1100	4	9.30e-04	64711	501	4.73e-01
GPT	661	9	1.50e-03	2003	8	6.99e-01
HEATR6	98	8	6.17e-04	2894	69	5.23e-01
MCM7	385	8	4.03e-04	3023	28	6.73e-01
METAP1D	556	6	1.27e-03	41095	442	1.25e-01
PCYOX1	245	2	1.01e-03	6244	88	4.32e-01
PIAS3	29	5	4.16e-04	34	1	9.17e-01
PIGP	226	3	4.14e-04	6035	49	8.13e-01
RECQL4	594	13	1.74e-04	4361	16	5.90e-01
RWDD4	291	1	7.73e-04	23355	141	2.62e-01
TFAP2B	495	4	9.89e-04	11301	131	7.63e-01
VWC2	757	5	5.61e-04	57463	576	1.00e+00
ZNF620	59	4	8.75e-04	3274	27	5.39e-01

Figure 3.8. SKAT-O Logistic Regression Results for High Tumor Grade

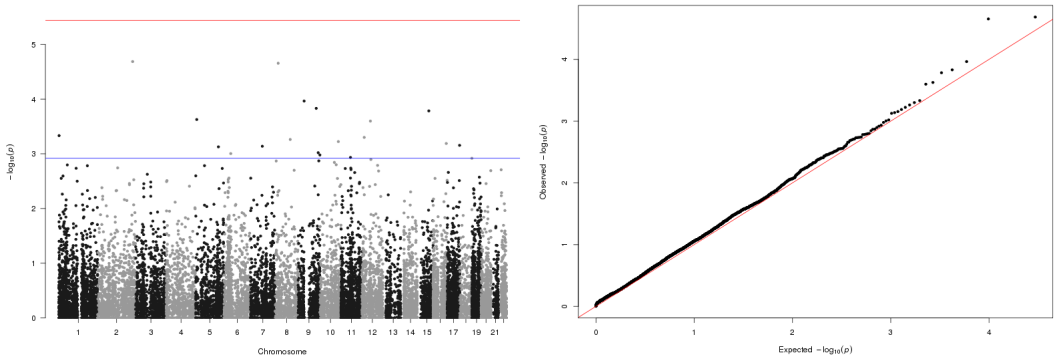


The red line represents a p-value threshold based on a Bonferroni correction of the effective number of tests, as calculated by the SKAT package; the blue line represents the p-value of the twentieth most significant gene

Table 3.11: Genes Suggestively Associated with High Tumor Grade in Primary Analysis

Gene	Minor Allele Count	Variants in Gene	p-value
C8orf37-AS1	1057	2	5.47e-04
CNTNAP3	628	1	1.09e-04
COL27A1	7713	26	1.48e-04
CTDP1	604	5	1.21e-03
GUCY2C	26	5	5.01e-04
HOXC4	1596	3	2.52e-04
NPLOC4	1863	4	7.00e-04
NSUN2	686	5	2.36e-04
PNLIPRP3	306	8	5.98e-04
PNPLA7	329	14	1.05e-03
RBM27	15	3	7.47e-04
SCG2	84	4	2.06e-05
SH2D4A	837	7	2.21e-05
SSC4D	2202	6	7.28e-04
STRA6	2276	10	1.65e-04
TAF1C	1623	15	6.48e-04
TCTE1	1314	11	9.92e-04
TMEM132A	886	9	1.17e-03
TMEM88B	172	2	4.66e-04
ZBTB43	14	4	9.55e-04

Figure 3.9. SKAT-O Logistic Regression Results for High Tumor Stage

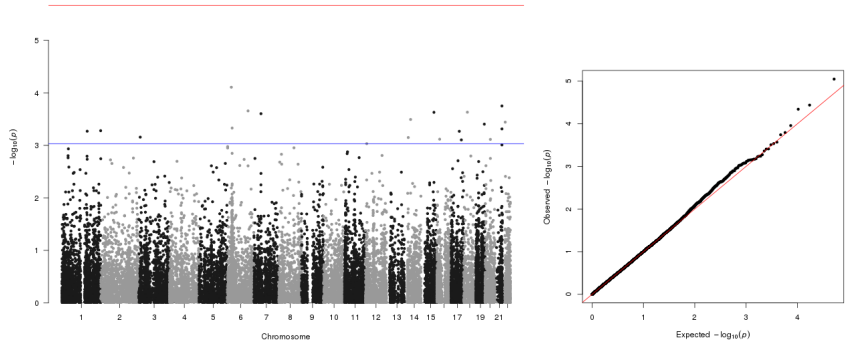


The red line represents a p-value threshold based on a Bonferroni correction of the effective number of tests, as calculated by the SKAT package; the blue line represents the p-value of the twentieth most significant gene

Table 3.12: Comparison of Genes Suggestively Associated with High Tumor Stage in Primary and Replication Analyses

	Exome Chip			TCGA		
Gene	Minor Allele Count	Variants in Gene	p-value	Minor Allele Count	Variants in Gene	p-value
C8orf37-AS1	1057	2	5.47e-04	525417	3137	5.96e-01
CNTNAP3	628	1	1.09e-04	42217	161	8.01e-01
COL27A1	7713	26	1.48e-04	165929	986	1.46e-01
CTDP1	604	5	1.21e-03	92110	465	1.00e+00
GUCY2C	26	5	5.01e-04	14059	333	5.68e-01
HOXC4	1596	3	2.52e-04	14994	116	2.28e-01
NPLOC4	1863	4	7.00e-04	96837	387	3.65e-01
NSUN2	686	5	2.36e-04	48493	236	2.07e-01
PNLIPRP3	306	8	5.98e-04	11371	200	6.80e-01
PNPLA7	329	14	1.05e-03	20244	89	1.00e+00
RBM27	15	3	7.47e-04	33619	222	6.18e-01
SCG2	84	4	2.06e-05	346	23	4.21e-01
SH2D4A	837	7	2.21e-05	47631	522	4.89e-01
SSC4D	2202	6	7.28e-04	27812	136	3.39e-01
STRA6	2276	10	1.65e-04	20707	136	2.85e-01
TAF1C	1623	15	6.48e-04	15310	98	4.25e-01
TCTE1	1314	11	9.92e-04	10972	116	3.86e-01
TMEM132A	886	9	1.17e-03	16624	65	1.31e-01
TMEM88B	172	2	4.66e-04	1058	11	7.67e-01
ZBTB43	14	4	9.55e-04	6467	111	6.02e-02

Figure 3.10. Single Marker Regression Cox regression Mortality Results for All Cases



The red line represents a p-value of $1.93 \cdot 10^{-6}$, and the blue line represents the p-value of the twentieth most significant SNV.

the three most significant genes identified as being associated with PR status, CD68 ($p = 8.33 \cdot 10^{-6}$) and MPDU1 ($p = 1.65 \cdot 10^{-5}$).

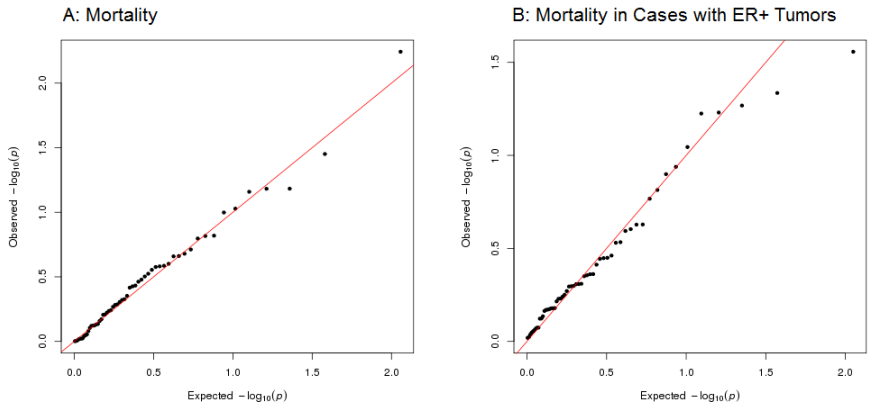
Tables 3.8 to 3.12 show the p-values of the top genes for each analysis in the primary analysis compared to their p-values in the analysis of the TCGA participants (tumor grade was not assessed in TCGA, so Table 3.11 shows only the p-values in the primary analysis). None of the genes that are considered suggestive of association in the primary analyses are significant at the Bonferroni corrected threshold in the TCGA replication analysis. With a less stringent threshold for replication of p<0.05, three genes were suggestively associated with tumor characteristic. MEIS3 is the gene with the smallest p-value in the primary ER analysis ($p = 1.40 \cdot 10^{-5}$) and nominally associated with ER status in the TCGA population ($p = 7.19 \cdot 10^{-3}$). PCDHA4 and UBN1 are in the top 20 genes associated with PR status in the primary analysis ($p = 7.66 \cdot 10^{-4}$ and $7.77 \cdot 10^{-4}$, respectively, and had a p-value smaller than 0.05 in the TCGA analysis ($p = 3.49 \cdot 10^{-2}$ and $1.68 \cdot 10^{-2}$).

Table 3.13: Variants Suggestively Associated with Mortality in Primary Analysis

CHR	Position	Variant Name	Gene	Minor Allele	Major Allele	MAF	Annotated Function*	HR (se)	p-value
6	35430686	exm541201	FANCE	G	A	0.055	nonsynonymous SNV	1.558 (0.155)	8.948E-06
20	1551564	exm1519096	SIRPB1	T	C	0.018	nonsynonymous SNV	1.933 (0.309)	3.649E-05
1	116310967	exm86861	CASQ2	T	C	0.287	nonsynonymous SNV	1.260 (0.072)	4.555E-05
8	19819724	exm686341	LPL	C	G	0.100	stopgain	1.371 (0.112)	1.104E-04
17	66364691	exm1348351	ARSG	C	G	0.429	nonsynonymous SNV	0.820 (0.043)	1.606E-04
6	35423886	exm541160	FANCE	C	T	0.014	nonsynonymous SNV	1.933 (0.340)	1.807E-04
10	108543337	exm2267041	SORCS1	A	G	0.413	intronic	0.820 (0.045)	2.696E-04
6	34730395	exm2257778	SNRPC	C	T	0.014	synonymous SNV	1.934 (0.352)	2.913E-04
3	52833219	rs2535629	ITIH3	G	A	0.347	intronic	0.813 (0.047)	3.111E-04
1	205318321	exm2250505	KLHDC8A	C	T	0.387	intronic	0.823 (0.045)	3.884E-04
2	87044316	rs6547705	CD8B	A	G	0.222	intronic	0.787 (0.053)	3.924E-04
2	179545859	exm247903	TTN	C	T	0.289	nonsynonymous SNV	1.219 (0.069)	4.364E-04
2	179432185	exm246313	TTN	A	G	0.288	nonsynonymous SNV	1.216 (0.069)	5.309E-04
7	123599845	exm654516	SPAM1	A	T	0.021	nonsynonymous SNV	1.704 (0.264)	5.757E-04
14	24458162	exm1091506	DHRS4L2	G	C	0.332	nonsynonymous SNV	1.204 (0.065)	5.770E-04
3	52874288	exm2273373	TMEM110	T	C	0.243	UTR3	0.800 (0.052)	5.913E-04
22	44324727	exm1615904	PNPLA3	C	G	0.233	nonsynonymous SNV	1.223 (0.072)	6.575E-04
16	17475645	rs7195703	XYLT1	T	C	0.455	intronic	1.199 (0.064)	6.763E-04
19	52249211	exm1497865	FPR1	G	T	0.206	synonymous SNV	0.793 (0.054)	6.781E-04
2	203765756	exm258210	WDR12	T	C	0.122	nonsynonymous SNV	0.744 (0.065)	6.985E-04

*From ANNOVAR
positions refer to the HG19 assembly
CHR=Chromosome
MAF=Minor Allele Frequency
HR=Hazard Ratio
se=Standard Error

Figure 3.11. QQ Plots of SKAT-O Cox regression Mortality Results for Genes Previously Reported as Associated with a Breast Cancer Phenotype



3.3.3 Single Marker Regression Analysis

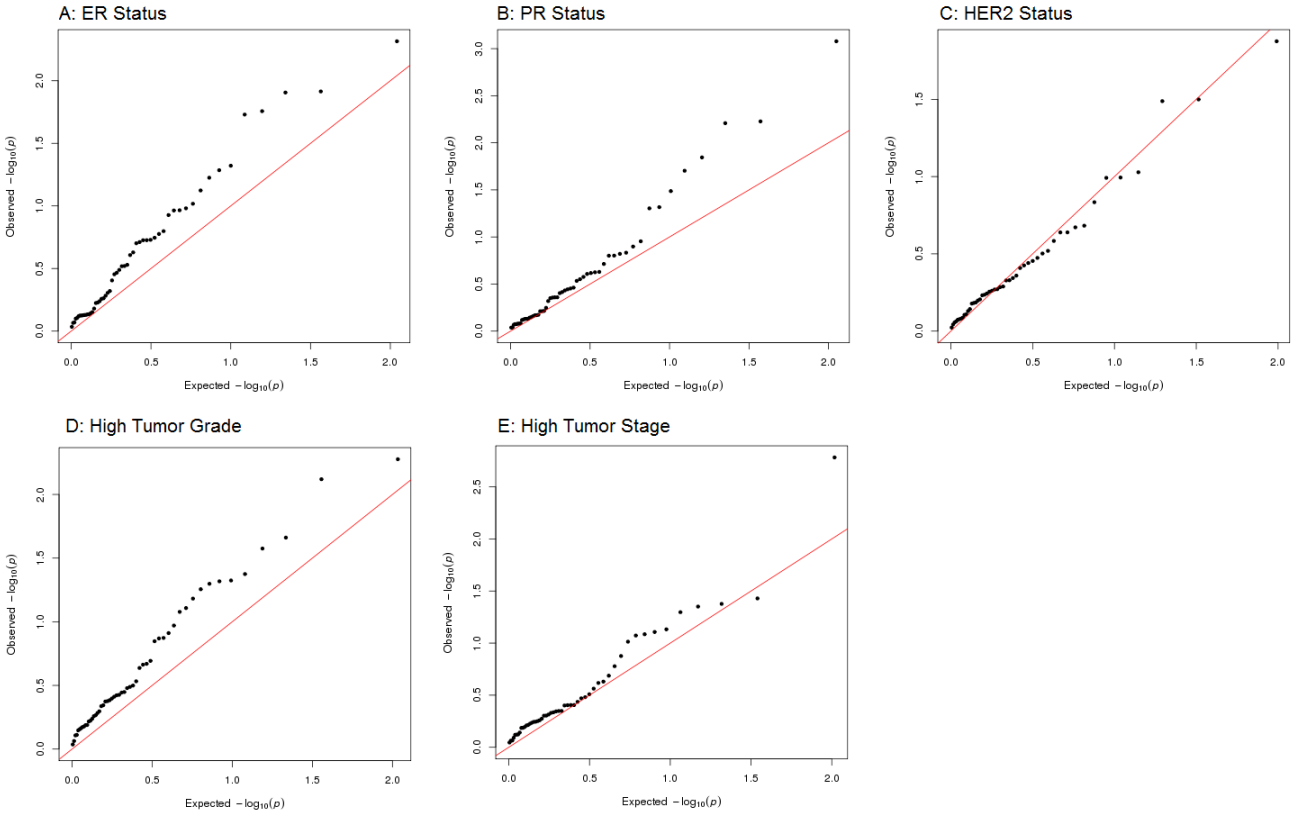
The single marker regression analysis of the association between each of the 25,938 common variants and mortality has a genomic inflation factor of 1.008 after controlling for the principal components. A summary of the association results is shown in Figure 3.10, and details of the twenty SNVs with the smallest p-values are shown in Table 3.13. None of the SNVs meet the pre-set threshold for statistical significance. The most significant SNV is located at chr6:35430686, a nonsynonymous SNV in exon of FANCE. The p-value of this SNV is $1.58 \cdot 10^{-5}$.

Results were similarly null for single marker regression analyses for mortality in ER+ tumors, ER status, PR status, HER2 status, high tumor grade, and high tumor stage.

3.3.4 Comparison with Previously Identified Breast Cancer Phenotype Loci

Previous GWASs identified 65 gene regions as being associated with breast cancer phenotypes with a p-value that was smaller than the genome-wide significance threshold of

Figure 3.12. QQ Plots of SKAT-O Logistic Regression Results for Tumor Characteristics for Genes Previously Reported as Associated with a Breast Cancer Phenotype



$5 \cdot 10^{-8}$. Fifty seven of these regions had variation in the primary data set, and the associations of these genes with mortality and tumor subtype are highlighted in Figures 3.11 and 3.12. These figures display quantile-quantile (QQ) plots of the p-values of these genes in the primary analyses for mortality (Figure 3.11 A), mortality in ER+ cases (Figure 3.11 B), ER status (Figure 3.12 A), PR status (Figure 3.12 B), HER2 status (Figure 3.12 C), high tumor grade (Figure 3.12 D), and high tumor stage (Figure 3.12 E).

While the analysis for mortality does not produce strong evidence that any of the genes that were previously reported as being associated with a breast cancer phenotype, QQ plots are qualitatively inflated for ER status, PR status, and tumor grade. However, only the analysis of PR status produces a gene whose p-value in the BCFR participants meets a Bonferroni corrected threshold of significance with $p < \frac{0.05}{57}$; $p = 8.8 \cdot 10^{-4}$: SLC4A7. This gene was reported in one previous study,[195] and one meta-analysis which included that original study,[119] both of which identified a variant in the 3 prime UTR of SLC4A7 as being associated with the risk of breast cancer in European women, with p-values of $2 \cdot 10^{-8}$ for the single study and $2 \cdot 10^{-30}$ for the meta analysis. This same variant was also identified by other studies, but not at a level that reached genome-wide-significance for that study. Variants in the adjacent gene of NEK10 were previously associated with risk, both in Chapter 2 and previous studies. The p-value for the adjacent gene, NEK10, is $6.19 \cdot 10^{-3}$ in the PR+ analysis of the primary study population.

3.3.5 Comparison with Previously Identified Mortality Loci

The above analysis was repeated using only the loci reported in NHGRI-EBI as being associated with mortality phenotypes. None of the loci that were previously identified as being associated with mortality with a genome-wide significant p-value were nominally ($p<0.05$) significant in any of the mortality analyses. Since so few statistically significant

results had been previously reported, this analysis was repeated using all loci listed in the NHGRI-EBI catalog as being associated at an significance level with breast cancer mortality (28 genes at 11 loci). This analysis demonstrated substantively similar null results for both mortality and mortality in ER+ breast cancer patients (no gene met a nominal significance threshold of 0.05).

3.4 Discussion

These analyses do not identify any gene regions in which variation is associated with mortality in breast cancer cases. These results represent the largest single study to investigate the relationship between genome-wide variation and breast cancer mortality, in terms of both the number of participants and the follow up time. The failure of single marker regression analyses to identify any loci in gene region as being associated with mortality provides complementary evidence that variation in gene regions is not responsible for a substantial portion of the variability in breast cancer mortality.

This analysis also does not provide evidence for an association between variation in gene regions and five traits that are indicators of tumor aggressiveness: ER status, PR status, HER2 status, high tumor grade, or high tumor stage with a significance that meets the genome-wide threshold.

While null results from gene-based tests are to be expected if rare variants make a large contribution to the trait,[40] these results, when taken together with previous genome-wide studies, suggest that variants that are associated with breast cancer mortality are either not common, not measured by the exome array, or, if they were measured by the exome array, that within any single gene, they collectively are responsible for a small fraction of mortality.

In agreement with other recent results,[196,197] this analysis does not find an association between variants in the CYP genes and survival in ER+ breast cancer patients, although the patients had an unknown tamoxifen treatment history. A previous genome-wide investigation suggests that among older women with known tamoxifen treatment, haplotypes that included CYP2D6 polymorphisms were associated with mortality.[145] CYP2D6 is highly polymorphic, but many of the polymorphisms that have been identified in the gene are of uncertain clinical importance. The exome array only measured four nonsynonymous variants on the exome array. While these variants likely would have tagged the effect of most truly causal variants within the gene, this may have limited the ability to detect an association between CYP2D6 and mortality

These analyses do identify a gene that was previously reported to be associated with breast cancer risk, SLC4A7, as suggestively associated with tumor PR status. While this association would need to be replicated, this suggests that women with variation in SLC4A7 may be particularly at risk for PR positive tumors. This information suggests that PR status, which is an independent indicator of treatment efficacy and ultimately survival,[198] may be driven by variation in SLC4A7. It is possible that this information can help to guide prophylactic treatment in women who are predisposed to the more lethal PR+ tumors.

These analyses also indicate that weighting by predicted functionality with the CADD scaled score can produce SKAT-O estimates that do not have inflated type I error rates, as the genomic inflation λ estimates were all near one. This suggests that the CADD scores are a valid way to incorporate *a priori* knowledge of genetic function into an analysis. Weighting by the CADD score also allowed for the inclusion of all measured variants in the gene rather than only rare variants, or only variants that were predicted to be functional.

This work suggests several next steps. At the present time, there is only a rudimentary ability to annotate non-exonic variants to genes, but this is a subject of much study. As the understanding of biological pathways improves, variants will be able to be connected

with a particular gene in ways that are more sophisticated than ANNOVAR's annotation. Regulatory variants that are not spatially near the genes that they regulate could be included in the analysis. At the same time next generation sequencing technologies are being widely implemented, which will increase the ability to detect rare variants. If any of these newly discovered or annotated variants are responsible for breast cancer mortality, their inclusion will improve the ability of gene-based tests to identify genes that direct the underlying cellular processes that confer this risk.

The participants of these studies are all of a homogeneous age (younger than 51 at diagnosis), ancestral background (European), and gender (women). As breast cancer affects people of all ages, ancestral backgrounds, and genders, additional SKAT-O analyses in populations with different characteristics will help to determine whether in these other populations, variation is associated with breast cancer mortality. Additional studies that include women of different ancestry backgrounds may help to resolve a long-running question about the determinants of differential mortality across ethnic groups.

Additional insight could also be gained by applying the analysis presented here to a population with known treatment regimens, as the effect of germline genetic variation may be heterogeneous across courses of treatment. Therefore, it would be fruitful to repeat the analyses with a sample of known, homogeneous treatment. In particular, given the still-unsettled relationship between variation in the CYP gene and survival, it would be of clinical interest to repeat the mortality analysis in patients with estrogen receptor positive tumors who were treated with tamoxifen.

There were some limitations to this analysis. The TCGA and BCFR samples differ in aspects that create challenges in using the TCGA data to replicate BCFR findings. The participants of the two studies were not well-matched on age (less than 750 TCGA participants and a small number of deaths in all TCGA analyses). Also, the TCGA participants were genotyped to measure variants that were largely common, and the presence of rare variation

was inferred through imputation, while the BCFR participants were genotyped with an exome array that targeted nonsynonymous variants in gene regions. While the CADD scores down-weighted intronic variants that were over sampled in the TCGA data and less likely to be causal, it would have been preferable to repeat the original analysis on a data set that interrogated more similar variants. It is possible that a larger secondary independent data set of age-matched patients that directly assayed the same rare variants that were captured by the primary data set may have been able to better replicate the BCFR analysis, and may have highlighted causal genes. As more data become publicly available and the methods to create gene-based tests from summary statistics improve, there will be a larger power to detect drivers of mortality.

Additionally, while large in comparison to previous single studies of mortality, the sample size here is modest by genome-wide standards. A larger sample size would have likely observed more rare variants, and could have more concretely demonstrated the association (or lack thereof) between mortality and variation in gene regions.

The analysis also would have been stronger if the data had included information on other treatment. Treatment is known to affect mortality, and may possibly interact with genetics, such that particular germline genetic variants may only affect mortality in the presence or absence of a particular kind of treatment. Similarly, tumor subtype is known to be a prognostic factor in breast cancer. While these analyses investigated mortality in ER+ cases specifically, the number of participants were not powered detect modest associations between germline genetic variation and morality within particular tumor subtypes.

In conclusion, this analysis suggests that variation assayed by the exome array does not explain a large portion of variation in mortality in early onset breast cancer cases. When combined with the evidence from family studies and heritability estimates that suggest that mortality does have a heritable component, this suggests that future work to identify variants associated with morality need to incorporate variation that is not assayed on the exome

array, and consider methods that allow for the detection of larger-than-gene pathways that have modest effect size on risk. This analysis does indicate that germline variation within SLC4A7 may predispose women with variation in that gene to a higher risk of PR+ tumors. This could help to design future preventative interventions that could be tailored specifically for the risk of tumors that over-express the progesterone receptor.

CHAPTER 4

ROLE OF GERMLINE GENETIC VARIATION IN PREDICTING RISK AND MORTALITY OF BREAST CANCER

4.1 Background

4.1.1 Non-genetic predictors of breast cancer risk and mortality

Breast cancer is the most frequently diagnosed cancer in women, with one in eight American women developing breast cancer over her lifetime.[1] Almost twenty five percent of women diagnosed with breast cancer eventually die of the disease,[2] and fear of recurrence and mortality lowers quality of life for women who are diagnosed.[3–6] Women who are diagnosed with breast cancer before the age of fifty (one in five of those diagnosed[2]) are more likely to die from breast cancer.[28] There is also evidence that tumors of women who develop breast cancer early are more likely to be driven by germline genetic variation than cancers that develop later in life.[50]

A strong prediction model of breast cancer risk and mortality would confer many clinical benefits. The ability to predict which women will develop breast cancer would identify low-risk women who could be screened less often, which would require less energy to be devoted to searching for symptoms of a disease they are unlikely to develop. Prediction models for breast cancer risk could also help to interpret an otherwise inconclusive screening result, and could reduce both unnecessary invasive procedures and unidentified tumors. In the context of breast cancer mortality, the ability to predict which women are at risk of death from the disease could identify high risk women who could benefit from more aggressive monitoring and treatment for high-risk subgroups. A strong prediction model would conversely also identify women who could pursue less aggressive treatments, which would reduce the morbidity associated with exposure to chemotherapies.[54]

Both breast cancer diagnosis and mortality have several known risk factors that are reproducibly associated with each trait. For both outcomes, existing prediction models are statistically significant and adequate for predicting the risk of a population. However their low discrimination makes them less relevant for individual risk decisions.[41] Identified risk factors for breast cancer risk include age, race, socioeconomic status, age at menarche, breast tissue characteristics, breastfeeding, reproductive history, hormone use, menopause history, alcohol use, body mass index, smoking history, and physical activity.[29,33,34,46,199–202] The effect of age on risk is not straightforward, as it interacts with other risk factors. For example, nulliparity and obesity are associated with decreased breast cancer risk in younger women, but increased risk later in life.[133,134] Prediction models for risk that incorporate these risk factors have modest predictive power. Current prediction models produce areas under the receiver operating characteristic curve (AUCs) between 0.6 and 0.7[46] (with AUCs that are significantly different than 0.5 interpreted as models that are better than chance).

For women who have already been diagnosed with breast cancer, their survival is associated with several factors that are ascertained as of the time of diagnosis, and others that develop over the course of the disease: age, race, socioeconomic status, treatment, tumor size, nodal status, grade, presence of metastases, estrogen receptor (ER) status, progesterone receptor (PR) status, HER2-positivity, gene profile, comorbidities, and the genetic aberrations of the tumor.[81,130,131,133,134,203] As with breast cancer risk, the risk factors for breast cancer mortality can also interact with age.[133] The most effective models predict breast cancer mortality with an AUC of approximately 0.7.[47]

4.1.2 Genetic Prediction

In addition to the non-genetic factors mentioned above, germline genetic variation is convincingly associated with breast cancer risk. Several high risk variants have been identified that are highly penetrant but rare in the overall population, including mutations found in BRCA1, BRCA2, and TP53.[33] Additional risk variants have been implicated by single marker regression analyses in genome-wide association studies (GWASs). These studies have identified 128 risk loci that are common, and affect breast cancer phenotypes with modest or moderate association strengths.[60] However, despite these successes, there still remains "missing heritability" in breast cancer, where variants that have been identified only contribute about half of the total risk due to genetics that is expected from family studies.[35,46,71,72,204]

The relationship between mortality and germline genetics is less clearly described than the relationship between risk and germline genetics, but several lines of evidence suggest that germline genetic variation contributes to mortality, including family studies,[35] animal studies,[150,173] highly penetrant uncommon variants such as those found in BRCA1 and BRCA2[205,206] (although their effect on mortality is counteracted by the susceptibility of these tumors to DNA-damaging chemotherapies[207]), candidate gene studies,[208] and single marker regression investigations (summarized in Table 4.1 and Table 4.2). Many of the genome-wide investigations have been undertaken with small sample sizes, and the variants that were highlighted have not been replicated widely, which has resulted in a lack of consensus on the validity of the individual variants identified by the single marker regression investigations.

Given the evidence that both risk and prognosis are influenced by germline genetics, prediction models that incorporate genetic variation would likely improve the ability to predict these two traits. However, genetic data posses several distinct characteristics that

Table 4.1: Genome-wide Studies of the Association between Germline Genetic Variation and Breast Cancer Mortality

Study Title	Year	Population Description	Outcome	N	Median Follow Up Time	Events	Variants	Replication Description	Findings
A Genome-Wide Association Study of Prognosis in Breast Cancer[91]	2010	Postmenopausal women with invasive breast cancer	Breast cancer specific survival	1145	6 years	93	528,252	Top 10 genotyped in 4335 women with invasive breast cancer with 38,148 years at risk	Nothing genome-wide significant
A Genome-wide Association Study Identifies Locus at 10q22 Associated with Clinical Outcomes of Adjuvant Tamoxifen Therapy for Breast Cancer Patients in Japanese[145]	2011	Japanese patients with hormone receptor-positive, invasive breast cancer receiving adjuvant tamoxifen therapy	Recurrence-free survival	240	7 years	30	470,796	Two independent sets of 105 and 117 cases	15 SNVs in the primary analysis; rs10509373 (chr10:76397814) replicated (combined $p = 1.26 \cdot 10^{-10}$)
Novel Genetic Markers of Breast Cancer Survival Identified by a Genome-Wide Association Study[170]	2012	Shanghai-resident Chinese women	Total mortality	1950	6 years	299	613,031	Top 49 associations replicated in 4160 Shanghai women with breast cancer; Top association examined in Nurses Health Study	rs3784099 (chr14:68283210; $p = 1.44 \cdot 10^{-8}$ in discovery only)
Identification of Inherited Genetic Variations Influencing Prognosis in Early-onset Breast Cancer[171]	2013	UK women aged 40 or younger at diagnosis	Breast cancer specific survival	536	4 years	236	487,496	Top 35 associations genotyped in 1,516 independent cases from the same early-onset cohort	Nothing genome-wide significant
Genome Wide Meta-Analysis Study for Identification of Common Variation Associated with Breast Cancer Prognosis[172]	2014	UK women aged 40 or younger at diagnosis, and Finish women of all ages	Breast cancer specific survival	1341	6 years	237	475,141, imputed to 7.5 million	1523 additional participants of the POSH study	Nothing genome-wide significant

positions refer to the HG19 assembly
Studies that published both single-study results and contributed to a meta-analysis will be represented twice

Table 4.2: Genome-wide Studies of the Association between Germline Genetic Variation and Breast Cancer Mortality (continued)

Study Title	Year	Population Description	Outcome	N	Median Follow Up Time	Events	Variants	Replication Description	Findings
Identification of Novel Genetic Markers of Breast Cancer Survival[173]	2015	Meta-analysis of studies in populations of European ancestry	Breast cancer specific survival	37,954	5 years	2900	200,000-700,000; imputed to 9 million	N/A	rs148760487 (chr2:162922103; $p = 1.5 \cdot 10^{-8}$) and 27 others in high LD; rs2059614 (chr11:125389528; $p = 1.3 \cdot 10^{-9}$ in ER-cases
Polymorphism at 19q13.41 Predicts Breast Cancer Survival Specifically after Endocrine Therapy[174]	2015	Meta analysis of UK women aged 40 or younger at diagnosis, and Finish women of all ages	Breast cancer specific survival	1341	7 years	547	486,478	Two independent data sets with 5011 patients	Nothing genome-wide significant
Prediction of Breast Cancer Survival Using Clinical and Genetic Markers by Tumor Subtypes[175]	2015	Incident breast cancer cases in Seoul, South Korea	Recurrence-free survival	1732	4 years	214	2,210,580 genotyped and imputed	Any SNVs identified with $p < 10^{-6}$ and MAF > .1, and any common variants in high ($r^2 > 0.4$) LD with them were genotyped in 1494 additional women from South Korea	Nothing genome-wide significant

positions refer to the HG19 assembly
Studies that published both single-study results and contributed to a meta-analysis will be represented twice

must be accounted for in prediction models: the number of variants to include can exceed the number of study participants; the predictors are often correlated due to linkage disequilibrium (LD) between the variants; the form that describes the relationship between variants and disease is unknown; the sparsity and distribution of the causal variants throughout the genome is unknown; and many of the putatively associated variants likely have small effect on the trait.[52] Two methods that can be appropriate for prediction given these challenges are polygenic risk scores (PRS) and restricted maximum likelihood estimates (REML) from linear mixed models (LMM) .

Polygenic risk scores multiply the per-allele risk (found in prior literature or a training set of individuals) for each test individual by the number of risk alleles at a locus, and sum this over each variant of interest to produce a score that reflects a test individual's risk of disease. While PRSs can be implemented in many ways,[209,210] in most studies, PRSs include a limited number of variants in the score, typically those that pass a significance threshold in association analyses in the training set. Before the advent of LMMs for prediction, some investigators were able to successfully create PRSs with a large number (<10,000) of genotyped variants.[211] However, in general, PRS predictions that use a large number of variants are often unstable,[97] and most PRSs now contain fewer than 100 variants.

Polygenic risk scores are not a preferred method for whole genome prediction. It has been shown that the prediction risk based on a per-allele odds ratio of a training set is likely to be substantially inaccurate for rare alleles.[129] There is evidence that predictions from PRS can be biased upwards,[212] and that LD structure can lead to inconsistent results. Even when unbiased, polygenic risk scores do not have much predictive power in a complex, non-Mendelian trait, for intuitive reasons:[52] first, the effect sizes of rare variants are poorly estimated, and therefore not able to be reflected in a PRS; and second, the threshold that is used to include variants is arbitrary by its nature. Many causal single nucleotide variants

(SNVs) may not meet the significance threshold, and lowering the significance threshold introduces many variants into the risk score which are not truly causal.[213] This can both obscure the effect of a truly causal variant, and also increase the possibility of a spurious associations driving a prediction.

PRSs are best suited to predict traits that have few causal variants of larger effects. While not conclusive, previous studies along with the analyses of Chapters 2 and 3 suggest that this is not the case in breast cancer risk or prognosis, but if it is, PRSs would be well-suited incorporate genetic data into a prediction model. PRSs have been implemented for breast cancer risk, but they have largely been poorly replicated[214–216] (perhaps due to the instability of the estimate), or produced a prediction that was statistically significant but not clinically meaningful.[217] These modest predictive powers are consistent with simulations that have shown that large sample sizes (10,000 or more participants) are often necessary to achieve enough power for PRSs to produce a statistically significant prediction for most genetic traits.[217]

Genetic similarity has also been translated into prediction by REML LMM models (GREML) that summarize genetic similarity in genetic relatedness matrices (GRMs). While possible upward biases of heritability estimation using REML models has been debated,[218–220] the current consensus is that they are largely accurate for prediction,[221,222] and under certain plausible assumptions these predictions are the best linear unbiased prediction (BLUP).[223] These methods begin with a training set of individuals with known disease status, and calculate the genetic similarity between each individual in that training set and an individual in the test set. The risk of this test individual is then computed as the weighted average of the case statuses of the training set, with the weights calculated as a transformation of the pairwise genetic similarity between the individuals.

GREML approaches are more appropriate than PRSs when the trait is highly polygenic.[42,224,225] If a GREML-BLUP is implemented with a trait that is driven by a small

number of causal variants, the prediction will have large variance, but be unbiased.[225] The Kriging method developed by Wheeler et al.[42] is equivalent to the BLUPs of GREML, but is motivated differently. The Kriging method extends GREML predictions to integrate more than one matrix of -omic similarity. This extension allows for the grouping of variants based on prior information that indicates that variants within a single group affect the trait under study in a similar manner. Separate GRMs are then constructed from the variants in each of the groups. If variants are grouped together in a manner that reflects true similarities of their underlying association with disease, the performance of the prediction can substantially increase.[42] The Kriging method is less model dependent than other GREMLs, in that the weights of the different GRMs are found by maximizing prediction performance (as measured by AUC for dichotomous outcomes[226]) rather than direct estimation.

In the Kriging framework, covariates are added linearly to each pairwise similarity vector before transformation into a similarity matrix.[223] Under certain assumptions, this is equivalent to regressing the outcome on the covariates, and then using the residuals for the phenotype of the Kriging procedure. The incorporation of non-genetic covariates allows for a final prediction model that represents both the genetic and non-genetic influences of breast cancer.

While the primary goal of prediction models is not to identify the specific variants that are associated with disease, the grouping allowed by the Kriging prediction method can be used to characterize the causal variants. Using common annotation software, it is possible to group variants based on whether it is common or rare, and by whether the variant has a particular predicted functionality. When these GRMs are used in whole genome prediction using Kriging, the magnitude of the weights on each GRM suggest whether the variation that drives breast cancer is located in variants that common or rare (or both), and whether that variation is likely to be found within variants of a particular type of predicted functionality. This knowledge will help to resolve long-running questions about the relative

importance of different portions of the genome. This analysis will help to design future studies that may aim to identify risk variants. *A priori*, variants that are rare are more likely to be associated with disease, as evolutionary constraints would likely keep them at a low frequency in the population. This insight motivates many whole-genome and whole-exome sequencing projects, which, although they are more expensive than studies that use an array-based technology, identify rare variation more effectively. However, there are many exceptions to this general rule, since for many diseases (including breast cancer), many variants that have been identified as associated with disease are prevalent in the general population.[80] In the case of breast cancer, it is unclear if the not-yet-discovered variation that is associated with disease is likely to be found in variation that is common or rare.

Similarly, the extent to which different classes of predicted variant functionality are likely to drive genetic association with breast cancer risk and prognosis is still unknown. *A priori*, variants that cause changes in the translation of amino acids are considered most likely to affect a trait, and this has justified many studies that utilize whole exome sequencing and exome arrays. However, genome-wide analyses have frequently produced new discoveries in variants that were thought to be "junk" DNA:[180,181] Variants near gene regions that do not directly cause changes in proteins are over-represented in GWAS results;[34,176,177] and intergenic variants often contribute to complex traits.[34] It is currently unclear whether, in the context of breast cancer, the missing heritability is driven by variants that will be identified by studies that focus only on variation in the exome.

An additional unanswered question is to what extent not-yet identified variants contribute to risk and prognosis of breast cancer. Due to the instability of polygenic risk scores, and the lack of a whole-genome based heritability estimate for breast cancer risk and prognosis, it is unclear whether the 100+ variants that have already been associated with breast cancer phenotypes drive the association with either diagnosis or mortality, or whether there are additional risk loci that are associated with either breast cancer trait.

In the context of Kriging, GRMs can be constructed so that each of these groups of variants can contribute to prediction with a separate strength. The optimal weights of each GRM will help to describe the relative importance of the variants that make up those GRMs. This has not yet been done in for any breast cancer phenotype.

4.1.3 Gaps in knowledge

With the preceding as background, this manuscript will predict breast cancer risk and prognosis for the first time using a Kriging framework in a way that will allow for variants with different predicted functionality and different prevalences to contribute to risk with different strengths. Given the already demonstrated polygenic nature of breast cancer risk, the possible polygenic nature of breast cancer prognosis, and the lack of success of previous polygenic risk scores, Kriging represents a promising method to predict breast cancer risk and mortality. No whole-genome prediction model has been reported for either breast cancer risk or prognosis, and the analyses of this manuscript will illuminate the genetic architecture of breast cancer and identify classes of variation that drive each trait.

When combined with non-genetic information, the prediction models may more accurately predict population-level risk and also improve upon current risk estimates. If successful, this will further a goal of precision medicine and produce individual prediction models that are clinically actionable. Given the high prevalence of breast cancer (45,000 early onset diagnoses each year in American women and 231,000 diagnoses in American women of all ages[1]), even a modest increase in the total ability to predict risk could potentially impact the interpretation of ambiguous screening results for many women,[52] and provide additional context for women who are considering other medical interventions that may increase their breast cancer risk, such as menopausal hormone therapy or hormonal assisted reproductive therapies.[53] In the context of breast cancer prognosis, a more accu-

Table 4.3: Characteristics of Studies Included in Analysis

Study Name	Study Location	Years Recruiting	Case Criteria	Control Criteria	Cases	Controls
Breast Cancer Family Registry	Australia	1992-2000	Living in the Melbourne and Sydney metro areas, family recruited from the Victoria and NSW cancer registries	Randomly selected from electoral rolls, matched to cases on age and city	561	119
Breast Cancer Family Registry	Northern California	1996-2003	SEER Cancer registry in the San Francisco metro area	Random digit dialing in study area, matched to cases on age and race/ethnicity	180	65
Breast Cancer Family Registry	Ontario	2001-2010	Ontario Cancer Registry	Random digit dialing in study area, matched to cases on age	574	154
Genetic Epidemiologic Study of Breast Cancer by Age 50	Germany	1992-1995	38 clinics in the Rhein-Neckar-Odenwald and Freiburg regions	Randomly selected from local population registries	516	483
Long Island Breast Cancer Study Project	New York	1996-1999	Nassau and Suffolk counties	Random digit dialing in study area, matched to cases on age	198	110
Seattle	Seattle, Washington	1990-1992	King, Pierce, and Snohomish counties; age less than 45 at diagnosis	Random digit dialing in study area, matched to cases on age and race	294	103

Cases and controls are numbers included in the analysis before quality control

rate prediction model would be able to better identify the estimated 60% of women who are treated with toxic chemotherapies who extract little to no survival benefit from the treatments,[55] while also identifying those at high risk of mortality who may want to be treated more aggressively.

4.2 Methods

4.2.1 Study Data

4.2.1.1 Participants

The participants for these analyses were selected from six ongoing studies designed to assess the risk factors associated with early onset breast cancer. Participants are women of European descent who were 51 years or younger at the time of their diagnosis (for cases)

or enrollment (for controls) and not known to carry pathogenic germline mutations in the genes BRCA1 or BRCA2. Details of the recruitment are found in Table 4.3. Three of the study sites (Australia, Northern California, and Ontario) were members of the Breast Cancer Family Registry (BCFR), whose methods have been described elsewhere.[63] Northern California and Ontario recruited through population-based registries, and Australia recruited through a mix of population and clinic-based outreach. Participants were also included from three population-based case-control studies not included in the BCFR consortium: the German Genetic Epidemiologic Study of Breast Cancer;[64] the Long Island Breast Cancer Study Project;[65] and the Seattle study.[66]

4.2.1.2 Genetic Data and Quality Control

Germline DNA was extracted from blood drawn from 3357 participants (2323 cases and 1034 controls). Genetic variation was measured using two Illumina array-based genotyping methods: (1) an exome array that that was designed to more closely interrogate often-rare variants in the gene regions, with particular emphasis on nonsynonymous variants, and (2) a GWAS array that was designed to interrogate common variation over the whole genome.

Two versions of the exome array were used: 1849 cases and 831 controls were genotyped on the Illumina HumanExome 12v1.0 chip, and 474 cases and 203 controls were genotyped on the Illumina HumanExome 12v1.1 chip. To improve the quantity and quality of available genomic DNA, the samples were whole genome amplified using the Qiagen Repli-G mini kit,[22] and were processed using 49 plates in two batches, following the manufacturer's protocol. TeCan Evo was used for automation. Raw data was processed by Genome Studio on 2010.3 software, and the no-call threshold was set at 0.15, per Illumina's recommendation for Infinium chips. Clustering was done using the Illumina sup-

plied cluster files. After keeping only variants that were on both chips, 238,524 variants were interrogated.

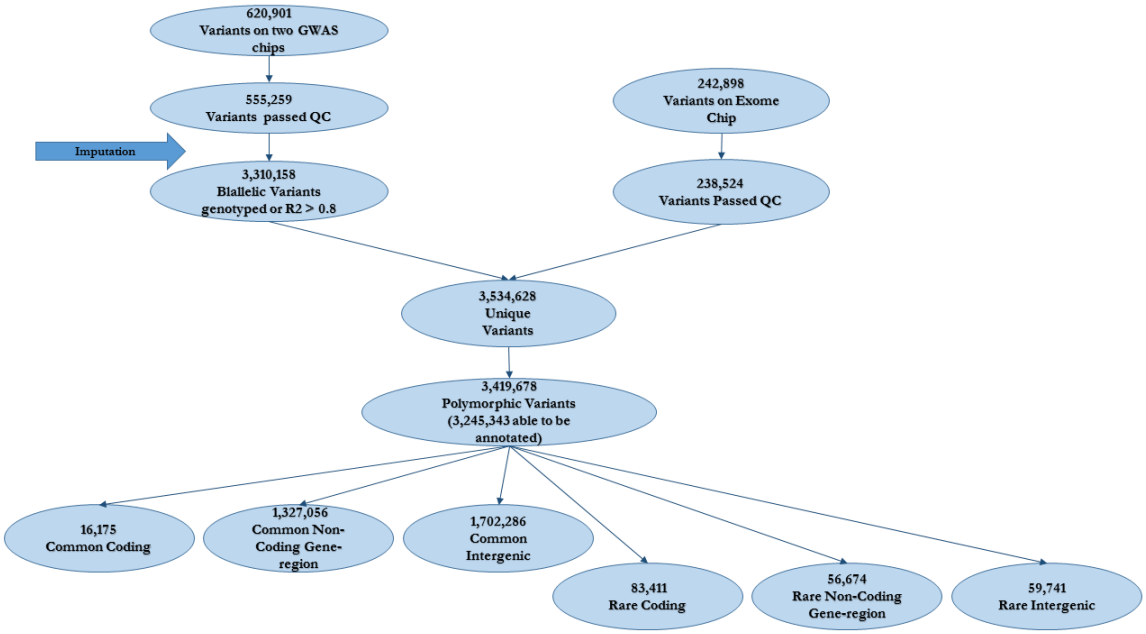
The quality control followed the protocol outline by Guo et al.[96] Since the accuracy of genotype calling from an exome array is slightly less than the accuracy of a genome-wide array of common variants,[227,228] and the variants from the exome array was of particular interest for this analysis, the quality control for the exome array was done separately from the quality control for the GWAS array, and then the two were combined. Participants were excluded for low genotyping rate (rate < 95%; 219 excluded), high heterozygosity (F statistic greater than three standard deviations from the mean, or heterozygosity greater than four standard deviations from the mean; 31 excluded), and one of each pair of duplicated genotypes (eight samples excluded: three replicates; five duplicates from the same center). Additionally, due to the family-based case ascertainment of some of the studies, seven participants were excluded whose genotypes were highly correlated (estimated relatedness from a GCTA-created genetic relatedness matrix greater than 0.4).[97] Variants were excluded from the analysis if they had a low call rate (rate < 95%; 4335 excluded), or if they were common variants (defined below) with Hardey-Weinberg equilibrium p-values of less than $2.5 \cdot 10^{-7}$ in controls (p = 0.05 Bonferroni corrected for 200,000 tests; 39 excluded). The final variant-level exclusions were the result of evidence that on some plates variants were unreliably assigned (a plate-by-plate single marker regression analysis found that in some cases genotype could predict plate). For these variant-plate combinations, variants were excluded for all participants on that plate if this GWAS p-value was smaller than $2.5 \cdot 10^{-7}$. As a result of this quality control step, 100 variant-plate combinations were set to missing.

Variation genome-wide was additionally measured for the same 3357 participants. The procedure to genotype and impute from variants assayed on the genome-wide array are detailed elsewhere,[50] and summarized on the left hand side of Figure 4.1. Briefly, the DNA

was genotyped using the Illumina 610-Quad and Cyto12 v2 BeadChips, and standard laboratory quality control procedures were applied. After quality control, 555,254 variants and 3333 participants were brought forward to imputation, which was implemented by the Michigan imputation server,[186] employing ShapeIt[187] to pre-phase the variants and minimac3 to impute.[188] In order to best impute rare variants,[189,190] the entire 1000 Genomes phase 3 release[132] was used for a reference panel. While it might have been optimal to combine the genotyped variants from both arrays before imputing, the non-imputed genome-wide genotype data was no longer available. However, the LD structure of rare variants differs from the LD structure of common variants,[99] and previous research suggests that few additional variants would have been imputed with high quality had the exome array variants also been included in the imputation.[227] Although variants with an imputation r^2 greater than 0.3 are generally considered adequate for association studies,[191,192] rare variants (which are of particular interest to this study) and have an estimated r^2 with genotyped variants that is more variable than common variants.[229] For these reasons, only imputed variants with an imputation r^2greater than 0.8 were kept, consistent with other classifications of "high quality" imputation[227,230,231] (3,310,158 variants).

After the quality control steps, 2869 participants were measured with both arrays. The post-quality control, post-imputed genotypes from the two arrays were combined. A small fraction of the variants were measured by both methods (14,054 variants), and of these, 124 variants were called differently for at least one participant. Of these 124 differences, 17 were variants that were genotyped on the genome-wide arrays, and the rest were imputed. Since more than 85% of the discordant calls were a result of imputation, in the cases where the two methods disagreed, the allele called from the genotyped exome array was used. After the quality control steps, 3,534,628 polymorphic variants were available for 2869 participants. Of these, 3,245,343 could have their expected functionality annotated by the

Figure 4.1. Variants Used in Primary Analysis



ANNOVAR software[98] and were retained for analysis. A schematic of the variants used in this analysis is shown in Figure 4.1.

4.2.2 Classification of Variants and Creation of the Genetic Relatedness Matrices

In previous work, a designation of "common" and "rare" variants roughly corresponded with their ability to be included in a GWAS-framework single marker regression study. Following this, a threshold of frequency equal to $\left(\frac{1}{2n}\right)^{\frac{1}{2}} = 0.0127$ was used to distinguish common variants from rare variants,[59] which resulted in 3,045,517 common variants and 199,826 rare variants. Three functional categories were also created: variants that are expected to cause a change in amino acid translation, variants that are located in gene regions but not associated with amino acid translation, and intergenic variants. To identify variants

Table 4.4: Genetic Relatedness Matrices Used in Prediction

Model	Genetic Relatedness Matrix	Variants in Matrix	Genotyped Variants	Imputed Variants
Model 1	All Variants	3,245,343	436,208	2,809,135
Model 2	Common	3,045,517	344,649	2,700,868
	Rare	199,826	91,559	108,267
Model 3	Protein Damaging	99,586	98,190	1,396
	Other Gene Region	1,383,730	165,731	1,217,999
	Intergenic	1,762,027	172,287	1,589,740
Model 4	Common Protein Damaging	16,175	14,912	1,263
	Common Other Gene Region	1,327,056	160,115	1,166,941
	Common Intergenic	1,702,286	169,622	1,532,664
	Rare Protein Damaging	83,411	83,278	133
	Rare Other Gene Region	56,674	5,616	51,058
	Rare Intergenic	59,741	2,665	57,076
Model 5	Previously Identified	2,787	1,554	1,233
	All Not Identified	3,242,556	434,654	2,807,902
Model 6	Previously Identified	2,787	1,554	1,233
	Not Identified Common Protein Damaging	16,171	14,909	1,262
	Not Identified Common Other Gene Region	1,324,780	158,630	1,166,150
	Not Identified Common Intergenic	1,701,779	169,556	1,532,223
	Not Identified Rare Protein Damaging	83,411	83,278	133
	Not Identified Rare Other Gene Region	56,674	5,616	51,058
	Not Identified Rare Intergenic	59,741	2,665	57,076

Only polymorphic variants are included in these counts

that were likely to cause a change in the translated amino acid, a functional category of damaging variants was created that included variants annotated by ANNOVAR as nonsynonymous, stop-loss, stop-gain, frameshift substitution, or nonframeshift substitution in an exon (99,586 variants). he final variant-level exclusions were the result of evidence that on some plates variants were unreliably assigned (a plate-by-plate single marker regression analysis found that in some cases genotype could predict plate). Variants near gene regions that did not directly cause changes in amino acid translation included all other variants that ANNOVAR annotated to genes, including introns, synonymous SNVs, UTRs, and variants within 1 kilobase of the start and stop sites (1,383,730 variants). Intergenic variants contained all other variants (1,762,027 intergenic variants).

In addition to minor allele frequency and expected functionality, a third grouping of variants was of interest: those that have been previously identified by other researchers as being associated with a breast cancer phenotype in a single marker regression framework. As of September 2016, 174 associations (128 unique SNVs) are listed in the NHGRI-EBI GWAS catalog[34] that connect germline genetic variation with a breast cancer phenotype with a p-value less than the genome-wide significant threshold of $5 \cdot 10^{-8}$. In order to also include in this group of variants other SNVs that may tag this previously known association well, the Broad Institute's SNAP program[232] was then used to identify a total of 2791 SNVs that were within 500kb of the original SNV, and in high LD with it ($R^2 > 0.8$ in the CEU 1000 Genomes[132] population). Out of this combination of previously identified SNVs and those in high LD with them, 2,787 were interrogated in the genetic data of the study population.

The above-described classifications were then used to create separate genetic relatedness matrices to test six different prediction models, summarized in Table 4.4. The Kriging method developed by Wheeler et al[42] follows Yang et al[61] and defines each element of the

GRM as the non-standardized and non-centered correlation between the genotypes of each individual:

$$\frac{1}{M}\sum_{l=1}^{M}\frac{\left(X_{il}^{G}-2p_l\right)\left(X_{jl}^{G}-2p_l\right)}{2p_l\left(1-p_l\right)}$$

with i and j denoting individuals, X_{il}^{G} the number of reference alleles of at marker l, p_l the frequency of the reference allele at marker l, and M being the number of genomic markers used in that GRM.

4.2.3 Prediction Models

Kriging was then used to predict two breast cancer phenotypes: case/control status (1998 cases and 871 controls), and ten year survival status (1903 cases with mortality information and 400 deaths from any cause before 10 years after diagnosis). Ten year survival was chosen because the low number of deaths by year five (229) would have resulted in an underpowered analysis.

Kriging prediction was implemented using the R package "omicKriging." For each iteration of the model, the Kriging formula was implemented using ten-fold cross validation to estimate a predicted value for each participant, with predicted values near zero indicating a low risk of breast cancer, and values near one indicating a high risk of breast cancer. These predictions were compared to that participant's actual breast cancer status to compute an AUC.

In order to produce a more stable estimate with valid confidence intervals, this procedure was repeated two hundred times (which simulations suggest is more than sufficient to produce a stable estimate given a sample > 1000[233,234]). The reported AUC is the mean of the two hundred calculated AUCs, and the 95% confidence intervals reported are the

2.5 and 97.5th percentiles of the replications. The models that contained multiple GRMs (Models 2-6), determine the optimal weighting by applying a grid search to each separate GRM in the model (the sum of the weights of the separate GRMS is constrained to equal one). The weights that produce the highest AUC are reported.

The genetic-only model that produced the highest AUC for risk was then used to predict overall cancer risk by incorporating non-genetic known risk factors. Non-genetic risk factors were available for 1903 cases and 855 controls, and included: age (although the cases and controls in this study were age-matched); socioeconomic status as captured by education (high school or less; some college; college degree or more) and marital status (married; single; previously married; other); smoking history (never, past, current); hormonal contraceptive use (ever, never); gravid (yes, no); number of pregnancies; age at menarche; and menopause (yes, no). Separate analyses (see Chapter 2) indicated that two principal components were also predictive of breast cancer risk, and these were also included.

In the mortality analysis, the model that produced the highest AUC for mortality would be used to predict overall cancer mortality by incorporating known clinical prognostic factors that were available for these participants: ER status, PR status, grade, and stage.

Uneven LD structure near causal SNVs can cause bias in heritability estimates from disproportionate tagging of the same SNV.[190] One method to avoid this bias is to construct GRMs out of pairwise independent SNVs (e.g.: 500 kilobase sliding window, moved forward 5kb at a time, remove variants with r2>0.8), or create multiple GRMs, stratified by local LD.[190] However, in the context of this analysis, either discarding variants or stratifying based on LD would obscure some of the relationships that were of interest. In order to see if these results were sensitive to this possible bias, the prediction model that used all annotated SNVs (Model 1) was repeated using GRMs that were stratified by local LD structure, as described in Yang,[190] and the results did not substantively change. This is

consistent with other results that show that the bias induced by uneven LD (while varying trait to trait) is typically low compared to the variance of the heritability estimation.[221,235]

4.2.4 Comparison with other methods

Heritability estimates and polygenic risk scores are two additional methods to describe the predictive power of genetics that are complementary to Kriging. To put the Kriging results in context with the results that are produced by these other methods, two additional analyses were completed for both phenotypes. The first complementary method, heritability, was computed using GCTA, using the GRM used in the Models 1 to estimate the heritability of risk (using a background prevalence of 8%) and ten year mortality (using our study-specific prevalence of 20%).

The second complementary method, polygenic risk scores, can be implemented either by cross-validating using a single data set,[217,236] or by using reported odds ratios of already-identified variants.[237] Due to the modest sample size of the participants in this study, the second method was chosen in order to reduce the variability of the estimated prediction. To create the polygenic risk score, the 128 unique SNVs that were listed in the NHGRI-EBI catalog as being associated with breast cancer phenotypes were further curated to keep only those that reported an odds ratio and were polymorphic in the participants (81 variants used in polygenic risk score). If variants were reported by multiple studies, the average of the reported odds ratios was used. In the analysis set, logs of the previously-reported odds ratio were multiplied by the number of risk alleles a person had at each locus, and summed across all 81 loci. This was transformed back to a predicted odds ratio for each person, and then compared to the actual status of the participant using AUC.

Table 4.5: Predictive Power and Optimal Weighting for Six Genetic-Only Predication Models of Breast Cancer Risk

Model	Optimal AUC (95% CI)	Optimal Weights
Model 1	0.570 (0.560-0.578)	
Model 2	0.573 (0.564-0.583)	common = 1.000 rare = 0.000
Model 3	0.578 (0.569-0.589)	protein damaging = 0.333 other gene region = 0.333 intergenic = 0.334
Model 4	0.580 (0.570-0.590)	rare protein damaging = 0.250 rare other gene region = 0.000 rare intergenic = 0.000 common protein damaging = 0.150 common other gene region = 0.300 common intergenic = 0.300
Model 5	0.609 (0.600-0.618)	previously discovered = 0.300 not yet discovered = 0.700
Model 6	0.618 (0.610-0.629)	previously discovered = 0.300 not yet discovered rare protein damaging = 0.175 not yet discovered rare other gene region = 0.000 not yet discovered rare intergenic = 0.000 not yet discovered common protein damaging = 0.105 not yet discovered common other gene region = 0.210 not yet discovered common intergenic = 0.210

4.3 Results

4.3.1 Risk

Table 4.5 summarizes the predictive power of each of the six genetic models and the weights that were used to achieve the optimal prediction. Model 1 (AUC: 0.570, 95% CI: 0.560-0.578), which considers all variants together and assumes that each variant follows the same normal risk distribution, is not as powerful of a predictive model as the models that allow different classes of variants to have different associations with risk. Prediction is improved by separating the genetic variants into both frequency and functional classes (Model 4 AUC: 0.580, 95% CI 0.570-0.590). Additional improvement is achieved

by allowing a separate GRM that contains the SNVs that have previously been associated with disease, and those in high LD with them (Model 5 AUC: 0.609, 95% CI 0.600-0.618). The optimal model is Model 6 (AUC: 0.618, 95% CI 0.610-0.629), which combines the rationale of Models 4 and 5.

A grid search of the weights for Model 2 finds that any weight given to the rare GRM produces a significantly lower AUC, and a grid search for Model 3 finds that giving approximately one third weights to each functional class is optimal. Differing from this $\frac{1}{3}/\frac{1}{3}/\frac{1}{3}$ split by more than 5% produces significantly lower AUCs.

In contrast, the analyses of Model 4, Model 5, and Model 6 produced optimal weights were not unique (other weights could have been used to produce substantively similar prediction metrics). In the analyses, Model 5 is optimized by a 30% weight on the GRM constructed of previously identified risk loci, but AUC for weights ranging from 5% to 80% is also possibly optimal. Similarly, analyses Model 4 and Model 6, the grid search revealed that as long as the weights on the GRMs made from rare intergenic and rare non-coding gene region variants were kept at zero, many other combinations of weights on the remaining four GRMs also produce an AUC with 95% confidence intervals that included the optimal AUC.

With those caveats, the optimal weights do suggest the relative importance of each of the variants that make up the GRM in predicting breast cancer risk. From Model 2, rare variants collectively have very little power to predict breast cancer risk, and when combined with Model 4, this can be refined to suggest that rare variants that do not cause changes in amino acid have very little predictive power. This suggests that rare amino acid-damaging variants have a different relationship with risk than other rare variants. Model 3 and Model 4 give strong evidence that the variants that are responsible for breast cancer risk are not located exclusively near gene regions. Model 5 and Model 6 suggest that there

Table 4.6: Characteristics of Participants in Risk Analysis

		All	Cases	Controls
N with Genetic Data		2869	1998	871
n with Genetic and Non-Genetic Data		2758	1903	855
Age	mean (sd)	41.3 (5.71)	41.4 (5.66)	41.2 (5.83)
Education	High School or Less (%)	710 (25.7)	542 (28.5)	168 (19.6)
	Some College (%)	1231 (44.6)	784 (41.2)	447 (52.3)
	Bachelors or More (%)	817 (29.6)	577 (30.3)	240 (28.1)
Marital Status	Married (%)	2146 (77.8)	1499 (78.8)	647 (75.7)
	Single (%)	241 (8.74)	155 (8.15)	86 (10.1)
	Previously Married (%)	347 (12.6)	239 (12.6)	108 (12.6)
	Other (%)	24 (0.87)	10 (0.525)	14 (1.64)
Smoking History	Never (%)	1296 (47)	902 (47.4)	394 (46.1)
	Past (%)	772 (28)	530 (27.9)	242 (28.3)
	Current (%)	690 (25)	471 (24.8)	219 (25.6)
Ever HC	n (%)	2379 (86.3)	1665 (87.5)	714 (83.5)
Ever Pregnant	n (%)	2299 (83.4)	1590 (83.6)	709 (82.9)
Number of Pregnancies	mean (sd)	2.14 (1.45)	2.16 (1.46)	2.09 (1.44)
Age at Menarche	mean (sd)	12.8 (1.5)	12.7 (1.47)	12.9 (1.56)
Post-Menopause	n (%)	607 (22)	494 (26)	113 (13.2)

HC: Hormonal Contraceptives
sd: standard deviation

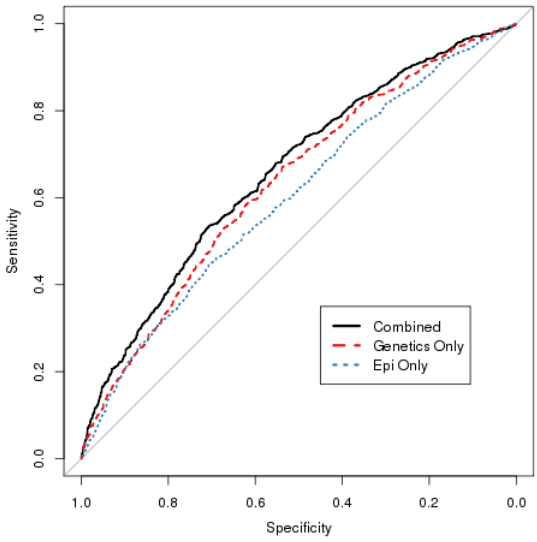
Table 4.7: Predictive Power of Models of Breast Cancer Risk

Model	Optimal AUC (95% CI)
Non-Genetic Risk Factors Alone	0.601 (0.579-0.623)
Genetics Alone	0.630 (0.622-0.637)
Combined	0.655 (0.649-0.660)

are still undiscovered variants that are responsible for breast cancer risk, and that these undiscovered variants are also found in all three functional categories.

The last analysis of breast cancer risk included covariates and combined their effect with the effect of genetic variation, using the weights found in Model 6. This analysis included the participants for whom non-genetic risk factors were also available (95% of participants; characteristics summarized in Table 4.6). The results of the genetic-only, non-genetic, and combined prediction models for these 2758 participants are summarized in Table 4.7 and Figure 4.2. Using a linear model, the non-genetic risk factors alone can

Figure 4.2. Optimal Predication Models of Breast Cancer Risk



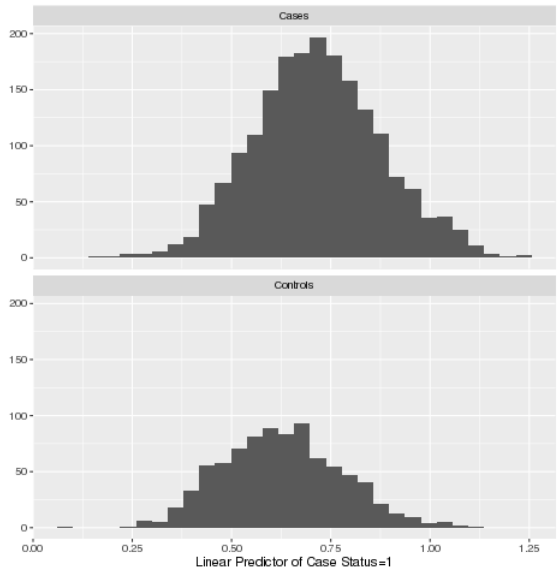
predict breast cancer risk in this population with an AUC of 0.601 (95% CI from 2000 bootstrap replicates: 0.579-0.623). The genetic only analysis in this subset of the population has superior predictive power than the non-genetic only analysis in this subset, and produces a prediction AUC of 0.630 (95% CI: 0.623-0.637). The prediction model that combined the optimal whole genome genetic information with the non-genetic risk factors was superior to both, with an AUC of 0.655 (95% CI: 0.649-0.660).

Figure 4.3 displays the distributions of the predicted risk of breast cancer in cases and controls from the combined model. This demonstrates that while the means of the two distributions are significantly different (0.627 for the controls and 0.717 for the cases; 95% CI for the difference in the means: 0.102-0.077), many women are still misclassified by this prediction model.

4.3.2 Prognosis

The risk of breast cancer mortality in cases was next predicted by the Kriging method in the 1903 cases of the primary analysis set for which mortality information was available.

Figure 4.3. Predicted Risk of Breast Cancer for Cases and Controls



Analyses indicate that germline genetic variation does not predict breast cancer mortality in this population. A preliminary analysis estimated the AUC of predictions that were computed using each of the twenty GRMs alone (without combining them with the non-parametric weights). The results of these predictions are summarized in Table 4.8. No GRM is able to predict breast cancer mortality with an AUC that was significantly different than 0.5. A grid search of reasonable weights for each of the six models (not shown) also indicates that the genetic information is unable to predict 10 year mortality from breast cancer.

As a comparison, the known non-genetic clinical prognostic risk factors of ER status, PR status, grade, and stage are available for 894 cases (Table 4.8), and predict breast cancer mortality at 10 years with an AUC of 0.691 (95% CI from 2000 bootstrap replications of 0.667-0.750).

Table 4.8: Predictive Power for Six Genetic-Only Predication Models of Breast Cancer Mortality

Model	Variants in GRM	AUC (95% CI)
Model 1	All Variants	0.493 (0.479-0.510)
Model 2	Common	0.499 (0.484-0.514)
	Rare	0.484 (0.469-0.500)
Model 3	Protein Damaging	0.487 (0.473-0.502)
	Other Gene Region	0.484 (0.471-0.499)
	Intergenic	0.504 (0.491-0.519)
Model 4	Rare Protein Damaging	0.489 (0.475-0.503)
	Rare Other Gene Region	0.485 (0.473-0.500)
	Rare Intergenic	0.497 (0.481-0.514)
	Common Protein Damaging	0.500 (0.488-0.513)
	Common Other Gene Region	0.490 (0.473-0.504)
	Common Intergenic	0.505 (0.490-0.519)
Model 5	Previously Discovered	0.505 (0.479-0.541)
	Not Yet Discovered	0.494 (0.479-0.507)
Model 6	Not Yet Discovered Rare Protein Damaging	0.489 (0.477-0.506)
	Not Yet Discovered Rare Other Gene Region	0.485 (0.469-0.503)
	Not Yet Discovered Rare Intergenic	0.497 (0.480-0.508)
	Not Yet Discovered Common Protein Damaging	0.499 (0.484-0.518)
	Not Yet Discovered Common Other Gene Region	0.489 (0.473-0.503)
	Not Yet Discovered Common Intergenic	0.507 (0.494-0.520)

Table 4.9: Characteristics of Participants in Mortality Analysis

		All	Alive	Died
N with Genetic Data		1903	1503	400
N with Genetic and Clinical Data		894	705	189
ER Positive	n (%)	518 (57.9)	411 (58.3)	107 (56.6)
PR Positive	n (%)	512 (57.3)	409 (58)	103 (54.5)
Grade	Well Differentiated	108 (12.1)	100 (14.2)	8 (4.23)
	Intermediate Differentiation	342 (38.3)	276 (39.1)	66 (34.9)
	Poor Differentiation	398 (44.5)	290 (41.1)	108 (57.1)
	Undifferentiated	46 (5.15)	39 (5.53)	7 (3.7)
Stage	1	341 (38.1)	310 (44)	31 (16.4)
	2	343 (38.4)	261 (37)	82 (43.4)
	3	157 (17.6)	109 (15.5)	48 (25.4)
	4	53 (5.93)	25 (3.55)	28 (14.8)

4.3.3 Comparison with other methods

Heritability estimates for risk and ten-year mortality were carried out using the variants from each of the trait's Model 1. Heritability estimates using all annotated variants for risk and mortality are 0.451 (standard error: 0.091) and 0.000002 (standard error: 0.2) respectively.

The polygenic risk score using the associations reported by the 81 SNVs previously reported as associated with breast cancer phenotypes. These analyses produce AUCs of 0.504 for risk (95% CI from 2000 bootstrap replications: 0.482-0.527) and 0.484 for mortality (95% CI: 0.453-0.516).

4.4 Discussion

This analysis demonstrates the usefulness of the Kriging method to use genome-wide germline genetic variation to predict early onset breast cancer risk. The Kriging model that combines the predictive power of limited non-genetic information with whole genome prediction predicts breast cancer risk with an AUC of 0.655 and is a significant improvement over the predictions from a polygenic risk score model. This is consistent with other studies that have found limited predictive power from the combination of variants that meet genome-wide p-value thresholds.[42,53,236] The heritability estimate, which is derived using similar techniques as Kriging, is also the first LMM-based heritability estimate for either breast cancer risk or prognosis, and the results suggest that risk is associated with germline genetic variation but prognosis is not.

This sample is composed of women who were younger than 51 at diagnosis. Since whole genome prediction has not yet been done in any other breast cancer studies, future research in older populations will be needed to investigate whether the Kriging method

produces similar risk estimates and insight in women who are diagnosed later, or whether younger women possess distinct genetic variants that drive breast cancer risk.

This analysis demonstrates that prediction is improved when the genome is partitioned into different classes of variation based on frequency and predicted functionality. This indicates that the underlying architecture may differ for common and rare variants (with common variants contributing more to risk), but may be similar for variants of different predicted functionality.

Compellingly, this analysis suggests breast cancer risk is associated with that variants tagged by this study but have not yet been identified. These variants span all functional categories and are both common and rare (although the rare causal variants appear to be concentrated in variants that cause changes to amino acid translation). Given the results of Chapter 2 (which found few exons where rare risk variants or risk variants of low effect clustered), and given the extensive previous research to look for common variation that is associated with breast cancer risk (which suggests that it is unlikely that common variants exist that are associated with risk with an OR greater than 2^{60}), subsequent studies will likely require very large sample sizes to identify individual variants that are associated with risk. These conclusions are consistent with recent analyses that suggest that there are few high-penetrance causal genes left to be discovered, and if they do exist, they likely exist in only a small number of families, and will not contribute much to population-level risk.[33]

Beyond suggesting that a large sample size would be needed to identify undiscovered variants that are associated with breast cancer risk, this analysis also can inform the methodologies that will be most efficient for subsequent studies of breast cancer risk. With the exception of rare variants that cause changes in amino acid transcription, rare variants collectively show little evidence of being associated with risk of breast cancer. It is possible that there is still predictive power in rare variants that are not well measured by the

genome-wide array or the exome array. However, the GCTA estimate of heritability using all annotated variants is statistically significant, and at 0.451, approaches the total heritability estimated from family studies.[35,71,72] Consequently, if unmeasured rare variation contributes to risk, it is likely to have a modest effect on overall prediction beyond what is tagged by the variants measured in this study. For this reason, while some variation that is associated with risk is likely to be uncovered by next generation sequencing (which can uncover novel rare variants), the overall impact on prediction of the variants discovered by those methods is likely to be small in breast cancer, and similar predictive power could have been achieved by array-based assays. However, since this analysis suggests that intergenic and non-coding variants contribute to risk, subsequent investigations using just exome arrays are unlikely to provide enough information to classify the genetic contribution of a women's risk of breast cancer.

In contrast to the risk analysis, this investigation does not find any compelling evidence that breast cancer prognosis is strongly driven by germline genetics. Two complementary methods, Kriging and polygenic risk scores, both produce null results, and the heritability estimate is consistent with mortality not being a heritable trait. This may be a result of a known downward bias of prediction that is calculated through GREML methods when variants that truly have multiple distributions that characterize their association with disease are included in the same GRM.[225] If that is the case in this analysis, and none of the constructed GRMs reflect the true classes of association between mortality and germline genetics, then the true prediction signal would be obscured. Sample size may also have been insufficient (the GCTA authors recommend a sample size of 3000 for heritability analyses,[238] and the estimate of heritability had a large standard error, which may be indicative of an underpowered analysis). However, the sample size for mortality was not unreasonably small, and previous applications of Kriging have found predictive power in sample sizes of 99.[42] The statistical assumptions of Kriging may also have been violated such as linearity in the

GRMs, independence of the GRMs, and modeling a hazard as a linear outcome. Mortality might be better predicted using methods that do not share these assumptions. However, the linear assumptions of Kriging are fairly robust to deviations from linearity,[225] and the null result presented here is consistent with the mostly null results found by previous single marker regression analyses[154,170–174,239] and polygenic risk score analyses.[240]

While these analyses do not rule out the possibility that germline genetic variation has an effect on breast cancer mortality, it does suggest limits on the genetic architecture of that association. If mortality were driven by variation in a small number (>10) of highly penetrant variants, or a limited number (>100) of variants of modest effect, Kriging would also have limited ability to detect that.[42] However, if that were the case, if they are present at a sufficient frequency to be included in the single marker regression analyses, they had a high probability of being identified by the analysis in Chapter 3 or previous single studies. The influence of germline genetics on mortality may also be mediated through genetic interactions, rather than a linear relationship, and in many circumstances this genetic architecture would not be well captures by either GREML methods or polygenic risk scores. Another possible explanation for the null results may be a limitation of the variants that were interrogated for this study. Since rare variants or copy number variants are less likely to be tagged by the variants in this study, it is possible that rare variants may affect mortality in a way that was not captured by the prediction models.

It is also possible that germline genetic variation is a predictor of mortality, but only for a subset of the cases, and this analysis was not designed or powered to detect any of these interactions. For example, germline genetic variation may have a larger influence on breast cancer mortality in populations that have different background risk factors such as age (this sample was young), ancestry (this sample was of European ancestry), or country of origin (this sample was recruited from affluent countries). It is also possible that genetic variation may be responsible for risk by way of an interaction with treatment, as

suggested by previous analyses.[198] Since treatment decisions were not available for these participants, we were unable to estimate any treatment-by-genetic interactions. Given the young age of these women, they may have been treated more aggressively compared to older women who may have had more comorbidities. It is possible that germline genetics has less of an impact on survival in the presence of aggressive surgery or treatment, than otherwise. Other interactions that could not have been detected are gene by gene interactions, particularly with highly penetrant but rare mutations in BRCA2. The women in this sample are not carriers of known pathogenic mutations in BRCA2. Recent research[33] indicates that BRCA2 mutations interact with other lower penetrance germline variation to produce worse outcomes. If that is the case, a prediction model using Kriging might be able to capture that association, but without any BRCA2 carriers in the study population, this could not be tested in this analysis.

This study has some limitations. The variation measured in this analysis is obtained from two array-based methods. Rare variation, which often does not have the same LD structure as common variation,[99,100] and therefore is tagged poorly by common SNVs, is mostly ascertained through imputation, and not interrogated comprehensively outside of gene regions. While this may underestimate the relative importance of rare variants, it is unlikely to affect overall prediction, since LMM heritability estimates (that are derived from similar methods to Kriging) indicate that the measured variation already accounts for most of the variation that is expected to be due to germline genetics. A second concern is that population stratification can induce upwards bias in prediction models using GREML methods,[220,222] when used to estimate the genetic component alone. This may upwardly bias the prediction for the genetic-only model (although there are low levels of population stratification in our sample), but would not upwardly bias the combined prediction.[42]

A third concern is related to the external validity of the results. As with all prediction models, these results may not produce prediction models that are accurate for women with

different characteristics than the study sample. This model would need to be verified in additional populations before being applied to them.

Tumor subtype is known to be a prognostic factor in breast cancer. While these analyses investigated mortality in ER+ cases specifically, the number of participants were not powered detect modest associations between germline genetic variation and morality within particular tumor subtypes.

In the context of breast cancer risk, the prediction method described here is an improvement on existing models. From an epidemiological perspective, the predictions are useful at the population level, and the understanding of the relative contributions of different classes of variants that is advanced by this analysis will help to better design future studies. From a clinical perspective, the model still has low levels of discrimination, but may be strong enough to be used in very specific scenarios, such as being used to augment the interpretation of screening tools such as mammography (which often return uncertain results), or to help individuals to decide their personal risk/benefit for other medical treatments that may increase the risk of breast cancer. In the context of breast cancer prognosis, these investigations support other lines of evidence that suggest that germline genetic variation does not strongly influence the prognosis of early onset breast cancer. While germline genetic variation may still influence mortality outcomes for some subsets of breast cancer patients, particularly for patients treated with specific systemic treatments, these investigations are unable to find evidence of this.

CHAPTER 5

CONCLUSIONS

5.1 Summary of Results

This thesis contributes to the field of cancer epidemiology through a thorough investigation into the genetic determinants of early onset breast cancer incidence and mortality. Chapters two and three present analyses that are designed to identify gene regions that harbor germline genetic variation that is associated with early onset breast cancer. Chapter two investigates this with respect to risk of developing early onset breast cancer in the general population, and chapter three investigates this with respect to the hazard of mortality for women who were diagnosed early in life. Chapter four presents analyses that predict a women's overall risk of both developing and dying from early onset breast cancer by incorporating whole-genome measures of variation.

The results of this thesis discovered novel risk loci which add meaningfully to the known genetic determinants of breast cancer, and the prediction model has significantly more predictive power than a model that uses only non-genetic risk factors. These insights represent a synthesis of multiple complementary methods, most of which had not been applied to any breast cancer phenotype.

The two complementary goals, identification and prediction, are investigated by analysis of genetic data of participants of existing studies that recruited women who developed breast cancer at a relatively young age. This population is not well-studied, and some non-genetic risk factors have opposite effects in early- and late-onset cases. While not conclusively able to reject this hypothesis, this investigation suggests that the genetic determinants of breast cancer do not systematically differ as a function of age of onset.

5.1.1 Identification of Genes Associated with Breast Cancer Risk

The analyses in Chapter 2 identify three genes in which variation is associated with risk of breast cancer: FGFR2 (discovery $p = 2.18 \cdot 10^{-5}$; replication $p < 10^{-30}$), NEK10 (discovery $p = 1.20 \cdot 10^{-3}$; replication $p < 10^{-30}$), and MKL1 (discovery $p = 2.62 \cdot 10^{-4}$; replication $p < 10^{-30}$). Previous genome-wide association studies (GWASs) had identified loci at each of these genes as being associated with breast cancer risk, but compellingly, conditional analyses indicate that the associations in the MKL1 and NEK10 genes are driven by risk loci are distinct from those previously reported. This suggests that there are risk loci whose combination of rareness or modest effect size cannot be identified by a single marker regression. The SKAT-O test does not directly calculate the magnitude of each of these genes' effect on risk. However, the results of the prediction analysis in Chapter 4 indicate that while their effect is statistically significant, it is likely was small.

Within breast cancer cases, the analysis of Chapter 3 also indicates that women with variation in SLC4A7, and to a lesser extent, the adjacent gene NEK10, are at a higher risk of developing progesterone receptor positive breast cancer (SLC4A7 $p = 8.8 \cdot 10^{-4}$, and contains a previously identified risk loci; NEK10 p=$6.19 \cdot 10^{-3}$). This suggests that future prevention efforts can be targeted to deliver chemoprevention that works through the progesterone receptor pathway to women who are most likely to benefit from it.

The analyses of Chapter 2 also characterize the sparsity of causal variants in genes that are identified as associated with risk. There have been few previously published descriptions of this characteristic of the genes that are responsible for breast cancer risk and prognosis, even though the sparsity of causal loci within genes dictates the optimal statistical method to identify those genes. The ρ mixing parameter indicates that the sequence kernel association test (SKAT) was a more appropriate test than the burden test for genes in which even a modest number of variants were measured (more than five variants), although there

were some exceptions. This suggests that many variants in those larger genes are not associated with risk, and that the directions of effect of the minor alleles that are causal can be both protective and deleterious. This is consistent with the single marker regression coefficients from Chapter 2 and previous research,[34] which report beta coefficients for the minor alleles that are both greater than and less than zero. This observation strongly suggests that in the case of breast cancer phenotypes, burden-style gene-based statistical approaches are not going to be optimally-powered to identify genes that are associated with risk, particularly if the variants are interrogated with sequencing techniques (the variants on the exome array used for this study are enriched for variants that were likely to be causal, and sequencing methods would likely detect many more not-associated variants, whose noise could further overwhelm burden-style tests). Since in some genes the burden style test was more appropriate than the SKAT test, omnibus tests such as the optimal SKAT (SKAT-O) that can detect genes in both sparsity scenarios will most likely be the optimal choice for future studies.

5.1.2 Whole Genome Prediction of Breast Cancer Risk

This thesis presents a prediction model of breast cancer risk in Chapter 4 that incorporates the effect of all measured germline genetic variation. The genetic data alone is able to predict breast cancer risk with an area under the receiver operating characteristic curve (AUC) of 0.618 (95% CI 0.610-0.629). When the influence of a limited set of non-genetic predictors is also incorporated, the combined model is able to predict breast cancer risk with an AUC of 0.655 (95% CI: 0.649-0.660). This combined model is a significant improvement over models that include only the genetic information or only the non-genetic risk factors.

The analyses of Chapter 4 also begin to characterize variants that are responsible for breast cancer risk. The prediction method that was implemented allows for groupings

of variants, and for the variants within the separate groups to each be characterized by separate distributions that describe their contribution to breast cancer risk. The analyses in Chapter 4 grouped variants by predicted functionality and rareness, and the weights on each of these groups represent the relative strength of these associations between the variants within that group with risk. These weights suggest that the variants that are responsible for breast cancer risk are annotated to all classes of predicted functionality. Variants that cause changes in amino acid translation, variants that are located within or near genes but do not cause changes in amino acid translation, and intergenic variants as a class have some power to predict breast cancer risk. The weights also are able to characterize the causal variants in terms of their rareness. The weights suggest that rare variants are largely not able to predict breast cancer risk, with the exception of rare variants that alter amino acid translation.

The analyses in Chapter 4 also suggest that there are still undiscovered variants that are responsible for breast cancer risk, and that these undiscovered variants are also found in all categories of annotated functionality. These findings have direct consequences for future studies of breast cancer risk that may attempt to identify novel risk loci. The results suggest that studies that exclusively measure variation with technologies such as whole-exome sequencing or exome arrays, which do not assay variants outside of the exons, will not capture the effect of all of the risk variants that are driving the whole genome predictive power. However, additional rare variants that can be measured are expected to contribute only small amounts to disease risk (see discussion in section 5.2). For this reason, whole-genome sequencing may not be an efficient use of resources compared to array-based methods. This implies that despite the gaps in their ability to interrogate rare variants, array based methods and imputation may continue to be a cost-effective way to identify novel risk loci.

These results together suggest that undiscovered variation that is associated with breast cancer risk within gene regions is likely to be characterized by one of three descriptions:

(1) of large effect size, but so rare as to not have much effect on population-level risk of disease; (2) common in the population, but increases the risk of breast cancer by a very small amount; (3) or both rare and of weak effect size. Undiscovered variation that lies outside of gene regions may be slightly more common or of larger effect than the undiscovered variation within gene regions, but is still and rare enough or of small enough effect size to not have been identified by previous single marker regression analyses, or influence the overall prediction model. This suggests that there is not much additional predictive power to be gained from their identification, and if they are rare with large effects (as is to be expected if they are under purifying selection[40]), then family-based studies may be more appropriate than population-based studies to identify them.

5.1.3 Genetic Determinants of Breast Cancer Prognosis

In contrast to the analyses of the genetic determinants of breast cancer development, the investigations of Chapters 3 and 4 do not find any compelling evidence that breast cancer mortality is strongly driven by germline genetics that could be measured by our study. Five complementary methods (single marker regression analyses, SKAT-O, Kriging, polygenic risk scores, and the heritability estimation) all find null results.

Mortality analysis in estrogen receptor positive patients did not find a significant association between any of the CYP genes and mortality. While recognizing that this null result was found in a modest sample size, interrogated a limited number of polymorphisms, and did not incorporate actual treatment information, this is consistent with other recent research that questions whether polymorphisms in the CYP genes translate into poorer outcomes for women whose metabolism of tamoxifen is affected by CYP polymorphisms.[196,197]

In terms of intermediate markers of prognosis, the analyses of Chapter 3 do find suggestive evidence that progesterone receptor status in cases is associated with variants in the solute carrier family 4 member 7 gene (SLC4A7). The SLC4A7 protein has a known role in neural sensory transmission, but it has been implicated in single marker regression analyses as being associated with both breast cancer and cardiovascular complex traits.[34] The nature of its role in cancer phenotypes has not yet been established.

5.1.4 Novel use of Methods

This thesis applies five complementary methods to investigate the relationship between germline genetic variation and the risk and prognosis of breast cancer: single marker regression associations of common variation in gene regions; SKAT-O associations of all variation in gene regions; whole-genome Kriging prediction; polygenic risk score prediction using previously associated loci; and whole-genome heritability estimation. For three of these methods (SKAT-O, Kriging, and heritability estimation), these analyses represent the first applications of those methods in the context of breast cancer.

The previous investigations into breast cancer risk that used gene-based tests all used burden-style analyses. Our analyses of the ρ mixing parameter of the risk analyses suggest that the assumptions of the burden test are not always reflective of the genetic architecture of breast cancer risk, and therefore the results of these studies may not be optimal.

In addition to the statistical methods, this thesis represent only the third study to directly interrogate rare variants and their association with either breast cancer risk or prognosis (this study and two previous ones measured rare variation using an exome array; whole-genome and whole-exome sequencing projects have not yet been completed). The success of the SKAT-O risk analyses in identifying genes that are suggestively associated with breast cancer risk suggests that better powered studies may find more such genes.

In the context of any application of gene-based tests, this thesis also represents the first ever use of the Combined Annotation Dependent Depletion (CADD) scores as weights for any phenotype. The investigations suggest that the CADD weights do improve power, and do not increase the rate of type I error. The CADD weights appear to have better performance than the often-used beta-transformation-of-minor-allele-frequency weights, and their use allows for all measured variation to be included in the analyses.

These analyses demonstrate the usefulness of the Kriging method to predict early onset breast cancer risk, and Kriging methods are a significant improvement in predictive power over the predictions from a polygenic risk score model. This is consistent with other studies that have found limited predictive power from the combination of variants that meet genome-wide p-value thresholds.[42,53,236]

5.2 Limitations

The analyses presented in this thesis have some limitations, which can be classified as being related to the study participants, the variants measured, and the analytical techniques.

5.2.1 Participants

The composition of the participants of the primary data used for this study, in particular their young age of onset, was both a strength and a weakness. Given the complex relationship between some non-genetic risk factors for breast cancer and age, the focus of the BCFR studies on women who developed breast cancer before menopause made it possible to directly investigate hypotheses about the differences and similarities of the genetic determinants of breast cancer by age. However, after the initial analysis there were no genes or loci that were so strongly related in the initial sample that they did not need to be confirmed in a second independent sample. Therefore, although a primary rationale for

enrolling women with an early onset of breast cancer was to be able to directly interrogate questions about the interaction between age and germline genetic risk factors, ultimately the analysis in Chapter 2 was only able to identify genes that were associated with breast cancer risk in women of all ages of onset.

The conclusions drawn from these analyses may have been more cohesive if the characteristics of the replication data sets better matched the characteristics of the discovery set. The participants in the replication data sets are of a different age. It is possible that a larger secondary independent data set of age-matched patients may have been able to better replicate the primary analysis. As more data become publicly available and the methods to create gene-based tests from summary statistics improve, there will be a better ability to match the characteristics of replication samples to those in the primary analysis, and there will be a greater ability to detect drivers of breast cancer risk and mortality.

The analyses are also limited by sample size. Although the analyses presented in Chapters 3 and 4 represent the largest single-study whole-genome investigation into breast cancer mortality, the number of participants is relatively modest, and may not be powered to detect some true associations. In particular, given the known relationship between tumor subtype and mortality, a future mortality studies may want to limit themselves to participants with homogeneous tumor subtypes in order to not introduce a possible source of noise into the analysis.

The analyses were also limited by the type of covariates that were collected, and the limited power to be able to detect interactions between environmental, tumor, and treatment characteristics and germline genetic variation. The BCFR participants were all of the same ancestry, and all recruited from OECD countries, it is possible that the diagnostic and treatment trajectories differed between the treatment sites. Treatment information was not available, and only limited diagnostic information was available, and even if further infor-

mation was known, the modest sample size of this study would not have been powered to detect differences in risk and mortality between those unmeasured confounders.

5.2.2 Variants Measured

The variants that were assayed in these studies likely did not tag all variants that are associated with breast cancer. In Chapters 2 and 3, the variation in gene regions was assayed using a genotyping array, which is only able to interrogate ~250,000 variants throughout gene regions, and can only detect the effect of causal variants that are in high linkage disequilibrium (LD) with a genotyped variant. Those variants were selected by Illumina because previous sequencing projects identified variation at those positions. It is almost certain that most individuals in this study are carriers of rare mutations that were not able to be interrogated. For this reason, sequencing of the whole gene region, which, unlike array based methods, does not require prior knowledge of variation at a locus to identify it, would have provided a more comprehensive analysis.

Similarly, in the whole genome prediction models of Chapter 4, whole-genome sequencing would have been preferable to the array-based ascertainment (augmented by imputation). Using the array based technologies, outside of gene regions, rare variation was almost exclusively inferred through imputation. Since rare variation often does not have the same LD structure as common variation,[99,100] and therefore often is tagged poorly by common single nucleotide variants (SNVs), it is almost certain that many rare variants existed in this sample that were not able to be included in the prediction model.

However, while this may result in an underestimate of the *relative* importance of rare variants in the Kriging model's optimal weights, it is unlikely to affect overall prediction, since the heritability estimate using all measured variants is already quite high, and of similar magnitude to estimates from family studies. Similarly, while the inclusion of more rare

variants may have identified additional genes in which variation is significantly associated with risk, the magnitude of the additional risk conferred by those genes is likely to be quite modest, or only affect a small number of women.

In addition to being limited by the technologies that were used to measure rare variation, the gene-based analyses of Chapter 2 are also limited due to the unbalanced ascertainment of rare variation that is a result of the uneven number of cases and controls. In an attempt to be well powered for a prognosis analysis, the risk analyses included 3479 cases and 973 controls. Since many of the rare variants were only observed in one participant, this imbalance in cases and controls resulted in a sample where rare variants were more likely to be seen in the cases, resulting in more power to detect rare deleterious rare variants over rare protective ones. This exacerbates the known bias in which single variant regressions of rare variants are known to be biased towards odds ratios larger than one.[129]

The conclusions drawn from these analyses may have been stronger if the variants measured by the replication data sets better matched the variants measured by the discovery set. In Chapters 2 and 3, the participants in the replication data sets have their genetic variation interrogated with genome-wide arrays and imputation rather than an exome array. As a result, the replication genetic data interrogated fewer rare variants in exons, but also was able to include many more common variants that were located elsewhere within gene regions (e.g.: introns). This came about because the exome array used in the discovery analysis targeted nonsynonymous SNVs, and other variants in gene regions were omitted due to the limited space on the array. While the CADD weights limit the effects of this ascertainment difference by effectively down-weighting the additional variants that were assayed in the replication sample, the different makeup of the replication genetic data is not ideal.

5.2.3 Analysis

During the quality control and analysis of the data, several decisions were made that included implicit assumptions that also present limitations for the interpretation of these results.

The quality control of the exome array excluded participants based on genotyping rate, gender mismatch, high heterozygosity, duplicated genotypes, principal component outliers, and participants whose genotypes were highly correlated. Ultimately, almost 10% of the participants who were genotyped were excluded, mostly due to high heterozygosity (52 excluded) or high estimated relatedness (126 excluded). While some of relatedness may be explained by family-based ascertainment of the breast cancer cases, both high heterozygosity and highly correlated genotypes may also be a marker for contamination between samples. There was also some evidence that on some plates variants were unreliably assigned, which also decreases confidence in the exome chip assay.

The variants that were used in the discovery sample to suggest that genes were suggestively associated with early onset breast cancer differed from the variants that were used in the GAME-ON/DRIVE analysis to attempt to replicate those signals. This decision complicates the interpretation of the suggestively associated genes of Chapter 2 and the largely null results of Chapter 3. In the suggestively associated genes identified by Chapter 2 as associated with breast cancer risk, an interpretation is that disruption within the gene regions of FGFR2, MKL1, and NEK10 is associated with risk. The discovery data set was able to identify the effect of this disruption that was caused or tagged by largely rare, nonsynonymous variants, and the replication data set was able to identify the effect of this disruption that was caused or tagged by common variants. If this interpretation is correct, than only genes in which both common and rare variation contributes to breast cancer risk could have been identified.

Additionally, Chapter 2, the method used to create a gene-based test differed between the discovery and replication samples. Being able to directly use SKAT-O in both analyses would have been preferable.

5.3 Next Steps

The conclusions of this thesis suggest several next steps. Given the modest sample size and the suggestive nature of the results in Chapter 2, the identified genes would have more robust evidence of association if their association is replicated in an independent set of cases and controls that are well matched in age and genetic ascertainment (or studies that interrogated all variants, such as sequencing studies). This would more comprehensively describe the role of rare variants in breast cancer risk.

It would also be fruitful to further investigate the genes that are identified in Chapter 2. While a search of gene expression databases indicate that they are expressed in breast tissue, the magnitude of the effect of polymorphisms on this gene on both the gene product and downstream phenotypes would further elucidate their role in breast cancer risk, and help to better describe the mechanism by which they increase that risk. Functional studies in model systems would help to further study the way by which these genes influence the progression of breast cancer.

As our understanding of the interactions within the genome improves, future analyses with this same study population may produce additional results. At the present time, there is only a rudimentary ability to annotate non-exonic variants to genes, but this is a subject of intense interest. As the understanding of biological pathways improves, variants will be able to be annotated to a particular gene in ways that are more sophisticated than ANNOVAR's annotation, which is used in these analyses. For example, regulatory variants that are not spatially near the genes that they regulate could be included in the analyses, and, if

these variants are responsible for breast cancer risk, their inclusion will improve the ability of gene-based tests to identify genes responsible for breast cancer.

The participants of these studies are all of a homogeneous age (younger than 51 at diagnosis), ancestral background (European), and gender (women). While ancestrally homogeneous samples have more power to detect the effect of rare variants (which are often population-specific), breast cancer affects people of all ages, ancestral backgrounds, and genders. For this reason, additional SKAT-O analyses in populations with different characteristics will help to determine whether the genes that harbor variation that is associated with breast cancer risk differ in their effect across populations. Genome-wide germline genetic variation is publically available for women of Latina,[241] African American,[242] Japanese,[243] and Chinese[243] ancestry for case control studies of women of all ages, and possible future collaborations could allow for the sharing of mortality information in the cases. The same analyses that were carried out in this thesis, when applied to different populations, may uncover additional insight into the genetic basis of differences in risk and mortality that are associated with these non-genetic traits.

Mortality may be affected by germline genetics in the context of particular treatment regimens. Therefore, it would be fruitful to repeat the analyses of Chapters 3 and 4 with a sample of known, homogeneous treatment. In particular, given the still-unsettled relationship between variation in the CYP gene and survival, it would be of clinical interest to repeat the mortality analyses of Chapter 3 on patients with estrogen receptor positive tumors who were treated with tamoxifen.

The success of the risk prediction model in Chapter 4 also suggests that incorporating all known non-genetic risk factors (rather than the limited set of non-genetic risk factors available for our study population) would produce a prediction model with even more predictive power, and may even produce a model whose predictive power is sufficient to be used in a clinical setting.

The largely null association between mortality and germline genetics is complicated by the results of family based studies,[35] which suggest that mortality does have heritable component, as first degree relatives have more similar mortality outcomes than would be expected by chance. These family studies may be biased by shared environment. However, the results of this thesis do not preclude a possible role for germline genetic variation in the survival of early onset breast cancer, but they do suggest some limits on the strengths of that association and the characteristics of the variants that drive it. To more comprehensively approach this question, future work to identify variants associated with morality would be able to investigate the genetic determinants of prognosis if the study (1) has many participants (both to be able to ascertain the existence of rare variants and also have statistical power to detect their association with disease), (2) incorporate more variation than is assayed on the exome array, (3) consider methods that allow for the detection of larger-than-gene pathways that have modest effect size on risk, and (4) enrolls participants with homogeneous and known treatment regimens.

5.4 Implications

These analyses identified three genes that are suggestively associated with breast cancer risk, and one that is suggestively associated with progesterone receptor status in cases. These all represent possible pharmacological targets for cancer chemoprevention. These analyses also developed a prediction model for breast cancer risk that improves upon existing methods of prediction, and is strong enough to be useful at the population level. From a clinical perspective, the model still has low levels of discrimination, but may be strong enough to be used in very specific scenarios, such as interpretation of ambiguous screening results, or to help individuals to understand their personal risk when considering other medical treatments that may increase the risk of breast cancer such as hormone replacement

therapy or hormone-assisted reproductive therapy. In the context of breast cancer prognosis, these investigations support other lines of evidence that suggest that for many women who are diagnosed with breast cancer, germline genetic variation does not strongly influence the risk of mortality. While germline genetic variation may still influence mortality outcomes for some subsets of breast cancer patients (and patients treated with specific systemic treatments are of particular interest in terms of patients who may have their mortality influenced by germline genetics), these investigations are unable to find evidence of this.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA: A Cancer Journal for Clinicians*, vol. 65, pp. 5–29, Jan. 2015.

[2] C. E. DeSantis, S. A. Fedewa, A. Goding Sauer, J. L. Kramer, R. A. Smith, and A. Jemal, "Breast cancer statistics, 2015: Convergence of incidence rates between black and white women," *CA: a cancer journal for clinicians*, vol. 66, pp. 31–42, Feb. 2016.

[3] L. B. Dunn, D. J. Langford, S. M. Paul, M. B. Berman, D. M. Shumay, K. Kober, J. D. Merriman, C. West, J. M. Neuhaus, and C. Miaskowski, "Trajectories of fear of recurrence in women with breast cancer," *Supportive Care in Cancer: Official Journal of the Multinational Association of Supportive Care in Cancer*, vol. 23, pp. 2033–2043, July 2015.

[4] C. Holmberg, "No one sees the fear: becoming diseased before becoming ill–being diagnosed with breast cancer," *Cancer Nursing*, vol. 37, pp. 175–183, June 2014.

[5] A. Mehnert, P. Berg, G. Henrich, and P. Herschbach, "Fear of cancer progression and cancer-related intrusive cognitions in breast cancer survivors," *Psycho-Oncology*, vol. 18, pp. 1273–1280, Dec. 2009.

[6] B. Jonsson and N. Wilking, "Prevention and the economic burden of breast cancer," tech. rep., GE Healthcare, 2013.

[7] C. Lu, M. Xie, M. C. Wendl, J. Wang, M. D. McLellan, M. D. M. Leiserson, K.-l. Huang, M. A. Wyczalkowski, R. Jayasinghe, T. Banerjee, J. Ning, P. Tripathi, Q. Zhang, B. Niu, K. Ye, H. K. Schmidt, R. S. Fulton, J. F. McMichael, P. Batra, C. Kandoth, M. Bharadwaj, D. C. Koboldt, C. A. Miller, K. L. Kanchi, J. M. Eldred, D. E. Larson, J. S. Welch, M. You, B. A. Ozenberger, R. Govindan, M. J. Walter, M. J. Ellis, E. R. Mardis, T. A. Graubert, J. F. Dipersio, T. J. Ley, R. K. Wilson, P. J. Goodfellow, B. J. Raphael, F. Chen, K. J. Johnson, J. D. Parvin, and L. Ding, "Patterns and functional implications of rare germline variants across 12 cancer types," *Nature Communications*, vol. 6, p. 10086, Dec. 2015.

[8] S. Chung, S.-K. Low, H. Zembutsu, A. Takahashi, M. Kubo, M. Sasa, and Y. Nakamura, "A genome-wide association study of chemotherapy-induced alopecia in breast cancer patients," *Breast cancer research: BCR*, vol. 15, no. 5, p. R81, 2013.

[9] L. R. Parham, L. P. Briley, L. Li, J. Shen, P. J. Newcombe, K. S. King, A. J. Slater, A. Dilthey, Z. Iqbal, G. McVean, C. J. Cox, M. R. Nelson, and C. F. Spraggs, "Comprehensive genome-wide evaluation of lapatinib-induced liver injury yields a single genetic signal centered on known risk allele HLA-DRB1*07:01," *The Pharmacogenomics Journal*, vol. 16, pp. 180–185, Apr. 2016.

[10] B. P. Schneider, L. Li, F. Shen, K. D. Miller, M. Radovich, A. O'Neill, R. J. Gray, D. Lane, D. A. Flockhart, G. Jiang, Z. Wang, D. Lai, D. Koller, J. H. Pratt, C. T. Dang, D. Northfelt, E. A. Perez, T. Shenkier, M. Cobleigh, M. L. Smith, E. Railey, A. Partridge, J. Gralow, J. Sparano, N. E. Davidson, T. Foroud, and G. W. Sledge, "Genetic variant predicts bevacizumab-induced hypertension in ECOG-5103 and ECOG-2100," *British Journal of Cancer*, vol. 111, pp. 1241–1248, Sept. 2014.

[11] M. P. Goetz, J. X. Sun, V. J. Suman, G. O. Silva, C. M. Perou, Y. Nakamura, N. J. Cox, P. J. Stephens, V. A. Miller, J. S. Ross, D. Chen, S. L. Safgren, M. J. Kuffel, M. M. Ames, K. R. Kalari, H. L. Gomez, A. M. Gonzalez-Angulo, O. Burgues, H. B. Brauch, J. N. Ingle, M. J. Ratain, and R. Yelensky, "Loss of heterozygosity at the CYP2d6 locus in breast cancer: implications for germline pharmacogenetic studies," *Journal of the National Cancer Institute*, vol. 107, Feb. 2015.

[12] C. B. Ambrosone, C.-C. Hong, and P. J. Goodwin, "Host Factors and Risk of Breast Cancer Recurrence: Genetic, Epigenetic and Biologic Factors and Breast Cancer Outcomes," *Advances in Experimental Medicine and Biology*, vol. 862, pp. 143–153, 2015.

[13] G. Absenger, J. Szkandera, M. Stotz, M. Pichler, T. Winder, T. Langsenlehner, U. Langsenlehner, H. Samonigg, W. Renner, and A. Gerger, "A common and functional gene variant in the vascular endothelial growth factor a predicts clinical outcome in early-stage breast cancer," *Molecular Carcinogenesis*, vol. 52 Suppl 1, pp. E96–102, Nov. 2013.

[14] M. J. Labonte, P. M. Wilson, D. Yang, W. Zhang, R. D. Ladner, Y. Ning, A. Gerger, P. O. Bohanes, L. Benhaim, R. El-Khoueiry, A. El-Khoueiry, and H.-J. Lenz, "The Cyclin D1 (CCND1) A870g polymorphism predicts clinical outcome to lapatinib and capecitabine in HER2-positive metastatic breast cancer," *Annals of oncology: official journal of the European Society for Medical Oncology / ESMO*, vol. 23, pp. 1455–1464, June 2012.

[15] G. Allegrini, L. Coltelli, P. Orlandi, A. Fontana, A. Camerini, A. Ferro, M. Cazzaniga, V. Casadei, S. Lucchesi, E. Bona, M. Di Lieto, I. Pazzagli, F. Villa, D. Amoroso, M. Scalese, G. Arrighi, S. Molinaro, A. Fioravanti, C. Finale, R. Triolo, T. Di Desidero, S. Donati, L. Marcucci, O. Goletti, M. Del Re, B. Salvadori, I. Ferrarini, R. Danesi, A. Falcone, and G. Bocci, "Pharmacogenetic interaction analysis of VEGFR-2 and IL-8 polymorphisms in advanced breast cancer patients treated with paclitaxel and bevacizumab," *Pharmacogenomics*, vol. 15, pp. 1985–1999, Dec. 2014.

[16] A. Rudolph, R. Hein, S. Lindström, L. Beckmann, S. Behrens, J. Liu, H. Aschard, M. K. Bolla, J. Wang, T. Truong, E. Cordina-Duverger, F. Menegaux, T. Brüning, V. Harth, GENICA Network, G. Severi, L. Baglietto, M. Southey, S. J. Chanock,

J. Lissowska, J. D. Figueroa, M. Eriksson, K. Humpreys, H. Darabi, J. E. Olson, K. N. Stevens, C. M. Vachon, J. A. Knight, G. Glendon, A. M. Mulligan, A. Ashworth, N. Orr, M. Schoemaker, P. M. Webb, kConFab Investigators, AOCS Management Group, P. Guénel, H. Brauch, G. Giles, M. García-Closas, K. Czene, G. Chenevix-Trench, F. J. Couch, I. L. Andrulis, A. Swerdlow, D. J. Hunter, D. Flesch-Janys, D. F. Easton, P. Hall, H. Nevanlinna, P. Kraft, J. Chang-Claude, and Breast Cancer Association Consortium, "Genetic modifiers of menopausal hormone replacement therapy and breast cancer risk: a genome-wide interaction study," *Endocrine-Related Cancer*, vol. 20, pp. 875–887, Dec. 2013.

[17] K. A. Pooley, S. E. Bojesen, M. Weischer, S. F. Nielsen, D. Thompson, A. Amin Al Olama, K. Michailidou, J. P. Tyrer, S. Benlloch, J. Brown, T. Audley, R. Luben, K.-T. Khaw, D. E. Neal, F. C. Hamdy, J. L. Donovan, Z. Kote-Jarai, C. Baynes, M. Shah, M. K. Bolla, Q. Wang, J. Dennis, E. Dicks, R. Yang, A. Rudolph, J. Schildkraut, J. Chang-Claude, B. Burwinkel, G. Chenevix-Trench, P. D. P. Pharoah, A. Berchuck, R. A. Eeles, D. F. Easton, A. M. Dunning, and B. G. Nordestgaard, "A genome-wide association scan (GWAS) for mean telomere length within the COGS project: identified loci show little association with hormone-related cancer risk," *Human Molecular Genetics*, vol. 22, pp. 5056–5064, Dec. 2013.

[18] J. L. Murray, P. Thompson, S. Y. Yoo, K.-A. Do, M. Pande, R. Zhou, Y. Liu, A. A. Sahin, M. L. Bondy, and A. M. Brewster, "Prognostic value of single nucleotide polymorphisms of candidate genes associated with inflammation in early stage breast cancer," *Breast Cancer Research and Treatment*, vol. 138, pp. 917–924, Apr. 2013.

[19] A. E. Teschendorff, A. Miremadi, S. E. Pinder, I. O. Ellis, and C. Caldas, "An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer," *Genome Biology*, vol. 8, no. 8, p. R157, 2007.

[20] M. Foedermayr, M. Sebesta, M. Rudas, A. S. Berghoff, R. Promberger, M. Preusser, P. Dubsky, F. Fitzal, M. Gnant, G. G. Steger, A. Weltermann, C. C. Zielinski, O. Zach, and R. Bartsch, "BRCA-1 methylation and TP53 mutation in triple-negative breast cancer patients without pathological complete response to taxane-based neoadjuvant chemotherapy," *Cancer Chemotherapy and Pharmacology*, vol. 73, pp. 771–778, Apr. 2014.

[21] V. Le Morvan, S. Litière, A. Laroche-Clary, S. Ait-Ouferoukh, R. Bellott, C. Messina, D. Cameron, H. Bonnefoi, and J. Robert, "Identification of SNPs associated with response of breast cancer patients to neoadjuvant chemotherapy in the EORTC-10994 randomized phase III trial," *The Pharmacogenomics Journal*, vol. 15, pp. 63–68, Feb. 2015.

[22] F. Jasmine, H. Ahsan, I. L. Andrulis, E. M. John, J. Chang-Claude, and M. G. Kibriya, "Whole-genome amplification enables accurate genotyping for microarray-based high-density single nucleotide polymorphism array," *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, vol. 17, pp. 3499–3508, Dec. 2008.

[23] C. K. Anders, D. S. Hsu, G. Broadwater, C. R. Acharya, J. A. Foekens, Y. Zhang, Y. Wang, P. K. Marcom, J. R. Marks, P. G. Febbo, J. R. Nevins, A. Potti, and K. L. Blackwell, "Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression," *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 26, pp. 3324–3330, July 2008.

[24] S. H. Ahn, B. H. Son, S. W. Kim, S. I. Kim, J. Jeong, S.-S. Ko, W. Han, and Korean Breast Cancer Society, "Poor outcome of hormone receptor-positive breast cancer at very young age is due to tamoxifen resistance: nationwide survival data in Korea–a report from the Korean Breast Cancer Society," *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 25, pp. 2360–2368, June 2007.

[25] N. S. El Saghir, M. Seoud, M. K. Khalil, M. Charafeddine, Z. K. Salem, F. B. Geara, and A. I. Shamseddine, "Effects of young age at presentation on survival in breast cancer," *BMC Cancer*, vol. 6, p. 194, 2006.

[26] S. A. Narod, "Breast cancer in young women," *Nature Reviews Clinical Oncology*, vol. 9, pp. 460–470, Aug. 2012.

[27] A. Bharat, R. L. Aft, F. Gao, and J. A. Margenthaler, "Patient and tumor characteristics associated with increased mortality in young women (<=40 years) with breast cancer," *Journal of Surgical Oncology*, vol. 100, pp. 248–251, Sept. 2009.

[28] SEER, "Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Fast Stats Database: SEER Survival - SEER 13 Regs Research Data, (1988-2012) National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2016, based on the November 2015 submission.."

[29] K. M. O'Brien, J. Sun, D. P. Sandler, L. A. DeRoo, and C. R. Weinberg, "Risk factors for young-onset invasive and in situ breast cancer," *Cancer causes & control: CCC*, vol. 26, pp. 1771–1778, Dec. 2015.

[30] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin,

D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, pp. 747–753, Oct. 2009.

[31] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nature Reviews. Genetics*, vol. 9, pp. 356–369, May 2008.

[32] T. Zemunik and V. Boraska, "Genetics of Type 1 Diabetes," *http://cdn.intechopen.com/pdfs-wm/24179.pdf*, 2016.

[33] D. G. Evans, S. Astley, P. Stavrinos, E. Harkness, L. S. Donnelly, S. Dawe, I. Jacob, M. Harvie, J. Cuzick, A. Brentnall, M. Wilson, F. Harrison, K. Payne, and A. Howell, *Improvement in risk prediction, early detection and prevention of breast cancer in the NHS Breast Screening Programme and family history clinics: a dual cohort study*. Programme Grants for Applied Research, Southampton (UK): NIHR Journals Library, 2016.

[34] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 9362–9367, June 2009.

[35] O. Bahcall, "Common variation and heritability estimates for breast, ovarian and prostate cancers," *Nature Genetics*, 2013.

[36] G. A. Millot, M. A. Carvalho, S. M. Caputo, M. P. Vreeswijk, M. A. Brown, M. Webb, E. Rouleau, S. L. Neuhausen, T. v. O. Hansen, A. Galli, R. D. Brandão, M. J. Blok, A. Velkova, F. J. Couch, A. N. Monteiro, and on behalf of the ENIGMA (Evidence-based Network for the Interpretation of Germline Mutant Alleles) Consortium Functional Assay Working Group, "A guide for functional analysis of BRCA1 variants of uncertain significance," *Human Mutation*, vol. 33, pp. 1526–1537, Nov. 2012.

[37] S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, D. C. Christiani, M. M. Wurfel, and X. Lin, "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies," *American Journal of Human Genetics*, vol. 91, pp. 224–237, Aug. 2012.

[38] M.-X. Li, H.-S. Gui, J. S. Kwan, and P. C. Sham, "GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure," *American Journal of Human Genetics*, vol. 88, pp. 283–293, Mar. 2011.

[39] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nature Genetics*, vol. 46, pp. 310–315, Mar. 2014.

[40] L. H. Uricchio, N. A. Zaitlen, C. J. Ye, J. S. Witte, and R. D. Hernandez, "Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants," *Genome Research*, vol. 26, pp. 863–873, July 2016.

[41] E. Amir, O. C. Freedman, B. Seruga, and D. G. Evans, "Assessing women at high risk of breast cancer: a review of risk assessment models," *Journal of the National Cancer Institute*, vol. 102, pp. 680–691, May 2010.

[42] H. E. Wheeler, K. Aquino-Michaels, E. R. Gamazon, V. V. Trubetskoy, M. E. Dolan, R. S. Huang, N. J. Cox, and H. K. Im, "Poly-Omic Prediction of Complex Traits: OmicKriging," *Genetic Epidemiology*, vol. 38, pp. 402–415, July 2014. arXiv: 1303.1788.

[43] D. L. Hertz and J. M. Rae, "One step at a time: CYP2d6 guided tamoxifen treatment awaits convincing evidence of clinical validity," *Pharmacogenomics*, vol. 17, pp. 823–826, June 2016.

[44] H. Zembutsu, "Pharmacogenomics toward personalized tamoxifen therapy for breast cancer," *Pharmacogenomics*, vol. 16, no. 3, pp. 287–296, 2015.

[45] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill, "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually," *Journal of the National Cancer Institute*, vol. 81, pp. 1879–1886, Dec. 1989.

[46] C. Meads, I. Ahmed, and R. D. Riley, "A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance," *Breast Cancer Research and Treatment*, vol. 132, pp. 365–377, Oct. 2011.

[47] A. I. Vazquez, Y. Veturi, M. Behring, S. Shrestha, M. Kirst, M. F. R. Resende, and G. d. l. Campos, "Increased Proportion of Variance Explained and Prediction Accuracy of Survival of Breast Cancer Patients with Use of Whole-Genome Multiomic Profiles," *Genetics*, vol. 203, pp. 1425–1438, July 2016.

[48] C. A. Haiman, Y. Han, Y. Feng, L. Xia, C. Hsu, X. Sheng, L. C. Pooler, Y. Patel, L. N. Kolonel, E. Carter, K. Park, L. L. Marchand, D. V. D. Berg, B. E. Henderson, and D. O. Stram, "Genome-Wide Testing of Putative Functional Exonic Variants in Relationship with Breast and Prostate Cancer Risk in a Multiethnic Population," *PLOS Genet*, vol. 9, p. e1003419, Mar. 2013.

[49] S. A. Haddad, E. A. Ruiz-Narváez, C. A. Haiman, L. E. Sucheston-Campbell, J. T. Bensen, Q. Zhu, S. Liu, S. Yao, E. V. Bandera, L. Rosenberg, A. F. Olshan, C. B. Ambrosone, J. R. Palmer, and K. L. Lunetta, "An exome-wide analysis of low frequency and rare variants in relation to risk of breast cancer in African American Women: the AMBER Consortium," *Carcinogenesis*, vol. 37, pp. 870–877, Sept. 2016.

[50] H. Ahsan, J. Halpern, M. G. Kibriya, B. L. Pierce, L. Tong, E. Gamazon, V. McGuire, A. Felberg, J. Shi, F. Jasmine, S. Roy, R. Brutus, M. Argos, S. Melkonian, J. Chang-Claude, I. Andrulis, J. L. Hopper, E. M. John, K. Malone, G. Ursin, M. D. Gammon, D. C. Thomas, D. Seminara, G. Casey, J. A. Knight, M. C. Southey, G. G. Giles, R. M. Santella, E. Lee, D. Conti, D. Duggan, S. Gallinger, R. Haile, M. Jenkins, N. M. Lindor, P. Newcomb, K. Michailidou, C. Apicella, D. J. Park, J. Peto, O. Fletcher, I. d. S. Silva, M. Lathrop, D. J. Hunter, S. J. Chanock, A. Meindl, R. K. Schmutzler, B. Müller-Myhsok, M. Lochmann, L. Beckmann, R. Hein, E. Makalic, D. F. Schmidt, Q. M. Bui, J. Stone, D. Flesch-Janys, N. Dahmen, H. Nevanlinna, K. Aittomäki, C. Blomqvist, P. Hall, K. Czene, A. Irwanto, J. Liu, N. Rahman, C. Turnbull, f. t. F. B. C. Study, A. M. Dunning, P. Pharoah, Q. Waisfisz, H. Meijers-Heijboer, A. G. Uitterlinden, F. Rivadeneira, D. Nicolae, D. F. Easton, N. J. Cox, and A. S. Whittemore, "A Genome-wide Association Study of Early-Onset Breast Cancer Identifies PFKM as a Novel Breast Cancer Gene and Supports a Common Genetic Spectrum for Breast Cancer at Any Age," *Cancer Epidemiology Biomarkers & Prevention*, vol. 23, pp. 658–669, Apr. 2014.

[51] E. R. Behr, E. Savio-Galimberti, J. Barc, A. G. Holst, E. Petropoulou, B. P. Prins, J. Jabbari, M. Torchio, M. Berthet, Y. Mizusawa, T. Yang, E. A. Nannenberg, F. Dagradi, P. Weeke, R. Bastiaenan, M. J. Ackerman, S. Haunso, A. Leenhardt, S. Kääb, V. Probst, R. Redon, S. Sharma, A. Wilde, J. Tfelt-Hansen, P. Schwartz, D. M. Roden, C. R. Bezzina, M. Olesen, D. Darbar, P. Guicheney, L. Crotti, U. Consortium, and Y. Jamshidi, "Role of common and rare variants in SCN10a: results from the Brugada syndrome QRS locus gene discovery collaborative study," *Cardiovascular Research*, vol. 106, pp. 520–529, June 2015.

[52] G. de los Campos, D. Gianola, and D. B. Allison, "Predicting genetic predisposition in humans: the promise of whole-genome markers," *Nature Reviews Genetics*, vol. 11, pp. 880–886, Dec. 2010.

[53] Maas P, Barrdahl M, Joshi AD, and et al, "BReast cancer risk from modifiable and nonmodifiable risk factors among white women in the united states," *JAMA Oncology*, May 2016.

[54] E. Maae, R. F. Andersen, K. D. Steffensen, E. H. Jakobsen, I. Brandslund, F. B. Sørensen, and A. Jakobsen, "Prognostic Impact of VEGFA Germline Polymorphisms in Patients with HER2-positive Primary Breast Cancer," *Anticancer Re-*

search, vol. 32, pp. 3619–3627, Sept. 2012.

[55] M. Schmidt, A. Victor, D. Bratzel, D. Boehm, C. Cotarelo, A. Lebrecht, W. Siggelkow, J. G. Hengstler, A. Elsäßer, M. Gehrmann, H.-A. Lehr, H. Koelbl, G. v. Minckwitz, N. Harbeck, and C. Thomssen, "Long-term outcome prediction by clinicopathological risk classification algorithms in node-negative breast cancer—comparison between Adjuvant!, St Gallen, and a novel risk algorithm used in the prospective randomized Node-Negative-Breast Cancer-3 (NNBC-3) trial," *Annals of Oncology*, vol. 20, pp. 258–264, Feb. 2009.

[56] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature Genetics*, vol. 43, pp. 491–498, May 2011.

[57] L. S. Chen, L. Hsu, E. R. Gamazon, N. J. Cox, and D. L. Nicolae, "An Exponential Combination Procedure for Set-Based Association Tests in Sequencing Studies," *The American Journal of Human Genetics*, vol. 91, pp. 977–986, Dec. 2012.

[58] P. C. Sham and S. M. Purcell, "Statistical power and significance testing in large-scale genetic studies," *Nature Reviews. Genetics*, vol. 15, pp. 335–346, May 2014.

[59] "Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants."

[60] Hindorff, LA, MacArther, J, Morales, J, Junkins, HA, Hall, PN, Klemm, AK, and Manolio, TA, "GWAS Catalog."

[61] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 1019, pp. 215–236, 2013.

[62] Y. Zhang, J. Long, W. Lu, X.-O. Shu, Q. Cai, Y. Zheng, C. Li, B. Li, Y.-T. Gao, and W. Zheng, "Rare coding variants and breast cancer risk: evaluation of susceptibility Loci identified in genome-wide association studies," *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, vol. 23, pp. 622–628, Apr. 2014.

[63] E. M. John, J. L. Hopper, J. C. Beck, J. A. Knight, S. L. Neuhausen, R. T. Senie, A. Ziogas, I. L. Andrulis, H. Anton-Culver, N. Boyd, S. S. Buys, M. B. Daly, F. P. O'Malley, R. M. Santella, M. C. Southey, V. L. Venne, D. J. Venter, D. W. West, A. S. Whittemore, D. Seminara, and Breast Cancer Family Registry, "The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary

and translational studies of the genetic epidemiology of breast cancer," *Breast cancer research: BCR*, vol. 6, no. 4, pp. R375–389, 2004.

[64] J. Chang-Claude, N. Eby, M. Kiechle, G. Bastert, and H. Becher, "Breastfeeding and breast cancer risk by age 50 among women in Germany," *Cancer causes & control: CCC*, vol. 11, pp. 687–695, Sept. 2000.

[65] M. D. Gammon, A. I. Neugut, R. M. Santella, S. L. Teitelbaum, J. A. Britton, M. B. Terry, S. M. Eng, M. S. Wolff, S. D. Stellman, G. C. Kabat, B. Levin, H. L. Bradlow, M. Hatch, J. Beyea, D. Camann, M. Trent, R. T. Senie, G. C. Garbowski, C. Maffeo, P. Montalvan, G. S. Berkowitz, M. Kemeny, M. Citron, F. Schnabe, A. Schuss, S. Hajdu, V. Vincguerra, G. W. Collman, and G. I. Obrams, "The Long Island Breast Cancer Study Project: description of a multi-institutional collaboration to identify environmental risk factors for breast cancer," *Breast Cancer Research and Treatment*, vol. 74, pp. 235–254, June 2002.

[66] D. M. Friedrichsen, K. E. Malone, D. R. Doody, J. R. Daling, and E. A. Ostrander, "Frequency of CHEK2 mutations in a population based, case-control study of breast cancer in young women," *Breast cancer research: BCR*, vol. 6, no. 6, pp. R629–635, 2004.

[67] E. Lee, H. Ma, R. McKean-Cowdin, D. V. D. Berg, L. Bernstein, B. E. Henderson, and G. Ursin, "Effect of Reproductive Factors and Oral Contraceptives on Breast Cancer Risk in BRCA1/2 Mutation Carriers and Noncarriers: Results from a Population-Based Study," *Cancer Epidemiology Biomarkers & Prevention*, vol. 17, pp. 3170–3178, Nov. 2008.

[68] G. B. Byrnes, M. C. Southey, and J. L. Hopper, "Are the so-called low penetrance breast cancer genes, ATM, BRIP1, PALB2 and CHEK2, high risk for women with strong family histories?," *Breast Cancer Research : BCR*, vol. 10, no. 3, p. 208, 2008.

[69] Y. Zhang, J. Long, W. Lu, X.-O. Shu, Q. Cai, Y. Zheng, C. Li, B. Li, Y.-T. Gao, and W. Zheng, "Rare coding variants and breast cancer risk: evaluation of susceptibility Loci identified in genome-wide association studies," *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, vol. 23, pp. 622–628, Apr. 2014.

[70] F. Lalloo, J. Varley, A. Moran, D. Ellis, L. O'Dair, P. Pharoah, A. Antoniou, R. Hartley, A. Shenton, S. Seal, B. Bulman, A. Howell, and D. G. R. Evans, "BRCA1, BRCA2 and TP53 mutations in very early-onset breast cancer with associated risks to relatives," *European Journal of Cancer*, vol. 42, pp. 1143–1150, May 2006.

[71] O. M. Sinilnikova, M.-G. Dondon, S. Eon-Marchais, F. Damiola, L. Barjhoux, M. Marcou, C. Verny-Pierre, V. Sornin, L. Toulemonde, J. Beauvallet, D. Le Gal, N. Mebirouk, M. Belotti, O. Caron, M. Gauthier-Villars, I. Coupier, B. Buecher, A. Lortholary, C. Dugast, P. Gesta, J.-P. Fricker, C. Noguès, L. Faivre, E. Luporsi, P. Berthet, C. Delnatte, V. Bonadona, C. M. Maugard, P. Pujol, C. Lasset, M. Longy, Y.-J. Bignon, C. Adenis, L. Venat-Bouvet, L. Demange, H. Dreyfus, M. Frenay, L. Gladieff, I. Mortemousque, S. Audebert-Bellanger, F. Soubrier, S. Giraud, S. Lejeune-Dumoulin, A. Chevrier, J.-M. Limacher, J. Chiesa, A. Fajac, A. Floquet, F. Eisinger, J. Tinat, C. Colas, S. Fert-Ferrer, C. Penet, T. Frebourg, M.-A. Collonge-Rame, E. Barouk-Simonet, V. Layet, D. Leroux, O. Cohen-Haguenauer, F. Prieur, E. Mouret-Fourme, F. Cornélis, P. Jonveaux, O. Bera, E. Cavaciuti, A. Tardivon, F. Lesueur, S. Mazoyer, D. Stoppa-Lyonnet, and N. Andrieu, "GENESIS: a French national resource to study the missing heritability of breast cancer," *BMC Cancer*, vol. 16, p. 13, 2016.

[72] F. S. M. Hilbers, M. P. G. Vreeswijk, C. J. van Asperen, and P. Devilee, "The impact of next generation sequencing on the analysis of breast cancer susceptibility: a role for extremely rare genetic variation?," *Clinical Genetics*, vol. 84, pp. 407–414, Nov. 2013.

[73] F. Lalloo and D. G. Evans, "Familial Breast Cancer," *Clinical Genetics*, vol. 82, pp. 105–114, Aug. 2012.

[74] J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, B. Go, S. Go, and o. b. o. t. N. E. S. Project, "Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes," *Science*, vol. 337, pp. 64–69, July 2012.

[75] B. Li and S. M. Leal, "Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data," *The American Journal of Human Genetics*, vol. 83, pp. 311–321, Sept. 2008.

[76] B. E. Madsen and S. R. Browning, "A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic," *PLOS Genet*, vol. 5, p. e1000384, Feb. 2009.

[77] K. Roeder, B. Devlin, and L. Wasserman, "Improving power in genome-wide association studies: weights tip the scale," *Genetic Epidemiology*, vol. 31, pp. 741–747, Nov. 2007.

[78] P. C. Sham and S. M. Purcell, "Statistical power and significance testing in large-scale genetic studies," *Nat Rev Genet*, vol. 15, pp. 335–346, May 2014.

[79] O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander, "Searching for missing heritability: Designing rare variant association studies," *Proceedings of the National Academy of Sciences*, vol. 111, pp. E455–E464, Jan. 2014.

[80] C. A. Cassa, M. Y. Tong, and D. M. Jordan, "Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals," *Human mutation*, vol. 34, pp. 1216–1220, Sept. 2013.

[81] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nat. Protocols*, vol. 4, pp. 1073–1081, June 2009.

[82] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, pp. 248–249, Apr. 2010.

[83] K. Wang, M. Li, and M. Bucan, "Pathway-Based Approaches for Analysis of Genomewide Association Studies," *American Journal of Human Genetics*, vol. 81, pp. 1278–1283, Dec. 2007.

[84] D. Curtis, A. E. Vine, and J. Knight, "A simple method for assessing the strength of evidence for association at the level of the whole gene," *Advances and applications in bioinformatics and chemistry: AABC*, vol. 1, pp. 115–120, 2008.

[85] H.-C. Yang, Y.-J. Liang, C.-M. Chung, J.-W. Chen, and W.-H. Pan, "Genome-wide gene-based association study," *BMC proceedings*, vol. 3 Suppl 7, p. S135, 2009.

[86] D. V. Zaykin, L. A. Zhivotovsky, P. H. Westfall, and B. S. Weir, "Truncated product method for combining P-values," *Genetic Epidemiology*, vol. 22, pp. 170–185, Feb. 2002.

[87] M.-X. Li, J. S. H. Kwan, and P. C. Sham, "HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis," *American Journal of Human Genetics*, vol. 91, pp. 478–488, Sept. 2012.

[88] J. Z. Liu, A. F. Mcrae, D. R. Nyholt, S. E. Medland, N. R. Wray, K. M. Brown, N. K. Hayward, G. W. Montgomery, P. M. Visscher, N. G. Martin, and S. Macgregor, "A Versatile Gene-Based Test for Genome-wide Association Studies," *The American Journal of Human Genetics*, vol. 87, pp. 139–145, July 2010.

[89] Y. Benjamini and Y. Hochberg, "Multiple Hypotheses Testing with Weights," *Scandinavian Journal of Statistics*, vol. 24, pp. 407–418, Sept. 1997.

[90] R. A. Fisher, *Statistical methods for research workers*. Edinburgh: Oliver and Boyd, 14th ed., revised and enlarged ed., 1970.

[91] E. M. Azzato, P. D. P. Pharoah, P. Harrington, D. F. Easton, D. Greenberg, N. E. Caporaso, S. J. Chanock, R. N. Hoover, G. Thomas, D. J. Hunter, and P. Kraft, "A genome-wide association study of prognosis in breast cancer," *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, vol. 19, pp. 1140–1143, Apr. 2010.

[92] W. Zhou, Y. Jiang, M. Zhu, D. Hang, J. Chen, J. Zhou, J. Dai, H. Ma, Z. Hu, G. Jin, J. Sha, and H. Shen, "Low-frequency nonsynonymous variants in FKBPL and ARPC1b genes are associated with breast cancer risk in Chinese women," *Molecular Carcinogenesis*, pp. n/a–n/a, Aug. 2016.

[93] O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander, "Searching for missing heritability: Designing rare variant association studies," *Proceedings of the National Academy of Sciences*, vol. 111, pp. E455–E464, Jan. 2014.

[94] R. Nanda, L. P. Schumm, S. Cummings, J. D. Fackenthal, L. Sveen, F. Ademuyiwa, M. Cobleigh, L. Esserman, N. M. Lindor, S. L. Neuhausen, and O. I. Olopade, "Genetic testing in an ethnically diverse cohort of high-risk women: a comparative analysis of BRCA1 and BRCA2 mutations in American families of European and African ancestry," *JAMA*, vol. 294, pp. 1925–1933, Oct. 2005.

[95] D. Huo, R. T. Senie, M. Daly, S. S. Buys, S. Cummings, J. Ogutha, K. Hope, and O. I. Olopade, "Prediction of BRCA Mutations Using the BRCAPRO Model in Clinic-Based African American, Hispanic, and Other Minority Families in the United States," *Journal of Clinical Oncology*, vol. 27, pp. 1184–1190, Mar. 2009.

[96] Y. Guo, J. He, S. Zhao, H. Wu, X. Zhong, Q. Sheng, D. C. Samuels, Y. Shyr, and J. Long, "Illumina human exome genotyping array clustering and quality control," *Nature protocols*, vol. 9, pp. 2643–2662, Nov. 2014.

[97] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: a tool for genome-wide complex trait analysis," *American Journal of Human Genetics*, vol. 88, pp. 76–82, Jan. 2011.

[98] G. Alves and Y.-K. Yu, "Accuracy Evaluation of the Unified P-Value from Combining Correlated P-Values," *PLoS ONE*, vol. 9, Mar. 2014.

[99] I. Mathieson and G. McVean, "Differential confounding of rare and common variants in spatially structured populations," *Nature Genetics*, vol. 44, pp. 243–246, Mar. 2012.

[100] E. Génin, S. Letort, and M.-C. Babron, "Population Stratification of Rare Variants," in *Assessing Rare Variation in Complex Traits*, pp. 227–237, Springer, 2015.

[101] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, pp. 904–909, Aug. 2006.

[102] "PLOS Genetics: Population Structure and Eigenanalysis."

[103] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *The American Journal of Human Genetics*, vol. 81, pp. 559–575, Sept. 2007.

[104] S. Purcell, "PLINK v1.07 http://pngu.mgh.harvard.edu/purcell/plink/," 2009.

[105] S. D. Turner, "qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots," *bioRxiv*, p. 005165, May 2014.

[106] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer Science & Business Media, Oct. 2009.

[107] R Core Team, "R: A Language and Environment for Statistical Computing, version3.2.2," 2015.

[108] S. Lee, C. Fuchsberger, S. Kim, and L. Scott, "An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies," *Biostatistics (Oxford, England)*, vol. 17, pp. 1–15, Jan. 2016.

[109] B. M. Kaminski, C. I. Amos, E. DeRycke, E. M. Gillanders, S. B. Gruber, B. E. Henderson, D. J. Hunter, P. K. Lepage, T. A. Sellers, and D. Seminara, "Abstract 78: Genetic Associations and Mechanisms in Oncology (GAME-ON): A network approach to post-GWAS research.," *Cancer Epidemiology Biomarkers & Prevention*, vol. 21, pp. 78–78, Nov. 2012.

[110] G. S. Dite, M. A. Jenkins, M. C. Southey, J. S. Hocking, G. G. Giles, M. R. E. McCredie, D. J. Venter, and J. L. Hopper, "Familial Risks, Early-Onset Breast Cancer, and BRCA1 and BRCA2 Germline Mutations," *Journal of the National Cancer Institute*, vol. 95, pp. 448–457, Mar. 2003.

[111] A. Hofman, M. M. B. Breteler, C. M. Duijn, H. L. A. Janssen, G. P. Krestin, E. J. Kuipers, B. H. C. Stricker, H. Tiemeier, A. G. Uitterlinden, J. R. Vingerling, and J. C. M. Witteman, "The Rotterdam Study: 2010 objectives and design update," *European Journal of Epidemiology*, vol. 24, pp. 553–572, Sept. 2009.

[112] M. Leu, K. Humphreys, I. Surakka, E. Rehnberg, J. Muilu, P. Rosenström, P. Almgren, J. Jääskeläinen, R. P. Lifton, K. O. Kyvik, J. Kaprio, N. L. Pedersen, A. Palotie, P. Hall, H. Grönberg, L. Groop, L. Peltonen, J. Palmgren, and S. Ripatti, "NordicDB: a Nordic pool and portal for genome-wide control data," *European Journal of Human Genetics*, vol. 18, pp. 1322–1326, Dec. 2010.

[113] J. Li, L. Eriksson, K. Humphreys, K. Czene, J. Liu, R. M. Tamimi, S. Lindström, D. J. Hunter, C. M. Vachon, and F. J. Couch, "Genetic variation in the estrogen metabolic pathway and mammographic density as an intermediate phenotype of breast cancer," *Breast Cancer Res*, vol. 12, no. 2, p. R19, 2010.

[114] J. P. A. Ioannidis, "Common Genetic Variants for Breast Cancer: 32 Largely Refuted Candidates and Larger Prospects," *Journal of the National Cancer Institute*, vol. 98, pp. 1350–1353, Oct. 2006.

[115] B. Frank, K. Hemminki, B. Wappenschmidt, A. Meindl, R. Klaes, R. K. Schmutzler, P. Bugert, M. Untch, C. R. Bartram, and B. Burwinkel, "Association of the CASP10 V410i variant with reduced familial breast cancer risk and interaction with the CASP8 D302h variant," *Carcinogenesis*, vol. 27, pp. 606–609, Mar. 2006.

[116] C. Turnbull, S. Ahmed, J. Morrison, D. Pernet, A. Renwick, M. Maranian, S. Seal, M. Ghoussaini, S. Hines, C. S. Healey, D. Hughes, M. Warren-Perry, W. Tapper, D. Eccles, D. G. Evans, Breast Cancer Susceptibility Collaboration (UK), M. Hooning, M. Schutte, A. van den Ouweland, R. Houlston, G. Ross, C. Langford, P. D. P. Pharoah, M. R. Stratton, A. M. Dunning, N. Rahman, and D. F. Easton, "Genome-wide association study identifies five new breast cancer susceptibility loci," *Nature Genetics*, vol. 42, pp. 504–507, June 2010.

[117] D. Flesch-Janys, T. Slanger, E. Mutschelknauss, S. Kropp, N. Obi, E. Vettorazzi, W. Braendle, G. Bastert, S. Hentschel, J. Berger, and J. Chang-Claude, "Risk of different histological types of postmenopausal breast cancer by type and regimen of menopausal hormone therapy," *International Journal of Cancer*, vol. 123, pp. 933–941, Aug. 2008.

[118] M. Ghoussaini, O. Fletcher, K. Michailidou, C. Turnbull, M. K. Schmidt, E. Dicks, J. Dennis, Q. Wang, M. K. Humphreys, C. Luccarini, C. Baynes, D. Conroy, M. Maranian, S. Ahmed, K. Driver, N. Johnson, N. Orr, I. dos Santos Silva, Q. Waisfisz, H. Meijers-Heijboer, A. G. Uitterlinden, F. Rivadeneira, Netherlands Collaborative Group on Hereditary Breast and Ovarian Cancer (hebon), P. Hall, K. Czene, A. Irwanto, J. Liu, H. Nevanlinna, K. Aittomäki, C. Blomqvist, A. Meindl, R. K. Schmutzler, B. Müller-Myhsok, P. Lichtner, J. Chang-Claude, R. Hein, S. Nickels, D. Flesch-Janys, H. Tsimiklis, E. Makalic, D. Schmidt, M. Bui, J. L. Hopper, C. Apicella, D. J. Park, M. Southey, D. J. Hunter, S. J. Chanock, A. Broeks, S. Verhoef, F. B. L. Hogervorst, P. A. Fasching, M. P. Lux, M. W. Beckmann, A. B.

Ekici, E. Sawyer, I. Tomlinson, M. Kerin, F. Marme, A. Schneeweiss, C. Sohn, B. Burwinkel, P. Guénel, T. Truong, E. Cordina-Duverger, F. Menegaux, S. E. Bojesen, B. G. Nordestgaard, S. F. Nielsen, H. Flyger, R. L. Milne, M. R. Alonso, A. González-Neira, J. Benítez, H. Anton-Culver, A. Ziogas, L. Bernstein, C. C. Dur, H. Brenner, H. Müller, V. Arndt, C. Stegmaier, Familial Breast Cancer Study (fbcs), C. Justenhoven, H. Brauch, T. Brüning, The Gene Environment Interaction of Breast Cancer in Germany (GENICA) Network, S. Wang-Gohrke, U. Eilber, T. Dörk, P. Schürmann, M. Bremer, P. Hillemanns, N. V. Bogdanova, N. N. Antonenkova, Y. I. Rogov, J. H. Karstens, M. Bermisheva, D. Prokofieva, E. Khusnutdinova, A. Lindblom, S. Margolin, A. Mannermaa, V. Kataja, V.-M. Kosma, J. M. Hartikainen, D. Lambrechts, B. T. Yesilyurt, G. Floris, K. Leunen, S. Manoukian, B. Bonanni, S. Fortuzzi, P. Peterlongo, F. J. Couch, X. Wang, K. Stevens, A. Lee, G. G. Giles, L. Baglietto, G. Severi, C. McLean, G. G. Alnæs, V. Kristensen, A.-L. Børrensen-Dale, E. M. John, A. Miron, R. Winqvist, K. Pylkäs, A. Jukkola-Vuorinen, S. Kauppila, I. L. Andrulis, G. Glendon, A. M. Mulligan, P. Devilee, C. J. van Asperen, R. A. E. M. Tollenaar, C. Seynaeve, J. D. Figueroa, M. Garcia-Closas, L. Brinton, J. Lissowska, M. J. Hooning, A. Hollestelle, R. A. Oldenburg, A. M. W. van den Ouweland, A. Cox, M. W. R. Reed, M. Shah, A. Jakubowska, J. Lubinski, K. Jaworska, K. Durda, M. Jones, M. Schoemaker, A. Ashworth, A. Swerdlow, J. Beesley, X. Chen, k. Investigators, A. O. C. S. Group, K. R. Muir, A. Lophatananon, S. Rattanamongkongul, A. Chaiwerawattana, D. Kang, K.-Y. Yoo, D.-Y. Noh, C.-Y. Shen, J.-C. Yu, P.-E. Wu, C.-N. Hsiung, A. Perkins, R. Swann, L. Velentzis, D. M. Eccles, W. J. Tapper, S. M. Gerty, N. J. Graham, B. A. J. Ponder, G. Chenevix-Trench, P. D. P. Pharoah, M. Lathrop, A. M. Dunning, N. Rahman, J. Peto, and D. F. Easton, "Genome-wide association analysis identifies three new breast cancer susceptibility loci," *Nature Genetics*, vol. 44, pp. 312–318, Mar. 2012.

[119] K. Michailidou, P. Hall, A. Gonzalez-Neira, M. Ghoussaini, J. Dennis, R. L. Milne, M. K. Schmidt, J. Chang-Claude, S. E. Bojesen, M. K. Bolla, Q. Wang, E. Dicks, A. Lee, C. Turnbull, N. Rahman, The Breast and Ovarian Cancer Susceptibility Collaboration, O. Fletcher, J. Peto, L. Gibson, I. dos Santos Silva, H. Nevanlinna, T. A. Muranen, K. Aittomäki, C. Blomqvist, K. Czene, A. Irwanto, J. Liu, Q. Waisfisz, H. Meijers-Heijboer, M. Adank, Hereditary Breast and Ovarian Cancer Research Group Netherlands (hebon), R. B. van der Luijt, R. Hein, N. Dahmen, L. Beckman, A. Meindl, R. K. Schmutzler, B. Müller-Myhsok, P. Lichtner, J. L. Hopper, M. C. Southey, E. Makalic, D. F. Schmidt, A. G. Uitterlinden, A. Hofman, D. J. Hunter, S. J. Chanock, D. Vincent, F. Bacot, D. C. Tessier, S. Canisius, L. F. A. Wessels, C. A. Haiman, M. Shah, R. Luben, J. Brown, C. Luccarini, N. Schoof, K. Humphreys, J. Li, B. G. Nordestgaard, S. F. Nielsen, H. Flyger, F. J. Couch, X. Wang, C. Vachon, K. N. Stevens, D. Lambrechts, M. Moisse, R. Paridaens, M.-R. Christiaens, A. Rudolph, S. Nickels, D. Flesch-Janys, N. Johnson, Z. Aitken, K. Aaltonen, T. Heikkinen, A. Broeks, L. J. V. Veer, C. E. van der Schoot, P. Guénel,

T. Truong, P. Laurent-Puig, F. Menegaux, F. Marme, A. Schneeweiss, C. Sohn, B. Burwinkel, M. P. Zamora, J. I. A. Perez, G. Pita, M. R. Alonso, A. Cox, I. W. Brock, S. S. Cross, M. W. R. Reed, E. J. Sawyer, I. Tomlinson, M. J. Kerin, N. Miller, B. E. Henderson, F. Schumacher, L. Le Marchand, I. L. Andrulis, J. A. Knight, G. Glendon, A. M. Mulligan, k. Investigators, A. O. C. S. Group, A. Lindblom, S. Margolin, M. J. Hooning, A. Hollestelle, A. M. W. v. d. Ouweland, A. Jager, Q. M. Bui, J. Stone, G. S. Dite, C. Apicella, H. Tsimiklis, G. G. Giles, G. Severi, L. Baglietto, P. A. Fasching, L. Haeberle, A. B. Ekici, M. W. Beckmann, H. Brenner, H. Müller, V. Arndt, C. Stegmaier, A. Swerdlow, A. Ashworth, N. Orr, M. Jones, J. Figueroa, J. Lissowska, L. Brinton, M. S. Goldberg, F. Labrèche, M. Dumont, R. Winqvist, K. Pylkäs, A. Jukkola-Vuorinen, M. Grip, H. Brauch, U. Hamann, T. Brüning, T. G. G. E. I. a. B. C. i. G. Network, P. Radice, P. Peterlongo, S. Manoukian, B. Bonanni, P. Devilee, R. A. E. M. Tollenaar, C. Seynaeve, C. J. v. Asperen, A. Jakubowska, J. Lubinski, K. Jaworska, K. Durda, A. Mannermaa, V. Kataja, V.-M. Kosma, J. M. Hartikainen, N. V. Bogdanova, N. N. Antonenkova, T. Dörk, V. N. Kristensen, H. Anton-Culver, S. Slager, A. E. Toland, S. Edge, F. Fostira, D. Kang, K.-Y. Yoo, D.-Y. Noh, K. Matsuo, H. Ito, H. Iwata, A. Sueta, A. H. Wu, C.-C. Tseng, D. V. D. Berg, D. O. Stram, X.-O. Shu, W. Lu, Y.-T. Gao, H. Cai, S. H. Teo, C. H. Yip, S. Y. Phuah, B. K. Cornes, M. Hartman, H. Miao, W. Y. Lim, J.-H. Sng, K. Muir, A. Lophatananon, S. Stewart-Brown, P. Siriwanarangsan, C.-Y. Shen, C.-N. Hsiung, P.-E. Wu, S.-L. Ding, S. Sangrajrang, V. Gaborieau, P. Brennan, J. McKay, W. J. Blot, L. B. Signorello, Q. Cai, W. Zheng, S. Deming-Halverson, M. Shrubsole, J. Long, J. Simard, M. Garcia-Closas, P. D. P. Pharoah, G. Chenevix-Trench, A. M. Dunning, J. Benitez, and D. F. Easton, "Large-scale genotyping identifies 41 new loci associated with breast cancer risk," *Nature Genetics*, vol. 45, pp. 353–361, Apr. 2013.

[120] A. Siddiq, F. J. Couch, G. K. Chen, S. Lindström, D. Eccles, R. C. Millikan, K. Michailidou, D. O. Stram, L. Beckmann, S. K. Rhie, C. B. Ambrosone, K. Aittomäki, P. Amiano, C. Apicella, A. B. C. T. B. Investigators, L. Baglietto, E. V. Bandera, M. W. Beckmann, C. D. Berg, L. Bernstein, C. Blomqvist, H. Brauch, L. Brinton, Q. M. Bui, J. E. Buring, S. S. Buys, D. Campa, J. E. Carpenter, D. I. Chasman, J. Chang-Claude, C. Chen, F. Clavel-Chapelon, A. Cox, S. S. Cross, K. Czene, S. L. Deming, R. B. Diasio, W. R. Diver, A. M. Dunning, L. Durcan, A. B. Ekici, P. A. Fasching, F. B. C. Study, H. S. Feigelson, L. Fejerman, J. D. Figueroa, O. Fletcher, D. Flesch-Janys, M. M. Gaudet, T. G. Consortium, S. M. Gerty, J. L. Rodriguez-Gil, G. G. Giles, C. H. v. Gils, A. K. Godwin, N. Graham, D. Greco, P. Hall, S. E. Hankinson, A. Hartmann, R. Hein, J. Heinz, R. N. Hoover, J. L. Hopper, J. J. Hu, S. Huntsman, S. A. Ingles, A. Irwanto, C. Isaacs, K. B. Jacobs, E. M. John, C. Justenhoven, R. Kaaks, L. N. Kolonel, G. A. Coetzee, M. Lathrop, L. L. Marchand, A. M. Lee, I.-M. Lee, T. Lesnick, P. Lichtner, J. Liu, E. Lund, E. Makalic, N. G. Martin, C. A. McLean, H. Meijers-Heijboer, A. Meindl, P. Miron, K. R. Monroe, G. W.

Montgomery, B. Müller-Myhsok, S. Nickels, S. J. Nyante, C. Olswold, K. Overvad, D. Palli, D. J. Park, J. R. Palmer, H. Pathak, J. Peto, P. Pharoah, N. Rahman, F. Rivadeneira, D. F. Schmidt, R. K. Schmutzler, S. Slager, M. C. Southey, K. N. Stevens, H.-P. Sinn, M. F. Press, E. Ross, E. Riboli, P. M. Ridker, F. R. Schumacher, G. Severi, I. d. S. Silva, J. Stone, M. Sund, W. J. Tapper, M. J. Thun, R. C. Travis, C. Turnbull, A. G. Uitterlinden, Q. Waisfisz, X. Wang, Z. Wang, J. Weaver, R. Schulz-Wendtland, L. R. Wilkens, D. V. D. Berg, W. Zheng, R. G. Ziegler, E. Ziv, H. Nevanlinna, D. F. Easton, D. J. Hunter, B. E. Henderson, S. J. Chanock, M. Garcia-Closas, P. Kraft, C. A. Haiman, and C. M. Vachon, "A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11," *Human Molecular Genetics*, vol. 21, pp. 5373–5384, Dec. 2012.

[121] M. Garcia-Closas, F. J. Couch, S. Lindstrom, K. Michailidou, M. K. Schmidt, M. N. Brook, N. Orr, S. K. Rhie, E. Riboli, H. S. Feigelson, L. Le Marchand, J. E. Buring, D. Eccles, P. Miron, P. A. Fasching, H. Brauch, J. Chang-Claude, J. Carpenter, A. K. Godwin, H. Nevanlinna, G. G. Giles, A. Cox, J. L. Hopper, M. K. Bolla, Q. Wang, J. Dennis, E. Dicks, W. J. Howat, N. Schoof, S. E. Bojesen, D. Lambrechts, A. Broeks, I. L. Andrulis, P. Guénel, B. Burwinkel, E. J. Sawyer, A. Hollestelle, O. Fletcher, R. Winqvist, H. Brenner, A. Mannermaa, U. Hamann, A. Meindl, A. Lindblom, W. Zheng, P. Devillee, M. S. Goldberg, J. Lubinski, V. Kristensen, A. Swerdlow, H. Anton-Culver, T. Dörk, K. Muir, K. Matsuo, A. H. Wu, P. Radice, S. H. Teo, X.-O. Shu, W. Blot, D. Kang, M. Hartman, S. Sangrajrang, C.-Y. Shen, M. C. Southey, D. J. Park, F. Hammet, J. Stone, L. J. V. Veer, E. J. Rutgers, A. Lophatananon, S. Stewart-Brown, P. Siriwanarangsan, J. Peto, M. G. Schrauder, A. B. Ekici, M. W. Beckmann, I. dos Santos Silva, N. Johnson, H. Warren, I. Tomlinson, M. J. Kerin, N. Miller, F. Marme, A. Schneeweiss, C. Sohn, T. Truong, P. Laurent-Puig, P. Kerbrat, B. G. Nordestgaard, S. F. Nielsen, H. Flyger, R. L. Milne, J. I. A. Perez, P. Menéndez, H. Müller, V. Arndt, C. Stegmaier, P. Lichtner, M. Lochmann, C. Justenhoven, Y.-D. Ko, The Gene ENvironmental Interaction and breast CAncer (GENICA) Network, T. A. Muranen, K. Aittomäki, C. Blomqvist, D. Greco, T. Heikkinen, H. Ito, H. Iwata, Y. Yatabe, N. N. Antonenkova, S. Margolin, V. Kataja, V.-M. Kosma, J. M. Hartikainen, R. Balleine, kConFab Investigators, C.-C. Tseng, D. V. D. Berg, D. O. Stram, P. Neven, A.-S. Dieudonné, K. Leunen, A. Rudolph, S. Nickels, D. Flesch-Janys, P. Peterlongo, B. Peissel, L. Bernard, J. E. Olson, X. Wang, K. Stevens, G. Severi, L. Baglietto, C. McLean, G. A. Coetzee, Y. Feng, B. E. Henderson, F. Schumacher, N. V. Bogdanova, F. Labrèche, M. Dumont, C. H. Yip, N. A. M. Taib, C.-Y. Cheng, M. Shrubsole, J. Long, K. Pylkäs, A. Jukkola-Vuorinen, S. Kauppila, J. A. Knight, G. Glendon, A. M. Mulligan, R. A. E. M. Tollenaar, C. M. Seynaeve, M. Kriege, M. J. Hooning, A. M. W. van den Ouweland, C. H. M. van Deurzen, W. Lu, Y.-T. Gao, H. Cai, S. P. Balasubramanian, S. S. Cross, M. W. R. Reed, L. Signorello, Q. Cai, M. Shah, H. Miao, C. W. Chan, K. S. Chia, A. Jakubowska, K. Jaworska, K. Durda, C.-N. Hsiung, P.-E. Wu,

J.-C. Yu, A. Ashworth, M. Jones, D. C. Tessier, A. González-Neira, G. Pita, M. R. Alonso, D. Vincent, F. Bacot, C. B. Ambrosone, E. V. Bandera, E. M. John, G. K. Chen, J. J. Hu, J. L. Rodriguez-Gil, L. Bernstein, M. F. Press, R. G. Ziegler, R. M. Millikan, S. L. Deming-Halverson, S. Nyante, S. A. Ingles, Q. Waisfisz, H. Tsimiklis, E. Makalic, D. Schmidt, M. Bui, L. Gibson, B. Müller-Myhsok, R. K. Schmutzler, R. Hein, N. Dahmen, L. Beckmann, K. Aaltonen, K. Czene, A. Irwanto, J. Liu, C. Turnbull, Familial Breast Cancer Study (fbcs), N. Rahman, H. Meijers-Heijboer, A. G. Uitterlinden, F. Rivadeneira, Australian Breast Cancer Tissue Bank (ABCTB) Investigators, C. Olswold, S. Slager, R. Pilarski, F. Ademuyiwa, I. Konstantopoulou, N. G. Martin, G. W. Montgomery, D. J. Slamon, C. Rauh, M. P. Lux, S. M. Jud, T. Bruning, J. Weaver, P. Sharma, H. Pathak, W. Tapper, S. Gerty, L. Durcan, D. Trichopoulos, R. Tumino, P. H. Peeters, R. Kaaks, D. Campa, F. Canzian, E. Weiderpass, M. Johansson, K.-T. Khaw, R. Travis, F. Clavel-Chapelon, L. N. Kolonel, C. Chen, A. Beck, S. E. Hankinson, C. D. Berg, R. N. Hoover, J. Lissowska, J. D. Figueroa, D. I. Chasman, M. M. Gaudet, W. R. Diver, W. C. Willett, D. J. Hunter, J. Simard, J. Benitez, A. M. Dunning, M. E. Sherman, G. Chenevix-Trench, S. J. Chanock, P. Hall, P. D. P. Pharoah, C. Vachon, D. F. Easton, C. A. Haiman, and P. Kraft, "Genome-wide association studies identify four ER negative-specific breast cancer risk loci," *Nature Genetics*, vol. 45, pp. 392–398, Apr. 2013.

[122] G. Palomba, A. Loi, E. Porcu, A. Cossu, I. Zara, M. Budroni, M. Dei, S. Lai, A. Mulas, N. Olmeo, M. T. Ionta, F. Atzori, G. Cuccuru, M. Pitzalis, M. Zoledziewska, N. Olla, M. Lovicu, M. Pisano, G. R. Abecasis, M. Uda, F. Tanda, K. Michailidou, D. F. Easton, S. J. Chanock, R. N. Hoover, D. J. Hunter, D. Schlessinger, S. Sanna, L. Crisponi, and G. Palmieri, "Genome-wide association study of susceptibility loci for breast cancer in Sardinian population," *BMC Cancer*, vol. 15, p. 383, 2015.

[123] A. Voorman, J. Brody, and T. Lumley, "skatMeta: Efficient meta analysis for the SKAT test," June 2013.

[124] "A global reference for human genetic variation : Nature : Nature Publishing Group."

[125] C. Dufour, H. Guenou, K. Kaabeche, D. Bouvard, A. Sanjay, and P. J. Marie, "FGFR2-Cbl interaction in lipid rafts triggers attenuation of PI3k/Akt signaling and osteoblast survival," *Bone*, vol. 42, pp. 1032–1039, June 2008.

[126] L. S. Moniz and V. Stambolic, "Nek10 mediates G2/M cell cycle arrest and MEK autoactivation in response to UV irradiation," *Molecular and Cellular Biology*, vol. 31, pp. 30–42, Jan. 2011.

[127] M. A. Scharenberg, R. Chiquet-Ehrismann, and M. B. Asparuhova, "Megakaryoblastic leukemia protein-1 (MKL1): Increasing evidence for an involvement in cancer progression and metastasis," *The International Journal of Biochemistry &*

Cell Biology, vol. 42, pp. 1911–1914, Dec. 2010.

[128] "Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants."

[129] C. R. King and D. L. Nicolae, "GWAS to Sequencing: Divergence in Study Design and Analysis," *Genes*, vol. 5, pp. 460–476, May 2014.

[130] L. A. Carey, C. M. Perou, C. A. Livasy, L. G. Dressler, D. Cowan, K. Conway, G. Karaca, M. A. Troester, C. K. Tse, S. Edmiston, S. L. Deming, J. Geradts, M. C. U. Cheang, T. O. Nielsen, P. G. Moorman, H. S. Earp, and R. C. Millikan, "Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study," *JAMA*, vol. 295, pp. 2492–2502, June 2006.

[131] M. P. Nilsson, L. Hartman, I. Idvall, U. Kristoffersson, O. T. Johannsson, and N. Loman, "Long-term prognosis of early-onset breast cancer in a population-based cohort with a known BRCA1/2 mutation status," *Breast Cancer Research and Treatment*, vol. 144, pp. 133–142, Feb. 2014.

[132] The 1000 Genomes Project Consortium, "A global reference for human genetic variation," *Nature*, vol. 526, pp. 68–74, Oct. 2015.

[133] I. Jatoi and W. F. Anderson, "Qualitative age interactions in breast cancer studies: a mini-review," *Future Oncology*, vol. 6, pp. 1781–1788, Nov. 2010.

[134] S. Kobayashi, H. Sugiura, Y. Ando, N. Shiraki, T. Yanagi, H. Yamashita, and T. Toyama, "Reproductive history and breast cancer risk," *Breast Cancer (Tokyo, Japan)*, vol. 19, pp. 302–308, Oct. 2012.

[135] P. Kumar and R. Aggarwal, "An overview of triple-negative breast cancer," *Archives of Gynecology and Obstetrics*, vol. 293, pp. 247–269, Feb. 2016.

[136] L. Tao, S. L. Gomez, T. H. M. Keegan, A. W. Kurian, and C. A. Clarke, "Breast Cancer Mortality in African-American and Non-Hispanic White Women by Molecular Subtype and Stage at Diagnosis: A Population-Based Study," *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, vol. 24, pp. 1039–1045, July 2015.

[137] I. Menashe, W. F. Anderson, I. Jatoi, and P. S. Rosenberg, "Underlying causes of the black-white racial disparity in breast cancer mortality: a population-based analysis," *Journal of the National Cancer Institute*, vol. 101, pp. 993–1000, July 2009.

[138] J. Kollias, C. W. Elston, I. O. Ellis, J. F. Robertson, and R. W. Blamey, "Early-onset breast cancer–histopathological and prognostic considerations.," *British Journal of*

Cancer, vol. 75, no. 9, pp. 1318–1323, 1997.

[139] A. M. Gonzalez-Angulo, K. Broglio, S.-W. Kau, Y. Eralp, J. Erlichman, V. Valero, R. Theriault, D. Booser, A. U. Buzdar, G. N. Hortobagyi, and B. Arun, "Women age < or = 35 years with primary breast carcinoma: disease features at presentation," *Cancer*, vol. 103, pp. 2466–2472, June 2005.

[140] Q. Hu, T. Luo, X. Zhong, P. He, T. Tian, and H. Zheng, "Application status of tamoxifen in endocrine therapy for early breast cancer," *Experimental and Therapeutic Medicine*, vol. 9, pp. 2207–2212, June 2015.

[141] H. B. Muss, "Factors Used to Select Adjuvant Therapy of Breast Cancer in the United States: an Overview of Age, Race, and Socioeconomic Status," *JNCI Monographs*, vol. 2001, pp. 52–55, Dec. 2001.

[142] M. Jamshidi, M. K. Schmidt, T. Dörk, M. Garcia-Closas, T. Heikkinen, S. Cornelissen, A. J. van den Broek, P. Schürmann, A. Meyer, T.-W. Park-Simon, J. Figueroa, M. Sherman, J. Lissowska, G. T. H. Keong, A. Irwanto, M. Laakso, S. Hautaniemi, K. Aittomäki, C. Blomqvist, J. Liu, and H. Nevanlinna, "Germline variation in TP53 regulatory network genes associates with breast cancer survival and treatment outcome," *International Journal of Cancer*, vol. 132, pp. 2044–2055, May 2013.

[143] Y.-Z. Jiang, K.-D. Yu, W.-T. Peng, G.-H. Di, J. Wu, G.-Y. Liu, and Z.-M. Shao, "Enriched variations in TEKT4 and breast cancer resistance to paclitaxel," *Nature Communications*, vol. 5, p. 3802, 2014.

[144] S.-Y. Lee, S.-A. Im, Y. H. Park, S. Y. Woo, S. Kim, M. K. Choi, W. Chang, J. S. Ahn, and Y.-H. Im, "Genetic polymorphisms of SLC28a3, SLC29a1 and RRM1 predict clinical outcome in patients with metastatic breast cancer receiving gemcitabine plus paclitaxel chemotherapy," *European Journal of Cancer (Oxford, England: 1990)*, vol. 50, pp. 698–705, Mar. 2014.

[145] K. Kiyotani, T. Mushiroda, T. Tsunoda, T. Morizono, N. Hosono, M. Kubo, Y. Tanigawara, C. K. Imamura, D. A. Flockhart, F. Aki, K. Hirata, Y. Takatsuka, M. Okazaki, S. Ohsumi, T. Yamakawa, M. Sasa, Y. Nakamura, and H. Zembutsu, "A genome-wide association study identifies locus at 10q22 associated with clinical outcomes of adjuvant tamoxifen therapy for breast cancer patients in Japanese," *Human Molecular Genetics*, vol. 21, pp. 1665–1672, Apr. 2012.

[146] M. Pande, M. L. Bondy, K.-A. Do, A. A. Sahin, J. Ying, G. B. Mills, P. A. Thompson, and A. M. Brewster, "Association between germline single nucleotide polymorphisms in the PI3k-AKT-mTOR pathway, obesity, and breast cancer disease-free survival," *Breast Cancer Research and Treatment*, vol. 147, pp. 381–387, Sept. 2014.

[147] T. Winder, G. Giamas, P. M. Wilson, W. Zhang, D. Yang, P. Bohanes, Y. Ning, A. Gerger, J. Stebbing, and H.-J. Lenz, "Insulin-like growth factor receptor polymorphism defines clinical outcome in estrogen receptor-positive breast cancer patients treated with tamoxifen," *The Pharmacogenomics Journal*, vol. 14, pp. 28–34, Feb. 2014.

[148] R. Pei, P. Wang, Y. Zhou, J. Zhang, T. Ouyang, B. Li, and Y. Xie, "Association of BRCA1 K1183r polymorphism with survival in BRCA1/2-negative chinese familial breast cancer," *Clinical Laboratory*, vol. 60, no. 1, pp. 47–53, 2014.

[149] J. J. Dorairaj, D. W. Salzman, D. Wall, T. Rounds, C. Preskill, C. A. W. Sullivan, R. Lindner, C. Curran, K. Lezon-Geyda, T. McVeigh, L. Harris, J. Newell, M. J. Kerin, M. Wood, N. Miller, and J. B. Weidhaas, "A germline mutation in the BRCA1 3'UTR predicts Stage IV breast cancer," *BMC cancer*, vol. 14, p. 421, 2014.

[150] L. Bai, H. H. Yang, Y. Hu, A. Shukla, N.-H. Ha, A. Doran, F. Faraji, N. Goldberger, M. P. Lee, T. Keane, and K. W. Hunter, "An Integrated Genome-Wide Systems Genetics Screen for Breast Cancer Metastasis Susceptibility Genes," *PLOS Genet*, vol. 12, p. e1005989, Apr. 2016.

[151] N.-H. Ha, J. Long, Q. Cai, X. O. Shu, and K. W. Hunter, "The Circadian Rhythm Gene Arntl2 Is a Metastasis Susceptibility Gene for Estrogen Receptor-Negative Breast Cancer," *PLoS genetics*, vol. 12, p. e1006267, Sept. 2016.

[152] L. S. Lindström, P. Hall, M. Hartman, F. Wiklund, H. Grönberg, and K. Czene, "Familial concordance in cancer survival: a Swedish population-based study," *The Lancet. Oncology*, vol. 8, pp. 1001–1006, Nov. 2007.

[153] M. Hartman, L. Lindström, P. W. Dickman, H.-O. Adami, P. Hall, and K. Czene, "Is breast cancer prognosis inherited?," *Breast cancer research: BCR*, vol. 9, no. 3, p. R39, 2007.

[154] E. M. Azzato, K. E. Driver, F. Lesueur, M. Shah, D. Greenberg, D. F. Easton, A. E. Teschendorff, C. Caldas, N. E. Caporaso, and P. D. Pharoah, "Effects of common germline genetic variation in cell cycle control genes on breast cancer survival: results from a population-based cohort," *Breast Cancer Research*, vol. 10, p. R47, 2008.

[155] P. A. Fasching, C. R. Loehberg, P. L. Strissel, M. P. Lux, M. R. Bani, M. Schrauder, S. Geiler, K. Ringleff, S. Oeser, S. Weihbrecht, R. Schulz-Wendtland, A. Hartmann, M. W. Beckmann, and R. Strick, "Single nucleotide polymorphisms of the aromatase gene (CYP19a1), HER2/neu status, and prognosis in breast cancer patients," *Breast Cancer Research and Treatment*, vol. 112, pp. 89–98, Nov. 2007.

[156] E. L. Goode, A. M. Dunning, B. Kuschel, C. S. Healey, N. E. Day, B. A. J. Ponder, D. F. Easton, and P. P. D. Pharoah, "Effect of Germ-Line Genetic Variation on Breast Cancer Survival in a Population-based Study," *Cancer Research*, vol. 62, pp. 3052–3057, June 2002.

[157] M. S. Udler, E. M. Azzato, C. S. Healey, S. Ahmed, K. A. Pooley, D. Greenberg, M. Shah, A. E. Teschendorff, C. Caldas, A. M. Dunning, E. A. Ostrander, N. E. Caporaso, D. Easton, and P. D. Pharoah, "Common germline polymorphisms in COMT, CYP19a1, ESR1, PGR, SULT1e1 and STS and survival after a diagnosis of breast cancer," *International Journal of Cancer*, vol. 125, pp. 2687–2696, Dec. 2009.

[158] S. Hughes, O. Agbaje, R. L. Bowen, D. L. Holliday, J. A. Shaw, S. Duffy, and J. L. Jones, "Matrix Metalloproteinase Single-Nucleotide Polymorphisms and Haplotypes Predict Breast Cancer Progression," *Clinical Cancer Research*, vol. 13, pp. 6673–6680, Nov. 2007.

[159] R. Fagerholm, B. Hofstetter, J. Tommiska, K. Aaltonen, R. Vrtel, K. Syrjäkoski, A. Kallioniemi, O. Kilpivaara, A. Mannermaa, V.-M. Kosma, M. Uusitupa, M. Eskelinen, V. Kataja, K. Aittomäki, K. von Smitten, P. Heikkilä, J. Lukas, K. Holli, J. Bartkova, C. Blomqvist, J. Bartek, and H. Nevanlinna, "NAD(P)H:quinone oxidoreductase 1 NQO1*2 genotype (P187s) is a strong prognostic and predictive factor in breast cancer," *Nature Genetics*, vol. 40, pp. 844–853, July 2008.

[160] B. Kaabi, G. Belaaloui, W. Benbrahim, K. Hamizi, M. Sadelaoud, W. Toumi, and H. Bounecer, "ADRA2a Germline Gene Polymorphism is Associated to the Severity, but not to the Risk, of Breast Cancer," *Pathology oncology research: POR*, vol. 22, pp. 357–365, Apr. 2016.

[161] P. Seibold, P. Schmezer, S. Behrens, K. Michailidou, M. K. Bolla, Q. Wang, D. Flesch-Janys, H. Nevanlinna, R. Fagerholm, K. Aittomäki, C. Blomqvist, S. Margolin, A. Mannermaa, V. Kataja, V.-M. Kosma, J. M. Hartikainen, D. Lambrechts, H. Wildiers, V. Kristensen, G. G. Alnæs, S. Nord, A.-L. Borresen-Dale, M. J. Hooning, A. Hollestelle, A. Jager, C. Seynaeve, J. Li, J. Liu, K. Humphreys, A. M. Dunning, V. Rhenius, M. Shah, M. Kabisch, D. Torres, H.-U. Ulmer, U. Hamann, J. M. Schildkraut, K. S. Purrington, F. J. Couch, P. Hall, P. Pharoah, D. F. Easton, M. K. Schmidt, J. Chang-Claude, and O. Popanda, "A polymorphism in the base excision repair gene PARP2 is associated with differential prognosis by chemotherapy among postmenopausal breast cancer patients," *BMC cancer*, vol. 15, p. 978, 2015.

[162] Y.-M. Jia, Y.-T. Xie, Y.-J. Wang, J.-Y. Han, X.-X. Tian, and W.-G. Fang, "Association of Genetic Polymorphisms in CDH1 and CTNNB1 with Breast Cancer Susceptibility and Patients' Prognosis among Chinese Han Women," *PloS One*, vol. 10, no. 8, p. e0135865, 2015.

[163] Y. Li, Y.-L. Chen, Y.-T. Xie, L.-Y. Zheng, J.-Y. Han, H. Wang, X.-X. Tian, and W.-G. Fang, "Association Study of Germline Variants in CCNB1 and CDK1 with Breast Cancer Susceptibility, Progression, and Survival among Chinese Han Women," *PLOS ONE*, vol. 8, p. e84489, Dec. 2013.

[164] R. Ugenskienė, D. Myrzaliyeva, R. Jankauskaitė, J. Gedminaitė, R. Jančiauskienė, E. Šepetauskienė, and E. Juozaitytė, "The contribution of SIPA1 and RRP1b germline polymorphisms to breast cancer phenotype, lymph node status and survival in a group of Lithuanian young breast cancer patients," *Biomarkers: Biochemical Indicators of Exposure, Response, and Susceptibility to Chemicals*, vol. 21, pp. 363–370, June 2016.

[165] Y. Sapkota, S. Ghosh, R. Lai, B. P. Coe, C. E. Cass, Y. Yasui, J. R. Mackey, and S. Damaraju, "Germline DNA Copy Number Aberrations Identified as Potential Prognostic Factors for Breast Cancer Recurrence," *PLOS ONE*, vol. 8, p. e53850, Jan. 2013.

[166] S. A. Narod, "Early-onset breast cancer: what do we know about the risk factors?," *Current Oncology*, vol. 18, no. 5, pp. 204–205, 2011.

[167] D. R. Brenner, N. T. Brockton, J. Kotsopoulos, M. Cotterchio, B. A. Boucher, K. S. Courneya, J. A. Knight, I. A. Olivotto, M. L. Quan, and C. M. Friedenreich, "Breast cancer survival among young women: a review of the role of modifiable lifestyle factors," *Cancer Causes & Control*, vol. 27, pp. 459–472, Mar. 2016.

[168] H. Brauch and M. Schwab, "Prediction of tamoxifen outcome by genetic variation of CYP2d6 in post-menopausal women with early breast cancer," *British Journal of Clinical Pharmacology*, vol. 77, pp. 695–703, Apr. 2014.

[169] L. Binkhorst, R. H. J. Mathijssen, A. Jager, and T. van Gelder, "Individualization of tamoxifen therapy: much more than just CYP2d6 genotyping," *Cancer Treatment Reviews*, vol. 41, pp. 289–299, Mar. 2015.

[170] X. O. Shu, J. Long, W. Lu, C. Li, W. Y. Chen, R. Delahanty, J. Cheng, H. Cai, Y. Zheng, J. Shi, K. Gu, W.-J. Wang, P. Kraft, Y.-T. Gao, Q. Cai, and W. Zheng, "Novel Genetic Markers of Breast Cancer Survival Identified by a Genome-Wide Association Study," *Cancer Research*, vol. 72, pp. 1182–1189, Mar. 2012.

[171] S. Rafiq, W. Tapper, A. Collins, S. Khan, I. Politopoulos, S. Gerty, C. Blomqvist, F. J. Couch, H. Nevanlinna, J. Liu, and D. Eccles, "Identification of inherited genetic variations influencing prognosis in early-onset breast cancer," *Cancer Research*, vol. 73, pp. 1883–1891, Mar. 2013.

[172] S. Rafiq, S. Khan, W. Tapper, A. Collins, R. Upstill-Goddard, S. Gerty, C. Blomqvist, K. Aittomäki, F. J. Couch, J. Liu, H. Nevanlinna, and D. Eccles,

"A Genome Wide Meta-Analysis Study for Identification of Common Variation Associated with Breast Cancer Prognosis," *PLOS ONE*, vol. 9, p. e101488, Dec. 2014.

[173] Q. Guo, M. K. Schmidt, P. Kraft, S. Canisius, C. Chen, S. Khan, J. Tyrer, M. K. Bolla, Q. Wang, J. Dennis, K. Michailidou, M. Lush, S. Kar, J. Beesley, A. M. Dunning, M. Shah, K. Czene, H. Darabi, M. Eriksson, D. Lambrechts, C. Weltens, K. Leunen, S. E. Bojesen, B. G. Nordestgaard, S. F. Nielsen, H. Flyger, J. Chang-Claude, A. Rudolph, P. Seibold, D. Flesch-Janys, C. Blomqvist, K. Aittomäki, R. Fagerholm, T. A. Muranen, F. J. Couch, J. E. Olson, C. Vachon, I. L. Andrulis, J. A. Knight, G. Glendon, A. M. Mulligan, A. Broeks, F. B. Hogervorst, C. A. Haiman, B. E. Henderson, F. Schumacher, L. L. Marchand, J. L. Hopper, H. Tsimiklis, C. Apicella, M. C. Southey, A. Cox, S. S. Cross, M. W. R. Reed, G. G. Giles, R. L. Milne, C. McLean, R. Winqvist, K. Pylkäs, A. Jukkola-Vuorinen, M. Grip, M. J. Hooning, A. Hollestelle, J. W. M. Martens, A. M. W. v. d. Ouweland, F. Marme, A. Schneeweiss, R. Yang, B. Burwinkel, J. Figueroa, S. J. Chanock, J. Lissowska, E. J. Sawyer, I. Tomlinson, M. J. Kerin, N. Miller, H. Brenner, A. K. Dieffenbach, V. Arndt, B. Holleczek, A. Mannermaa, V. Kataja, V.-M. Kosma, J. M. Hartikainen, J. Li, J. S. Brand, K. Humphreys, P. Devilee, R. A. E. M. Tollenaar, C. Seynaeve, P. Radice, P. Peterlongo, B. Bonanni, P. Mariani, P. A. Fasching, M. W. Beckmann, A. Hein, A. B. Ekici, G. Chenevix-Trench, R. Balleine, k. Investigators, K.-A. Phillips, J. Benitez, M. P. Zamora, J. I. A. Perez, P. Menéndez, A. Jakubowska, J. Lubinski, K. Jaworska-Bieniek, K. Durda, U. Hamann, M. Kabisch, H. U. Ulmer, T. Rüdiger, S. Margolin, V. Kristensen, S. Nord, D. G. Evans, J. E. Abraham, H. M. Earl, L. Hiller, J. A. Dunn, S. Bowden, C. Berg, D. Campa, W. R. Diver, S. M. Gapstur, M. M. Gaudet, S. E. Hankinson, R. N. Hoover, A. Hüsing, R. Kaaks, M. J. Machiela, W. Willett, M. Barrdahl, F. Canzian, S.-F. Chin, C. Caldas, D. J. Hunter, S. Lindstrom, M. García-Closas, P. Hall, D. F. Easton, D. M. Eccles, N. Rahman, H. Nevanlinna, and P. D. P. Pharoah, "Identification of Novel Genetic Markers of Breast Cancer Survival," *Journal of the National Cancer Institute*, vol. 107, p. djv081, May 2015.

[174] S. Khan, R. Fagerholm, S. Rafiq, W. Tapper, K. Aittomäki, J. Liu, C. Blomqvist, D. Eccles, and H. Nevanlinna, "Polymorphism at 19q13.41 Predicts Breast Cancer Survival Specifically after Endocrine Therapy," *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, vol. 21, pp. 4086–4096, Sept. 2015.

[175] N. Song, J.-Y. Choi, H. Sung, S. Jeon, S. Chung, S. K. Park, W. Han, J. W. Lee, M. K. Kim, J.-Y. Lee, K.-Y. Yoo, B.-G. Han, S.-H. Ahn, D.-Y. Noh, and D. Kang, "Prediction of breast cancer survival using clinical and genetic markers by tumor subtypes," *PloS One*, vol. 10, no. 4, p. e0122413, 2015.

[176] D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, and N. J. Cox, "Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS," *PLOS Genet*, vol. 6, p. e1000888, Apr. 2010.

[177] A. J. Schork, W. K. Thompson, P. Pham, A. Torkamani, J. C. Roddey, P. F. Sullivan, J. R. Kelsoe, M. C. O'Donovan, H. Furberg, T. T. a. G. Consortium, T. B. D. P. G. Consortium, T. S. P. G. Consortium, N. J. Schork, O. A. Andreassen, and A. M. Dale, "All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs," *PLOS Genet*, vol. 9, p. e1003449, Apr. 2013.

[178] H. Chen, T. Lumley, J. Brody, N. L. Heard-Costa, C. S. Fox, L. A. Cupples, and J. Dupuis, "Sequence kernel association test for survival traits," *Genetic Epidemiology*, vol. 38, pp. 191–197, Apr. 2014.

[179] Z.-Z. Tang and D.-Y. Lin, "Meta-analysis for Discovering Rare-Variant Associations: Statistical Methods and Software Programs," *The American Journal of Human Genetics*, vol. 97, pp. 35–53, July 2015.

[180] T. E. P. Consortium, "The ENCODE (ENCyclopedia Of DNA Elements) Project," *Science*, vol. 306, pp. 636–640, Oct. 2004.

[181] E. Pennisi, "ENCODE Project Writes Eulogy for Junk DNA," *Science*, vol. 337, pp. 1159–1161, Sept. 2012.

[182] P. A. Fasching, P. D. P. Pharoah, A. Cox, H. Nevanlinna, S. E. Bojesen, T. Karn, A. Broeks, F. E. van Leeuwen, L. J. van't Veer, R. Udo, A. M. Dunning, D. Greco, K. Aittomäki, C. Blomqvist, M. Shah, B. G. Nordestgaard, H. Flyger, J. L. Hopper, M. C. Southey, C. Apicella, M. Garcia-Closas, M. Sherman, J. Lissowska, C. Seynaeve, P. E. A. Huijts, R. A. E. M. Tollenaar, A. Ziogas, A. B. Ekici, C. Rauh, A. Mannermaa, V. Kataja, V.-M. Kosma, J. M. Hartikainen, I. L. Andrulis, H. Ozcelik, A.-M. Mulligan, G. Glendon, P. Hall, K. Czene, J. Liu, J. Chang-Claude, S. Wang-Gohrke, U. Eilber, S. Nickels, T. Dörk, M. Schiekel, M. Bremer, T.-W. Park-Simon, G. G. Giles, G. Severi, L. Baglietto, M. J. Hooning, J. W. M. Martens, A. Jager, M. Kriege, A. Lindblom, S. Margolin, F. J. Couch, K. N. Stevens, J. E. Olson, M. Kosel, S. S. Cross, S. P. Balasubramanian, M. W. R. Reed, A. Miron, E. M. John, R. Winqvist, K. Pylkäs, A. Jukkola-Vuorinen, S. Kauppila, B. Burwinkel, F. Marme, A. Schneeweiss, C. Sohn, G. Chenevix-Trench, kConFab Investigators, D. Lambrechts, A.-S. Dieudonne, S. Hatse, E. van Limbergen, J. Benitez, R. L. Milne, M. P. Zamora, J. I. A. Pérez, B. Bonanni, B. Peissel, B. Loris, P. Peterlongo, P. Rajaraman, S. J. Schonfeld, H. Anton-Culver, P. Devilee, M. W. Beckmann, D. J. Slamon, K.-A. Phillips, J. D. Figueroa, M. K. Humphreys, D. F. Easton, and M. K. Schmidt, "The role of genetic breast cancer susceptibility variants as prognostic factors," *Human Molecular Genetics*, vol. 21, pp. 3926–3939, Sept. 2012.

[183] M. Baum, J. A. Dossett, J. S. Patterson, F. G. Smiddy, A. Wilson, D. Richards, D. M. Brinkley, K. Mcpherson, R. D. Rubens, B. A. Stoll, J. C. Lea, and S. H. Ellis, "Improved Survival Amongst Patients Treated with Adjuvant Tamoxifen After Masectomy for Early Breast Cancer," *The Lancet*, vol. 322, p. 450, Aug. 1983.

[184] H. J. Burstein, S. Temin, H. Anderson, T. A. Buchholz, N. E. Davidson, K. E. Gelmon, S. H. Giordano, C. A. Hudis, D. Rowden, A. J. Solky, V. Stearns, E. P. Winer, and J. J. Griggs, "Adjuvant Endocrine Therapy for Women With Hormone Receptor–Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update," *Journal of Clinical Oncology*, p. JCO.2013.54.2258, May 2014.

[185] "Broad Institute Birdsuite," *http://archive.broadinstitute.org/mpg/birdsuite/birdseed.html*.

[186] "Michigan Imputation Server," *https://imputationserver.sph.umich.edu/index.html*, Aug. 2016.

[187] O. Delaneau, J. Marchini, and J.-F. Zagury, "A linear complexity phasing method for thousands of genomes," *Nature Methods*, vol. 9, pp. 179–181, Feb. 2012.

[188] C. Fuchsberger, G. R. Abecasis, and D. A. Hinds, "minimac2: faster genotype imputation," *Bioinformatics*, vol. 31, pp. 782–784, Mar. 2015.

[189] A. R. Wood, J. R. B. Perry, T. Tanaka, D. G. Hernandez, H.-F. Zheng, D. Melzer, J. R. Gibbs, M. A. Nalls, M. N. Weedon, T. D. Spector, J. B. Richards, S. Bandinelli, L. Ferrucci, A. B. Singleton, and T. M. Frayling, "Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation," *PloS One*, vol. 8, no. 5, p. e64343, 2013.

[190] J. Yang, A. Bakshi, Z. Zhu, G. Hemani, A. A. E. Vinkhuyzen, S. H. Lee, M. R. Robinson, J. R. B. Perry, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, The LifeLines Cohort Study, T. Esko, L. Milani, R. Mägi, A. Metspalu, A. Hamsten, P. K. E. Magnusson, N. L. Pedersen, E. Ingelsson, N. Soranzo, M. C. Keller, N. R. Wray, M. E. Goddard, and P. M. Visscher, "Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index," *Nature Genetics*, vol. 47, pp. 1114–1120, Oct. 2015.

[191] P. I. W. d. Bakker, M. A. R. Ferreira, X. Jia, B. M. Neale, S. Raychaudhuri, and B. F. Voight, "Practical aspects of imputation-driven meta-analysis of genome-wide association studies," *Human Molecular Genetics*, vol. 17, pp. R122–R128, Oct. 2008.

[192] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes," *Genetic Epidemiology*, vol. 34, pp. 816–834, Dec. 2010.

[193] Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. van Duijn, "GenABEL: an R library for genome-wide association analysis," *Bioinformatics (Oxford, England)*, vol. 23, pp. 1294–1296, May 2007.

[194] L. C. Karssen, C. M. van Duijn, and Y. S. Aulchenko, "The GenABEL Project for statistical genomics," *F1000Research*, vol. 5, p. 914, May 2016.

[195] O. Fletcher, N. Johnson, N. Orr, F. J. Hosking, L. J. Gibson, K. Walker, D. Zelenika, I. Gut, S. Heath, C. Palles, B. Coupland, P. Broderick, M. Schoemaker, M. Jones, J. Williamson, S. Chilcott-Burns, K. Tomczyk, G. Simpson, K. B. Jacobs, S. J. Chanock, D. J. Hunter, I. P. Tomlinson, A. Swerdlow, A. Ashworth, G. Ross, I. dos Santos Silva, M. Lathrop, R. S. Houlston, and J. Peto, "Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study," *Journal of the National Cancer Institute*, vol. 103, pp. 425–435, Mar. 2011.

[196] M. M. Regan, B. Leyland-Jones, M. Bouzyk, O. Pagani, W. Tang, R. Kammler, P. Dell'Orto, M. O. Biasi, B. Thürlimann, M. B. Lyng, H. J. Ditzel, P. Neven, M. Debled, R. Maibach, K. N. Price, R. D. Gelber, A. S. Coates, A. Goldhirsch, J. M. Rae, and G. Viale, "CYP2d6 Genotype and Tamoxifen Response in Postmenopausal Women with Endocrine-Responsive Breast Cancer: The Breast International Group 1-98 Trial," *Journal of the National Cancer Institute*, vol. 104, pp. 441–451, Mar. 2012.

[197] M. F. Yazdi, S. Rafieian, M. Gholi-Nataj, M. H. Sheikhha, T. Nazari, and H. Neamatzadeh, "CYP2d6 Genotype and Risk of Recurrence in Tamoxifen Treated Breast Cancer Patients," *Asian Pacific journal of cancer prevention: APJCP*, vol. 16, no. 15, pp. 6783–6787, 2015.

[198] V.-J. Bardou, G. Arpino, R. M. Elledge, C. K. Osborne, and G. M. Clark, "Progesterone Receptor Status Significantly Improves Outcome Prediction Over Estrogen Receptor Status Alone for Adjuvant Endocrine Therapy in Two Large Breast Cancer Databases," *Journal of Clinical Oncology*, vol. 21, pp. 1973–1979, May 2003.

[199] C. J. Bradley, C. W. Given, and C. Roberts, "Race, socioeconomic status, and breast cancer treatment and survival," *Journal of the National Cancer Institute*, vol. 94, pp. 490–496, Apr. 2002.

[200] K. N. Anderson, R. B. Schwab, and M. E. Martinez, "Reproductive risk factors and breast cancer subtypes: a review of the literature," *Breast Cancer Research and Treatment*, vol. 144, pp. 1–10, Feb. 2014.

[201] P. D. Terry and T. E. Rohan, "Cigarette Smoking and the Risk of Breast Cancer in Women A Review of the Literature," *Cancer Epidemiology Biomarkers & Prevention*, vol. 11, pp. 953–971, Oct. 2002.

[202] L. Swanson, "On the predictive accuracy of transport to Iles de la Madeline.".

[203] I. Soerjomataram, M. W. J. Louwman, J. G. Ribot, J. A. Roukema, and J. W. W. Coebergh, "An overview of prognostic factors for long-term survivors of breast cancer," *Breast Cancer Research and Treatment*, vol. 107, pp. 309–330, Feb. 2008.

[204] P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, and K. Hemminki, "Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland," *The New England Journal of Medicine*, vol. 343, pp. 78–85, July 2000.

[205] X. Zhong, Z. Dong, H. Dong, J. Li, Z. Peng, L. Deng, X. Zhu, Y. Sun, X. Lu, F. Shen, X. Su, L. Zhang, Y. Gu, and H. Zheng, "Prevalence and Prognostic Role of BRCA1/2 Variants in Unselected Chinese Breast Cancer Patients," *PloS One*, vol. 11, no. 6, p. e0156789, 2016.

[206] L. Bordeleau, S. Panchal, and P. Goodwin, "Prognosis of BRCA-associated breast cancer: a summary of evidence," *Breast Cancer Research and Treatment*, vol. 119, p. 13, Sept. 2009.

[207] P. J. Goodwin, K.-A. Phillips, D. W. West, M. Ennis, J. L. Hopper, E. M. John, F. P. O'Malley, R. L. Milne, I. L. Andrulis, M. L. Friedlander, M. C. Southey, C. Apicella, G. G. Giles, and T. A. Longacre, "Breast cancer prognosis in BRCA1 and BRCA2 mutation carriers: an International Prospective Breast Cancer Family Registry population-based cohort study," *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 30, pp. 19–26, Jan. 2012.

[208] J. Lei, A. Rudolph, K. B. Moysich, S. Rafiq, S. Behrens, E. L. Goode, P. P. D. Pharoah, P. Seibold, P. A. Fasching, I. L. Andrulis, V. N. Kristensen, F. J. Couch, U. Hamann, M. J. Hooning, H. Nevanlinna, U. Eilber, M. K. Bolla, J. Dennis, Q. Wang, A. Lindblom, A. Mannermaa, D. Lambrechts, M. García-Closas, P. Hall, G. Chenevix-Trench, M. Shah, R. Luben, L. Haeberle, A. B. Ekici, M. W. Beckmann, J. A. Knight, G. Glendon, S. Tchatchou, G. I. G. Alnæs, A.-L. Borresen-Dale, S. Nord, J. E. Olson, E. Hallberg, C. Vachon, D. Torres, H.-U. Ulmer, T. Rüdiger, A. Jager, C. H. M. van Deurzen, M. M. A. Tilanus-Linthorst, T. A. Muranen, K. Aittomäki, C. Blomqvist, S. Margolin, V.-M. Kosma, J. M. Hartikainen, V. Kataja, S. Hatse, H. Wildiers, A. Smeets, J. Figueroa, S. J. Chanock, J. Lissowska, J. Li, K. Humphreys, K.-A. Phillips, kConFab Investigators, S. Linn, S. Cornelissen, S. A. J. van den Broek, D. Kang, J.-Y. Choi, S. K. Park, K.-Y. Yoo, C.-N. Hsiung, P.-E. Wu, M.-F. Hou, C.-Y. Shen, S. H. Teo, N. A. M. Taib, C. H. Yip, G. F. Ho, K. Matsuo, H. Ito, H. Iwata, K. Tajima, A. M. Dunning, J. Benitez, K. Czene, L. E. Sucheston, T. Maishman, W. J. Tapper, D. Eccles, D. F. Easton, M. K. Schmidt, and J. Chang-Claude, "Assessment of variation in immunosuppressive pathway genes reveals TGFBR2 to be associated with prognosis of estrogen receptor-negative breast

cancer after chemotherapy," *Breast cancer research: BCR*, vol. 17, p. 18, 2015.

[209] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, pp. 301–320, Apr. 2005.

[210] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[211] International Schizophrenia Consortium, S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, and P. Sklar, "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder," *Nature*, vol. 460, pp. 748–752, Aug. 2009.

[212] A. V. Rubanovich and N. N. Khromov-Borisov, "Genetic risk assessment of the joint effect of several genes: Critical appraisal," *Russian Journal of Genetics*, vol. 52, no. 7, pp. 757–769, 2016.

[213] C. C. A. Spencer, Z. Su, P. Donnelly, and J. Marchini, "Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip," *PLOS Genet*, vol. 5, p. e1000477, May 2009.

[214] S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, P. Sklar, S. M. P. (Leader), D. M. Ruderfer, A. McQuillin, D. W. Morris, C. T. O'Dushlaine, A. Corvin, P. A. Holmans, M. C. O'Donovan, S. Macgregor, H. Gurling, D. H. R. Blackwood, N. J. Craddock, M. Gill, C. M. Hultman, G. K. Kirov, P. Lichtenstein, W. J. Muir, M. J. Owen, C. N. Pato, E. M. Scolnick, D. S. Clair, P. S. (Leader), N. M. Williams, L. Georgieva, I. Nikolov, N. Norton, H. Williams, D. Toncheva, V. Milanova, E. F. Thelander, P. Sullivan, C. T. O'Dushlaine, E. Kenny, E. M. Quinn, K. Choudhury, S. Datta, J. Pimm, S. Thirumalai, V. Puri, R. Krasucki, J. Lawrence, D. Quested, N. Bass, C. Crombie, G. Fraser, S. L. Kuan, N. Walker, K. A. McGhee, B. Pickard, P. Malloy, A. W. Maclean, M. V. Beck, M. T. Pato, H. Medeiros, F. Middleton, C. Carvalho, C. Morley, A. Fanous, D. Conti, J. A. Knowles, C. P. Ferreira, A. Macedo, M. H. Azevedo, A. N. Kirby, M. A. R. Ferreira, M. J. Daly, K. Chambert, F. Kuruvilla, S. B. Gabriel, K. Ardlie, and J. L. Moran, "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder," *Nature*, vol. 460, pp. 748–752, Aug. 2009.

[215] J. S. Witte and T. J. Hoffmann, "Polygenic Modeling of Genome-Wide Association Studies: An Application to Prostate and Breast Cancer," *OMICS: A Journal of Integrative Biology*, vol. 15, pp. 393–398, Feb. 2011.

[216] S. Bayraktar, P. A. Thompson, S.-Y. Yoo, K.-a. Do, A. A. Sahin, B. K. Arun, M. L. Bondy, and A. M. Brewster, "The relationship between eight GWAS-identified

single-nucleotide polymorphisms and primary breast cancer outcomes," *The Oncologist*, vol. 18, no. 5, pp. 493–500, 2013.

[217] F. Dudbridge, "Power and Predictive Accuracy of Polygenic Risk Scores," *PLOS Genet*, vol. 9, p. e1003348, Mar. 2013.

[218] G. Bhatia, A. Gusev, P.-R. Loh, H. K. Finucane, B. J. Vilhjalmsson, S. Ripke, S. W. G. o. t. P. G. Cons, S. Purcell, E. Stahl, M. Daly, T. R. d. Candia, S. H. Lee, B. M. Neale, M. C. Keller, N. A. Zaitlen, B. Pasaniuc, N. Patterson, J. Yang, and A. L. Price, "Subtle stratification confounds estimates of heritability from rare variants," *bioRxiv*, p. 048181, Apr. 2016.

[219] S. K. Kumar, M. W. Feldman, D. H. Rehkopf, and S. Tuljapurkar, "Response to Commentary on "Limitations of GCTA as a solution to the missing heritability problem"," *bioRxiv*, Feb. 2016.

[220] S. K. Kumar, M. W. Feldman, D. H. Rehkopf, and S. Tuljapurkar, "Limitations of GCTA as a solution to the missing heritability problem," *Proceedings of the National Academy of Sciences*, vol. 113, pp. E61–E70, Jan. 2016.

[221] J. Yang, S. H. Lee, N. R. Wray, M. E. Goddard, and P. M. Visscher, "GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs," *Proceedings of the National Academy of Sciences*, vol. 113, pp. E4579–E4580, Aug. 2016.

[222] E. R. Gamazon and D. S. Park, "SNP-based heritability estimation: measurement noise, population stratification and stability," *bioRxiv*, Feb. 2016.

[223] N. A. C. Cressie, *Statistics for spatial data*. Wiley series in probability and mathematical statistics, New York: Wiley, rev. ed ed., 1993.

[224] R. Makowsky, N. M. Pajewski, Y. C. Klimentidis, A. I. Vazquez, C. W. Duarte, D. B. Allison, and G. d. l. Campos, "Beyond Missing Heritability: Prediction of Complex Traits," *PLOS Genet*, vol. 7, p. e1002051, Apr. 2011.

[225] X. Zhou, P. Carbonetto, and M. Stephens, "Polygenic Modeling with Bayesian Sparse Linear Mixed Models," *PLoS Genet*, vol. 9, p. e1003264, Feb. 2013.

[226] S. A. Gagliano, A. D. Paterson, M. E. Weale, and J. Knight, "Assessing models for genetic prediction of complex traits: a comparison of visualization and quantitative methods," *BMC Genomics*, vol. 16, p. 405, 2015.

[227] H.-F. Zheng, J.-J. Rong, M. Liu, F. Han, X.-W. Zhang, J. B. Richards, and L. Wang, "Performance of Genotype Imputation for Low Frequency and Rare Variants from the 1000 Genomes," *PLOS ONE*, vol. 10, p. e0116487, Jan. 2015.

[228] T.-J. Park, L. Heo, S. Moon, Y. J. Kim, J. H. Oh, S. Han, and B.-J. Kim, "Practical Calling Approach for Exome Array-Based Genome-Wide Association Studies in Korean Population," *International Journal of Genomics*, vol. 2015, 2015.

[229] T. J. Hoffmann and J. S. Witte, "Strategies for Imputing and Analyzing Rare Variants in Association Studies," *Trends in Genetics*, vol. 31, pp. 556–563, Oct. 2015.

[230] E. Kreiner-Møller, C. Medina-Gomez, A. G. Uitterlinden, F. Rivadeneira, and K. Estrada, "Improving accuracy of rare variant imputation with a two-step imputation approach," *European Journal of Human Genetics*, vol. 23, pp. 395–400, Mar. 2015.

[231] L. Li, Y. Li, S. R. Browning, B. L. Browning, A. J. Slater, X. Kong, J. L. Aponte, V. E. Mooser, S. L. Chissoe, J. C. Whittaker, M. R. Nelson, and M. G. Ehm, "Performance of Genotype Imputation for Rare Variants Identified in Exons and Flanking Regions of Genes," *PLOS ONE*, vol. 6, p. e24945, Sept. 2011.

[232] A. D. Johnson, R. E. Handsaker, S. L. Pulit, M. M. Nizzari, C. J. O'Donnell, and P. I. W. d. Bakker, "SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap," *Bioinformatics*, vol. 24, pp. 2938–2939, Dec. 2008.

[233] S. Borra and A. Di Ciaccio, "Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods," *Computational Statistics & Data Analysis*, vol. 54, pp. 2976–2989, Dec. 2010.

[234] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics*, vol. 21, pp. 3301–3307, Aug. 2005.

[235] D. Speed, G. Hemani, M. R. Johnson, and D. J. Balding, "Response to Lee et al.: SNP-Based Heritability Analysis with Dense Data," *American Journal of Human Genetics*, vol. 93, pp. 1155–1157, Dec. 2013.

[236] N. Mavaddat, P. D. P. Pharoah, K. Michailidou, J. Tyrer, M. N. Brook, M. K. Bolla, Q. Wang, J. Dennis, A. M. Dunning, M. Shah, R. Luben, J. Brown, S. E. Bojesen, B. G. Nordestgaard, S. F. Nielsen, H. Flyger, K. Czene, H. Darabi, M. Eriksson, J. Peto, I. dos Santos-Silva, F. Dudbridge, N. Johnson, M. K. Schmidt, A. Broeks, S. Verhoef, E. J. Rutgers, A. Swerdlow, A. Ashworth, N. Orr, M. J. Schoemaker, J. Figueroa, S. J. Chanock, L. Brinton, J. Lissowska, F. J. Couch, J. E. Olson, C. Vachon, V. S. Pankratz, D. Lambrechts, H. Wildiers, C. V. Ongeval, E. v. Limbergen, V. Kristensen, G. G. Alnæs, S. Nord, A.-L. Borresen-Dale, H. Nevanlinna, T. A. Muranen, K. Aittomäki, C. Blomqvist, J. Chang-Claude, A. Rudolph, P. Seibold, D. Flesch-Janys, P. A. Fasching, L. Haeberle, A. B. Ekici, M. W. Beckmann, B. Burwinkel, F. Marme, A. Schneeweiss, C. Sohn, A. Trentham-Dietz, P. Newcomb, L. Titus, K. M. Egan, D. J. Hunter, S. Lindstrom, R. M. Tamimi, P. Kraft, N. Rahman, C. Turnbull, A. Renwick, S. Seal, J. Li, J. Liu, K. Humphreys, J. Benitez, M. P.

Zamora, J. I. A. Perez, P. Menéndez, A. Jakubowska, J. Lubinski, K. Jaworska-Bieniek, K. Durda, N. V. Bogdanova, N. N. Antonenkova, T. Dörk, H. Anton-Culver, S. L. Neuhausen, A. Ziogas, L. Bernstein, P. Devilee, R. A. E. M. Tollenaar, C. Seynaeve, C. J. v. Asperen, A. Cox, S. S. Cross, M. W. R. Reed, E. Khusnutdinova, M. Bermisheva, D. Prokofyeva, Z. Takhirova, A. Meindl, R. K. Schmutzler, C. Sutter, R. Yang, P. Schürmann, M. Bremer, H. Christiansen, T.-W. Park-Simon, P. Hillemanns, P. Guénel, T. Truong, F. Menegaux, M. Sanchez, P. Radice, P. Peterlongo, S. Manoukian, V. Pensotti, J. L. Hopper, H. Tsimiklis, C. Apicella, M. C. Southey, H. Brauch, T. Brüning, Y.-D. Ko, A. J. Sigurdson, M. M. Doody, U. Hamann, D. Torres, H.-U. Ulmer, A. Försti, E. J. Sawyer, I. Tomlinson, M. J. Kerin, N. Miller, I. L. Andrulis, J. A. Knight, G. Glendon, A. M. Mulligan, G. Chenevix-Trench, R. Balleine, G. G. Giles, R. L. Milne, C. McLean, A. Lindblom, S. Margolin, C. A. Haiman, B. E. Henderson, F. Schumacher, L. L. Marchand, U. Eilber, S. Wang-Gohrke, M. J. Hooning, A. Hollestelle, A. M. W. v. d. Ouweland, L. B. Koppert, J. Carpenter, C. Clarke, R. Scott, A. Mannermaa, V. Kataja, V.-M. Kosma, J. M. Hartikainen, H. Brenner, V. Arndt, C. Stegmaier, A. K. Dieffenbach, R. Winqvist, K. Pylkäs, A. Jukkola-Vuorinen, M. Grip, K. Offit, J. Vijai, M. Robson, R. Rau-Murthy, M. Dwek, R. Swann, K. A. Perkins, M. S. Goldberg, F. Labrèche, M. Dumont, D. M. Eccles, W. J. Tapper, S. Rafiq, E. M. John, A. S. Whittemore, S. Slager, D. Yannoukakos, A. E. Toland, S. Yao, W. Zheng, S. L. Halverson, A. González-Neira, G. Pita, M. R. Alonso, N. Álvarez, D. Herrero, D. C. Tessier, D. Vincent, F. Bacot, C. Luccarini, C. Baynes, S. Ahmed, M. Maranian, C. S. Healey, J. Simard, P. Hall, D. F. Easton, and M. Garcia-Closas, "Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants," *Journal of the National Cancer Institute*, vol. 107, p. djv036, May 2015.

[237] Y. Shieh, D. Hu, L. Ma, S. Huntsman, C. C. Gard, J. W. T. Leung, J. A. Tice, C. M. Vachon, S. R. Cummings, K. Kerlikowske, and E. Ziv, "Breast cancer risk prediction using a clinical risk model and polygenic risk score," *Breast Cancer Research and Treatment*, vol. 159, pp. 513–525, Aug. 2016.

[238] "GCTA-GREML Power Calculator," *http://cnsgenomics.com/shiny/gctaPower/*, 2016.

[239] R. Fagerholm, M. K. Schmidt, S. Khan, S. Rafiq, W. Tapper, K. Aittomäki, D. Greco, T. Heikkinen, T. A. Muranen, P. A. Fasching, W. Janni, R. Weinshilboum, C. R. Loehberg, J. L. Hopper, M. C. Southey, R. Keeman, A. Lindblom, S. Margolin, A. Mannermaa, V. Kataja, G. Chenevix-Trench, kConFab Investigators, D. Lambrechts, H. Wildiers, J. Chang-Claude, P. Seibold, F. J. Couch, J. E. Olson, I. L. Andrulis, J. A. Knight, M. García-Closas, J. Figueroa, M. J. Hooning, A. Jager, M. Shah, B. J. Perkins, R. Luben, U. Hamann, M. Kabisch, K. Czene, P. Hall, D. F. Easton, P. D. P. Pharoah, J. Liu, D. Eccles, C. Blomqvist, and H. Nevanlinna, "The SNP rs6500843 in 16p13.3 is associated with survival specifically among chemotherapy-

treated breast cancer patients," *Oncotarget*, vol. 6, pp. 7390–7407, Apr. 2015.

[240] A. Pirie, Q. Guo, P. Kraft, S. Canisius, D. M. Eccles, N. Rahman, H. Nevanlinna, C. Chen, S. Khan, J. Tyrer, M. K. Bolla, Q. Wang, J. Dennis, K. Michailidou, M. Lush, A. M. Dunning, M. Shah, K. Czene, H. Darabi, M. Eriksson, D. Lambrechts, C. Weltens, K. Leunen, C. van Ongeval, B. G. Nordestgaard, S. F. Nielsen, H. Flyger, A. Rudolph, P. Seibold, D. Flesch-Janys, C. Blomqvist, K. Aittomäki, R. Fagerholm, T. A. Muranen, J. E. Olsen, E. Hallberg, C. Vachon, J. A. Knight, G. Glendon, A. M. Mulligan, A. Broeks, S. Cornelissen, C. A. Haiman, B. E. Henderson, F. Schumacher, L. Le Marchand, J. L. Hopper, H. Tsimiklis, C. Apicella, M. C. Southey, S. S. Cross, M. W. Reed, G. G. Giles, R. L. Milne, C. McLean, R. Winqvist, K. Pylkäs, A. Jukkola-Vuorinen, M. Grip, M. J. Hooning, A. Hollestelle, J. W. Martens, A. M. van den Ouweland, F. Marme, A. Schneeweiss, R. Yang, B. Burwinkel, J. Figueroa, S. J. Chanock, J. Lissowska, E. J. Sawyer, I. Tomlinson, M. J. Kerin, N. Miller, H. Brenner, K. Butterbach, B. Holleczek, V. Kataja, V.-M. Kosma, J. M. Hartikainen, J. Li, J. S. Brand, K. Humphreys, P. Devilee, R. A. Tollenaar, C. Seynaeve, P. Radice, P. Peterlongo, S. Manoukian, F. Ficarazzi, M. W. Beckmann, A. Hein, A. B. Ekici, R. Balleine, K.-A. Phillips, kConFab Investigators, J. Benitez, M. P. Zamora, J. I. A. Perez, P. Menéndez, A. Jakubowska, J. Lubinski, J. Gronwald, K. Durda, U. Hamann, M. Kabisch, H. U. Ulmer, T. Rüdiger, S. Margolin, V. Kristensen, S. Nord, NBCS Investigators, D. G. Evans, J. Abraham, H. Earl, C. J. Poole, L. Hiller, J. A. Dunn, S. Bowden, R. Yang, D. Campa, W. R. Diver, S. M. Gapstur, M. M. Gaudet, S. Hankinson, R. N. Hoover, A. Hüsing, R. Kaaks, M. J. Machiela, W. Willett, M. Barrdahl, F. Canzian, S.-F. Chin, C. Caldas, D. J. Hunter, S. Lindstrom, M. Garcia-Closas, F. J. Couch, G. Chenevix-Trench, A. Mannermaa, I. L. Andrulis, P. Hall, J. Chang-Claude, D. F. Easton, S. E. Bojesen, A. Cox, P. A. Fasching, P. D. Pharoah, and M. K. Schmidt, "Common germline polymorphisms associated with breast cancer-specific survival," *Breast cancer research: BCR*, vol. 17, p. 58, 2015.

[241] L. Fejerman, G. K. Chen, C. Eng, S. Huntsman, D. Hu, A. Williams, B. Pasaniuc, E. M. John, M. Via, C. Gignoux, S. Ingles, K. R. Monroe, L. N. Kolonel, G. Torres-Mejía, E. J. Pérez-Stable, E. G. Burchard, B. E. Henderson, C. A. Haiman, and E. Ziv, "Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas," *Human Molecular Genetics*, vol. 21, pp. 1907–1917, Apr. 2012.

[242] Y. Feng, D. O. Stram, S. K. Rhie, R. C. Millikan, C. B. Ambrosone, E. M. John, L. Bernstein, W. Zheng, A. F. Olshan, J. J. Hu, R. G. Ziegler, S. Nyante, E. V. Bandera, S. A. Ingles, M. F. Press, S. L. Deming, J. L. Rodriguez-Gil, J. R. Palmer, O. I. Olopade, D. Huo, C. A. Adebamowo, T. Ogundiran, G. K. Chen, A. Stram, K. Park, K. A. Rand, S. J. Chanock, L. Le Marchand, L. N. Kolonel, D. V. Conti, D. Easton, B. E. Henderson, and C. A. Haiman, "A comprehensive examination of breast cancer risk loci in African American women," *Human Molecular Genetics*,

vol. 23, pp. 5518–5526, Oct. 2014.

[243] W. Zheng, B. Zhang, Q. Cai, H. Sung, K. Michailidou, J. Shi, J.-Y. Choi, J. Long, J. Dennis, M. K. Humphreys, Q. Wang, W. Lu, Y.-T. Gao, C. Li, H. Cai, S. K. Park, K.-Y. Yoo, D.-Y. Noh, W. Han, A. M. Dunning, J. Benitez, D. Vincent, F. Bacot, D. Tessier, S.-W. Kim, M. H. Lee, J. W. Lee, J.-Y. Lee, Y.-B. Xiang, Y. Zheng, W. Wang, B.-T. Ji, K. Matsuo, H. Ito, H. Iwata, H. Tanaka, A. H. Wu, C.-c. Tseng, D. Van Den Berg, D. O. Stram, S. H. Teo, C. H. Yip, I. N. Kang, T. Y. Wong, C.-Y. Shen, J.-C. Yu, C.-S. Huang, M.-F. Hou, M. Hartman, H. Miao, S. C. Lee, T. C. Putti, K. Muir, A. Lophatananon, S. Stewart-Brown, P. Siriwanarangsan, S. Sangrajrang, H. Shen, K. Chen, P.-E. Wu, Z. Ren, C. A. Haiman, A. Sueta, M. K. Kim, U. S. Khoo, M. Iwasaki, P. D. P. Pharoah, W. Wen, P. Hall, X.-O. Shu, D. F. Easton, and D. Kang, "Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls," *Human Molecular Genetics*, vol. 22, pp. 2539–2550, June 2013.