A STUDY OF FAMILIAL BREAST CANCER:
IDENTIFYING ADDITIONAL BREAST
CANCER SUSCEPTIBILITY LOCI

by

Kristina Lisa Allen-Brady

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Medical Informatics
The University of Utah
May 2006
To the Graduate Council of the University of Utah:

I have read the dissertation of Kristina Lisa Allen-Brady in its final form and have found that (1) its format, citations, and bibliographic style are consistent and acceptable; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the supervisory committee and is ready for submission to The Graduate School.

Date

Approved for the Major Department

Joyce Mitchell
Chair/Dean

Approved for the Graduate Council

David S. Dean of The Graduate School
ABSTRACT

Breast cancer is a serious public health concern and despite intensive research, the etiology of breast cancer is poorly understood. Known risk factors explain only a small proportion of breast cancer, and yet familial aggregation, recent segregation analyses, and studies of low penetrance genes suggest that additional breast cancer susceptibility loci exist and perhaps act in a polygenic fashion. Here three strategies and their fulfillment, to aid in identification of additional breast cancer predisposition loci, are described. The first strategy, identification of a new breast cancer phenotype, was proposed to identify a homogenous subgroup of cases that could be clustered into a single gene or a minimal number of genes disorder. We show that the histological subtype lobular breast cancer, identified through a familial aggregation study, increases risk of breast cancer and specifically lobular breast cancer in lobular cancer families compared to risk associated with any type of breast cancer. Hence, study of lobular breast cancer families may be useful in locating additional breast cancer genes. Second, novel association methods that are adaptable to large pedigree resources, qualitative or quantitative data types, and the ability to perform a variety of statistical tests, correcting for correlations between related individuals, was proposed. This strategy was fulfilled with the development of the pedigree-based association tool, PedGenie. Lastly, strategies to improve the reproducibility of results and study design including use of tagging single nucleotide polymorphisms (tSNPs) to identify underlying haplotype and linkage disequilibrium
structure, and control of potential population stratification were proposed. These concepts were incorporated into a study of five candidate genes involved in the DNA repair pathway. Linkage disequilibrium structure and identification of tSNPs for each gene are reported. For the gene XRCC4, we show that two 4-locus haplotypes were significantly associated with age at diagnosis of breast cancer (one earlier diagnosis and one later diagnosis) and that two 2-locus haplotypes were significantly associated with breast cancer risk. Each proposed strategy and its fulfillment in the Ph.D. dissertation aid in a small measure to the identification of additional breast cancer loci, which ultimately will benefit women worldwide.
This dissertation is dedicated to Ryan Brady, a loving husband who has supported me in every way possible over the last four years to accomplish this Ph.D. I also dedicate this work to my grandmother, Margit Sponbeck, and to a committee member, Dr. Lisa Cannon-Albright, for their courage and determination in facing their breast cancer diagnoses.
# TABLE OF CONTENTS

ABSTRACT ........................................................................................................ iv  
LIST OF TABLES ............................................................................................. x  
ACKNOWLEDGEMENTS ................................................................................ xii  
1. INTRODUCTION ...................................................................................... 1  
   High-penetrance breast cancer susceptibility genes ........................................ 3  
   Low-penetrance breast cancer susceptibility genes and a polygenic model ....... 5  
   Segregation analysis suggests the existence of breast cancer susceptibility loci .. 7  
   Use of linkage analysis to identify breast cancer susceptibility loci .......... 9  
   Strategies to identify additional breast cancer susceptibility genes .......... 10  
   A new, informative phenotype: lobular breast cancer ................................. 18  
   A novel association package: PedGenie ..................................................... 18  
   Improving the reproducibility of results and study design ......................... 19  
 References .................................................................................................... 22  
2. LOBULAR BREAST CANCER: EXCESS FAMILIALITY OBSERVED IN THE UTAH POPULATION DATABASE ............................................. 31  
   Material and methods ................................................................................ 33  
      Utah Cancer Registry ................................................................. 33  
      Histology ................................................................................. 33  
      Statistical analysis ................................................................. 34  
   Results .................................................................................................... 34  
   Discussion ............................................................................................. 36  
   Conclusion ............................................................................................ 39  
 References .................................................................................................... 39  
3. A CAUTIONARY NOTE ON THE APPROPRIATENESS OF USING A LINKAGE RESOURCE FOR AN ASSOCIATION STUDY ................................. 42  
   Abstract ................................................................................................. 43  
   Background ............................................................................................ 43  
   Methods ................................................................................................ 44  
      Genotype data .............................................................................. 44  
      Phenotype data ............................................................................... 44
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>Demographics of females with breast cancer in the Utah Cancer Registry (UCR) and the Utah Population Database (UPDB) with at least three generations of genealogy data available</td>
<td>35</td>
</tr>
<tr>
<td>II.</td>
<td>Risk of breast cancer among relatives of any breast cancer cases in the UPDB</td>
<td>35</td>
</tr>
<tr>
<td>III.</td>
<td>Risk of breast cancer among relatives of lobular breast cancer cases</td>
<td>37</td>
</tr>
<tr>
<td>IV.</td>
<td>Risk of breast cancer among relatives of in situ and invasive lobular breast cancer cases</td>
<td>37</td>
</tr>
<tr>
<td>V.</td>
<td>Genealogical index of familiality analysis for breast cancer</td>
<td>37</td>
</tr>
<tr>
<td>1.</td>
<td>Linkage disequilibrium across 50 replicates</td>
<td>46</td>
</tr>
<tr>
<td>2.</td>
<td>Power: association findings for the region containing gene b22</td>
<td>46</td>
</tr>
<tr>
<td>3.</td>
<td>False Positives: association findings for regions containing genes b14 and b18</td>
<td>46</td>
</tr>
<tr>
<td>4.1</td>
<td>Comparison of PedGenie to the standard distribution using simulated data for independent individuals and nuclear families</td>
<td>64</td>
</tr>
<tr>
<td>4.2</td>
<td>Results for PedGenie chi-square statistic and Slager and Schaid Armitage trend test using GAW12 data</td>
<td>66</td>
</tr>
<tr>
<td>4.3</td>
<td>Characteristics of breast cancer cohorts</td>
<td>67</td>
</tr>
<tr>
<td>4.4</td>
<td>Association of each NBS1 tSNP with breast cancer status and age at diagnosis of breast cancer</td>
<td>69</td>
</tr>
<tr>
<td>4.5</td>
<td>Association of NBS1 by composite genotype in the Case-Control cohort</td>
<td>71</td>
</tr>
</tbody>
</table>
4.6 Association of NBS1 haplotypes and breast cancer in the Case-Control cohort

1. Characteristics of SNPs analyzed

2. Haplotypes with frequency >0.01, LD group characterization and tSNPs selected using Utah genotyping data

3. Comparison of LD groups for the Utah breast cancer cases and controls with Applied Biosystems (ABI) data

4. Haplotypes with frequency >0.01, LD group characterization and tSNP selected using data from Applied Biosystems

6.1 Single locus at a time associations of XRCC4 tSNPs with breast cancer risk and for cases, age at diagnosis

6.2 XRCC4 haplotypes observed among 94 unrelated individuals

6.3 Association of 2-locus XRCC4 tSNP haplotypes with breast cancer risk and for cases, age at diagnosis

6.4 Association of selected 4-locus XRCC4 haplotypes and age at diagnosis of breast cancer

6.5 Association of selected 4-locus XRCC4 haplotypes and breast cancer risk
ACKNOWLEDGMENTS

This work was supported by grants from the National Library of Medicine (T15 LM0724) and the Susan G. Komen Breast Cancer Foundation (DISS0201521). I sincerely thank Dr. Nicola J. Camp, my committee chair, for her instruction, insights, encouragement, and assistance in bringing this dissertation and all of its subparts to a completion. I also recognize and thank the other committee members, Doctors Lisa Cannon-Albright, Reed Gardner, David Goldgar, and John Ward, for their support. I appreciate the assistance of Kim Nguyen (Genetic Epidemiology) and Michael Hoffman (Family and Preventive Medicine) for their help in the laboratory. I thank Helaman Escobar (Director of Sequencing and Genomics) and Michael Klein (Genomics) from the Core Resource Facilities, University of Utah, for use of their equipment and assistance on this project. I thank Jathine Wong for her programming assistance of PedGenie. Data collected for this publication were assisted by the Utah Cancer Registry supported by National Institutes of Health, Contract NO1-PC-35141, Surveillance, Epidemiology and End Results (SEER) Program, with additional support from the Utah Department of Health and the University of Utah. Partial support for all datasets within the Utah Population Database (UPDB) was provided by the University of Utah Huntsman Cancer Institute.
INTRODUCTION

Despite improvements in cancer screening and novel and more effective treatment strategies, breast cancer remains one of the largest public health problems for women worldwide. It is the most common cancer diagnosed in women and second only to lung cancer as a cause of cancer related death in women.\(^1\) The American Cancer Society (http://www.cancer.org) estimates that in the United States in 2005, 211,240 new breast cancers will be diagnosed (32% of all cancers diagnosed in women) and 40,410 individuals will die from breast cancer (15% of all cancer deaths for women). Breast cancer incidence rates have steadily increased over the past 10-15 years, and rates are expected to continue to increase due to the overall growth and aging of the population.\(^2\) Although breast cancer mortality rates are declining 2.3% per year\(^1\) due to improvements in diagnosis and treatment, the current and future burden of new breast cancer cases will keep breast cancer a major public health problem for many years to come.

Breast cancer is thought to be attributable to a combination of environmental and genetic risk factors. Breast cancer risk increases with female gender, increasing age, and higher socioeconomic status (see review \(^3\)). Geographic location plays a role as well as women in the United States and Western countries have a 5- to 10-fold increased risk of cancer compared to low risk regions such as the Far East, Africa and South America.\(^4\) The contribution of genetic and environmental factors to breast cancer
is also evident in studies of migrants. Non-Caucasian migrants to the United States tend to have breast cancer rates that are higher than their country of origin, but less than those of Caucasians, while Caucasian migrants to the U.S. approach native-born Caucasians within the first generation.\textsuperscript{5,6} Differences in breast cancer rates by race also exist, even within the United States. African Americans have a lower overall incidence, but a higher overall mortality compared to Caucasian Americans, even after controlling for the more advanced tumor stage, higher-risk tumor biology, and earlier age at diagnosis often observed in African American women with breast cancer (see review \textsuperscript{7}). Reproductive factors strongly influence risk due most likely to the lifetime exposure of various endogenous sex hormones,\textsuperscript{8} including early age at menarche (< 12 years), late age at menopause (>54 years), nulliparity and older age at first birth, whereas breast-feeding and the selective estrogen receptor modulators (SERMs) such as tamoxifen and raloxifene\textsuperscript{9} can decrease risk. Exogenous hormone exposure from hormonal replacement therapy (HRT)\textsuperscript{10} and recent use of oral contraceptives\textsuperscript{11,12} also increase breast cancer risk but only moderately. The lifetime cumulative exposure of estrogen is often cited as a major cause of breast cancer;\textsuperscript{13} however, whether estrogen hormone exposure is directly causal or functions as a promoter of breast cancer\textsuperscript{14} is not known. Lifestyle also impacts breast cancer risk. Dietary consumption of alcohol,\textsuperscript{15} well-done meats,\textsuperscript{16} and a high fat diet\textsuperscript{17} increase risk moderately. Obesity and weight gain increase risk in the postmenopausal woman; for each 5 kg of weight gain since the lowest adult weight, breast cancer risk increases by 8%\textsuperscript{18} whereas caloric restriction confers protection.\textsuperscript{19} Increased physical activity has been found to be associated with a decreased breast cancer risk; longer duration provides the most benefit.\textsuperscript{20,21} Other
factors that have been shown to be associated with increased breast cancer include ionizing radiation exposure in childhood, history of benign breast disease, high mammographic breast density, increased bone density in postmenopausal women, tall stature in postmenopausal women, antibiotic use, heavier birth weight and earlier age at peak growth. As the pathophysiology of most breast cancer cases remains elusive, for many of the epidemiological factors listed above, whether they are truly causally related to breast cancer or merely associated with a causal factor remains to be determined.

High-penetrance breast cancer susceptibility genes

The role of genetics is also central to breast cancer etiology. Often as a precursor to a genetic study, family aggregation studies have shown that family history is well-recognized as a strong risk factor for female breast cancer. Risk varies from an increase of approximately 1.8 to 3.3 fold by having a first degree relative with breast cancer, with increased effects noted if the relative was diagnosed < 45 years of age, and can increase up to 3.9 fold by having 3 first-degree relatives with breast cancer.

The study of families with multiple cases of breast cancer has been very helpful in the identification of several high-penetrance genes that are associated with hereditary breast cancer. Identification of truncating mutations in the coding sequence of BRCA1, for example, was made possible with multiple cases of early-onset breast cancer cases in several large families that also included ovarian cancer cases. Using positional cloning methods, BRCA1 on chromosome 17q was identified in Utah in five of eight families presumed to be segregating BRCA1 susceptibility alleles. Germline mutations in BRCA1 have since been identified in 15-20% of women with a family history of
breast cancer, with penetrance estimates of breast cancer to age 70 years ranging from 36-69% among a group of high-risk Ashkenazi Jewish women (see review). As Ashkenazi Jewish women are at high-risk for breast cancer due to a restricted number of mutant BRCA1 alleles, they are an ideal population to estimate BRCA1 penetrance functions, whereas in other populations the large number of possible mutants makes very difficult if not cost prohibitive to obtain good penetrance estimates.

Another high-penetrance breast cancer susceptibility gene, BRCA2 was discovered in families with a high incidence of male breast cancer that did not carry the BRCA1 mutation. The lifetime breast cancer risk for BRCA2 among high-risk Ashkenazi Jewish women ranges between 26-74% (see review), and as the Ashkenazi Jewish population also have a limited number of BRCA2 mutations they are an ideal population to study for BRCA2 penetrance estimates. Germline mutations in BRCA1 and BRCA2 are estimated to account for about 15-30% of families with multiple cases of breast cancer but less than 5% of all breast cancers cases. Other high-penetrance genetic mutations found in individuals with rare genetic syndromes have also resulted in an increased risk of breast cancer. Germline mutations in the gene p53 were found in families with the Li-Fraumeni cancer susceptibility syndrome, an autosomal dominant disorder resulting in an increased risk of various cancers including breast, sarcoma, leukemia, and brain cancers. Germline mutations in PTEN gene have been found to be associated with Cowden syndrome, a rare hereditary syndrome characterized by an increased risk of breast and thyroid cancer, polyps and gastrointestinal cancers, and ovarian cancer. Other rare genetic disorders that increase risk of breast cancer include the Nijmegen breakage syndrome with mutations in the NBS1 gene, the Peutz-Jeghers
syndrome with mutations in the *LKB1* gene, and possibly heterozygous carriers of mutations in the *ATM* gene which result in the autosomal recessive disorder ataxia telangiectasia (see review \(^45\)). Together, these rare syndromes account for less than 1% of all hereditary breast cancers.\(^46\)

**Low-penetrance breast cancer susceptibility genes and a polygenic model**

Although much effort has been expended since 1994 when *BRCA1* and *BRCA2* were first discovered, it is apparent that much of familial breast cancer cannot be explained by mutations in these known more common high-penetrance genes. In several studies of high risk breast cancer families a substantial percentage were not found to be due to mutations in *BRCA1* or *BRCA2*.\(^47\)-\(^49\) A population-based series of early-onset breast cancer cases looking for mutations in *BRCA1* and *BRCA2* showed that <10% of the 617 cases diagnosed < 45 years of age were attributable to mutations in *BRCA1* or *BRCA2*.\(^50\) Hereditary breast cancers due to mutations in high-risk, high-penetrance genes account for about 5-10% of all breast cancer cases; familial breast cancer cases, or families with more than one breast cancer case but no known genetic cause, account for approximately 20% of all breast cancer cases; and the remaining approximately 70% of breast cancer cases are considered sporadic.\(^50,51\) Many have searched for a *BRCA3* gene and despite advances in high-throughput DNA processing and sequencing of the human genome it remains elusive. It has been postulated that some cases of breast cancer may be due to a small number of highly penetrant mutations, such as *BRCA1* and *BRCA2*, whereas other cases may be due a large number of low penetrant mutations acting in combination and in conjunction with various
environmental exposures. Pharoah et al. have proposed a polygenic model for breast cancer in which susceptibility to breast cancer is conferred by a large number of genes; the risk associated with any one gene is small but the effects are multiplicative. As the number of susceptibility genes increases, the number of risk groups increases until the risk within the population resembles a normal distribution. Currently, many are searching in biological pathways thought to be involved in breast cancer for polymorphic variants that may moderately increase risk of breast cancer and also are fairly common in the population. One such example is the identification of the low penetrant but fairly common founder mutation (1100delC) in the CHK2 gene, a gene activated in the DNA repair pathway and important in regulation of cell-cycle checkpoints, repair of double strand breaks through phosphorylation of BRCA1 and induction of apoptosis through phosphorylation of p53. A truncating mutation in CHK2 was found to increase risk of breast cancer in a European population; the mutation was found in 1.1% of the general population, 1.4% of breast cancer cases with no family history of breast cancer, 4.2% of breast cancer cases (diagnosed ≤ 60 years) with a family history of multiple breast cancer cases not attributable to either BRCA1 or BRCA2, and 13.5% of cases with a family history of male breast cancer. Although additional studies in Finland, the USA, and the UK have not found the 1100delC variant to be a risk factor for male breast cancer and other studies from the United States and Australia have found the variant to be rare in their population, still other studies have confirmed the increased risk in families with multiple cases of female breast cancer.
There are other examples of low-penetrance breast cancer susceptibility genes that have been demonstrated to have moderate effects on breast cancer. Most polymorphisms studied to date have examined biological pathways postulated to be involved in breast carcinogenesis, including DNA repair, metabolism, cell signaling, steroid hormone, etc. Unfortunately, just as was illustrated with \textit{CHK2}, many polymorphisms are published that significantly impact risk of breast cancer in certain ethnic populations, but replication in other ethnic cohorts or larger cohorts fails to confirm the results. Whether some of these polymorphisms will ultimately be viewed as low-penetrance susceptibility alleles remains to be determined. Examples of such breast cancer susceptibility polymorphisms that have been reported to increase breast cancer risk are available.\textsuperscript{3,45}

\textbf{Segregation analysis suggests the existence of breast cancer susceptibility loci}

In the quest to identify breast cancer susceptibility loci, several segregation analysis studies, a method to fit a genetic model to the phenotype data of family members from a particular cohort, have been performed.\textsuperscript{30,60,61} Claus et al., the most widely cited segregation analysis prior to the discovery of \textit{BRCA1} and \textit{BRCA2}, using the Cancer and Steroid Hormone (CASH) study data, found that the familial breast cancer data could best be explained by a rare autosomal dominant allele (minor allele frequency=0.0033) that was also dependent on a woman’s age.\textsuperscript{30} Breast cancer cases between the ages of 20-29 years were most likely to carry this high-risk allele (36%), while cases 80 years or older were least likely (1%). Cumulative lifetime risk of carriers of the allele was estimated to be 92%. This segregation model was later
validated with the identification of the autosomal dominantly inherited breast cancer susceptibility genes \textit{BRCA1} and \textit{BRCA2}, suggesting that segregation analyses may be very helpful in the pursuit of additional breast cancer genes. Two more recent segregation analyses have tried to account for \textit{BRCA1} and \textit{BRCA2} in their models. Antoniou et al. ascertained 1,484 women in the United Kingdom diagnosed with breast cancer before 55 years and tested them for \textit{BRCA1}/2 mutations.\textsuperscript{61} Genetic models, to assess effects of \textit{BRCA1}, \textit{BRCA2}, a hypothetical \textit{BRCA3}, and a polygenic component, were built using information on breast and ovarian history in first-degree relatives and \textit{BRCA1}/2 status in the probands. The best-fitting major gene model was a recessive model with disease allele frequency of 0.24 and cumulative risk to age 70 years of 42%. After adjusting for parity, however, the polygenic model fit best. Reanalysis of the data showed similar results\textsuperscript{62} and these data were used as the basis for the polygenic model by Pharoah et al.\textsuperscript{52} described earlier. In another segregation analysis, Cui et al. ascertained 858 Australian women with breast cancer diagnosed <40 years and conducted single-locus and two-locus segregation analyses using three-generation family histories.\textsuperscript{60} These women were also tested for \textit{BRCA1}/2 mutations. The most parsimonious model after excluding known carriers of mutations in the \textit{BRCA1}/2 genes was a recessive model with an allele frequency of 0.0653 and probability of disease for carriers equal to 86% by age 50 and 99% by age 70. A residual dominantly inherited risk was also observed, in addition to that observed for \textit{BRCA1}/2 genes. The susceptibility allele frequency in the dominant model, after excluding \textit{BRCA1} and \textit{BRCA2} mutation carriers from the analysis, was 0.0011 with a penetrance estimate of 48% by age 70 years. These segregation analyses would suggest that additional breast
cancer susceptibility loci exist, particularly one or more that is inherited in a recessive fashion.

Use of linkage analysis to identify breast cancer susceptibility loci

The standard technique to identify breast cancer susceptibility loci is the traditional linkage method. Several reports have been published that indicate significant linkage of the breast cancer phenotype to various regions of the genome. One such chromosomal region is the 8p12-p22 where two separate studies have found significant results. Linkage analysis in eight French breast cancer families found a multipoint log of odds (LOD) score of 2.51 with two markers on chromosome 8p, and linkage analysis of two large German breast cancer families showed a multipoint LOD score of 3.30 at two other markers between the two markers in the French study. Loss of heterozygosity has also been observed in multiple studies within this chromosomal region; in both sporadic and familial tumors and particularly in invasive tumors. However, confirmation of these linkage findings was not supported in a study of 23 families with at least three breast cancer cases with a low probability of being attributable to BRCA1/2 and diagnosed < 60 years from the UK, Germany, Holland and the USA. This study concluded that the 8p12-p22 region is unlikely to make a substantial contribution to the predisposition of familial breast cancer. A similar scenario is described for the chromosomal region 13q21 where significant evidence for linkage was found using 77 families with multiple cases of breast cancer from Scandinavia, but failed to be confirmed in a subsequent study using 128 high-risk families of Western European ancestry. Study of common, low-penetrance alleles in genes is difficult as illustrated in these two examples. Implicit in the definition of a
A low-penetrance gene is that the penetrance of such a gene will likely be of lower penetrance than BRCA1 or BRCA2 and hence it is harder to find. Furthermore, as breast cancer is most likely a genetically heterogeneous disease, finding multiple genes, each contributing a small effect, will be difficult using traditional linkage techniques.

**Strategies to identify additional breast cancer susceptibility genes**

**An informative phenotype**

What is needed to assist in the identification of additional breast cancer susceptibility loci? One of the most pressing needs is an informative phenotype that can be used to identify genetically homogenous breast cancer cases. Prior to the discovery of the BRCA1 and BRCA2 genes, it was well recognized that ovarian cancer and male breast cancer were components of a breast cancer syndrome. This allowed for targeted ascertainment of early-onset breast cancer families that also included ovarian and male breast cancer. Currently in the quest to find additional genes, identification of families with multiple cases of breast cancer that are not attributable to either BRCA1 or BRCA2 is difficult. The larger the family, the more likely breast cancer is due to either BRCA1 or BRCA2 mutation and if it is not due to either of these two genes, the more likely it is due to multiple susceptibility alleles or locus heterogeneity. Familial breast cancer is thought to be easier to study as it is expected to be more homogenous than sporadic cases of breast cancer. Ideally, new ways to identify families most likely due to a single gene disorder or explained by a simpler genetic mechanism are needed.

Numerous studies have collected information on the molecular profile of BRCA1, BRCA2 and other non-BRCA1/2 familial tumors to distinguish their phenotypes. According to these studies, the BRCA1 tumors are less frequently
estrogen receptor negative (ER-), progesterone receptor negative (PR-), androgen receptor negative (AR-), BCL2 negative, P27^Kip1 negative, ERBB2 negative, and lymph node positive; but they are more frequently P-cadherin negative, cytokeratin (CK) 5/6 negative, and p53 positive. Most BRCA1 tumors are high grade at diagnosis, characterized by a greater number of mitoses, a greater degree of cellular pleomorphism, and the histological subtypes of medullary and atypical medullary carcinomas are over-represented in this cohort. BRCA2 and non-BRCA1/2 familial tumors are more heterogenous and hence more difficult to cleanly classify. BRCA2 tumors are also more frequently high grade, more likely to have a smooth non-infiltrative border, and less likely to have tubule formation compared to sporadic tumors. Non-BRCA1/2 familial tumors are characterized as being lower grade, fewer mitoses, less nuclear pleomorphism, and more likely to be of invasive lobular histological type compared to BRCA1/2 tumors. They are also less frequently p53 and ERBB2 positive than sporadic tumors. Recently, further classification of breast cancer subtypes has been performed using gene expression microarrays. Five main subtypes of breast cancer have been defined: two “luminal-like,” one “basal-like,” one ERBB2 over-expressing, and one normal breast tissue-like. Most BRCA1 tumors have been classified as “basal-like,” whereas BRCA2 and the non-BRCA1/2 tumors are more difficult to classify into a single class, however, BRCA2 tumors appear to fit better in the “luminal-like” cohorts. Although the idea that breast tumor subtypes can be represented as biologically distinct disease entities is appealing, there is great difficulty classifying non-BRCA1/2 tumors using expression arrays, particularly in light of the limited reproducibility of results across microarray platforms and researchers, the
background noise associated with microarray assays, and the complexity of describing biological pathways by clustering genes with similar expression patterns (see review\textsuperscript{80}). Much work remains to use microarray expression platforms to distinctively identify a non-\textit{BRCA1}/2 phenotype; whether microarray expression assays are the best mechanism to identify a new breast cancer phenotype is not currently known.

Novel association methods

Association analyses are considered superior to linkage techniques for finding genes of small effect. Also, analyzing familial breast cancer cases increases the power to detect those effects. Hence, a useful tool in the quest to identify additional breast cancer susceptibility loci is the development of novel association methods/packages that perform valid analyses when multiple related cases are considered in an analysis. Such methods need to be flexible in terms of the number and choice of subjects (i.e., unrelated case-control data or large family resources) and the number of markers that can be analyzed, as well as the ability to process haplotypes and interactions between and within various loci, and to incorporate covariate data. Strategies for overcoming the penalty of testing multiple hypotheses from multiple tests of low-penetrance alleles must also be sought, as even in the presence of a true causal locus or true interactions between loci the overall power of the study may not be sufficient to discover true associations. Currently, additional low penetrance susceptibility alleles are being searched among common sequence variants in the human genome, principally single nucleotide polymorphisms (SNPs). There are now over nine million SNPs listed in public databases, with more than one million typed in population samples and used to determine allele frequencies.\textsuperscript{81,82} SNP genotyping technology is now at the point that it
is sufficiently accurate and reasonably priced that genome-wide SNP genotyping assays are feasible, hence allowing testing of large collections of markers simultaneously. Tools to analyze this vast amount of data are necessary. Results from studies utilizing SNPs have the additional advantage that association may be detected if the SNP itself is causally associated with the phenotype or if it is in linkage disequilibrium with a causative locus.

Strategies for testing interactions will be very important. For breast cancer, Dr. Bruce Ponder suggested that over 200 common low penetrance alleles with a frequency of 0.01 would be needed to account for the approximate 2-fold increased risk observed among relatives of breast cancer subjects. If the allele frequency increased to 0.3, then between 13-18 such variants would be needed. As each of these common variants might have a small individual effect, the individual association for any particular locus may be weak. Hence, effects may only be seen if interactions among loci are tested. Recently, Marchini et al. using a simulated genome-wide association data, investigated various analytical approaches to detect gene-gene interactions. They found, as would be expected, that the single locus at a time approach was the most powerful for loci that had both independent main effects as well as interactions with other loci. They also investigated a pair-wise testing approach and found it worked best when the loci effects were entirely dependent on a combination of alleles. The authors concluded that the best strategy for detecting multiple loci is to test locus-by-locus first, followed by pair-wise testing of loci reaching a very modest nominal level of significance. As the Marchini et al. study focused on simulated data and investigated only single locus and
pair-wise combination analytical techniques, additional strategies that are robust to multiple models of gene effects and based on empirical data are necessary.

Statistical tests for analysis of family based data types are often limited by the pedigree size, relationship structures, and/or genotype configuration that are permitted. Some statistical tests such as the transmission disequilibrium test (TDT), a family based test in which the pattern of alleles transmitted from a heterozygous parent to an affected child is considered, are limited in that the parent-offspring trios that can be studied require genotype data on an affected offspring as well as his/her two parents, one of whom must be heterozygous for the marker of interest. Limitations also exist for other family based tests. Other family-based tests require the calculation of an inherited by descent (IBD) parameter, a quantitative measure of relatedness between individuals. Current methods that calculate this parameter limit the size of the family that can be analyzed which inhibits its use on large genealogical databases. Many of the family based methods are limited in the range of statistics or type of data (i.e., quantitative or qualitative) that can be analyzed. Ideally, analysis of all family data that is available is most efficient and informative. Opportunities exist to develop family based methods that are flexible in terms of the statistics that can be performed and family size that can be studied.

Improving the reproducibility of results and study design

In addition to the identification of a new, informative breast cancer phenotype and the development of new association methods/tools, the discovery of novel breast cancer susceptibility loci will be greatly benefited by strategies to improve the reproducibility of results and study design. Failure to replicate results is of serious
concern and as described above, few low-penetrance breast cancer loci have been identified unequivocally. For example, Dunning et al. performed a systematic review of genetic polymorphisms from 46 published studies and examined their effect on breast cancer risk.\textsuperscript{86} Twelve studies reported statistically significant results, but none of these significant results appeared in more than one study. A meta-analysis of all 46 studies found only three SNPs (i.e., \textit{CYP19} (TTTA)\textsubscript{n} polymorphism, \textit{GSTP1} Ile105Val polymorphism, \textit{TP53} Arg72Pro polymorphism) to be significantly associated with breast cancer and all with a magnitude of effect less than 2.5-fold increased risk. A more rigorous study design will greatly aid in the reproducibility of results. One of the major limitations of most of these published polymorphism studies is that too few cases and controls were included, such that there was inadequate statistical power to detect true associations. Dunning et al. reported that only 10 out 46 (~22\%) of the published literature analyzed were of sufficient sample size to be able to detect a modest allele frequency of 0.2, with 90\% power at a 5\% significance threshold. Studies of sufficient sample size are essential in order to find novel breast cancer susceptibility loci. Along with this problem of inadequate sample size is the problem of multiple testing that occurs as multiple SNPs are analyzed across the genome. Analysis of a large number of polymorphisms renders a multiple testing penalty such that any positive association evidence will most likely be lost. New strategies for dealing with multiple testing across multiple SNPs, that may be in linkage disequilibrium with each other, and hence not independent and thus not requiring a strict Bonferroni correction, are imperative (e.g., Nyholt\textsuperscript{87}).
Another problem that impacts reproducibility of association study results is that many published studies report the results for only a single SNP per gene or only a single locus at a time. A single SNP simply cannot account for the genetic variance across a gene and the single locus at a time approach incurs the multiple testing penalty described above. It has been suggested that the use of haplotypes in association studies may have increased power over single-allele studies as more of the genetic variance is captured. Characterization of haplotypes across genes is thus an important first step prior to any genetic association analysis. Furthermore, genotyping costs may limit the selection of SNPs. To decrease costs and decrease redundancy, a reduced set of tagging-SNPs based on the linkage disequilibrium structure may be used; tSNPs will identify all common haplotypes within a region of high linkage disequilibrium. Identification of tSNPs greatly aids the efficiency and decreases the cost of genetic association studies.

Population stratification is another potential problem for association studies involving case control subjects and can lead to failure to replicate results. Population stratification is essentially a confounder in genetic association studies; differences in ethnicity result in differences in allele frequency at all loci that differentiate these groups whether the alleles are causally related to disease or not. Although genetic studies often use ethnicity matching to overcome the population stratification bias, hidden ethnic differences or cryptic relatedness may exist and affect the validity of association results. Undetected population stratification can result in either false positive signals or false negative signals. Although determining whether population stratification is a major threat or not to genetic association studies has been intensively
debated (see \textsuperscript{89,90}), more recent reports have shown that even modest levels of population stratification should not be ignored. Marchini et al.\textsuperscript{91} showed, using \textasciitilde 15,000 SNPs in three different ethnic populations, that population stratification increases with sample size. Even correcting for population stratification through a series of unlinked markers (i.e., genomic control) may not accomplish adequate correction if too few markers are selected. Furthermore, population stratification may exist in even carefully matched study designs. A recent analysis of population structure using genealogic and genetic data from Iceland indicated that substructure exists even in this relatively homogenous genetic isolate population.\textsuperscript{92} The use of family based controls circumvents the problem of population stratification as within a family even hidden ethnicity is perfectly matched; however, family studies are less powerful than case-control methods.\textsuperscript{93}

As summarized only briefly here in this review of the breast cancer literature, breast cancer most likely results from a combination of genetic and environmental factors. Several high-penetrance breast cancer predisposition genes have been identified, but additional low-penetrance breast cancer predisposing loci most likely exist, acting in a polygenic fashion to increase risk of breast cancer. Finding these low penetrance genetic variants is proving to be very difficult and a few strategies for increasing the likelihood of success have been proposed. As part of the doctoral dissertation, each of these proposed strategies, although by no means comprehensive, has been incorporated with the goal of furthering our knowledge about the pathophysiology of breast cancer and perhaps aiding in the earlier diagnosis of breast cancer.
A new, informative phenotype: lobular breast cancer

The first strategy presented here was the need to identify an informative breast cancer phenotype that might be used to cluster families with excess breast cancer, not attributable to BRCA1 or BRCA2. Using the Utah Population Database (UPDB), we studied excess familiality by histological subtypes of breast cancer. We focused on a strict definition of the histological subtype lobular breast cancer, since some previous studies indicated that the lobular breast cancer subtype resulted in an increased risk of breast cancer to relatives. We found that relatives of lobular breast cancer probands are at increased risk of breast cancer and specifically lobular breast cancer. More interestingly, the lobular breast cancer phenotype is less likely to be attributable to a BRCA1 or BRCA2 mutation. Studies involving lobular breast cancer families may be very useful in the identification of additional breast cancer susceptibility loci. Our results are published in the International Journal of Cancer.

A novel association package: PedGenie

The second strategy presented was the need to develop flexible association analysis techniques that could handle case control and pedigree data without restrictions on size or relationship structure, quantitative and qualitative data, haplotype data, and inclusion of covariates. As part of the dissertation project, the association package entitled PedGenie was developed. PedGenie is a software tool for association testing of qualitative and quantitative traits in pedigrees of any size, including loops that performs Monte Carlo simulations to assess significance. Specifically, PedGenie was expanded to include tests for the transmission disequilibrium test (TDT) for qualitative
data and quantitative data (QTDT). A web site for PedGenie was developed (see http://bioinformatics.med.utah.edu). Two manuscripts describing the functionality of PedGenie, comparing it to other association packages, and illustrating its use have been written. The first manuscript using data from the Genetic Analysis Workshop (GAW13) in addition to describing the functionality of PedGenie, illustrated the problems associated with using a linkage resource of simulated microsatellite markers and testing it using an empirical chi-squared test from PedGenie. This manuscript was published in BMC Genetics. The second manuscript, submitted to BMC Bioinformatics, 2005, shows the expansion of PedGenie’s functionality to handle both quantitative and qualitative data, and classical tests of association as well as tests of linkage disequilibrium (i.e., the TDT based tests). Data from the NBS1 gene is used for illustration in this manuscript.

Improving the reproducibility of results and study design

The final strategy proposed was multifaceted and provided suggestions to improve the reproducibility of published results and study design. Incorporating these suggestions for an improved study design, a grant was submitted and obtained from the Susan G. Komen Breast Cancer Foundation providing the funds to genotype 1,040 individuals from 139 high-risk Utah breast cancer pedigrees for five genes in the DNA repair pathway (ATM, MRE11, XRCC4, NBS1 and RAD50). These families were selected because their rates of breast cancer exceeded the population rate of breast cancer determined using the Utah Population Database, and their breast cancer most likely was not attributable to either a BRCA1 or BRCA2 mutation. The use of family data, accounting for all known genealogical relationships, prevented population
substructure from biasing our results. Genes were selected from the DNA repair pathway, as they are good candidates for a role in polygenic disease; an inability to respond or properly repair DNA damage leads to genetic instability, which may increase the rate of cancer development. All of the laboratory experiments, including finding the samples in freezers, sample dilution, DNA extraction where needed, preliminary testing of samples by PCR, actual genotyping, and quality control, were performed as part of the Ph.D. dissertation. Once the genotyping was completed, the linkage disequilibrium structure and haplotype architecture of these five candidate genes for breast cancer was characterized, and tSNPs capturing >90% of the intragenetic variation for each gene were identified. These results are published in *BMC Cancer*. The identified tSNPs were genotyped on our entire study population (n = 464 cases and n = 576 controls) and associations with breast cancer status and age at diagnosis analyzed using PedGenie. Interesting and consistent results were found for several of the tSNPs in *XRCC4*. A manuscript of the *XRCC4* results is contained in this Ph.D. dissertation. Hypotheses of interactions between all five (i.e., *ATM, MRE11, XRCC4, NBS1* and *RAD50*) are also being pursued and manuscripts are being prepared which are complementary, but beyond the scope of this dissertation.

Here the complexity of the study of breast cancer has been outlined. Although much has been learned in the past few decades about the genetic susceptibility to breast cancer, many questions remain unanswered. Some of these questions and their answers are addressed in this dissertation. Of course, other questions have arisen from the work presented here and are suggested in the Conclusion section of the dissertation. The
personal and societal costs of breast cancer are staggering and as further knowledge about breast cancer is elucidated, women, their families, and all of society will benefit.
References


LOBULAR BREAST CANCER: EXCESS FAMILIALITY OBSERVED IN
THE UTAH POPULATION DATABASE


Published: International Journal of Cancer 2005; 117: 655-661
Family history of breast cancer (BC) is a strong predictor for developing female BC. Whether this excess familiality differs within morphological BC subgroups remains unclear. We assessed the risk of lobular breast cancer (LOB) and any BC among relatives of probands with LOB. We used the Utah Population Database (UPDB) to estimate familial relative risks (FRR) as well as average relatedness, using the genealogical index of familiality (GIF) statistic. The UPDB, a population-based resource, links genealogical data from over 2 million individuals to the Utah Cancer Registry. Consistent with other studies, analysis of all BC cases showed significantly increased risk of BC to relatives (first-degree relative [FDR]: FRR = 1.83, 95% confidence interval [CI] = 1.75–1.90). Morphology-specific risks showed that relatives of LOB probands had an increased risk of LOB (FDR: FRR = 4.51, 95% CI = 2.79–6.89) and an increased risk of any BC (FDR: FRR = 2.47, 95% CI = 2.12–2.85); both measures were significantly greater than the all BC FRR estimates, and surpassed even generalized early-onset BC risk. GIF analyses corroborated the FRR results and indicated that the excess relatedness of LOB cases extended to third-degree relatives. Our findings suggest that LOB has a heritable component and may represent a genetically homogeneous form of BC. Pedigrees with excess LOB may be useful in isolating additional BC predisposition genes. Relatives of women with LOB are at higher risk for BC than relatives of other BC subtypes; a more rigorous BC screening regime may be warranted for these individuals.

Key words: breast cancer; lobular; familiality

Family history is well-recognized as a strong risk factor for female breast cancer. First-degree female relatives of women with breast cancer have a 2-to 3-fold increased risk of breast cancer compared to the general population risk. The magnitude of breast cancer risk among relatives of a woman with breast cancer increases with her earlier age at onset, the number of other affected relatives in her family and their relationship to her. In 2 separate meta-analyses of 52 and 74 breast cancer studies, the increased risk associated with one affected first-degree relative ranged from 1.8–3.3-fold depending on whether the relative was a mother or a sister and whether diagnosis occurred at <50 years of age. Breast cancer risk estimates varied up to 3.9-fold by having 3 affected first-degree relatives.

Breast cancer is a mix of morphologically distinct cancers. Attempts to determine whether the increased risk for breast cancer varies by morphological subtype have met with limited success in the past. Before the discovery of the breast cancer predisposing gene BRCA1, it was reported that relatives of those with medullary breast carcinomas were at increased risk of breast cancer. BRCA1 mutations have since been shown to be associated with an excess of medullary and atypical medullary cancers, and more recently, the basal epithelial subtype. Although histology was not useful for localization of the BRCA1 gene, it is possible that histology may be useful in defining cohorts where breast cancer predisposition loci can more easily be identified. Increased breast cancer risk associated with morphology-specific subtypes, other than medullary and basal epithelial subtypes, is less clear. We and others have found some evidence for an increased familiality of lobular breast cancer.

Abbreviations: BC, breast cancer; CI, confidence interval; COK, coefficient of kinship; Exp, expected; FDR, first-degree relative; FRR, familial relative risk; GIF, genealogical index of familiality; ICD-O, International Classification of Diseases of Oncology; LOB, lobular breast cancer; Obs, observed; SEER, Surveillance, Epidemiology, and End Results; SDR, second-degree relative; UCR, Utah Cancer Registry; UPDB, Utah Population Database.

Grant sponsor: National Library of Medicine; Grant number: T15 LM0724; Grant sponsor: University of Utah; Grant sponsor: Susan G. Komen; Grant number: DSS0201521.

*Correspondence to: Genetic Epidemiology, 391 Chipeta Way, Suite D, Salt Lake City, UT 84108. Fax: 1801-581-6052, 1801-581-4297. E-mail: kristina.allen@hsc.utah.edu

Received 5 November 2004. Accepted after revision 24 March 2005

DOI 10.1002/ijc.21236

Published online 31 May 2005 in Wiley InterScience (www.interscience.wiley.com).
Cannon-Albright et al. observed that lobular cancer showed a significant excess relatedness using the Utah Population Database (UPDB). Claus et al. in the CASH study of early onset breast cancers (i.e., ages 20–54 years) found that cases with lobular carcinoma in situ were significantly more likely to have a mother or sister affected with any type of breast cancer than cases with other tumor types, but this same trend was not seen for relatives of those with invasive lobular cancer. Erdreich et al. found that lobular carcinoma in situ occurred more frequently in younger women and showed a greater frequency of bilaterality, but not an increased family history risk. Rosan et al. found that women with lobular cancer were more likely to have at least one sister affected with any type of breast cancer than were women with other histologic types of cancer.

Hemminki and Granstrom, using the Swedish Cancer Registry, reported recently that the risk of any type of breast cancer for mothers and sisters was independent of the morphologic type of cancer of the female probands. Many of these studies are hindered by small sample size and reliance on the proband’s self-report of breast cancer diagnoses among relatives, which may be limited by recall bias. Furthermore, due to a limited number of morphology-specific relative pairs, all previous studies, excluding ours, focused on the risk of any type of breast cancer among relatives rather than a morphology-specific risk.

The UPDB resource has grown significantly since our initial preliminary report of an excess of familiality for lobular cancer. The UPDB consists of genealogical data from over 2 million individuals that has been record linked to the Utah Cancer Registry (UCR). The large number of observations in this database coupled with extensive genealogy data allows us to examine and clarify the familial nature of lobular breast cancer. We explore in more depth the familiality of female lobular breast cancer. We estimate relative risks of lobular cancer and all breast cancer for first-degree and second-degree relatives of women with lobular breast cancer. We explore the familiality of lobular in situ and invasive lobular cancer and extend our findings to early onset cancers (i.e., ≤50 years). We also present the familial clustering (measured by average relatedness) of the various morphology-specific subtypes using the genealogical index of familiality (GIF).

Material and methods

Our study was approved by the University of Utah Institutional Review Board. To protect the privacy of individuals within the database, names and other identifying pieces of information were not available to the authors of our study.

The genealogical portion of the Utah Population Database (UPDB) resource consists of computerized genealogy data for more than 2 million individuals who are descendants of approximately 10,000 largely unrelated founding Utah pioneers. Data for the genealogy database came from family group sheets, which include names, places and dates of birth and death of a husband, wife, their parents, and their children. These group sheets have been linked to extended family pedigrees, which in some cases descend to nine generations. For our study we only considered individuals who had at least 3 generations of genealogy data to increase the consistency of information across individuals and to better match cases and controls. The Utah population represented in this database have been shown, to be similar to other Caucasian populations of Northern and Western European descent with normal levels of inbreeding.

Other information regarding the UPDB is available.

Utah Cancer Registry

The Utah Cancer Registry (UCR) is a population-based cancer registry that has been part of a statewide breast cancer registry program since 1966, but includes cancer records dating back to 1952. In 1973, the UCR became part of the Surveillance, Epidemiology, and End Results (SEER) Program and as such, all cancers except basal and squamous carcinomas of the skin were required to be reported to the Registry. As one of the SEER sites, the UCR must follow rigorous guidelines to ensure accuracy of its data. Every reported cancer case requires a pathology report with follow-up accomplished through manual review of hospital records and hospital site visits. There are over 86,000 SEER registry records that have been linked to the genealogy records in the UPDB. Approximately 60% of cancer records in the UCR link to a UPDB genealogy record; record linking rates are slightly lower for females due to name changes that reduce the probability of successful linkage. Linkage of UCR information to the UPDB genealogy records occurred August 2003 and includes cancer cases through December 2002. Many of the descendants of the original pioneers were members of the Church of Jesus Christ of Latter-day Saints (LDS) and abided by its proscriptions against consumption of coffee, tea, alcohol and tobacco. Cancer rates in Utah have been shown to be low compared to the U.S. Third National Cancer Survey due largely to reduced smoking-related cancers.

Histology

Cancer records in the UCR are coded by disease site according to the International Classification of Diseases of Oncology (ICD-0) and contain information on age at diagnosis, histology (including behavior codes), stage, grade and survival of patients. Lobular carcinoma was defined by ICD-0 morphology code 8520. Lobular in situ was defined with behavior = 2 and invasive lobular carcinoma was defined with behavior = 3. The diagnosis of mixed lobular and ductal breast cancer was not included in our study. For individuals with multiple primary diagnoses of breast
cancer, only the initial diagnosis was considered to avoid double counting. Some women (n = 6 in the entire UCR and n = 4 in the UPDB who also had 3 generations of genealogy data) had a concurrent diagnosis of both lobular carcinoma in situ and invasive lobular carcinoma; however breast laterality differed between the diagnoses. To avoid bias, those with a concurrent diagnosis were excluded from the lobular carcinoma in situ and invasive lobular carcinoma analyses. Due to the limited number of lobular carcinoma in situ cases, analyses involving early-onset lobular in situ cancer (i.e., ≤50 years) were not carried out.

**Statistical analysis**

**Familial relative risk.** Familial relative risk (FRR) is an estimation of the risk of disease among relatives of a set of probands compared to the general population. As diseases with a heritable component are expected to occur at higher frequency among relatives of diseased probands than among random controls, this measure of familial risk provides a way of assessing the significance of the familial clustering of a particular type of cancer. FRR in our study is estimated as the number of observed cancers (O) among the relatives of probands divided by the expected number of cancers (E) among the relatives using internal population rates, or FRR = O/E.

The number of observed breast cancers (O) in relatives are counted, without duplication, for any first-degree relative (FDR) and any second-degree relative (SDR) of the probands. For matching purposes, relatives of probands are placed into 64 distinct cohorts based on year of birth (5-year birth cohorts), female gender and place of birth (Utah versus non-Utah). As Utah cancer records are considered complete only since 1973, affected status depends on the year of cancer diagnosis and place of diagnosis. We calculate cohort-specific rates of breast cancer using the total number of breast cancer cases per cohort divided by the total number of individuals in the UPDB who are in that cohort.

The expected number of cancers (E) among a defined set of relatives is calculated using the formula E = \( \frac{R_c \times N_r \times C/N_i}{i} \) (for i between 1 and 64) where \( R_c \), \( N_r \), and \( C/N_i \) are the numbers of relatives of probands, the number of individuals in the UPDB, and the number of breast cancer cases in the UPDB, respectively, in the i-th cohort group. Approximate 95% confidence intervals (CI) and hypothesis tests of the null hypothesis FRR = 1.0 are constructed, assuming that the number of cancers found among the relatives follows a Poisson distribution with mean E.

**Genealogical index of familiality.** The Genealogical Index of Familiality (GIF) was developed to measure the degree of familial clustering of cancers specifically in the UPDB. \(^9,30\) The GIF provides a broader measure of familiality than the FRR, as it considers the relatedness of all affected, not only close relatives. The GIF test of hypothesis of excess relatedness compares the average relatedness of a set of cases to that of matched controls, testing the hypothesis that the cases are more related than would be expected for a random set of individuals of the same age and gender from this population. The GIF quantifies the average relatedness between all pairs of cases within the UPDB using the Malecot coefficient of kinship \(^28\) (COK): the GIF statistic represents the average COK between all cases, and is multiplied by \( 10^v \) for ease of presentation. The COK is defined as the probability that randomly selected homologous genes from 2 individuals are identical by descent from a common ancestor. This is calculated by counting the number of paths of common descent; with each path contributing an exponent of \( \frac{1}{2} \) to the total coefficient of kinship value. The value of the exponent is the number of people in the path connecting a pair of cases. For example, the coefficient of kinship between a parent and offspring is \( \left( \frac{1}{2} \right)^0 \) or 0.25 and between siblings is \( \left( \frac{1}{2} \right)^1 \) or 0.125.

To avoid bias in the amount of information available on cases and controls, we also require controls to have at least 3 generations of genealogy data available. Controls are randomly selected from the UPDB and matched by cohort, as described earlier. For a set of matched controls, the GIF statistic is calculated precisely as described above for cases. Because the distribution of kinship and the GIF statistic for the controls will vary depending on the controls selected, 1,000 different control sets are selected for each case GIF calculation and a control GIF is calculated for each control set. A mean control GIF based on the 1,000 controls sets is calculated. The distribution of the 1,000 control GIF provides an empirical distribution from which the case GIF is assessed for statistical significance. A number of different cancer studies have used the GIF previously.

For the GIF analyses, each morphology-specific subtype was considered. The impact of age at diagnosis was evaluated by stratifying the population into those cases ≤50 years of age at diagnosis and those >50 years at diagnosis. Lobular carcinoma in situ was excluded from the GIF analyses due to the small number of cases.

**Results**

Table 1 shows the number of females with primary breast cancer listed by morphology-specific type in the entire Utah Cancer Registry data set and those breast cancer cases analyzed in our study (i.e., all breast cancer cases in the UPDB for whom at least three generations of pedigree data were available). There were 22,519 breast cancer diagnoses reported to the Utah Cancer Registry since its inception, with genealogy data available for 12,996 cases (57.7%). Lobular breast cancer accounted for 6.5% (1453 cases) of the total breast cancers in Utah and genealogy was available for 54.6%. The lobular breast cancer frequency present in the UPDB and UCR is comparable to other large cancer registries. Comparing median age at diagnosis, median survival rates, and lobular cancer;
TABLE I – DEMOGRAPHICS OF FEMALES WITH BREAST CANCER IN THE UTAH CANCER REGISTRY (UCR) AND THE UTAH POPULATION DATABASE (UPDB) WITH AT LEAST THREE GENERATIONS OF GENEALOGY DATA AVAILABLE

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Age (years)</th>
<th>UCR</th>
<th>Median age at diagnosis (years)</th>
<th>Median survival (months)</th>
<th>UPDB</th>
<th>Median age at diagnosis (years)</th>
<th>Median survival (months)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>All breast cancer</td>
<td>22.51</td>
<td>100.0</td>
<td>61</td>
<td>79</td>
<td>12.99</td>
<td>100.0</td>
<td>62</td>
<td>81</td>
</tr>
<tr>
<td>Early onset ≤50</td>
<td>6,130</td>
<td>27.2</td>
<td>44</td>
<td>98</td>
<td>3,276</td>
<td>25.2</td>
<td>44</td>
<td>96</td>
</tr>
<tr>
<td>Lobular</td>
<td>1,453</td>
<td>6.5</td>
<td>61</td>
<td>80</td>
<td>793</td>
<td>6.1</td>
<td>62</td>
<td>79</td>
</tr>
<tr>
<td>Early onset ≤50</td>
<td>402</td>
<td>1.8</td>
<td>46</td>
<td>123.5</td>
<td>201</td>
<td>1.6</td>
<td>45</td>
<td>121</td>
</tr>
<tr>
<td>in situ</td>
<td>274</td>
<td>1.2</td>
<td>52</td>
<td>126</td>
<td>146</td>
<td>1.2</td>
<td>53</td>
<td>122.5</td>
</tr>
<tr>
<td>Invasive</td>
<td>1,273</td>
<td>5.2</td>
<td>63</td>
<td>73</td>
<td>643</td>
<td>5.3</td>
<td>64</td>
<td>73</td>
</tr>
<tr>
<td>Early onset ≤50</td>
<td>277</td>
<td>1.2</td>
<td>46</td>
<td>1-6</td>
<td>139</td>
<td>1.1</td>
<td>45</td>
<td>107</td>
</tr>
</tbody>
</table>

1 Each morphological category lists unique individuals

TABLE II – RISK OF BREAST CANCER AMONG RELATIVES OF ANY BREAST CANCER CASES IN THE UPDB

<table>
<thead>
<tr>
<th>Cancer in proband</th>
<th>Cancer in relative</th>
<th>Relative of interest</th>
<th>Relatives, n</th>
<th>Observed</th>
<th>Expected</th>
<th>FRR</th>
<th>95%CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any breast cancer</td>
<td>Any breast</td>
<td>First-degree</td>
<td>102,235</td>
<td>2,286</td>
<td>1,282.6</td>
<td>1.83</td>
<td>1.75-1.90</td>
</tr>
<tr>
<td>Any breast, early onset (≤50 years)</td>
<td>Any breast</td>
<td>Second-degree</td>
<td>284,053</td>
<td>2,416</td>
<td>1,897.0</td>
<td>1.27</td>
<td>1.22-1.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>First-degree</td>
<td>22,333</td>
<td>527</td>
<td>218.17</td>
<td>2.42</td>
<td>2.21-2.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>59,610</td>
<td>608</td>
<td>417.43</td>
<td>1.46</td>
<td>1.34-1.58</td>
</tr>
</tbody>
</table>

1 Obs. observed; Exp. expected; FRR, familial relative risk estimate. 2 All values significant at p < 0.05

Histology between the UCR and the UPDB shows that the breast cancer cases we analyzed were representative of the total Utah breast cancer population.

Table II lists the familial relative risk estimates for breast cancer and early onset breast cancer (i.e., ≤50 years) overall within the UPDB. First-degree and second-degree female relatives of breast cancer patients have a significantly elevated risk of breast cancer, consistent with previous reports. Breast cancer risk among both first- and second-degree relatives of early onset breast cancer probands is significantly higher than the risk among first- and second-degree relatives of all breast cancer cases diagnosed at any age.

The risk of breast cancer among relatives of lobular cancer probands is shown in Table III. First-degree relatives of lobular cancer probands, including early-onset probands, were found to have a significantly increased risk for any type of breast cancer (FRR = 2.47, 95% CI = 2.12-2.85; early onset FRR = 3.31, 95% CI = 2.40-4.46), and in particular, an increased risk for specifically lobular cancer (FRR = 4.51, 95% CI = 2.79-6.89; early onset FRR = 6.92, 95% CI = 2.54-15.07). These risk estimates for first-degree relatives of lobular cancer probands significantly exceeded the familial risks estimated for any breast cancer in first-degree relatives of any breast cancer cases (FRR = 1.83, 95% CI = 1.75-1.90, see Table II). Strikingly, the risk of lobular cancer for first-degree relatives of lobular cancer probands (FRR = 4.51, 95% CI = 2.79-6.89) was also significantly greater than the risk of breast cancer among probands with any early-onset breast cancer (FRR = 2.42, 95% CI = 2.21-2.63, see Table II). Risk estimates for any breast cancer among second-degree relatives of lobular cancer probands were also significantly increased, but they were not statistically significantly different from estimates for any breast cancer in second-degree relatives of any breast cancer cases (see Table II).

Because the set of lobular breast cancers includes in situ and invasive cancers, we calculated familial relative risk estimates separately for each subtype.
determine if the risks differed between the types. Table IV shows the results of the breast cancer risk among relatives of lobular carcinoma \textit{in situ} and invasive lobular cancer probands. First-degree relatives of both \textit{in situ} and invasive lobular cancer probands had an increased risk of any type of breast cancer that was significantly greater than that presupposed for relatives of any breast cancer cases. An increased risk of specifically lobular carcinoma and lobular carcinoma \textit{in situ} among relatives of lobular carcinoma \textit{in situ} cases was not observed. However, first-degree relatives of invasive lobular cancer probands had an increased risk of lobular cancer and specifically an increased risk of invasive lobular cancer, both of which exceeded the risk associated with having a family history of breast cancer. The risk of lobular cancer among relatives of invasive lobular cancer cases exceeded even the risk observed for relatives of those with any type of early-onset breast cancer. Second-degree relatives of invasive lobular cancer had a significantly increased risk of any breast cancer, but this risk did not exceed that associated with a family history of breast cancer.

Table V shows the GIF results for the various morphology-specific breast cancers. Excess relatedness was seen for the lobular, invasive lobular, and the all breast cancer groups. The relatedness of the lobular breast cancer cases, in particular invasive lobular cancer cases, was greater among those diagnosed at or before the age of 50 years as compared to those diagnosed after the age of 50 years.

A graphical display of the extent of excess familiality represented in the GIF results for lobular cancer is shown in Figure 1. The overall case GIF and control GIF have been partitioned into the relative contributions from relationships of differing genetic distances or path length. The overall case GIF and control GIF are the sums of the various contributions from each path length. As indicated in the graph, lobular cancer cases were observed to have more close and more distant genetic relationships than expected. The case GIF exceeded the control GIF out to a path length of 4, indicating that lobular cancer cases are more related than controls out to a third-degree relative relationship.

**Discussion**

The Utah Population Database is a unique resource, combining genealogy data with complete cancer registry data, which makes it possible to examine the familial nature of breast cancer in a large population. In particular, it is large enough to allow morphology-specific subtypes of breast cancer to be studied. We have examined the morphology-specific risks of lobular breast cancer among relatives of lobular cancer probands.

Our data suggest that relatives of lobular cancer probands are at increased risk of any type of breast cancer and specifically lobular breast cancer, with the lobular cancer risk significantly surpassing that for general early-onset breast cancer. Although the majority of the increased breast cancer risk can be ascribed to first-degree relatives, second-degree and even third-degree relatives were also at increased risk. For first-degree relatives of lobular cancer probands, the risk of breast cancer and specifically lobular cancer are greater than that presupposed by having a relative with any type of breast cancer. Other previous studies have shown that first-degree relatives of lobular cancer probands were at increased risk of breast cancer, but our study is the first to report familial relative risk estimates of specifically lobular breast cancer among relatives of lobular breast cancer probands, including early-onset subgroups.

Further investigation of the subdivisions of lobular cancer, namely, lobular \textit{in situ} and invasive lobular cancer cases, showed that morphology specific risk extended to invasive lobular cancer but not lobular \textit{in situ} cancer. A possible explanation for the lack of lobular carcinoma \textit{in situ} specific risk to relatives is the difficulty detecting lobular carcinoma \textit{in situ} by mammogram. The risk of any type of breast cancer among relatives of both invasive lobular cancer and lobular carcinoma \textit{in situ} probands were both significantly greater than the risk observed for having a family history of breast cancer. Previous studies, such as the CASH study of early-onset breast cancer cases (i.e., ages 20–54) showed an increased risk of breast cancer among relatives attributable to lobular \textit{in situ} cancer cases but not invasive lobular cancer cases. Lobular carcinoma \textit{in situ} typically occurs in premenopausal women, which is approximately 10–15 years younger than when invasive breast carcinoma occurs. As our study design did not include age restrictions, a more comprehensive population-based analysis could be achieved.

The implications of our results suggest that as female relatives in families with lobular cancer have an increased risk of any type of breast cancer, a more rigorous screening regime may be warranted for these individuals. Recent reports have shown that MRI screening seems to be more sensitive than mammography in detecting breast cancer. However, MRI screening for breast cancer has been shown to have only moderate specificity and thus is recommended only for those with increased inherited susceptibility. As a family history that includes lobular breast cancer increases breast cancer susceptibility, and as lobular breast cancer is notoriously difficult to detect by either physical exam or mammography, the use of MRI screening in lobular breast cancer families may be beneficial.
### TABLE III - RISK OF BREAST CANCER AMONG RELATIVES OF LOBULAR BREAST CANCER CASES

<table>
<thead>
<tr>
<th>Cancer in proband</th>
<th>Cancer in relative</th>
<th>Relative of interest</th>
<th>Relatives, n</th>
<th>Observed</th>
<th>Expected</th>
<th>FRR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lobular</td>
<td>Any breast</td>
<td>First-degree</td>
<td>6,144</td>
<td>186</td>
<td>75.5</td>
<td>2.47</td>
<td>2.12-2.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>19,047</td>
<td>162</td>
<td>117.3</td>
<td>1.38</td>
<td>1.18-1.61</td>
</tr>
<tr>
<td>Lobular</td>
<td>First-degree</td>
<td>6,144</td>
<td>21</td>
<td>4.67</td>
<td>4.51</td>
<td>2.79</td>
<td>6.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>19,047</td>
<td>7</td>
<td>6.2</td>
<td>1.13</td>
<td>0.45-2.32</td>
</tr>
<tr>
<td>Lobular, early onset (≤50 years)</td>
<td>Any breast</td>
<td>First-degree</td>
<td>1375</td>
<td>43</td>
<td>13.0</td>
<td>3.34</td>
<td>2.40-4.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>6</td>
<td>41</td>
<td>25.2</td>
<td>1.63</td>
<td>1.17-2.21</td>
</tr>
<tr>
<td>Lobular</td>
<td>First-degree</td>
<td>1,375</td>
<td>6</td>
<td>0.87</td>
<td>6.92</td>
<td>2.54</td>
<td>15.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>3,718</td>
<td>5</td>
<td>1.28</td>
<td>1.56</td>
<td>0.19-5.64</td>
</tr>
</tbody>
</table>

Values are significant at p < 0.05. - Result is significantly greater than the risk of breast cancer among relatives of probands with any type of breast cancer (see Table II). - Result is significantly greater than the risk of breast cancer among relatives of probands with early-onset breast cancer (see Table II). - Value masked to ensure confidentiality.

### TABLE IV - RISK OF BREAST CANCER AMONG RELATIVES OF _IN SITU_ AND _INVASIVE_ LOBULAR BREAST CANCER CASES

<table>
<thead>
<tr>
<th>Cancer in proband</th>
<th>Cancer in relative</th>
<th>Relative of interest</th>
<th>Relatives, n</th>
<th>Observed</th>
<th>Expected</th>
<th>FRR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lobular <em>in situ</em></td>
<td>Any breast</td>
<td>First-degree</td>
<td>1,107</td>
<td>36</td>
<td>12.3</td>
<td>2.93</td>
<td>2.05-4.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>3,480</td>
<td>23</td>
<td>23.05</td>
<td>1.00</td>
<td>0.63-1.50</td>
</tr>
<tr>
<td>Lobular <em>in situ</em></td>
<td>First-degree</td>
<td>1,107</td>
<td>5</td>
<td>0.77</td>
<td>3.89</td>
<td>0.80</td>
<td>0.30-13.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>3,480</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Lobular <em>in situ</em></td>
<td>First-degree</td>
<td>1,107</td>
<td>0</td>
<td>0.160.21</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>3,480</td>
<td>0</td>
<td>63.4</td>
<td>9.49</td>
<td>0</td>
</tr>
<tr>
<td>Lobular <em>invasive</em></td>
<td>Any breast</td>
<td>First-degree</td>
<td>5,026</td>
<td>151</td>
<td>3.9</td>
<td>2.38</td>
<td>2.02-2.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>15,645</td>
<td>142</td>
<td>5.1</td>
<td>1.30</td>
<td>1.26-1.76</td>
</tr>
<tr>
<td>Lobular <em>invasive</em></td>
<td>First-degree</td>
<td>5,026</td>
<td>7</td>
<td>4.1</td>
<td>1.39</td>
<td>0.56</td>
<td>0.28-2.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>15,645</td>
<td>7</td>
<td>8.9</td>
<td>4.53</td>
<td>2.97-6.74</td>
</tr>
<tr>
<td>Lobular <em>invasive</em>, early onset (≤50 yrs)</td>
<td>Any breast</td>
<td>First-degree</td>
<td>948</td>
<td>28</td>
<td>0.890.49</td>
<td>3.16</td>
<td>2.10-4.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>2,554</td>
<td>30</td>
<td>0.75</td>
<td>1.76</td>
<td>1.19-2.51</td>
</tr>
<tr>
<td>Lobular <em>invasive</em>, early onset (≤50 yrs)</td>
<td>First-degree</td>
<td>948</td>
<td>5</td>
<td>6.76</td>
<td>1.84</td>
<td>1.31</td>
<td>1.48-17.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>2,554</td>
<td>5</td>
<td>2.24</td>
<td>0.27</td>
<td>0.87-8.11</td>
</tr>
<tr>
<td>Lobular <em>invasive</em>, early onset (≤50 yrs)</td>
<td>First-degree</td>
<td>948</td>
<td>5</td>
<td>6.18</td>
<td>1.27</td>
<td>1.87</td>
<td>1.27-18.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Second-degree</td>
<td>2,554</td>
<td>5</td>
<td>2.65</td>
<td>0.32</td>
<td>0.32-9.58</td>
</tr>
</tbody>
</table>

Values are significant at p < 0.05. - Result is significantly greater than the risk of breast cancer among relatives of probands with any type of breast cancer (see Table II). - Result is significantly greater than the risk of breast cancer among relatives of probands with early-onset breast cancer (see Table II). - Value masked to ensure confidentiality.

### TABLE V - GENEALOGICAL INDEX OF FAMILIARITY ANALYSIS FOR BREAST CANCER

<table>
<thead>
<tr>
<th>Cancer in proband</th>
<th>Age (years)</th>
<th>Case GIF</th>
<th>Control GIF</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>All breast cancer</td>
<td>2.43</td>
<td>2.33</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td>&gt;50</td>
<td>1.87</td>
<td>1.73</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>Lobular</td>
<td>2.77</td>
<td>2.25</td>
<td>0.019</td>
<td></td>
</tr>
<tr>
<td>&gt;50</td>
<td>4.44</td>
<td>1.78</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>Lobular invasive</td>
<td>2.89</td>
<td>2.30</td>
<td>0.026</td>
<td></td>
</tr>
<tr>
<td>&gt;50</td>
<td>4.5</td>
<td>1.76</td>
<td>0.031</td>
<td></td>
</tr>
</tbody>
</table>

for earlier detection of breast cancer. Prospective studies will be needed to validate our conclusions.

Another implication of these results is that families with excess lobular cancer are more homogeneous and therefore might be advantageous to use in genetic studies. Two known breast cancer predisposing genes studies to localize novel breast cancer predisposing that may contribute to the excess familiality of lobular cancer observed are E-cadherin and BRCA2. To explore the possibility that these genes might be candidate lobular breast cancer susceptibility genes in the Utah high risk
The genealogical index of familiality (GIF) for lobular cancer cases and controls is shown. Each path length or coefficient of kinship exponent contributes to the overall GIF. The sum of each contribution equals the GIF score.

In a post-hoc analysis of our data, we did not observe a significant excess of stomach cancer among first-degree relatives of our lobular cancer probands (data not shown). Invasive lobular carcinoma tends to occur more frequently in BRCA2 associated breast cancers than BRCA1 tumor types, however rates of invasive lobular carcinoma attributable to a BRCA2 mutation do not differ significantly from sporadic cases. We examined the risk of other BRCA2 type cancers including prostate, ovarian and pancreatic cancers as well as melanoma among first-degree relatives of our lobular cancer probands in the UPDB. Only prostate cancer was found to be in significant excess, but these rates did not differ statistically from a similar comparison of first-degree relatives of ductal cancer probands (data not shown). Although we cannot rule out the possibility that some of the observed lobular clusters may be attributable to an E-cadherin or BRCA2 mutation, it is likely that other lobular breast cancer predisposition genes exist.

A number of genetic and molecular differences have already been observed between lobular carcinoma and the more common ductal carcinoma. Differences have been noted in the expression of a number of proteins including VEGF, cathepsin D, cytokeratin 8, vimentin, thrombospondin and cyclin A as well as E cadherin. Loss of chromosome 16q, the site of the E-cadherin gene, is much more common in invasive lobular cancer than in invasive ductal cancer. Recent microarray analyses comparing invasive lobular cancer and invasive ductal cancers suggest that genes involved in cell adhesion, lipid/fatty acid metabolism, immune/defense/stress response, electron transport, and nucleosome assembly are differentially expressed. Further subdivisions of invasive lobular cancer itself may also be warranted as observed by Zhao et al. in a microarray analysis of invasive lobular and invasive ductal cancer. Approximately half of their invasive lobular cancer cases closely resembled invasive ductal cancer cases whereas the other half differed in their global transcription programs. Certainly, further studies are needed to explore the genetic and molecular characteristics of lobular breast cancer, including addition of known breast cancer risk factors.

We are aware of several potential biases that should be discussed. First, whether morphology should be used to define cohorts with excess familiality remains a quandary. Breast tumors may be of a single histological type or a composite of types, with both benign and malignant cells present in the same tumor. Further, pathologist interobserver agreement for the diagnosis of various morphologic types is not optimal. Despite these limitations, tumor histology is an accepted standard for classifying patients for treatment regimens and histology codes are readily available for study. Second, it should be remembered that familial aggregation of cancers could be due to shared environmental factors as well as inherited genetic factors. However, excess familiality extending to second-and even third-degree relatives, as observed in our study for lobular cancer, provides evidence in
support of genetic factors playing an important role. Third, breast cancer incidence and mortality rates in Utah are lower than elsewhere in the USA. This most likely is due to increased parity and earlier age at first birth, which are more common in Utah. Both of these factors have been associated with reduced breast cancer rates. Our results need to be validated by other investigators using an adequate sample size to explore these morphological subgroups. Finally, comparing our current findings with our previous study in 1994, we have found that the average relatedness of cases (i.e., GIF values) in the current study are lower than reported previously. The UPDB has increased in size, resulting in lower average relatedness of the entire resource: mean relatedness values calculated in the current study should more accurately reflect the true population mean. Furthermore, the lobular cancer definition (ICD-0); that was utilized in this current study is more restrictive and this will contribute to the differences noted.

Conclusion
Our study provides a unique population-based assessment of the familiality of the morphology-specific subtypes of lobular breast cancer. We have shown that there is an excess familial clustering of lobular cancer. The high-risk lobular cancer pedigrees present in the UPDB, which contributed to the observed results, provide an excellent resource for future genetic studies. Furthermore, as families with lobular cancer also have an increased risk of any type of breast cancer that exceeds general breast cancer familial risk, a more rigorous screening regime, such as MRI screening, may be warranted for these individuals.

Acknowledgements We thank the 3 anonymous reviewers for their constructive comments. We also thank S. Backus for providing computer support for this project. Data collection for this publication was supported by a grant from the National Library of Medicine T15 L140724 (K.A.-B.); a University of Utah funding incentive seed grant (to L.A.C.A.) and a Susan G. Komen research grant DINOS021523 (to N.J.C.). Data collected for this publication was assisted by the Utah Cancer Registry supported by National Institutes of Health Contract NO1-PC-35141 Surveillance, Epidemiology and End Results (SEER) Program; with additional support from the Utah Department of Health and the University of Utah. Partial support for all datasets within the Utah Population Database (UPDB) was provided by the University of Utah Huntsman Cancer Institute.

References


A CAUTIONARY NOTE ON THE APPROPRIATENESS OF USING A LINKAGE RESOURCE FOR AN ASSOCIATION STUDY

Kristina Allen-Brady, James M. Farnham, Jeff Weiler, and Nicola J. Camp

Published: BMC Genetics 2003; 4 Suppl 1:S89
A cautionary note on the appropriateness of using a linkage resource for an association study

Kristina Allen-Brady*, James M Farnham, Jeff Weiler and Nicola J Camp

Address: Department of Medical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, USA and Eccles Health Sciences Library, University of Utah, Salt Lake City, Utah, USA
Email: Kristina Allen-Brady - kristina.allen@hsc.utah.edu; James M Farnham - jim@genepi.med.utah.edu; Jeff Weiler - j.weiler@ncc.utah.edu; Nicola J Camp - nicki@genepi.med.utah.edu

Published: December 2003
BMC Genetics 2003, 4(Suppl 1) S89
This article is available from http://www.biomedcentral.com/1471-2156/4/s1/S89

Abstract

**Background:** Utilizing a linkage resource for association analysis requires consideration both of the marker data used and correlations among relatives in pedigrees. We previously developed a method for association testing in pedigrees. We applied our method to 50 replicates of microsatellite data surrounding five genes involved in high-density lipoprotein (HDL) in the Genetic Analysis Workshop 13 (GAW13) simulated data and examined association with HDL as well as linkage disequilibrium (LD) between markers.

**Results:** Although no association was intentionally simulated, we found significant evidence of weak LD between microsatellite markers (flanking/5 cM from the genes), in some but not all replicates. This level of LD compared well to that observed in the real GAW13 Framingham data. Only one region had sufficient replicates to assess power, and this was low (12.5–20.8%). More power was attained using all individuals and accounting for relationships, compared with one independent individual/pedigree, although this was not significant due to small sample sizes. Not accounting for relatedness inflated statistical significance (p < 0.0001).

**Conclusion:** A correction for dependence is necessary in association studies to avoid an inflation of significance probabilities. Our results further illustrate that use of microsatellite marker data is not an effective approach for association testing.

**Background**

Association and linkage disequilibrium analyses are tools that are often used to fine-map and refine promising linkage findings. These tools are particularly effective because they increase power to detect genes with small effects [1]. Many pedigree resources ascertained for linkage studies exist. However, the use of a linkage resource for an association analysis requires not only a methodology to address the issue of correlations among relatives in extended pedigrees but also consideration of the appropriateness of using microsatellite marker data previously genotyped for linkage analyses.

There is great potential for the utilization of pedigree resources for association analyses. Our group and others have developed methodologies to enable association analyses to be performed in extended pedigrees [2,3]. If multiple related individuals are included in a study without accounting for the relationships among them, there is an underestimate of the variance, which leads to an increased probability of a type I error. Slager and Schaid [2] approached the problem from a theoretical perspective and derived the correction necessary for a case-control association test. Using identity by descent (IBD) probabilities from a linkage analysis on marker data using GENEHUNTER [4], their method corrects the variance estimate to produce a correct significance assessment. Unfortunately, in moderate to large pedigrees, the calculation of the necessary IBD
probabilities cannot be done, since GENEHUNTER is limited in the pedigree size it can analyze. Our method can analyze pedigrees of arbitrary size by using an empirical approach to provide a valid statistical test for association [3].

For each simulation, our program uses a gene-drop to generate null genotype configurations for a given set of individuals in pedigrees. The gene-drop requires the assignment of genotypes to the pedigree founders, based on allele frequencies estimated from the study population. This is followed by the use of Mendelian inheritance probabilities to determine the genotypes of all descendants. The resulting genotype configuration on each pedigree represents a possible configuration under the null hypothesis of no association between allele and disease. Using the simulated genotypes for individuals for whom real data are available and the true phenotype data, the statistic of interest is calculated (for example, in this study a chi-squared test for independence). The resulting statistic is from the null distribution because it was derived from data under the null hypothesis of no association. This procedure is repeated many times, creating an empirical null distribution. The real genotype data for the same individuals and real phenotype data are then used to calculate the observed statistic. This observed statistic is then compared with the empirical distribution to determine significance. The method can be applied to pedigrees of any structure and size and to any statistic of interest.

Association analyses are based on either testing the true disease-causing variant itself, or alleles at a marker in linkage disequilibrium with the true variant. Since linkage disequilibrium (LD) generally extends over very short distances (typically < 100 kb, [5]) and since markers with multiple alleles impose multiple testing, markers for association analyses are usually chosen to be intragenic single-nucleotide polymorphisms (SNPs). In contrast, linkage markers are chosen to be highly polymorphic to maximize linkage informativeness and usually are at a resolution of 5–10 cM. Hence, marker choice for the two methods is distinct.

The Genetic Analysis Workshop 13 (GAW13) simulated data were modeled on the real GAW13 Framingham Heart Study data, as a genomic search with average resolution of approximately 9.5 cM, and hence no LD, or association, was intentionally simulated. To investigate the value, if any, of using linkage microsatellite marker data for association studies, we chose to analyze the simulated high-density lipoprotein (HDL) phenotype (dichotomized to the two extreme quartiles: high versus low quartile) and microsatellite markers either flanking or within approximately 5 cM of the true location of five genes involved in determination of baseline simulated HDL, as indicated in the ‘answers’.

Here we report the results of our empirical simulation method using a chi-squared association analysis for genotype data in the GAW13 simulated data. Although no association was intentionally included in the GAW13 simulated data, we report the underlying linkage disequilibrium (LD) between microsatellite markers and the ability of our empirical association method to detect power and false-positive signals. As a comparison we also report the LD in the Framingham data for the same location as one of the simulated baseline HDL genes.

Methods
The first 50 replicates of the simulated GAW13 data with missing genotype and phenotype values were utilized in this study. There were 330 families, ranging in size from 7 to 84 members, in each of the 50 replicates studied.

Genotype data
Eleven baseline genes (b12, ... , b22) contribute to the HDL phenotype. We chose to study five of these genes (b13, b14, b16, b18, and b22), selected to represent a spectrum of percent contribution to the HDL trait (see Table 1). For each gene of interest, microsatellite markers either flanking (and < 15 cM), or within ~5 cM of the true gene location were studied, resulting in two to four markers per gene region being analyzed. The average marker resolution ranged from 3.3 cM (b22) to 14.3 cM (b14). To reduce multiple testing, we selected only the three highest frequency alleles from each microsatellite marker for our analyses, and analyzed genotype distributions considering both dominant and recessive modes of inheritance.

Phenotype data
For all analyses, we used the maximum HDL measurement for each individual across the longitudinal study period as the trait of interest. We selected the covariates sex, age (at maximum HDL), BMI (using mean height and weight across the study period), smoking (ever/never), alcohol (ever/never), and fasting glucose (ever/taller > 126 mg/dl).

We considered the analysis of the HDL data in three different ways. First, we adjusted the maximum HDL value using the generating equation as provided in the ‘answers’ (GAW13_HDL(PED)). Second, we used our own linear regression equation to adjust HDL (LR_HDL) using the above listed covariates, where for each replicate we performed a linear regression with these covariates as independent variables to determine the regression coefficients. Third, we performed an ‘independent’ (IND) analysis by selecting only the first individual from each family with data and also using the GAW13 ‘answers’ (GAW13_HDL(IND)).

Linkage disequilibrium
The estimating haplotype (EH) program [6] was used to determine maximum-likelihood estimates of LD in all 50 replicates for markers in all five regions. Comparisons between all pair-wise combinations of
the microsatellite markers in the regions of interest were performed. The difference between the maximum-likelihood values calculated by EH for the haplotype frequencies under H₁ (allelic association allowed) and for the haplotype frequencies under H₀ (no association) were used to calculate the raw disequilibrium, from which D, the proportion of maximum possible disequilibrium, was determined. Significance was determined by the EH program.

To compare results of the simulated data to the GAW13 Framingham data, D for the three most common alleles at the first four markers on chromosome 4, corresponding to the exact position of the four markers surrounding gene h22 in the simulated data set, were analyzed.

It should be noted that the simulated data were generated to contain linkage to baseline HDL genes, but not LD or association with particular gene variants in the candidate genes. However, linkage and association differ only in the fact that the former is a phenomenon of loci and the latter of alleles at loci. In replicates where LD exists across markers, the simulated linkage creates a scenario equivalent to allelic heterogeneity for association. While certainly not ideal, allelic heterogeneity is a reasonable model for real data in which multiple common variants within a gene may increase disease susceptibility.

**Statistical analyses**

In all analyses we compared the highest and lowest quartiles of the HDL phenotype of interest. We used the chi-squared statistic for independence considering both dominant and recessive models as our statistic of interest. For each analysis 10,000 simulations were generated to create the empirical null distribution. To illustrate the necessity to correct for relatedness, we compared the results from GAW13_HDL(PED) with an analysis where we ignored genealogy and included all individuals (ALL) with data in the pedigrees without any correction for relatedness. A paired t-test for the average of the -ln(p) across markers over the 50 replicates was used to compare the results.

We report the number of replicates showing LD at various levels of significance for each gene region. Only one region surrounding gene h22 had sufficient replicates with significant LD to assess power. For this region we also show the percentage of the replicates with significant LD for which a significant association was found (p < 0.01) for each of the three analysis types, indicating a power estimate for each. Two regions (surrounding b14 and b18) had sufficient replicates without LD (p > 0.5) to assess false-positive findings. For these regions we show the proportion of replicates without LD for which significant associations were found (p < 0.01).

All regression analyses (for LR_HDL) and classic chi-square analyses (for GAW13_HDL(IND)) were performed using STATA 6.0 (College Station, Texas). Fisher's exact p-values are reported for the GAW13_HDL(IND) analyses, where appropriate.

**Results**

For the 'independent' (IND) analysis the average sample size per replicate with data was 328.7 individuals and ranged between 324 and 330 individuals. For all other analyses using all individuals with available data, the average sample size per replicate was 1672.6 individuals and ranged between 1627 and 1715. More than 95% of families within the 50 replicates had two or more members per family with data.

**Empirical method versus all individuals/pedigree without correction**

A comparison of the empirical method using GAW13_HDL(PED) with the uncorrected GAW13_HDL(ALL) showed extreme statistical significance (t = -43.86, p < 0.0001), suggesting that not accounting for correlations among family members substantially underestimates the variance and inflates the significance, as expected.

**Linkage disequilibrium**

Although no association between microsatellite markers and the underlying genes was intentionally simulated, for every gene studied significant LD (p < 0.05) was present in at least one replicate (Table 1). The region surrounding gene h22 had the most replicates with significant LD, with nearly half (24/50) of the replicates analyzed indicating LD.

For the GAW13 Framingham data on chromosome 4 (equivalent to those markers surrounding h22 in the simulated data), values of D ranging from 9.7 to 47.6 were observed. These values compared well to those found for the simulated data (D = 9.4–31.8).

**Power: association findings for the region containing b22**

As the region containing b22 was the only one with significant LD present in sufficient replicates (24/50), we chose to further study power of the association analyses for only this region. Table 2 shows the number of replicates where a significant association (p < 0.01) was found using the various analyses. Power was low for all comparisons and ranged from 12.5–20.8%. Power was higher for analyses that corrected for covariates and that used pedigree data (average sample size = 1672.6, power = 20.8%), rather than 'independent' data (sample size = 328.7, power = 12.5%), although these differences are not statistically significant due to a small sample size (n = 24). However, 20.8 % is significantly different from 0.05 (type I error rate) as assessed by binomial distribution theory with n = 24 (p = 0.006).

**False positives: association findings for regions containing b14 and b18**

The number of false-positive results was assessed in
Table 1: Linkage disequilibrium across 50 replicates.

<table>
<thead>
<tr>
<th>Genes</th>
<th>b13</th>
<th>b14</th>
<th>b16</th>
<th>b22</th>
<th>b18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contribution to HDL trait</td>
<td>0.1</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
<td>0.005</td>
</tr>
<tr>
<td>Average marker resolution</td>
<td>5.52</td>
<td>14.33</td>
<td>7.14</td>
<td>3.30</td>
<td>5.45</td>
</tr>
<tr>
<td>No. reps with LD (p &lt; 0.05)</td>
<td>8</td>
<td>2</td>
<td>8</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>No. reps with LD (0.05 &lt; p &lt; 0.1)</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>No. reps with LD (0.1 &lt; p &lt; 0.2)</td>
<td>19</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>No. reps ‘without’ LD (p &gt; 0.5)</td>
<td>3</td>
<td>32</td>
<td>6</td>
<td>0</td>
<td>21</td>
</tr>
</tbody>
</table>

Table 2: Power: association findings for the region containing gene b22.

<table>
<thead>
<tr>
<th>Method of Analysis</th>
<th>No. Replicates Indicating Association</th>
<th>% Replicates with LD at p &lt; 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAW13_HDL (PED)</td>
<td>5/24</td>
<td>20.8</td>
</tr>
<tr>
<td>GAW13_HDL (IND)</td>
<td>3/24</td>
<td>12.5</td>
</tr>
<tr>
<td>LR_HDL (PED)</td>
<td>5/24</td>
<td>20.8</td>
</tr>
</tbody>
</table>

Table 3: False Positives: association findings for regions containing genes b14 and b18.

<table>
<thead>
<tr>
<th>Method of Analysis</th>
<th>No. Replicates Indicating Association</th>
<th>% Replicates with LD at p &lt; 0.05 (proportion)</th>
</tr>
</thead>
<tbody>
<tr>
<td>b14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAW13_HDL (PED)</td>
<td>4/32</td>
<td>0.125</td>
</tr>
<tr>
<td>GAW13_HDL (IND)</td>
<td>3/32</td>
<td>0.094</td>
</tr>
<tr>
<td>LR_HDL (PED)</td>
<td>4/32</td>
<td>0.125</td>
</tr>
<tr>
<td>b18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAW13_HDL (PED)</td>
<td>1/21</td>
<td>0.048</td>
</tr>
<tr>
<td>GAW13_HDL (IND)</td>
<td>1/21</td>
<td>0.048</td>
</tr>
<tr>
<td>LR_HDL (PED)</td>
<td>2/21</td>
<td>0.095</td>
</tr>
</tbody>
</table>

gene regions for which sufficient replicates were available and for which LD was not evident (p > 0.5). Two regions surrounding genes b14 and b18 had sufficient replicates (32/50 and 21/50, respectively) to assess false-positive findings. Results are shown in Table 3. By inspection, the number of replicates for which false positives were found differed at most by one across analyses. The observed rates of false-positive signals ranged from 0.048 (1/21) to 0.125 (4/32) and, due to small sample sizes, are not significantly different from each other, or from 0.05. These results are, of course, purely observational. To accurately assess false-positive rates thousands of replicates would be necessary.

Conclusions

Association-based analyses using extended pedigree data require a method that accounts for correlations among relatives. The substantial proportion of multiple related individuals per pedigree in this study necessitates such a correction. In this study, we have shown, as expected, that not correcting for the relatedness of family members resulted in a sharp inflation of the significance probabilities. A valid test can be derived by sampling only one individual from each pedigree or by using a correction method, such as our empirical approach, which determines the empirical significance accounting for the relatedness of individuals. Our results are consistent with the hypothesis that the increased sample size in using multiple individuals from extended pedigrees and accounting for their relatedness can increase power.

Although no association was intentionally simulated in this data set, we found significant LD between microsatellite markers for several gene regions, particularly the region containing gene b22, which we chose to study power. The result that the b22 region contained the most LD is perhaps not unexpected because markers in that region were at the highest density (average 3.30 cM). In fact, the frequency of detectable LD in replicates correlated
reasonably well with available marker density (see Table 1). However, the D' values were low, although approximately equal to those obtained in the real GAW13 Framingham data. Considering only the h22 region and replicates with significant LD, power was found to be low (≤21%). Again, perhaps this is not unexpected because LD values were low and h22 contributed only 1% to the genetic variance of HDL.

Considering all five gene regions studied, 80% (4/5) of the regions did not exhibit a high frequency of LD over the 50 replicates. Furthermore, as pointed out above, even for the one region with significant LD, power was low. These results could be extrapolated to association analyses in the real GAW13 Framingham data where LD was comparable and where small effect genes are expected. This indicates caution in the interpretation of any positive association findings in those data.

Microsatellite markers are not an ideal marker type for association testing. We selected the nearest two to four markers for each gene (each either flanking or within ~5 cM of the gene of interest). However, 5 cM represents a large distance between markers and true gene variants for association testing. Testing single-nucleotide polymorphism (SNP) variants within the gene, which are more likely to be in strong LD with disease-causing variants, would have been a more powerful situation.

A limitation of our analysis is that not all alleles were analyzed. Only the three highest frequency alleles for each marker were used to both increase the chance of an allele being observed sufficiently for association testing and to reduce multiple testing problems. This modified allele system is not biased for investigating LD, since alleles were not grouped according to their respective disequilibria. Further, we used the same allele definition for LD estimation and association, thus they are directly relevant to one another. We may, however, have missed positive association findings by not testing all alleles. A further limitation is that these data were not simulated to specifically contain association, and our power will be limited by the situation equivalent to allelic heterogeneity, which has been shown to reduce power in association-based analyses [7]. In addition, our study assessed the extreme quartiles for simulated HDL, and not the quantitative measure that was available. Our empirical approach, although theoretically completely general for any statistic, is not currently developed to analyze quantitative measures. This dichotomization may have reduced power to detect association.

In conclusion, our results indicate that the empirical approach makes efficient use of data collected from extended pedigrees. Further, this study illustrates the lack of power available by using linkage microsatellite markers, and strongly cautions against doing so for association testing.

Acknowledgments
This work was supported by grant T15LM0724 from the National Library of Medicine (which supports KA-B as an NLM Fellow) and by grant R03 CA099844-01 from the National Cancer Institute (NJC).

References
PEDGENIE: AN ANALYSIS APPROACH FOR GENETIC ASSOCIATION TESTING IN EXTENDED PEDIGREES AND GENEALOGIES OF ARBITRARY SIZE

Kristina Allen-Brady, Jathine Wong, Nicola J. Camp

Submitted to BMC Bioinformatics November 2005
Abstract

Family-based designs have been suggested for association testing in complex diseases because they are robust to potential population stratification and admixture. Here, we present a general approach to perform association analyses in pedigrees of arbitrary size and structure to test genetic markers and qualitative or quantitative traits. Our software, PedGenie, uses Monte Carlo significance testing to provide a valid test for related individuals that can be applied to any test statistic, including transmission disequilibrium statistics. Single locus at a time, composite genotype tests, and haplotype analyses may all be performed. We illustrate the validity and functionality of PedGenie using simulated and real data sets. For the real data set, associations of tagging-single nucleotide polymorphisms (tSNPs) in the DNA repair gene, \textit{NBS1}, and breast cancer are analyzed in high-risk Utah breast cancer families. The results from PedGenie were shown to be valid both for accurate \textit{p}-value calculations and consideration of pedigree structure in the simulated data set. A significant association with breast cancer was observed with the \textit{NBS1} tSNP rs709816 for carriage of the rare allele (OR = 1.61, 95\% CI = 1.10-2.35, \textit{p}=0.019). In conclusion, PedGenie is a flexible and valid statistical tool that is intuitively simple to understand, makes efficient use of all the data available from pedigrees without requiring trimming, and is flexible to the types of tests to which it can be applied. Further, our analyses of real data indicate \textit{NBS1} may play a role in the genetic etiology of breast cancer.
Background

Family-based tests of allelic association have received much attention recently due in part to their robustness against a bias that may arise in population data. Population based association studies may be hindered by genetic stratification that exists even within relatively homogeneous populations,\(^1\),\(^2\) resulting in spurious associations that markedly increase with sample size.\(^3\),\(^4\) Family-based association studies that focus on the preferential transmission from parents to affected offspring are robust to the effects of population stratification, admixture, and nonrandom mating.\(^5\)-\(^7\) and hence avoid the bias of population stratification. Additional advantages of family-based designs include error checking of genotype data, the capacity to detect parent-of-origin effects, and the improved efficiency and cost effectiveness of being able to perform association studies on cases previously ascertained for linkage studies.\(^8\)

Use of family data for allelic association testing includes the additional challenge of accounting for correlations between related individuals. Without correction for the genetic dependence of related individuals, an underestimate of the variance of a desired statistic and an increase in type I error will result.\(^9\) A number of family-based association tests that appropriately account for correlations between related individuals have been developed, many of which are extensions of the transmission disequilibrium test (TDT) (See review of family based methods\(^10\)). The TDT was originally proposed as a test for linkage disequilibrium in family trios, consisting of two parents (at least one heterozygous) and an affected offspring.\(^11\) The TDT has been extended to allow for the use of siblings instead of parents\(^12\)-\(^16\) (see also review\(^17\)); multiallelic markers,\(^18\)-\(^21\) and extensions to quantitative traits and
covariates.\textsuperscript{17,22-27} Most TDT based tests, however, are limited, as they focus on only small nuclear families and/or a particular genotype configuration to be valid tests.

Extensions of association and linkage disequilibrium tests to multi-generational pedigrees also exist.\textsuperscript{25,28-37} However, many of these methods restrict analyses to either dichotomous dependent variables\textsuperscript{28,29,35,37} or quantitative traits.\textsuperscript{22,30,32,36} For those methods that analyze quantitative traits, some require the trait to be normally distributed.\textsuperscript{22,30,32} Others limit the size of a pedigree that can be studied because they require calculation of inherited by descent (IBD) parameters;\textsuperscript{30,32,35} although the pedigree size for which IBD can be determined is increasing,\textsuperscript{38} pedigree size, for particularly large extended pedigrees or genealogical resources, still remains a limiting factor. Other methods decompose large pedigrees into nuclear families;\textsuperscript{4,25,28,29,31} if these methods are applied to a small sample of large pedigrees, this will lead to a substantial loss of information and power, and may risk inflation of false positive results.

A number of large and sometimes complex genealogy-based populations are being ascertained to study complex traits, including the Hutterites, the Pima Indians, Sardinia isolates, as well as studies in Iceland and Utah. Utilizing all cases in a pedigree or genealogy for an association study rather than just one case or nuclear family per pedigree (if one adopts an ‘independence’ protocol) increases power by increasing the effective sample size.\textsuperscript{9,37} Ideally, analyses utilizing all available information on pedigree structure are most informative and efficient.

We previously proposed a method that could perform valid simple case-control association analyses on extended pedigrees of arbitrary size and structure using an
empirical approach.⁹ Here, we present an expanded more versatile software tool, PedGenie that incorporates a number of common association and transmission disequilibrium statistics for qualitative and quantitative data. PedGenie accounts for correlations between related individuals, based on the original pedigree structure. A Monte Carlo approach is employed to generate an empirical null distribution from which significance for a statistic of interest can be determined. Pedigrees of arbitrary size or structure, from singletons (single, independent individuals) to entire genealogies with loops can all be analyzed simultaneously. Furthermore, our method is able to analyze multi-locus data either as composite genotypes (multiple insult hypothesis, where phase is not considered) and haplotypes (where phase is important), assuming that the correct haplotypes have been assigned to pedigree members. We illustrate the validity and utility of PedGenie through its application to simulated data as well as test the association between two tagging-SNPs in the DNA repair gene, \textit{NBS1}, and breast cancer.

\textbf{Implementation}

Our software tool, PedGenie, was written in Java version 1.4 and is freely available (see Availability and requirements). We begin with a description of its functionality for inheritance of a biallelic marker, although all analyses can be performed with multiallelic markers, followed by an explanation of how PedGenie handles haplotype data. Assume that we have a resource of large extended pedigrees containing cases with a particular disease. Although the genealogy of individuals may be known, typically genotype and perhaps phenotype data will be available only on individuals near the bottom of the pedigree, yet relationships between individuals are
still known. Under the null hypothesis of no association, the genotype of an affected individual in the pedigree is independent of their disease status. Based on this null hypothesis, new genotypes for individuals in the pedigree may be generated, which we term a "null genotypic configuration"; both relationships within the pedigree and the missing data structure are maintained in the null genotypic configuration. New traits, however, are not generated for individuals, thus also maintaining the original phenotypic familial correlations. Through generating multiple null genotypic configurations and calculating a statistic of interest for each, an empirical null distribution can be created for the statistic and this null distribution can be used to determine significance of an observed statistic (calculated from the real genotypic data).

An outline of steps for PedGenie is as follows. First, allele frequencies for the markers of interest are estimated from the data. These can be estimated by four different methods: genotyped founders only, all genotyped individuals, all founders with statistically inferred genotypes and user-defined. For the first option, allele frequency estimation is made using data from genotyped founders only; hence it is representative of the general population from which the study sample was obtained but requires that many of the founders have genotype data available. The second option uses all genotyped individuals, ignoring dependence between relatives, for which the point estimate is unbiased. The third option statistically infers the founder genotypes using maximum likelihood estimation, and then calculates allele frequencies from the inferred (or actual, if observed) genotypes of all the founders. We recommend statistically inferring founder genotypes if there are a small to moderate number of relatively large pedigrees. If there are a large number of relatively small
pedigrees and the number of genotyped founders in the resource is limited, we recommend the ‘all’ option, which is the current default in PedGenie. The final option allows a user to enter allele frequencies.

Second, alleles are assigned to the pedigree founders randomly, but in proportion to these allele frequencies, and a Mendelian gene-drop is performed. That is, the gene-drop is performed independent of trait information. For example, alleles A and B will be transmitted from a parent with genotype AB to their children with equal probability. The resulting null genotype configuration on each pedigree therefore represents a possible configuration under the null hypothesis of no association between allele and disease.

Third, the statistic of interest is calculated using the null genotype configuration and the true phenotype data. This resulting statistic, $S_n$, is from the null distribution since it was derived from data simulated under the null hypothesis. Steps 2 and 3 are repeated multiple (N) times, and the series of null statistics stored. Hence an empirical null distribution is created for the statistic of interest, conditional on the particular pedigree and phenotype structures contained in the resource.

Finally, the observed statistic, $S_o$, is computed based on the true genotype and phenotype data using the same statistic of interest. This observed statistic is then compared to the empirical null distribution to determine significance as follows:

$$p = \sum_{i=1}^{N} \left[ I(S_i) \right] / N$$

where:

$$I(S_i) = \begin{cases} 1 & \text{if } S_i \geq S_o \\ 0 & \text{otherwise} \end{cases}$$

The specified null hypothesis is rejected if the $p$-value is less than or equal to the required level of type I error ($\alpha$). The accuracy of the empirical $p$-value increases with
the size of the empirical null distribution simulated. An N of 2,000 gives a 95% confidence interval around \( \alpha = 0.05 \) with a width of 0.02 under the null hypothesis. It is possible that the statistic of interest cannot be calculated, such as when the data are too sparse (e.g., an inability to calculate an odds ratio due to a zero count in a contingency table). PedGenie provides information on the number of simulations for which a statistic can be calculated. If the number of calculated statistics is less than the total number of simulations (i.e., N), this would suggest sparse data for that particular analysis and caution is advised when interpreting the results.

To match the information content of the real data to that of the simulated data, we limit calculation of the statistic of interest in the simulated data to only those individuals with genotype data in the observed sample. In the gene-drop procedure, genotype information is initially simulated for all members of a pedigree; however, those individuals for whom observed genotypic data were not available for a specific locus are reset to missing. In this way, the missing data structure is captured.

**Composite genotype and haplotype-based analyses**

Testing multiple loci simultaneously either as composite genotypes or haplotypes is similar to that described above with a few exceptions. For both the composite genotype and haplotype analyses, haplotype frequencies are entered in place of allele frequencies for the gene drop. The gene drop proceeds as above, except haplotypes rather than alleles are dropped through the pedigrees. PedGenie allows the user to enter recombination rates (i.e., \( \theta \)) between markers and these values are used to determine recombinant events for generation of the empirical null distribution. For composite genotype testing and haplotype testing, user-defined population haplotype
frequency estimates are required. The difference between composite genotype and haplotype analyses is that phase information on the observed genotype data is required for the haplotype analysis but is not required for the composite genotype analysis. For haplotypes, PedGenie expects the pedigree genotype data to be ordered. A number of pedigree based haplotype methods are being developed. However, most of these haplotype methods are currently unable to provide both haplotype frequencies and individual haplotype assignment on large pedigrees with large amounts of missing data, assuming linkage disequilibrium between multiple markers; (see also hence additional work in this area is required.

Statistics

Our approach is general, such that an empirical null distribution may be computed for any statistic of interest. PedGenie version 1.2 currently incorporates statistics for both dichotomous and quantitative data outcomes. For data involving a dichotomous variable outcome, the implemented statistics are: basic genotype-based or allele-based chi-square statistics (for arbitrary number of categories that are user-specified), an odds ratio statistic both for allele counts and genotype data (with 95% confidence intervals determined from the empirical distribution), and a chi-square trend statistic with user-defined weights. In addition to these classical tests of association, our method is also able to test transmission-disequilibrium on large extended pedigrees, including the TDT statistic for trios, sibships, and combined trios/sibships. For quantitative dependent variable outcomes, a standardized difference in means statistic and an overall analysis of variance (ANOVA) are implemented. PedGenie also can test quantitative TDT statistics based on methods of Allison (TDT_{Qs}), Rabinowitz, and
Monks. Covariate data can be incorporated into these quantitative TDT models. The quantitative TDT statistics are made available by interfacing PedGenie with the freely available QTDT software. In brief, the real data and each simulated genotypic null distribution are communicated to the QTDT package, and results are parsed and summarized over all simulations within PedGenie and an empirical p-value is calculated as defined above.

**Defining genotype, allele, and haplotype groups**

PedGenie has been developed to be general with respect to how genotypes, alleles, combinations of genotypes, or haplotypes are grouped together for an analysis. In the simplest case, PedGenie is able to analyze data by a single locus at a time approach, comparing for example, in a biallelic system, carriers of a rare allele ‘2’ (genotype = 2/2 or 1/2 or 2/1) to individuals with the wild type genotype (1/1). User-defined weights can be assigned to these genotype groups and analyzed using any of the above listed statistical tests. The number of separate groups compared is user-defined. Similarly, multiple alleles may be considered as the unit for analysis and grouped together as necessary. For example, in a multiallelic system, allele 1 could be a group, allele 2 could be another group, and all other alleles could be considered a third group. For composite genotype tests (involving genotypes at multiple loci), PedGenie allows the user to define groups based on genotypes across multiple loci. For example, a user may wish to analyze individuals who are carriers of a rare allele at locus 1 and homozygous at locus 2 compared to individuals with all other genotype combinations. For haplotype-based tests, analogous to the allele tests, the user can define multiple haplotypes in a single group and compare this group to other groups of haplotypes.
Similarly to the single locus case, the number of groups to be compared remains user-defined.

Single locus genotype, composite genotype, and haplotypes may be tested using any of the statistics available, including TDT, sibTDT, and the combined TDT for dichotomous traits. However, QTDT analyses are limited to the capabilities of the QTDT package.

Functionality of PedGenie

We illustrate the validity and functionality of PedGenie using three different data sets. The first two data sets are simulated data to demonstrate that the techniques employed by PedGenie are valid and robust. In the final data set, we illustrate PedGenie’s functionality and ability to handle large, extended pedigrees with real data for breast cancer and two tagging-SNPs in the NBS1 gene.

Validation of statistics with simulated independent data

The purpose of this first data set was simply to illustrate that in set of independent individuals and a set of independent nuclear families that PedGenie computes an empirical \( p \)-value that corresponds to those from the appropriate standard statistical distribution. The standard distribution is defined as the distribution that typically would be used for a particular test (e.g., the standard distribution for the chi-square statistical test is the chi-square distribution).

In the first data set, we simulated data for a biallelic marker with a minor allele frequency of 0.2. We generated a set of 2,000 independent cases and 2,000 independent controls as well as 4,000 independent nuclear families with two offspring. For the
nuclear families, we assumed that the two parents were independent and the offspring, one of whom was affected, inherited alleles in a Mendelian fashion. No association between the biallelic marker and affected status was simulated. We also generated a quantitative trait for all individuals that was randomly assigned and normally distributed with a mean of 50 and a standard deviation of 10. The 2,000 independent case and control data were used to validate the chi-square, chi-squared trend, odds ratio, the standardized difference in means statistic and ANOVA. The 4,000 independent nuclear family data were used to test the TDT, sib-TDT, combined trio/sib TDT, and the combined trio/sib QTDT method by Monks. Each of these validations was run 1,000 times on PedGenie using a different initial random number seed. For each PedGenie analysis, the empirical null distribution and p-value were determined from a sample size of 2,000 null configurations and the allele frequency estimation method ‘all’.

Validation of inheritance with simulated pedigree data

The second data set was to illustrate that PedGenie appropriately accounts for relationships within a pedigree structure when performing the gene drop. We used simulated data obtained from the 12th Genetic Analysis Workshop (GAW12) and compared empirical p-values from PedGenie to an exact pedigree-based method proposed by Slager and Schaid based on the Armitage trend association statistic.

The Slager and Schaid method based on the Armitage test for trend accounts for relatedness of individuals by measuring a trend in proportions according to a general measure of genetic dosage, \( x \), where \( x \) is a vector of weights for each genotype. The Armitage test for trend degenerates to the standard chi-squared test for independence assuming a dominant or recessive mode of inheritance when \( x = (0, 1, 0) \) or \( x = (0, 0, 1) \).
1), respectively, where the three indices in the vector \( \mathbf{x} \) represent the wild type, heterozygote and homozygote genotypes. The Slager and Schaid method, in brief, accounts for correlations between relatives by correcting the variance estimate using a correlation matrix that is a function of posterior inherited by descent (IBD) sharing probabilities, estimated by GENEHUNTER.\(^{50}\) However, prior probabilities using kinship coefficients may also be substituted into the correlation matrix. In this analysis, we compared PedGenie to the Slager and Schaid method incorporating prior probabilities, as these are the probabilities that are sampled using a Mendelian gene-drop.

The GAW12 simulated dataset provided complete phenotypic and genotypic data for 23 extended pedigrees of 1,000 living individuals. Complex relationships between covariate data (quantitative traits, a disease liability, and age-at-onset) and gene sequence variants were also simulated.\(^{48}\) We selected a single replicate (Replicate 42) and analyzed the association of all variants with a minor allele frequency \( \geq 0.01 \) in the \( MG1 \) gene and \( Q1 \), a quantitative trait; 24\% of the variance of \( Q1 \) was attributable to \( MG1 \). The effects of other covariates, including age, sex, and an environmental component (EF1) were regressed out of \( Q1 \) using linear regression prior to the analyses. Individuals in the top tertile (333 individuals) for the residual genetic component of \( Q1 \) were designated cases, and those in the bottom tertile (\( N=333 \)) controls. For these analyses, we considered a dominant and recessive mode of inheritance for each variant studied. We compared results from PedGenie using the chi-square trend statistic, with weights designed to test a dominant and recessive model, to the Slager and Schaid exact method using prior IBD probabilities in the correlation matrix. The empirical null
distribution and \( p \)-value were determined from a sample size of 2,000 null configurations and the allele frequency estimation method ‘all’.

Testing of the \( NBS1 \) gene and breast cancer

The functionality and ability of PedGenie to handle large extended pedigrees was illustrated using individuals selected from 139 Utah high-risk breast cancer pedigrees of Northern and Western European descent, with family size ranging from 1 (a single individual) to 1,195 individuals in a single pedigree; however, typically only individuals at the bottom of each pedigree were genotyped. Individuals were selected because they belonged to pedigrees with rates of breast cancer exceeding the population rate, and their breast cancer most likely was not attributable to a \( BRCA1/2 \) mutation, due to either negative \( BRCA1/2 \) test results for case(s) within the family or a low probability of a \( BRCA1/2 \) mutation based on the number of breast cancer cases present and/or ages at diagnosis of breast cancer within the family. Breast cancer diagnosis information was obtained from either medical records for the subjects or the Utah Cancer Registry. Previously we characterized the linkage disequilibrium (LD) structure across the \( NBS1 \) gene and identified two tagging-SNPs (tSNPs),\(^5\) and these two tSNPs have been genotyped on our entire study population (see\(^5\) for genotyping details). Where possible, Mendelian inheritance was verified; samples with inheritance incompatibilities were either re-genotyped and/or set to missing if they could not be resolved. Here, we analyzed associations with these two tSNPs, rs12680687 (minor allele frequency = 0.28) and rs709816 (minor allele frequency = 0.45), and breast cancer status as well as for breast cancer cases, age-at-diagnosis.
From the 139 breast cancer pedigrees, two cohorts were defined: a Case-Control cohort and a Nuclear Family cohort. The Case-Control cohort was used to illustrate the more standard association-based tests available in PedGenie, while the Nuclear Family cohort was used to illustrate the transmission/disequilibrium statistics. The Case-Control cohort was composed of 236 breast cancer cases matched to 236 controls based on birth year (within 5 years), female gender, and age-at-diagnosis, such that the control was cancer free at the age the case was diagnosed. Controls were selected to be as distantly related to any other case or control as possible to increase power, and as old as possible while remaining breast cancer free. Despite selecting the most distantly related individuals as possible, it should be noted that the Case-Control cohort contains related individuals, and these relationships for the purpose of association testing must be taken into account. For the Nuclear Family cohort, we selected 39 parent/affected offspring trios, with the nontransmitted alleles from the two parents serving as controls, and 167 sibships each containing at least one affected and one unaffected sibling, with the unaffected sibling(s) serving as control(s). Blood samples were collected on all subjects and all individuals gave informed consent. This study was approved by the University of Utah Institutional Review Board.

We illustrate single locus analyses for both tSNPs, composite genotype tests across both loci in order to test a multiple insult hypothesis, and haplotype-based analyses. Haplotype frequencies of the two tSNPs for the composite genotype and haplotype analyses were determined using 94 unrelated breast cancer cases (n=47) and controls (n=47) selected from the 139 breast cancer pedigrees using an expectation-maximization algorithm. Given that these tSNPs are only ~16 kb apart, the
recombination fraction in PedGenie was set to zero for the gene-drop. For haplotype analyses, phased genotypes were estimated using an expectation maximization (EM) algorithm. Only haplotypes that could be assigned to an individual with >80% probability were accepted. As the EM algorithm is designed for unrelated individuals, all assigned haplotypes were checked for segregation within a family wherever possible. Haplotypes that were incompatible within a family were set to zero.

We accounted for multiple comparisons of the data by correcting for two tSNPs tested on the \textit{NBS1} gene and two phenotypes (i.e., breast cancer status and age-at-diagnosis). Thus, our significance threshold was set to $p < 0.0125 (0.05/4)$. The empirical null distribution and $p$-value were determined from a sample size of 2,000 null configurations and the allele frequency estimation method ‘GeneCounter’.

Results

Validation of statistics with simulated independent data

Table 4.1 shows the results comparing PedGenie $p$-values to those derived from the standard distributions for simulated data. Using both independent case-control and nuclear family data, the empirical $p$-value results from PedGenie compared well to results from the standard distribution, illustrating that for independent samples, the Monte Carlo simulation employed by PedGenie is valid. Several $p$-values from PedGenie were estimated precisely to three decimal places when compared to the standard distribution, the majority were within 0.001, and none were significantly different from that expected from the standard distribution.
Table 4.1. Comparison of PedGenie to the standard distribution using simulated data for independent individuals and nuclear families

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Chi Square(^2)</th>
<th>Chi Square Trend(^2)</th>
<th>HET vs WT(^2)</th>
<th>HOM vs WT(^2)</th>
<th>HET vs WT(^2)</th>
<th>HOM vs WT(^2)</th>
<th>ANOVA(^2)</th>
<th>Trio TDT(^2)</th>
<th>Sib TDT(^3)</th>
<th>Comb TDT(^3)</th>
<th>QTDT(^4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value standard</td>
<td>0.292</td>
<td>0.288</td>
<td>0.968</td>
<td>0.949</td>
<td>0.149</td>
<td>0.130</td>
<td>0.017</td>
<td>2.463</td>
<td>1.667</td>
<td>1.563</td>
<td>1.77</td>
</tr>
<tr>
<td>distribution</td>
<td>0.864</td>
<td>0.592</td>
<td>0.632</td>
<td>0.754</td>
<td>0.882</td>
<td>0.896</td>
<td>0.983</td>
<td>0.117</td>
<td>0.096</td>
<td>0.118</td>
<td>0.077</td>
</tr>
<tr>
<td>95% CI standard</td>
<td>-</td>
<td>-</td>
<td>0.846-1.307</td>
<td>0.682-1.320</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>distribution</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mean empirical p-value</td>
<td>0.865</td>
<td>0.591</td>
<td>0.633</td>
<td>0.749</td>
<td>0.882</td>
<td>0.896</td>
<td>0.983</td>
<td>0.117</td>
<td>0.087</td>
<td>0.118</td>
<td>0.078</td>
</tr>
<tr>
<td>(SD)</td>
<td>0.008</td>
<td>0.012</td>
<td>0.021</td>
<td>0.021</td>
<td>0.007</td>
<td>0.007</td>
<td>0.003</td>
<td>0.008</td>
<td>0.006</td>
<td>0.007</td>
<td>0.006</td>
</tr>
<tr>
<td>Empirical p-range</td>
<td>0.844-0.891</td>
<td>0.550-0.625</td>
<td>0.572-0.695</td>
<td>0.676-0.826</td>
<td>0.860-0.903</td>
<td>0.870-0.913</td>
<td>0.972-0.991</td>
<td>0.095-0.108</td>
<td>0.097-0.109</td>
<td>0.055-0.057</td>
<td>0.096</td>
</tr>
<tr>
<td>95% CI Empirical</td>
<td>-</td>
<td>-</td>
<td>0.846-1.107</td>
<td>0.685-1.314</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(Mean values)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

1 For each statistical test, PedGenie was run 1,000 times to compute the average and range of empirical p-values. The standard distribution is the standard distribution used for the particular statistical test (e.g., the p-value reported for the Chi-square test is from the chi-square distribution). HOM = homozygous for minor allele, HET = heterozygous, WT = wild type.

2 Test run using 2,000 independent, unrelated cases and 2,000 independent, unrelated controls

3 Test run using 4,000 independent nuclear families, composed of either trios (2 parents and one affected offspring), sib-pairs (one affected and one unaffected), or for the combined analysis, a combination of both trios (2,000) and sib-pairs (2,000).

4 QTDT analysis run using method of Monks et al.\(^{17}\)
Validation of inheritance with simulated pedigree data

Table 4.2 shows the results comparing the chi-square trend test (weighted to provide dominant and recessive tests) from PedGenie to the Slager and Schaid Armitage trend method using GAW12 simulated pedigree data. Overall, PedGenie compared well. Again, a p-value to three decimal places matched precisely between PedGenie and the Slager and Schaid method, and other results were not significantly different. There was a tendency for the significance results from PedGenie to be more conservative. The Armitage test for trend asymptotically follows a chi-squared distribution with one degree of freedom and some appreciable discrepancies between the Armitage statistic significance probabilities, assuming the asymptotic distribution, and the exact Binomial probabilities have been noted for the lower end of the distribution. Overall, these results illustrate that PedGenie accounts for relationships between individuals in an appropriate manner.

Testing of the \textit{NBS1} gene and breast cancer

For the breast cancer real data set, overall there were 3,762 individuals from 139 high-risk breast cancer pedigrees. The median family size was 9; however, one family had 1,195 individuals and spanned 8 generations. In Table 4.3 we show the number of individuals with complete genotype data for the \textit{NBS1} gene and either the age-at-diagnosis (for cases) or the last known age that a control was cancer free for the Case Control and Nuclear Family cohorts. Blood samples and ultimately genotype data for the \textit{NBS1} gene were available for 1,034 individuals.
### Table 4.2. Results for PedGenie chi-square statistic and Slager and Schaid Armitage trend test using GAW12 data

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>MOI</th>
<th>Minor Allele Freq</th>
<th>D^3</th>
<th>r^2</th>
<th>PedGenie chi-square p-value</th>
<th>PedGenie Mean (SD)</th>
<th>Slager &amp; Schaid Trend p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>5782</td>
<td>D</td>
<td>0.092</td>
<td>--</td>
<td>--</td>
<td>&lt;0.001^4</td>
<td>0.010</td>
<td>2.06E-13</td>
</tr>
<tr>
<td>5782</td>
<td>R</td>
<td>0.092</td>
<td>--</td>
<td>--</td>
<td>0.232 (0.010)</td>
<td>0.205-0.263</td>
<td>0.213</td>
</tr>
<tr>
<td>5007</td>
<td>D</td>
<td>0.093</td>
<td>1.00</td>
<td>1.00</td>
<td>&lt;0.001^4</td>
<td>0.021-0.042</td>
<td>3.42E-13</td>
</tr>
<tr>
<td>5007</td>
<td>R</td>
<td>0.093</td>
<td>1.00</td>
<td>1.00</td>
<td>0.232 (0.010)</td>
<td>0.021-0.042</td>
<td>0.213</td>
</tr>
<tr>
<td>4848</td>
<td>D</td>
<td>0.40</td>
<td>1.00</td>
<td>0.064</td>
<td>0.030 (0.004)</td>
<td>0.663-0.723</td>
<td>0.688</td>
</tr>
<tr>
<td>4848</td>
<td>R</td>
<td>0.40</td>
<td>1.00</td>
<td>0.064</td>
<td>0.693 (0.010)</td>
<td>0.021-0.042</td>
<td>0.688</td>
</tr>
<tr>
<td>11146</td>
<td>D</td>
<td>0.029</td>
<td>0.84</td>
<td>0.003</td>
<td>0.219 (0.009)</td>
<td>0.021-0.042</td>
<td>0.212</td>
</tr>
<tr>
<td>11146</td>
<td>R</td>
<td>0.029</td>
<td>0.84</td>
<td>0.003</td>
<td>_5</td>
<td>_5</td>
<td>_5</td>
</tr>
</tbody>
</table>

PedGenie results for the chi-square test averaged over 1000 runs are compared to the Slager and Schaid Armitage test for trend using prior coefficient of kinship probabilities in the correlation matrix. The ‘answer’ in this simulated data set is SNP 5782, dominant model.

^2 Mode of inheritance. D=dominant R=recessive

^3 D^3 and r^2: linkage disequilibrium measures, calculated between SNP 5782 and each subsequent SNP.

^4 The significance threshold of PedGenie is limited by the number of simulations used to create the empirical null distribution (i.e., 2,000). For this result, all statistics based on observed values were less than the statistics based on simulated values.

^5 Result could not be calculated due to sparse data.
Table 4.3. Characteristics of breast cancer cohorts

<table>
<thead>
<tr>
<th></th>
<th>Case-Control Cohort</th>
<th>Nuclear Family Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number with complete genotype data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cases</td>
<td>Controls</td>
</tr>
<tr>
<td></td>
<td>231</td>
<td>235</td>
</tr>
<tr>
<td>Age³ (mean ± SD)</td>
<td>58.3 ± 13.0</td>
<td>74.8 ± 14.7</td>
</tr>
</tbody>
</table>

¹The Case-Control cohort originally was composed of 236 cases and 236 controls. Genotyping could not be completed for five cases and one control.

²The Trios cohort consists of 39 two-parent/one affected offspring sets. Two trio sets had the same parents and some of the parents also had breast cancer. The Sib cohort included 167 unique sibships composed of at least one affected and one unaffected sibling. Some sibships had more than one breast cancer case and more than one unaffected sibling.

³The age for cases is the age-at-diagnosis, whereas the age for controls is the last attained age at which the control was known to be breast cancer free or the age at which the control died due to causes other than breast cancer.
Table 4.4 reports all of the statistics that can be run by PedGenie for each of the two NBS1 tSNPs tested separately. We observed significant results for rs709816 under a dominant model for the chi-square \((p=0.015)\) and odds ratio tests \((\text{OR}=1.61, \text{95\% CI: (1.10, 2.35), } p=0.019)\). In particular, it can be seen that the majority of the difference observed in the dominant model for breast cancer status was due to heterozygous individuals compared to the homozygous wild-type individuals \((\text{odds ratio}=1.77, \text{95\% CI: (1.16, 2.72), } p=0.006)\). No significant results were observed for the age-at-diagnosis data and no significant results were observed using the TDT based tests with Nuclear Family data for either affected status or age-at-diagnosis phenotypes.

The composite genotype results for the NBS1 tSNPs are reported in Table 4.5 using the Case-Control cohort. None of the various combinations of inheritance models across the two loci met the \(p < 0.0125\) significance criteria.

Table 4.6 illustrates the haplotype results from PedGenie for NBS1, again using the Case-Control cohort. When all haplotypes were entered separately into the analysis, a single haplotype \((1-2)\) was found to be nominally significant \((p=0.03)\) when compared to the most common haplotype \((1-1)\) for breast cancer status. When this haplotype \((1-2)\) was tested in a dominant mode versus all other haplotypes, the mode of inheritance found to be significant in the single locus analysis, the result was nonsignificant \((\text{odds ratio}=0.55, \text{95\% CI: (0.30-1.04), } p=0.062)\). The linkage disequilibrium (LD), measured by \(D^*\), between the two tSNPs was 0.775. This would suggest that the nominal effect observed for the single haplotype \((1-2)\) analyses is most likely due to the LD between the two tSNPs and the effects of the association for tSNP rs709816 alone.
### Table 4.4. Association of each NBS1 tSNP with breast cancer status and age at diagnosis of breast cancer

**a. Case-Control cohort**

<table>
<thead>
<tr>
<th>Weight</th>
<th>Statistic</th>
<th>rs12680687</th>
<th></th>
<th>rs709816</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistic (Emp 95% CI)</td>
<td><em>p</em></td>
<td></td>
<td>Statistic (Emp 95% CI)</td>
<td><em>p</em></td>
</tr>
<tr>
<td>Dominant</td>
<td>Chi-Square¹</td>
<td>1.42</td>
<td>0.26</td>
<td>6.35</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>Odds Ratio¹</td>
<td>1.25</td>
<td>0.25</td>
<td>1.61</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>0.68</td>
<td>0.51</td>
<td>-0.11</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>Recessive</td>
<td>Chi-Square¹</td>
<td>2.48</td>
<td>0.13</td>
<td>0.076</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Odds Ratio¹</td>
<td>0.56</td>
<td>0.18</td>
<td>0.93</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>0.51</td>
<td>0.68</td>
<td>-0.47</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>Additive</td>
<td>Chi-Square Trend¹</td>
<td>0.086</td>
<td>0.80</td>
<td>2.73</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>Odds Ratio HET vs. WT¹</td>
<td>1.43</td>
<td>0.079</td>
<td>1.77</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>HOM vs. WT¹</td>
<td>0.64</td>
<td>0.31</td>
<td>1.22</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>Means HET vs. WT²</td>
<td>0.80</td>
<td>0.44</td>
<td>0.937</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>HOM vs. WT²</td>
<td>-0.34</td>
<td>0.74</td>
<td>-0.42</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>ANOVA²</td>
<td>0.41</td>
<td>0.68</td>
<td>0.10</td>
<td>0.91</td>
</tr>
<tr>
<td>Allele tests</td>
<td>Chi-Square¹</td>
<td>0.09</td>
<td>0.79</td>
<td>2.99</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>Odds Ratio¹</td>
<td>1.047</td>
<td>0.78</td>
<td>1.27</td>
<td>0.088</td>
</tr>
<tr>
<td></td>
<td>0.77</td>
<td>1.42</td>
<td>(0.97, 1.67)</td>
<td>0.888</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Means²</td>
<td>-8.19</td>
<td>0.31</td>
<td>-4.05</td>
<td>0.61</td>
</tr>
</tbody>
</table>
b. Nuclear Family cohort

<table>
<thead>
<tr>
<th>Statistic</th>
<th>rs12680687 Statistic</th>
<th>p</th>
<th>rs709816 Statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trio TDT(^1)</td>
<td>0.15</td>
<td>0.72</td>
<td>0.04</td>
<td>0.83</td>
</tr>
<tr>
<td>Sib TDT(^1)</td>
<td>1.52</td>
<td>0.095</td>
<td>0.45</td>
<td>0.58</td>
</tr>
<tr>
<td>Combined TDT(^1)</td>
<td>-1.20</td>
<td>0.23</td>
<td>-0.41</td>
<td>0.70</td>
</tr>
<tr>
<td>Trio QTDT(^2)</td>
<td>(-^3)</td>
<td>(-)</td>
<td>(-^3)</td>
<td>(-)</td>
</tr>
<tr>
<td>Combined QTDT(^2)</td>
<td>0.92</td>
<td>0.37</td>
<td>0.69</td>
<td>0.51</td>
</tr>
</tbody>
</table>

\(^1\)Breast cancer status was used as the dependent variable

\(^2\)Age-at-diagnosis was used as the dependent variable

\(^3\)The Trio QTDT requires a minimum of 30 trio sets that contain at least one heterozygous parent, complete genotype data on the parents, and age-at-diagnosis of the affected offspring. We had only 25 probands that met all of these criteria; hence the test could not be run.
Table 4.5. Association of *NBS1* by composite genotype in the Case-Control cohort

<table>
<thead>
<tr>
<th>Model</th>
<th>Odds ratio (Emp 95% CI)</th>
<th>p</th>
<th>Means test</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dom-Dom</td>
<td>1.23</td>
<td>0.28</td>
<td>0.30</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>(0.85, 1.79)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rcc-Rcc</td>
<td>0.49</td>
<td>0.13</td>
<td>0.005</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>(0.18, 1.29)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rcc-Dom</td>
<td>0.56</td>
<td>0.16</td>
<td>-0.59</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>(0.25, 1.27)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dom-Rec</td>
<td>0.90</td>
<td>0.77</td>
<td>-0.79</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>(0.48, 1.71)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results for this table are derived from requiring specific inheritance models for the two tSNPs for *NBS1*. Dom = dominant, Rec = recessive.
Table 4.6. Association of $NBS^I$ haplotypes and breast cancer in the Case-Control cohort

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Freq$^1$</th>
<th>Odds ratio</th>
<th>$p$</th>
<th>Emp 95% CI</th>
<th>Means test</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>0.63</td>
<td>Reference</td>
<td>-</td>
<td>-</td>
<td>Reference</td>
<td>-</td>
</tr>
<tr>
<td>2-2</td>
<td>0.21</td>
<td>1.25</td>
<td>0.25</td>
<td>(0.86, 1.82)</td>
<td>0.31</td>
<td>0.76</td>
</tr>
<tr>
<td>1-2</td>
<td>0.12</td>
<td>1.73</td>
<td>0.03</td>
<td>(1.04, 2.99)</td>
<td>-1.079</td>
<td>0.28</td>
</tr>
<tr>
<td>2-1</td>
<td>0.037</td>
<td>1.13</td>
<td>0.74</td>
<td>(0.53, 2.39)</td>
<td>-0.10</td>
<td>0.92</td>
</tr>
</tbody>
</table>

$^1$Haplotype frequency determined using a sample of 94 independent case-control subjects.
Discussion

This paper presents an approach and software (PedGenie) for association testing that incorporates family data of any size or structure, allowing for a mixture of structures from singletons to large extended pedigrees and genealogies to be analyzed together. The approach can be used to generate valid tests for any statistic. It is intuitively simple to understand and efficiently utilizes all information available within a family resource.

Another advantage of PedGenie is that it does not require a quantitative trait or the test statistic itself to have a known distribution. Thus, many of the inflexibilities of traditional tests can be avoided. For instance, the QTDT software requires that the quantitative trait be normally distributed. In contrast, using PedGenie to calculate the empirical \( p \)-value for an observed QTDT statistic does not necessitate this assumption. The traditional TDT and sib-TDT require that a single trio or a single-sibship be selected from each pedigree to be valid tests of both linkage and association; the null hypothesis for these traditional TDT tests is no association or no linkage. In contrast, PedGenie, which is able to test multiple nuclear families at a time within a pedigree, tests the combined null hypothesis of no association and no linkage for the TDT based tests using all available data in the pedigree.

Here we have demonstrated PedGenie’s ability to perform valid classical tests of dichotomous outcome as well as tests involving quantitative trait data. Multiallelic markers can be analyzed, with user-defined groups, if desired. Tests of composite genotype data and haplotype analyses can all be performed, including composite genotype and haplotype TDT statistics.
We illustrated PedGenie’s validity by examining its ability to simulate correctly from the empirical null distribution a \( p \)-value that corresponds well to the standard distribution for a particular statistical test. We have also shown that PedGenie accounts for relationships within a pedigree similar to an exact method. Finally, we illustrated PedGenie’s functionality and ability to assess associations between the outcome variables breast cancer status and breast cancer age-at-diagnosis with tSNPs in the \( NBS1 \) gene in some large Utah high-risk breast cancer families. In these real data, we found a significant association for a common variant in the \( NBS1 \) gene (rs709816, \( p=0.019 \)) indicating a potential role in breast cancer status, particularly for heterozygous carriers of this tSNP (\( p=0.006 \)). As the tSNP rs709816 results in a synonymous (silent) mutation in exon 7 of the \( NBS1 \) gene, the most likely causal allele for the association observed between the \( NBS1 \) gene and breast cancer is in linkage disequilibrium with this particular SNP.

Prior evidence for the involvement of the \( NBS1 \) gene with breast cancer is limited. Homozygous truncating germline mutations in the \( NBS1 \) gene (~90% due to 657del5 mutation on exon 6) result in the Nijmegen chromosomal breakage syndrome (NBS), a rare autosomal recessive disorder that includes an increased susceptibility to lymphoid malignancy. Heterozygous carriers of the 657del5 mutation have been suggested in a series of Polish case-control studies to be associated with an increased risk of breast cancer.\(^{53-55} \) Although the \( NBS1 \) 657del5 Slavic founder mutation is relatively rare even in Poland,\(^{54, 55} \) it contributes to only limited fraction of breast cancer cases in other Eastern European countries.\(^{56, 57} \) Several other distinct \( NBS1 \) mutations of non-Slavic origin have been described,\(^{58-60} \) but their association with breast cancer
has not been studied in depth. A limited number of common SNP variants across the
*NBS1* gene have been studied for their association with breast cancer, but no
significant differences between breast cancer cases and controls were observed. In our
study, we systematically identified tagging-SNPs for the *NBS1* gene, tested the tSNPs in
high-risk breast cancer families that are relatively homogeneous, and accounted for all
known relationships in our statistical tests. The *NBS1* protein, as it is involved in
variety of processes including sensing DNA damage, DNA double strand break repair,
telomere maintenance, and cell cycle checkpoint regulation, hence it is a likely
candidate for a breast cancer susceptibility gene. It follows that replication of our
results and further studies involving other *NBS1* variants are essential.

Empirical methods are often criticized for the computation time required to
calculate an empirical *p*-value, however this does not seem to be a serious problem with
PedGenie. Analysis of all our 139 pedigrees (3,761 total individuals) for the Case-
Control cohort for two tSNPs tested separately, 13 tests, and 2000 simulations each
using the allele frequency estimation method ‘all’ required 4 minutes, 30 seconds using
a Dell Precision 650n 2x2.8 GHz Xeon workstation. The same analysis using the allele
frequency estimation of ‘GeneCounter’ took 24 minutes, 8 seconds. The majority of
analysis time required for PedGenie is spent generating the null genotypic
configurations.

Family-based association methods have the disadvantage of being less efficient
than the traditional case-control study. The TDT test, for example, has an efficiency of
2/3 for father-mother-child trios compared to case-control studies and this increases
the cost of a family-based design. This disadvantage argues strongly that new resources
for association should concentrate on ascertaining independent individuals. However, for already ascertained resources, as we have here for breast cancer, pedigree-based methods are a reasonable choice and valid methods, such as those described here, are required.

In conclusion, our empirical method provides a valid approach to perform association and TDT testing in pedigrees of any size for both qualitative and quantitative data. Our software, PedGenie, which currently implements a broad range of statistical association-based statistics is freely available from our website.

Availability and requirements

Project name: PedGenie version 1.2
Project home page: http://bioinformatics.med.utah.edu/PedGenie
Operating system: Platform independent
Programming language: Java 1.4 and Java 1.5 (for GeneCounter method)
Other requirements: Java 1.5 or higher
License: none required
Any restrictions to use by nonacademics: none

Author’s contributions

KAB carried out the genotyping for the NBS1 gene, performed the statistical analyses, participated in the design of the software, and drafted the manuscript. JW performed the programming of PedGenie and participated in the design of the software. NJC conceived of the methodology, directed the design of the software, and helped to draft the manuscript.
Acknowledgements

Kristina Allen-Brady is an NLM fellow, supported by NLM grant T15 LM0724. This research was supported by a dissertation research grant for Kristina Allen-Brady (DISS0201521, to NJC) and an NIH NCI grant CA 098364 (to NJC). We appreciate the assistance of Kim Nguyen (Genetic Epidemiology) and Michael Hoffman (Family and Preventive Medicine) for their help in the laboratory. We also thank Helaman Escobar (Director of Sequencing and Genomics) and Michael Klein (Genomics) from the Core Resource Facilities, University of Utah, for use of their equipment and assistance on this project. We acknowledge the programming assistance for early work on PedGenie by John Elliot. Data collected for this publication were assisted by the Utah Cancer Registry supported by National Institutes of Health, Contract NOI-PC-35141, Surveillance, Epidemiology and End Results (SEER) Program, with additional support from the Utah Department of Health and the University of Utah. Partial support for all datasets within the Utah Population Database (UPDB) was provided by the University of Utah Huntsman Cancer Institute.
References


38. SimWalk2. [http://watson.hgen.pitt.edu/docs/simwalk2.html]


45. Pedigree Analysis Package for Java. [http://hasstedt.genetics.utah.edu/jpap/]


47. QTDT. [http://www.sph.umich.edu/csg/abecasis/QTDT/]


52. SNPHAP. [http://www-gene.cimr.cam.ac.uk/clayton/software]


CHARACTERIZATION OF THE LINKAGE DISEQUILIBRIUM STRUCTURE AND IDENTIFICATION OF TAGGING-SNPS IN FIVE DNA REPAIR GENES

Kristina Allen-Brady & Nicola J. Camp

Published: BMC Cancer 2005, 5:99
Characterization of the linkage disequilibrium structure and identification of tagging-SNPs in five DNA repair genes

Kristina Allen-Brady* and Nicola J Camp

Background: Characterization of the linkage disequilibrium (LD) structure of candidate genes is the basis for an effective association study of complex diseases such as cancer. In this study, we report the LD and haplotype architecture and tagging-single nucleotide polymorphisms (tSNPs) for five DNA repair genes: ATM, MRE11A, XRCC4, NBS1 and RAD50.

Methods: The genes ATM, MRE11A, and XRCC4 were characterized using a panel of 94 unrelated female subjects (47 breast cancer cases, 47 controls) obtained from high-risk breast cancer families. A similar LD structure and tSNP analysis was performed for NBS1 and RAD50, using publicly available genotyping data. We studied a total of 61 SNPs at an average marker density of 10 kb. Using a matrix decomposition algorithm, based on principal component analysis, we captured >90% of the intragenetic variation for each gene.

Results: Our results revealed that three of the five genes did not conform to a haplotype block structure (MRE11A, RAD50 and XRCC4). Instead, the data fit a more flexible LD group paradigm, where SNPs in high LD are not required to be contiguous. Traditional haplotype blocks assume recombination is the only dynamic at work. For ATM, MRE11A and XRCC4 we repeated the analysis in cases and controls separately to determine whether LD structure was consistent across breast cancer cases and controls. No substantial difference in LD structures was found.

Conclusion: This study suggests that appropriate SNP selection for an association study involving candidate genes should allow for both mutation and recombination, which shape the population-level genomic structure. Furthermore, LD structure characterization in either breast cancer cases or controls appears to be sufficient for future cancer studies utilizing these genes.

Abstract

Background: Characterization of the linkage disequilibrium (LD) structure of candidate genes is the basis for an effective association study of complex diseases such as cancer. In this study, we report the LD and haplotype architecture and tagging-single nucleotide polymorphisms (tSNPs) for five DNA repair genes: ATM, MRE11A, XRCC4, NBS1 and RAD50.

Methods: The genes ATM, MRE11A, and XRCC4 were characterized using a panel of 94 unrelated female subjects (47 breast cancer cases, 47 controls) obtained from high-risk breast cancer families. A similar LD structure and tSNP analysis was performed for NBS1 and RAD50, using publicly available genotyping data. We studied a total of 61 SNPs at an average marker density of 10 kb. Using a matrix decomposition algorithm, based on principal component analysis, we captured >90% of the intragenetic variation for each gene.

Results: Our results revealed that three of the five genes did not conform to a haplotype block structure (MRE11A, RAD50 and XRCC4). Instead, the data fit a more flexible LD group paradigm, where SNPs in high LD are not required to be contiguous. Traditional haplotype blocks assume recombination is the only dynamic at work. For ATM, MRE11A and XRCC4 we repeated the analysis in cases and controls separately to determine whether LD structure was consistent across breast cancer cases and controls. No substantial difference in LD structures was found.

Conclusion: This study suggests that appropriate SNP selection for an association study involving candidate genes should allow for both mutation and recombination, which shape the population-level genomic structure. Furthermore, LD structure characterization in either breast cancer cases or controls appears to be sufficient for future cancer studies utilizing these genes.

Background

Candidate gene association studies are a powerful study design for complex diseases such as cancer. Advances in association studies have been furthered by the recent discovery of single nucleotide polymorphisms (SNPs): their vast density throughout the genome, ease of genotyping and moderate cost contribute greatly to their utility. Association testing is efficient when the SNPs being analyzed represent the entire genetic variation of the gene. It has been suggested that nearby SNPs are organized into regions of high linkage disequilibrium (LD)
separated by short segments of very low LD [1-6]. In Caucasians, high LD regions may vary in length from a few kb to >300 kb [2,6,7]. Regions of high LD contain redundant information and can be reduced to smaller subsets of tagging-SNPs (tSNPs)[8], such that tSNPs identity all common haplotypes within the region of high LD. A number of algorithms have been proposed to define regions of high LD and tSNPs [4,8-14]. Thus far, no consensus of which algorithm is best has been achieved. Several studies have suggested the utility of matrix decomposition algorithms [12,13,15-17]. One advantage of these algorithms is that SNPs in high LD are not required to be contiguous nor mutually exclusive, a flexibility that is necessary for analyzing small genomic regions and rare variants. Further, these methods are stable with regards to marker density, minor allele frequency, analysis window, and possible analysis window length [18].

Growing evidence appears to suggest that tumorigenesis is a multi-step process of genetic alterations that transform a normal human cell into a malignant derivative [19]. The ability of a cell to maintain genomic stability through DNA repair mechanisms is essential to prevent tumor initiation and progression. A number of different types of cancer have been attributed to defective DNA repair including xeroderma pigmentosum [20], hereditary non-polyposis colorectal cancer [21], and breast cancer due to mutations in *BRCA1* and *BRCA2* as well as other DNA repair genes (e.g., *ATM*, *TP53* and *CHEK2*)[22]. Many published candidate gene association studies involving DNA repair genes and cancer risk have assessed risk by examining a single SNP per gene or a single locus at a time analysis approach. Unfortunately, the former approach is often inadequate in comprehensively accounting for the genetic variation of a gene, and the latter incurs multiple testing corrections, which usually eliminate all or most of the association evidence found. It has been suggested that use of haplotypes in association studies may have increased power over single-allele studies [8]. Descriptions of haplotype diversity and LD structure as well as identification of potential tSNPs will be key for success in candidate gene association studies.

Here we describe haplotypes, LD structure and potential tSNPs in five DNA repair breast cancer susceptibility genes: *ATM*, *MRE11A*, *NBS1*, *RAD50*, and *XRCC4*. We used a matrix decomposition algorithm based on a method of principal components analysis [13]; this method does not require SNPs to be in contiguous block structure. Characterization of the LD structure and tSNPs are necessary for the design of future effective association studies.

**Methods**

**Subjects**

This study is part of a larger study involving 139 high-risk Caucasian breast cancer families, defined as high risk because cancer rates in these families were significantly higher than the general population rate determined using the Utah Population Database (UPDB) [23-25]. All breast cancer cases in the larger cohort met at least one of the following criteria: 1) their family tested negative for a *BRCA1* or *BRCA2* mutation, 2) the case themselves tested negative for the same *BRCA1* or *BRCA2* mutation that was present in their family, or 3) their family had a low probability of carrying a *BRCA1* or *BRCA2* mutation based on the number of breast cancer cases present in the family and/or ages at diagnosis of breast cancer within the family. Therefore, all breast cancer cases in the larger study had a low residual probability of their cancer being due to mutations in *BRCA1* or *BRCA2*. Breast cancer diagnosis information was obtained from medical records for the subject or the Utah Cancer Registry.

For this LD characterization study, we selected a panel of 94 individuals (47 female breast cancer cases and 47 female controls), chosen randomly from separate kindreds to ensure independence. Both cases and controls were chosen such that comparisons of LD structure could be made between the groups. The sample size of 188 chromosomes is larger than generally used for this type of study [26-29], but inadequate for an association analysis. This current study is not a case-control study and associations with disease were not assessed.

Blood samples were collected on all subjects and all individuals signed consent to participate this study. This study was approved by the University of Utah Institutional Review Board.

**Genes and SNP selection**

For each gene of interest (i.e., *ATM*, *MRE11A*, *NBS1*, *RAD50* and *XRCC4*), all SNPs available from Applied Biosystems[30], within each gene and the flanking ±10 kb on either side, that had been validated to have a minor allele frequency greater than 0.01 in Caucasians were selected. For *ATM* (on chromosome 11q22-2q23), which spans approximately 143 kb and contains 64 exons, 14 SNPs were studied with a SNP resolution of 1 SNP/10,489 bp. For *MRE11A* (11q21), which spans approximately 76 kb and contains 20 exons, 11 SNPs were studied with a SNP resolution of 1 SNP/8539 bp. For *NBS1* (8q21), which contains 16 exons and spans about 51 kb, 5 SNPs were studied with a SNP resolution of 1 SNP/8256 bp. The *RAD50* gene (5q31) spans approximately 87 kb contains 25 exons, and we studied 10 SNPs at a resolution of 1 SNP/10,533 bp. Finally, for *XRCC4* (5q13-2q14) with 8 exons and approximately 276 kb in length, we studied 21 SNPs at a resolution of 1 SNP/13,198 bp. The vast majority of the SNPs studied were intronic (see Table 1).
Table 1: Characteristics of SNPs analyzed

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP Code</th>
<th>SNP ID</th>
<th>Base Change</th>
<th>Position†</th>
<th>MAF‡</th>
<th>ABI reported MAF§</th>
<th># bp from the most 5' SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATM</td>
<td>A1</td>
<td>rs228589</td>
<td>T/A</td>
<td>Flanking</td>
<td>0.45</td>
<td>0.33</td>
<td>0</td>
</tr>
<tr>
<td>ATM</td>
<td>A2</td>
<td>rs228591</td>
<td>G/A</td>
<td>mRNA-utr</td>
<td>0.45</td>
<td>0.33</td>
<td>4125</td>
</tr>
<tr>
<td>ATM</td>
<td>A3</td>
<td>rs641605</td>
<td>T/C</td>
<td>Intron</td>
<td>0.45</td>
<td>0.33</td>
<td>29,981</td>
</tr>
<tr>
<td>ATM</td>
<td>A4</td>
<td>rs228599</td>
<td>A/G</td>
<td>Intron</td>
<td>0.44</td>
<td>0.31</td>
<td>14,452</td>
</tr>
<tr>
<td>ATM</td>
<td>A5</td>
<td>rs600931</td>
<td>T/C</td>
<td>Intron</td>
<td>0.45</td>
<td>0.35</td>
<td>24,127</td>
</tr>
<tr>
<td>ATM</td>
<td>A6</td>
<td>rs228592</td>
<td>A/C</td>
<td>Intron</td>
<td>0.45</td>
<td>0.33</td>
<td>29,981</td>
</tr>
<tr>
<td>ATM</td>
<td>A7</td>
<td>rs646477</td>
<td>T/C</td>
<td>Intron</td>
<td>0.43</td>
<td>0.33</td>
<td>49,974</td>
</tr>
<tr>
<td>ATM</td>
<td>A8</td>
<td>rs1003623</td>
<td>T/C</td>
<td>Intron</td>
<td>0.45</td>
<td>0.33</td>
<td>59,374</td>
</tr>
<tr>
<td>ATM</td>
<td>A9</td>
<td>rs609326</td>
<td>C/T</td>
<td>mRNA-utr, intron</td>
<td>0.45</td>
<td>0.32</td>
<td>64,926</td>
</tr>
<tr>
<td>ATM</td>
<td>A10</td>
<td>rs645485</td>
<td>G/A</td>
<td>Intron</td>
<td>0.45</td>
<td>0.32</td>
<td>75,655</td>
</tr>
<tr>
<td>ATM</td>
<td>A11</td>
<td>rs563281</td>
<td>A/G</td>
<td>Intron</td>
<td>0.45</td>
<td>0.31</td>
<td>88,861</td>
</tr>
<tr>
<td>ATM</td>
<td>A12</td>
<td>rs227061</td>
<td>A/G</td>
<td>mRNA-utr, intron</td>
<td>0.45</td>
<td>0.34</td>
<td>112,175</td>
</tr>
<tr>
<td>ATM</td>
<td>A13</td>
<td>rs227062</td>
<td>A/G</td>
<td>mRNA-utr, intron</td>
<td>0.45</td>
<td>0.33</td>
<td>112,175</td>
</tr>
<tr>
<td>ATM</td>
<td>A14</td>
<td>rs562311</td>
<td>A/G</td>
<td>Flanking</td>
<td>0.45</td>
<td>0.36</td>
<td>146,861</td>
</tr>
<tr>
<td>MRE1</td>
<td>M1</td>
<td>rs646130</td>
<td>T/C</td>
<td>Flanking</td>
<td>0.3</td>
<td>0.39</td>
<td>0</td>
</tr>
<tr>
<td>MRE1</td>
<td>M2</td>
<td>rs491404</td>
<td>G/C</td>
<td>Flanking</td>
<td>0.3</td>
<td>0.4</td>
<td>9192</td>
</tr>
<tr>
<td>MRE1</td>
<td>M3</td>
<td>rs10831227</td>
<td>G/A</td>
<td>Intron</td>
<td>0.36</td>
<td>0.36</td>
<td>16,336</td>
</tr>
<tr>
<td>MRE1</td>
<td>M4</td>
<td>rs601341</td>
<td>A/G</td>
<td>Intron</td>
<td>0.34</td>
<td>0.36</td>
<td>18,956</td>
</tr>
<tr>
<td>MRE1</td>
<td>M5</td>
<td>rs564715</td>
<td>T/C</td>
<td>Flanking</td>
<td>0.3</td>
<td>0.36</td>
<td>28,556</td>
</tr>
<tr>
<td>MRE1</td>
<td>M6</td>
<td>rs556477</td>
<td>A/G</td>
<td>Intron</td>
<td>0.3</td>
<td>0.36</td>
<td>40,656</td>
</tr>
<tr>
<td>MRE1</td>
<td>M7</td>
<td>rs1805365</td>
<td>A/G</td>
<td>Intron</td>
<td>0.02</td>
<td>0.02</td>
<td>61,721</td>
</tr>
<tr>
<td>MRE1</td>
<td>M8</td>
<td>rs680695</td>
<td>A/G</td>
<td>Intron</td>
<td>0.34</td>
<td>0.36</td>
<td>72,913</td>
</tr>
<tr>
<td>MRE1</td>
<td>M9</td>
<td>rs1009455</td>
<td>C/G</td>
<td>mRNA-utr</td>
<td>0.02</td>
<td>0.0111</td>
<td>85,033</td>
</tr>
<tr>
<td>MRE1</td>
<td>M10</td>
<td>rs1009456</td>
<td>C/A</td>
<td>locus-region, mRNA-utr</td>
<td>0.01</td>
<td>0.02</td>
<td>87,401</td>
</tr>
<tr>
<td>NBS1</td>
<td>N1</td>
<td>rs12680887</td>
<td>G/T</td>
<td>Flanking</td>
<td>0.09</td>
<td>0.06</td>
<td>93,946</td>
</tr>
<tr>
<td>NBS1</td>
<td>N2</td>
<td>rs709816</td>
<td>A/G</td>
<td>Coding-syncton</td>
<td>-</td>
<td>0.28</td>
<td>0</td>
</tr>
<tr>
<td>NBS1</td>
<td>N3</td>
<td>rs18057590</td>
<td>C/T</td>
<td>Flanking</td>
<td>-</td>
<td>0.39</td>
<td>23,313</td>
</tr>
<tr>
<td>NBS1</td>
<td>N4</td>
<td>rs741778</td>
<td>C/T</td>
<td>Intron</td>
<td>-</td>
<td>0.36</td>
<td>33,415</td>
</tr>
<tr>
<td>NBS1</td>
<td>N5</td>
<td>rs1805841</td>
<td>C/G</td>
<td>Intron</td>
<td>-</td>
<td>0.45</td>
<td>14,282</td>
</tr>
<tr>
<td>RAD50</td>
<td>R1</td>
<td>rs2522406</td>
<td>G/A</td>
<td>Flanking</td>
<td>-</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>RAD50</td>
<td>R2</td>
<td>rs2244012</td>
<td>C/T</td>
<td>Intron</td>
<td>-</td>
<td>0.19</td>
<td>12,116</td>
</tr>
<tr>
<td>RAD50</td>
<td>R3</td>
<td>rs2299015</td>
<td>T/G</td>
<td>Intron</td>
<td>-</td>
<td>0.15</td>
<td>12,388</td>
</tr>
<tr>
<td>RAD50</td>
<td>R4</td>
<td>rs2299014</td>
<td>T/G</td>
<td>Intron</td>
<td>-</td>
<td>0.41</td>
<td>14,290</td>
</tr>
<tr>
<td>RAD50</td>
<td>R5</td>
<td>rs2706577</td>
<td>A/G</td>
<td>Intron</td>
<td>-</td>
<td>0.01</td>
<td>50,388</td>
</tr>
<tr>
<td>RAD50</td>
<td>R6</td>
<td>rs2301713</td>
<td>C/T</td>
<td>Intron</td>
<td>-</td>
<td>0.19</td>
<td>62,877</td>
</tr>
<tr>
<td>RAD50</td>
<td>R7</td>
<td>rs204703</td>
<td>G/A</td>
<td>Flanking</td>
<td>-</td>
<td>0.23</td>
<td>83,149</td>
</tr>
<tr>
<td>RAD50</td>
<td>R8</td>
<td>rs2240032</td>
<td>C/T</td>
<td>Intron</td>
<td>-</td>
<td>0.18</td>
<td>88,018</td>
</tr>
<tr>
<td>RAD50</td>
<td>R9</td>
<td>rs1809025</td>
<td>C/T</td>
<td>Flanking</td>
<td>-</td>
<td>0.19</td>
<td>103,700</td>
</tr>
<tr>
<td>RAD50</td>
<td>R10</td>
<td>rs2066960</td>
<td>C/A</td>
<td>Flanking</td>
<td>-</td>
<td>0.17</td>
<td>105,326</td>
</tr>
<tr>
<td>XRC4</td>
<td>X1</td>
<td>rs1993948</td>
<td>T/A</td>
<td>Flanking</td>
<td>0.46</td>
<td>0.47</td>
<td>0</td>
</tr>
<tr>
<td>XRC4</td>
<td>X2</td>
<td>rs1478485</td>
<td>G/A</td>
<td>mRNA-utr</td>
<td>0.47</td>
<td>0.45</td>
<td>8247</td>
</tr>
<tr>
<td>XRC4</td>
<td>X3</td>
<td>rs11951257</td>
<td>T/C</td>
<td>Intron</td>
<td>0.47</td>
<td>0.45</td>
<td>31,031</td>
</tr>
<tr>
<td>XRC4</td>
<td>X4</td>
<td>rs10045104</td>
<td>T/C</td>
<td>Intron</td>
<td>0.43</td>
<td>0.42</td>
<td>40,982</td>
</tr>
<tr>
<td>XRC4</td>
<td>X5</td>
<td>rs6452526</td>
<td>C/T</td>
<td>Intron</td>
<td>0.47</td>
<td>0.43</td>
<td>64,351</td>
</tr>
<tr>
<td>XRC4</td>
<td>X6</td>
<td>rs1382369</td>
<td>G/A</td>
<td>Intron</td>
<td>0.47</td>
<td>0.43</td>
<td>69,149</td>
</tr>
<tr>
<td>XRC4</td>
<td>X7</td>
<td>rs1382368</td>
<td>C/T</td>
<td>Intron</td>
<td>0.47</td>
<td>0.41</td>
<td>78,795</td>
</tr>
<tr>
<td>XRC4</td>
<td>X8</td>
<td>rs1382363</td>
<td>C/T</td>
<td>Intron</td>
<td>0.47</td>
<td>0.42</td>
<td>80,292</td>
</tr>
<tr>
<td>XRC4</td>
<td>X9</td>
<td>rs13180316</td>
<td>G/A</td>
<td>Intron</td>
<td>0.23</td>
<td>0.36</td>
<td>87,173</td>
</tr>
<tr>
<td>XRC4</td>
<td>X10</td>
<td>rs17141420</td>
<td>A/T</td>
<td>Intron</td>
<td>0.47</td>
<td>0.44</td>
<td>98,452</td>
</tr>
<tr>
<td>XRC4</td>
<td>X11</td>
<td>rs2731861</td>
<td>T/C</td>
<td>Intron</td>
<td>0.47</td>
<td>0.45</td>
<td>112,984</td>
</tr>
<tr>
<td>XRC4</td>
<td>X12</td>
<td>rs2662238</td>
<td>G/A</td>
<td>Intron</td>
<td>0.46</td>
<td>0.45</td>
<td>127,027</td>
</tr>
<tr>
<td>XRC4</td>
<td>X13</td>
<td>rs1039786</td>
<td>C/T</td>
<td>Intron</td>
<td>0.46</td>
<td>0.45</td>
<td>127,761</td>
</tr>
<tr>
<td>XRC4</td>
<td>X14</td>
<td>rs963248</td>
<td>T/C</td>
<td>Intron</td>
<td>0.19</td>
<td>0.16</td>
<td>161,614</td>
</tr>
<tr>
<td>XRC4</td>
<td>X15</td>
<td>rs301276</td>
<td>G/A</td>
<td>Intron</td>
<td>0.23</td>
<td>0.23</td>
<td>175,451</td>
</tr>
<tr>
<td>XRC4</td>
<td>X16</td>
<td>rs35268</td>
<td>T/C</td>
<td>Intron</td>
<td>0.16</td>
<td>0.13</td>
<td>216,216</td>
</tr>
<tr>
<td>XRC4</td>
<td>X17</td>
<td>rs301296</td>
<td>T/C</td>
<td>Intron</td>
<td>0.16</td>
<td>0.16</td>
<td>230,675</td>
</tr>
<tr>
<td>XRC4</td>
<td>X18</td>
<td>rs301289</td>
<td>C/T</td>
<td>Intron</td>
<td>0.17</td>
<td>0.17</td>
<td>233,955</td>
</tr>
</tbody>
</table>
Table 1: Characteristics of SNPs analyzed (Continued)

<table>
<thead>
<tr>
<th>Locus</th>
<th>X19</th>
<th>rs2386275</th>
<th>G/A</th>
<th>Intron</th>
<th>0.09</th>
<th>0.12</th>
<th>270,260</th>
</tr>
</thead>
<tbody>
<tr>
<td>XRCC4</td>
<td>X20</td>
<td>rs2891980</td>
<td>T/C</td>
<td>Intron</td>
<td>0.09</td>
<td>0.13</td>
<td>270,383</td>
</tr>
<tr>
<td>XRCC4</td>
<td>X21</td>
<td>rs1056503</td>
<td>T/G</td>
<td>Coding-synon</td>
<td>0.09</td>
<td>0.12</td>
<td>276,697</td>
</tr>
</tbody>
</table>

* Base change listed as Major allele / Minor allele
† Position obtained from the University of California, Santa Cruz Genome Browser http://genome.ucsc.edu/cgi-bin/hgGateway; Flanking = within 10 kb of either side of gene; Locus region = variation in region of gene, but not in transcript; mRNA-utr = variation in transcript, but not in coding region interval
‡ MAF = minor allele frequency using our panel of 94 breast cancer case and control subjects
§ Applied Biosystems reported minor allele frequency in Caucasians
|| Corrected value. Applied Biosystems acknowledged error in reported minor allele frequency of 0.49 on their web site, but it has not been updated.
** NBS1 and RAD50 were not genotyped in the current study. All analyses for these two genes were performed using the raw genotype data freely available online from Applied Biosystems. Base change obtained from University of California, Santa Cruz Genome Browser.

Genotyping

For the ATM and XRCC4 all SNPs that met the above criteria were genotyped on our panel of 94 subjects. For MRE11A, one SNP repeatedly failed to amplify (rs10831224) and was removed from the study.

Genomic DNA was isolated and purified using standard phenol/chloroform DNA extraction. SNP genotyping was performed using the fluorogenic 5' nuclease TaqMan Assay [31] (Applied Biosystems). The TaqMan Assay requires TaqMan PCR Master Mix (Applied Biosystems), which we used according to manufacturer’s instructions, yielding a final volume of 5 μl per well. PCR amplification was also performed according to the Applied Biosystems protocol. The 7900HT Sequence Detection System (Applied Biosystems) was used to measure each fluorescent dye-labeled probe specific for each allele studied and results were analyzed with the Sequence Detection Software (Applied Biosystems).

Haplotype structure and tSNP selection

Haplotypes and haplotype frequencies were estimated from unphased genotype data using an expectation-maximization algorithm. SNPHAP[32]. SNPHAP uses a maximum-likelihood program to predict nullithous haplotypes. Haplotypes with a frequency of at least 0.01 were analyzed using a two-step PCA method[13]. This method does not require that groups of SNPs be contiguous along a DNA fragment and also allows SNPs to be present in more than one group. In step I, LD groups are determined. In brief, the PCA method extracts factors (LD groups) to capture ≥ 90% of the genetic diversity. An LD group is defined as those SNPs that load onto the same factor. In step II, tSNPs are selected for each LD group. Each LD group is considered separately and the PCA method again extracts factors. tSNPs are chosen as the SNPs with the highest factor loading. When a number of SNPs load equally well on an LD group, these can all be considered potential tSNPs. Under such circumstances, we selected the single SNP that performed best in the genotyping assay. This was done in order to minimize errors in allele calls.

We compared our genotype data for ATM, MRE11A, and XRCC4 with genotyping data for these same genes obtained from Applied Biosystems (ABI)[30] on 45 Caucasians. We found good concordance in allele frequencies between the data sets. Further, we applied the same LD characterization to both data sets and found excellent concordance in the LD groups and potential tSNPs (see Results). We therefore characterized LD groups and tSNPs for NBS1 and RAD50 using the genotyping data available online.

We also examined whether differences existed between LD group structure and tSNP selection when cases and controls were considered separately. This analysis could only be performed for ATM, MRE11A, and XRCC4.

Results

Characteristics of the SNPs studied are listed in Table 1. Minor allele frequencies from our 94 subjects compared well with those listed by Applied Biosystems[30]. Despite the very low minor allele frequencies in some of the SNPs studied, we observed heterozygosity for all SNPs genotyped.

Table 2 lists the haplotypes with a frequency > 0.01 obtained from SNPHAP, and the LD group designation and the tSNPs that were selected using the PCA method, for ATM, MRE11A, and XRCC4. Haplotypes are reported using the standard convention of designating the major allele as ‘1’ and the minor allele as ‘2’, in order to more easily spot occurrences of the minor allele. Please see Table 1 for the corresponding base pair change. For ATM, 7
Table 2: Haplotypes with frequency > 0.01, LD group characterization and tSNPs

selected using Utah genotyping data

### a. ATM

<table>
<thead>
<tr>
<th>LD Group and tSNP Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 1 1 1 1 1 1 1 1 1 1 1 0.54</td>
</tr>
<tr>
<td>2 2 1 1 1 1 1 1 1 1 1 1 1 0.28</td>
</tr>
</tbody>
</table>

### b. MRE11A

<table>
<thead>
<tr>
<th>LD Group and tSNP Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 1 1 1 1 1 1 1 1 1 1 1 1 0.35</td>
</tr>
</tbody>
</table>

### c. XRCC4

<table>
<thead>
<tr>
<th>LD Group and tSNP Designation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1 1 1 1 1 1 1 1 1 1 1 1 1 0.25</td>
</tr>
</tbody>
</table>

* Analysis considers the total panel of 94 individuals together
† tSNP selected / group

haplotypes overall were observed and 5 had a frequency > 0.01. Using the PCA method, a single LD group was identified, encompassing the entire gene and accounting for 98.8% of the genetic variance across the gene. From this single LD group, a single tSNP (A13) was selected.

For MRE11A, we observed 9 haplotypes in total and 6 with frequency > 0.01. From the PCA analysis, four LD groups were identified based on these 6 haplotypes with a frequency > 0.01, and accounted for 99.1% of the genetic variance. The LD groups did not conform to haplotype blocks. SNP M4 separated LD group 1 into two parts and M8 separated LD group 2. Each LD group was represented by a single tSNP, such that the tSNP set contained 4 tSNPs (M6, M10, M11, and M14). For XRCC4, we observed 26 haplotypes overall, 13 of which had a frequency > 0.01. From the PCA method, four LD groups were observed which accounted for 97.2% of the variance. Similarly to MRE11A, the LD groups were not contiguous blocks. LD group 1 was divided by X9 and LD group 2 was divided by X15. Each of the LD groups could be represented by a single SNP resulting in the tSNP set (X2, X9, X14, and X21).
Table 3: Comparison of LD groups for the Utah breast cancer cases and controls with Applied Biosystems (ABI) data*

<table>
<thead>
<tr>
<th>Gene</th>
<th>Group</th>
<th>Utah breast cancer case/ control SNPs</th>
<th>Utah potential tSNPs</th>
<th>Utah % variance captured/ group</th>
<th>ABI SNPs</th>
<th>ABI potential tSNPs</th>
<th>ABI % variance captured/ group</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRE11</td>
<td>1</td>
<td>M1, M2, M3, M5, M6</td>
<td>M1, M2, M3, M5, M6</td>
<td>100%</td>
<td>M1, M2, M3, M5, M6</td>
<td>M1, M2, M3, M5, M6</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>M7, M9, M10</td>
<td>M10</td>
<td>64.3%</td>
<td>M7, M9, M10</td>
<td>M7, M9, M10</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>M11</td>
<td>M11</td>
<td>100%</td>
<td>M11</td>
<td>M11</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>M4, M8</td>
<td>M4, M8</td>
<td>82.2%</td>
<td>M4, M8</td>
<td>M4, M8</td>
<td>83.9%</td>
</tr>
<tr>
<td>XRC4</td>
<td>1</td>
<td>X1-X8, X10-X13</td>
<td>X2-X3, X5-X8, X10-X11</td>
<td>95.3%</td>
<td>X2, X3, X10, X11</td>
<td>X2, X3, X10, X11</td>
<td>96.0%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>X14, X16-X18</td>
<td>X14</td>
<td>91.6%</td>
<td>X14, X16, X18</td>
<td>X14, X17, X18</td>
<td>93.5%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>X19-X21</td>
<td>X19-X21</td>
<td>100%</td>
<td>X19-X21</td>
<td>X19-X21</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>X9, X15</td>
<td>X9, X15</td>
<td>97.4%</td>
<td>X9, X15</td>
<td>X9, X15</td>
<td>98.8%</td>
</tr>
</tbody>
</table>

*We used Applied Biosystems’ validated SNP genotype data for 45 Caucasian subjects.

Table 4: Haplotypes with frequency>0.01, LD group characterization and tSNP selected using data from Applied Biosystems*

a. NBS1

<table>
<thead>
<tr>
<th>N1</th>
<th>N2</th>
<th>N3†</th>
<th>N4†</th>
<th>N5†</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.55</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.26</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.03</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.03</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.03</td>
</tr>
</tbody>
</table>

LD Group and tSNP Designation

2†

b. RAD50

<table>
<thead>
<tr>
<th>R1†</th>
<th>R2</th>
<th>R3†</th>
<th>R4†</th>
<th>R5</th>
<th>R6†</th>
<th>R7</th>
<th>R8</th>
<th>R9</th>
<th>R10</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.50</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.21</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0.11</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0.009§</td>
</tr>
</tbody>
</table>

LD Group and tSNP Designation

2†

*We used Applied Biosystems’ validated SNP genotype data for 45 Caucasian subjects.
† Allelic designations have been changed from that listed by ABI to conform to the convention 1 = common allele, 2 = rare allele.
‡ tSNP selected / group § The haplotype with a frequency 0.009 was also analyzed to allow inclusion of rare variants at R1 and R5.
Table 3 shows the LD groups and tSNPs for ATM, MRE11A, and XRCC4 using our panel of 94 subjects and using the 45 Caucasian subjects from Applied Biosystems [30]. For these three genes, we observed the same number of LD groups containing precisely the same SNPs for both data sets. The difference between the results was in the number of potential tSNPs for each LD group. For the majority of LD groups, the potential tSNPs using Applied Biosystems data were a subset of those from our sample. This is perhaps expected, because our sample size was more than double their size and is therefore capable of better resolution.

Table 4 lists the haplotypes, LD group designation, potential tSNPs, and tSNP selected per group for NBS1 and RAD50 using the Applied Biosystems' data. For NBS1, 6 haplotypes overall were observed and all 6 haplotypes had a frequency > 0.01. Using the PCA method, two LD groups were identified and accounted for 93.8% of the variance. Two tSNPs were sufficient to tag these groups (N1, N2). However, N5 could replace N2 with no reduction in the variance explained. For the RAD50 gene, in order to include two available rare SNPs in the analysis, we lowered the haplotype acceptance threshold to 0.009. We observed a total of 14 haplotypes, 10 with a frequency greater than 0.01. Using the PCA method, we identified three LD groups, which accounted for 91.5% of the variance. Similarly to MRE11A and XRCC4, the LD groups for RAD50 were not contiguous blocks. Three tSNPs were sufficient to tag the groups (R1, R3, and R10), although R5 could replace R1 and R6 could replace R3 with no loss of variance explained.

For ATM, MRE11A, and XRCC4, we compared haplotypes and LD structure between the breast cancer cases and controls. For ATM and XRCC4 no difference in the LD structure was observed when cases and controls were analyzed separately. For the MRE11A gene differences in LD structure were noted. However, these were minor and likely attributable to small sample size since the differences were driven by 3 rare haplotypes (frequency = 0.02).

Discussion

Identification of the most informative markers to use in a large-scale association analysis for studies of complex disease, such as breast cancer, is critical to the success of the study. The key to this process is to select SNPs that are most informative about the underlying haplotype structure in a population of interest. As haplotype based designs have been suggested as being more powerful than the single-allele approach for association studies [8], a haplotype-based approach should result in more accurate and definitive findings. In this study, we have described haplotypes and characterized the LD structure of the ATM, MRE11A, and XRCC4 genes using a panel of 94 subjects, including breast cancer cases from high-risk breast cancer families as well as controls. Further, we identified tSNPs that can be used in future haplotype-based association studies. A similar analysis was performed for NBS1 and RAD50 using publicly available genotype data. We identified, using Principal Components Analysis [13], a single LD group for ATM, four noncontiguous LD groups for MRE11A, two LD groups for NBS1, three noncontiguous LD groups for RAD50, and four noncontiguous LD groups for XRCC4. In each case, the LD groups captured greater than 90% of the variance of the total SNPs available from Applied Biosystems across the gene. Furthermore for each gene, we present tSNPs that could be selected to represent the gene.

It is of interest that the LD structure for three of these five DNA repair genes did not conform to the haplotype block model, that is, that the LD groups did not contain contiguous SNPs. This was true whether the genotyping data came from our own study or from Applied Biosystems. Although we did not directly sequence these genes to identify all possible variants, the discontinuity we observed illustrates that the underlying LD structure cannot conform to contiguous haplotype blocks. A more flexible LD group representation (as supported under principle components analysis) fit the data better and appears to be stable to differences in minor allele frequency. Similar findings of a complex pattern of LD structure were recently reported in a high-resolution study of the ELAC2 gene [15]. Our results suggest that when studying small genomic regions and low frequency variants (<0.2), mutation is an important dynamic in LD structure, and the simple recombination-only model used in classical haplotype block methods does not fit the data well and hence will lead to a poor selection of tSNPs.

Due to the stability of the results for ATM, MRE11A and XRCC4, we pursued two additional DNA repair genes of interest (i.e., NBS1 and RAD50). Applied Biosystems provides freely-available genotyping data for four ethnically diverse populations of 45 subjects in each, therefore, even with limited funds, the haplotype structure and selection of tSNPs can be estimated for a study prior to any genotyping costs. However, caution must be used if this option is exercised as one's population must be one of Applied Biosystems' ethnic cohorts (i.e., Caucasian, African American, Chinese, or Japanese) and our experience is that occasionally errors exist in the data.

Of the genes studied here, only ATM has previously been studied in any depth for LD structure. The reason that ATM has received so much attention is that patients with the recessive disease ataxia-telangiectasia, due to a mutation in the ATM gene, have a 100-fold increased risk of cancer [33,34] and obligate heterozygous carriers of ATM mutations may have an increased risk of cancer, particularly breast cancer [35-39], although this finding is controversial [40,41]. Extensive LD across the ATM gene has...
Some limitations are inherent in this study and must be pointed out. First, we did not sequence our genes of interest and thus all of the genetic diversity within these genetic regions may not be captured. Our results must be interpreted in light of this. The gold standard is to identify all variants within a gene and select a subset of tSNPs from this set. It would be interesting to evaluate the robustness of our findings using sequence data. However, the SNPs examined were relatively evenly spaced, on the order of 1 SNP every 10 kb, and our results are important as they illustrate how smaller budget studies can best select tSNPs. Second, our sample size was modest (188 chromosomes), although larger than other previous studies examining LD and tSNPs [26-29]. Finally, haplotype block and haplotype-tagging SNP analyses have been suggested to only be reliable when markers are dense, otherwise marker sets have considerable loss of information [50]. This result may extend to PCA methods, however, the matrix decomposition algorithm used has been suggested to be stable with regards to varying levels of marker density [18].

Conclusion

In conclusion, we have described haplotypes, linkage disequilibrium structure, and identified tSNPs from all available Applied Biosystems’ validated SNPs in ATM, MRE11A, NBS1, RAD50, and XRCC4 genes in a Caucasian population. As has been found for other genes, we identified LD structures that did not conform to contiguous haplotype block structures. This illustrates the importance of using flexible methods, such as matrix decomposition, that allow for multiple population dynamics such as recombination, mutation and selection. Although the gold standard for SNP characterization across a candidate gene is sequencing to identify all variants, we describe a low-budget means to characterize the LD structure and select tSNPs using publicly available data. Comprehensive characterization of the LD structure at genes of interest will be essential for future effective association studies.

Electronic database information

The data from the 94 breast cancer case and control subjects for these tables is publicly available at http://bioninformatics.med.utah.edu under Supplemental Materials to Publication. On request from Dr. Nicola Camp a user-name and password to access the data will be given.

Competing interests

The author(s) declare that they have no competing interests.

Authors’ contributions

KAB assisted in the study design, performed the genotyping, and drafted the manuscript. NJC conceived of the study and its design and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Kristina Allen-Brady is an NLM fellow, supported by NLM grant T15 LM0724. This research was supported by a dissertation research grant from the Susan G. Komen Breast Cancer Foundation for Kristina Allen-Brady (DISS0201521, to NJC).
and an NIH NCI grant CA 093644 to N.J.C. We appreciate the assistance of Kim Nguyen (Genetic Epidemiology) and Michael Hoffman (Family and Preventive Medicine) for their help in the laboratory. We also thank Helaman Escobar (Director of Sequencing and Genomics) and Michael Klein (Genomics) from the Core Resource Facilities, University of Utah, for use of their equipment and assistance on this project. Data collected for this publication was assisted by the Utah Cancer Registry supported by National Institutes of Health, Contract NOI-PC-35141, Surveillance, Epidemiology and End Results (SEER) Program, with additional support from the Utah Department of Health and the University of Utah. Partial support for all datasets within the Utah Population Database (UPDB) was provided by the University of Utah Huntsman Cancer Institute.

References


49. HapMap [http://www.hapmap.org/]

A ROLE FOR XRCC4 IN AGE AT DIAGNOSIS AND
BREAST CANCER RISK IN NON-BRCA1/2,
HIGH-RISK BREAST CANCER CASES

Kristina Allen-Brady & Nicola J. Camp

Submitted to Cancer Epidemiology Biomarkers & Prevention

December 2005
Abstract

Genetic polymorphisms in DNA repair genes influence the ability to repair damaged DNA, and un-repaired or improperly repaired DNA may lead to genetic instability and carcinogenesis. We evaluated the role of four tagging-SNPs (tSNPs) in the DNA repair gene, XRCC4, and its association with breast cancer risk and age at diagnosis of breast cancer in 139 high risk, non-BRCA1/2 breast cancer cases and controls from Utah pedigrees (cases = 464, controls = 576). We observed a significant association for two 4-locus tSNP haplotypes and age at diagnosis. Carriage of one haplotype was associated with later diagnosis (haplotype frequency = 0.039, mean age at diagnosis = 67.17 years, p=0.001) and carriage of the other was associated with earlier diagnosis (haplotype frequency 0.214, mean age at diagnosis = 54.04 years, p=0.0085). For breast cancer risk, two 2-locus tSNP haplotypes explained the observed association as well as extended 4-locus haplotypes. The two 2-locus haplotypes were nominally associated with breast cancer risk, one for reduced risk (OR= 0.57 (95% CI: 0.36-0.90), p=0.014) and one for increased risk (OR=1.30 (95% CI: 1.02-1.67), p=0.033). Moreover, one of our tSNPs was in high linkage disequilibrium (D’=0.81) with an XRCC4 SNP found to be significantly associated with breast cancer in Taiwan [Cancer Res 2003; 63:2440-2446], hence confirming their findings. Our results suggest that XRCC4 may play a role in the age at diagnosis and risk of breast cancer in non-BRCA1/2, heritable breast cancer cases.
Introduction

Although repair of double strand DNA breaks (DSB) is part of normal physiologic processes including meiotic recombination and immunoglobulin gene rearrangement (V(D)J recombination), DSB may also arise through toxic lesions by exposure to DNA damaging agents, such as ionizing radiation or radiomimetic compounds, or produced by the collapse of replication forks when the replication machinery encounters a single-stranded break. DSB lesions left un-repaired can threaten genome integrity and may, through the process of activating cell-cycle checkpoints, result in cell death. Un-repaired or improperly repaired DSB may lead to genomic rearrangements, destabilizing the genome, and eventually cancer.

In eukaryotic cells, DSB are repaired by two different pathways: homologous recombination (HR) and non-homologous end joining (NHEJ). HR, in brief, requires a homologous template on the sister chromatid in order to fix a break, whereas NHEJ, considered more error prone, reseals the two free DNA ends without the need of a template. DNA DSB repair pathways are of particular etiological importance during tumorigenesis, particularly breast cancer tumorigenesis. A number of high-penetrant mutations in genes involved in DSB repair have been found to be involved in breast cancer, including BRCA1, BRCA2, and ATM (see review). As high penetrant mutations explain only a small percentage of breast cancer, recent efforts have focused on common variants in DNA repair pathways and nominally significant results have been observed. Unfortunately, most of these studies have focused on a limited number of single nucleotide polymorphisms (SNPs) in the DNA repair pathways without regard to whether these common variants capture all or most of the underlying genetic variation across the
gene. A more thorough approach is to study tagging-SNPs (tSNPs), which are specifically selected to represent the majority of the underlying genetic variability.18

As the BRCA1 and BRCA2 genes play an important role in HR repair and these genes strongly predispose to cancer, much emphasis has been placed on the HR pathway to find additional breast cancer susceptibility genes.19,20 Recent evidence, however, suggests that BRCA1 may also play a role in the NHEJ pathway and hence the NHEJ pathway may also contribute to breast cancer pathophysiology.21,22 Animal studies have shown that BRCA1-deficient mouse embryonic fibroblasts were significantly more likely to have reduced NHEJ activity.23,24 In a recent report, Bau et al. found that breast cancer risk was associated with the joint effect of having a later age at first pregnancy, a higher number of high-risk genotypes in NHEJ genes, and carriage of glycine at the BRCA1 Glu1039Gly polymorphism. They further found using breast cancer cell lines that the preciseness of the NHEJ capacity was higher in BRCA1 expressing cell lines (MCF-7 cells) compared to cell lines with defective BRCA1 expression (HCC1937). These studies would suggest that the HR and NHEJ pathways might not be as distinct as was previously thought and further study of the genes involved in these pathways will be essential.

Because of the potential contribution of the NHEJ pathway to breast cancer and the need for a more rigorous analysis of genes involved in this pathway, the aim of the current study was to determine whether common polymorphic variants in one of the NHEJ pathway genes, XRCC4, is associated with breast cancer. Here we report the genotypic and haplotypic association of XRCC4 with breast cancer risk and age at diagnosis for breast cancer cases from high-risk, non-BRCA1/2, Utah breast cancer
families. Association studies using familial breast cancer cases can increase the power to detect rare low penetrance variants over that of unselected breast cancer cases.25

Materials and Methods

Subjects

Subjects for this study were selected from 139 non-\textit{BRCA1}\slash{}2, high-risk Utah breast cancer pedigrees of northern European descent. These families were originally selected because their rates of breast cancer exceeded the population rate of breast cancer determined using the Utah Population Database, a database linking genealogy data to the Utah Cancer Registry.26\textendash{}28 Breast cancer cases within these families had a low probability of being attributable to a \textit{BRCA1}\slash{}2 mutation because either breast cancer case(s) within the family tested negative for a \textit{BRCA1}\slash{}2 mutation, or the number and constellation of breast cancer cases present and/or ages of diagnosis of breast cancer in the family were unlikely to be due to a \textit{BRCA1}\slash{}2 mutation. Breast cancer case status and age at diagnosis were obtained from either medical records for the subject or the Utah Cancer Registry (UCR). As Utah is part of the Surveillance, Epidemiology, and End Results (SEER) Program, all breast cancers must be reported to the UCR, making it a reliable information source.

We selected nuclear families conducive for transmission\slash{}disequilibrium testing, composed of parent-affected female offspring trios and, when parental blood was unavailable, female sibships containing at least one affected and one unaffected sibling. These individuals were identified from the already-ascertained 139 Utah non-\textit{BRCA1}\slash{}2 high-risk breast cancer pedigrees described above. We supplemented the nuclear families by adding the remaining female non-\textit{BRCA1}\slash{}2 breast cancer cases from the breast cancer
pedigree resource for whom blood was available, together with a matched control based on birth year (within 5 years), female gender, and age at diagnosis, such that the control was cancer free at the age the case was diagnosed. These matched controls were also chosen from the breast cancer pedigree resource and were selected to be as distantly related to any other matched case and control as possible to maximize power, and as old as possible to ensure that they were less likely to become a case. In total the sample size contained all non-\textit{BRCA1/2} breast cancer cases with samples available (464 cases), and 576 controls (composed of unaffected parents from 39 parent-affected offspring trios, sisters from 167 sibships, and 236 matched controls). These subjects were part of pedigrees, which ranged in size from 1 (only 1 individual selected from a pedigree) to 1,195 individuals in a single pedigree, although typically only individuals at the bottom of each pedigree had data. All the subjects studied gave informed consent. This study was approved by the University of Utah Institutional Review Board.

\textbf{tSNP determination and genotyping of subjects}

We previously characterized the linkage disequilibrium structure, haplotype architecture, and identified four tSNPs that captured 97.2\% of the intragenic variation across the \textit{XRCC4} gene.\textsuperscript{29} In brief, we evaluated 21 SNPs from Applied Biosystems Assays-on-Demand across \textit{XRCC4} at a resolution of 1 SNP/13,198 bp using 94 unrelated individuals. Using a principle components analysis (PCA) method,\textsuperscript{30} we observed four linkage disequilibrium groups leading to the identification of four tSNPs, one for each group. The four tSNPs used for this study were rs1478485, rs13180316, rs963248, and rs1056503, which we will refer to as X1, X2, X3 and X4, respectively.
These four tSNPs were genotyped on the entire study population (N=1,040), using the same genotyping procedure as that used for the tSNP determination (see \textsuperscript{31} for genotyping details). For quality control, six individuals were duplicated across all plates. Analysis required that the quality control samples across plates have matching genotype assignments. Where possible, Mendelian inheritance was verified; samples with inheritance incompatibilities were either re-genotyped and/or set to missing if they could not be resolved.

Statistical analysis

As all of our subjects were selected from pedigrees and many of them are related, a correction is required for the genetic dependence between them. Without correction, an underestimate of the variance and an increase in the type I error rate may result. We utilized PedGenie,\textsuperscript{32} a freely-available tool developed by our group, to perform association testing between genetic markers and qualitative and quantitative traits in pedigree data of any size or structure.\textsuperscript{33} PedGenie accounts for the relatedness of individuals using a Monte Carlo approach to significance testing, whereby an empirical null distribution is generated and is used to determine the significance of an observed result. PedGenie performs classical tests of association and transmission disequilibrium for a single locus at a time analysis as well as phased haplotype data. Association tests were performed using data on all individuals in the sample (1,040 total individuals) and transmission tests were restricted to the subsample with relevant structure (39 trios and 167 sib-ships).

We tested associations for breast cancer status and age at diagnosis for breast cancer subjects with our four \textit{XRCC4} tSNPs. In all analyses, the base variant with the
minor allele frequency was considered as the allele of interest at each locus. We examined each of the four tSNPs independently and in multilocus haplotypes. For single locus analyses, the allele frequency estimation method ‘GeneCounter’ in PedGenie was used, such that simulations are based on allele frequencies that are statistically inferred for founder genotypes using maximum likelihood estimation. For haplotype analyses, PedGenie requires haplotype frequencies and recombination fractions between loci to be defined, so that the appropriate linkage disequilibrium structure across the gene is maintained when generating the empirical null distribution. Haplotype frequencies were determined from the panel of 94 unrelated individuals used for tSNP determination using an expectation maximization (EM) algorithm. The recombination fractions between each of the four tSNPs were set to zero as the distance between the SNPs was small (range: ~74 -115 kb).

Haplotype testing requires phase information for all genotyped subjects. Although a number of pedigree based haplotype methods are beginning to be developed, none are able to provide individual haplotype assignment on large pedigrees with large amounts of missing data, assuming linkage disequilibrium between multiple markers (see also ). We inferred haplotype phase information for all genotyped subjects using the EM algorithm, ignoring relationships. With an assumption of zero recombination, this is unbiased. We inferred haplotypes that could be assigned to an individual with >80% probability. All assigned haplotypes were checked for segregation within families wherever possible. Haplotypes that were incompatible within the family were set to zero.
All analyses were performed using PedGenie. Breast cancer risk was assessed using the odds ratio statistic. The quantitative age at diagnosis phenotype was assessed using a difference in means Z statistic and analysis of variance (ANOVA) statistic. Over-transmission of alleles and haplotypes for both breast cancer risk and age at diagnosis was assessed using the transmission/disequilibrium test statistic (TDT) and the quantitative TDT statistic on all individuals who were part of nuclear families (i.e., trios or sibships). For all analyses, the empirical null distribution and p-values from PedGenie were determined from a sample size of 2,000 null configurations.

To account for multiple testing, and realizing that all of the tests performed and the loci considered were not independent, we report all nominal findings ($p<0.05$) as interesting, and have considered a probability threshold of $p \leq 0.006$, accounting for 8 tests (4 tSNPs and 2 phenotypes), as significant.

**Results**

The total sample size for this study was 1,040 individuals, 464 cases and 576 controls. For the nuclear family subset, there were 39 two-parent/one affected offspring trio sets including 52 breast cancer cases, as some parents were also cases, and 167 unique sibships containing 182 cases and 275 controls. The mean ± standard deviation for age at diagnosis was 52.6 ± 12.6 years for the 219 of the 234 nuclear family cases for which diagnosis data were available. In the complement sample (i.e., matched case-control cohort), we originally matched the remaining 236 cases with 236 controls; however all genotyping assays failed for three cases resulting in 233 cases and 236 matched controls. The mean age at diagnosis for these cases was 58.2 ± 12.9 years. The overall average age at diagnosis was 55.6 ± 13.0 years in the complete study.
Table 6.1 shows the results for each of the four tSNPs analyzed separately for association with breast cancer risk, and for breast cancer cases, the age at diagnosis. Two tSNPs achieved significance for single locus tests for age at diagnosis (X1 and X2). A significant association was found for breast cancer cases homozygous for allele A at tSNP X1 (recessive model) resulting in a later age at diagnosis of cancer (p=0.003, mean age at diagnosis 59.1 years) compared to homozygote and wild type individuals (mean age at diagnosis 54.6 years). Tagging-SNP X2 showed an interesting result for carriage of allele A (dominant model, p=0.01), with the result driven by the heterozygotes who had significantly earlier age at diagnosis of cancer (p=0.004, mean age at diagnosis 53.4 years) compared to wild type individuals (mean age at diagnosis 57.0 years). For breast cancer risk, heterozygotes for tSNP X2 also approached nominal significance (p=0.079).

The TDT results were nominally significant for over-transmission of the rare allele A at X1 to breast cancer cases (TDT = 5.44, p = 0.02) determined using the trio TDT statistic; however, the sibTDT and combined TDT were nonsignificant. All other trio, sib and combined TDT statistics were nonsignificant for all other single locus analyses. For the quantitative TDT statistic, only the combined trio and sibship cohort had sufficient sample size to perform the analyses. The results, however, were nonsignificant across all single locus tests.

Table 6.2 lists all of the phased haplotypes and their frequencies that were observed in 94 unrelated individuals using the four \textit{XRCC4} tSNPs. Ten haplotypes ranging in frequency from 0.4% to 39.7% were observed. Assessment of pairwise linkage disequilibrium (LD) between the tSNPs showed weak and negative absolute
### Table 6.1. Single locus at a time associations of XRCC4 tSNPs with breast cancer risk and for cases, age at diagnosis

<table>
<thead>
<tr>
<th>Locus</th>
<th>Breast Cancer Risk</th>
<th>Age at diagnosis</th>
<th>OR (95% CI)</th>
<th>p</th>
<th>Means statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1 (rs1478485)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOM¹</td>
<td>0.98 (0.75, 1.27)</td>
<td>0.892</td>
<td>0.73</td>
<td>0.470</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REC²</td>
<td>1.08 (0.79, 1.48)</td>
<td>0.596</td>
<td>3.01</td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HET³ vs. WT⁴</td>
<td>0.95 (0.71, 1.27)</td>
<td>0.735</td>
<td>-0.24</td>
<td>0.818</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOM⁵ vs. WT</td>
<td>1.05 (0.74, 1.47)</td>
<td>0.767</td>
<td>2.44</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANOVA⁶</td>
<td>-</td>
<td>-</td>
<td>4.29</td>
<td>0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X2 (rs13180316)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOM</td>
<td>1.23 (0.97, 1.56)</td>
<td>0.081</td>
<td>-2.56</td>
<td>0.010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REC</td>
<td>1.05 (0.64, 1.74)</td>
<td>0.837</td>
<td>0.66</td>
<td>0.503</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HET vs. WT</td>
<td>1.25 (0.98, 1.59)</td>
<td>0.079</td>
<td>-2.88</td>
<td>0.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOM vs. WT</td>
<td>1.15 (0.69, 1.90)</td>
<td>0.590</td>
<td>-0.03</td>
<td>0.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANOVA</td>
<td>-</td>
<td>-</td>
<td>4.19</td>
<td>0.013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X3 (rs963248)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOM</td>
<td>1.01 (0.78, 1.30)</td>
<td>1.00</td>
<td>-0.78</td>
<td>0.441</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REC</td>
<td>0.86 (0.43, 1.72)</td>
<td>0.649</td>
<td>0.92</td>
<td>0.353</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HET vs. WT</td>
<td>1.02 (0.79, 1.33)</td>
<td>0.903</td>
<td>-1.09</td>
<td>0.283</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOM vs. WT</td>
<td>0.87 (0.43, 1.74)</td>
<td>0.664</td>
<td>0.81</td>
<td>0.405</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANOVA</td>
<td>-</td>
<td>-</td>
<td>1.22</td>
<td>0.291</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X4 (rs1056503)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOM</td>
<td>1.08 (0.81, 1.45)</td>
<td>0.583</td>
<td>0.52</td>
<td>0.611</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REC</td>
<td>0.94 (0.23, 3.92)</td>
<td>0.951</td>
<td>0.64</td>
<td>0.558</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HET vs. WT</td>
<td>1.09 (0.81, 1.47)</td>
<td>0.550</td>
<td>0.41</td>
<td>0.690</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOM vs. WT</td>
<td>0.96 (0.22, 4.11)</td>
<td>0.971</td>
<td>0.67</td>
<td>0.539</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANOVA</td>
<td>-</td>
<td>-</td>
<td>0.22</td>
<td>0.806</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

¹Dominant model (carriage of rare allele versus non-carriage); ²Recessive model (homozygous for the rare allele vs other genotypes); ³Heterozygous genotype; ⁴Wild type (homozygous for common allele); ⁵Homzygous for rare allele; ⁶ANOVA comparison includes three groups, one for each genotype (homozygous, heterozygous, and wild type). Interesting results are shown in bold.
Table 6.2. XRCC4 haplotypes observed among 94 unrelated individuals

<table>
<thead>
<tr>
<th>Haplotype Code</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>Haplotype Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>0.397</td>
</tr>
<tr>
<td>H2</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>0.214</td>
</tr>
<tr>
<td>H3</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>0.144</td>
</tr>
<tr>
<td>H4</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>0.127</td>
</tr>
<tr>
<td>H5</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>0.039</td>
</tr>
<tr>
<td>H6</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>0.028</td>
</tr>
<tr>
<td>H7</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>0.021</td>
</tr>
<tr>
<td>H8</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>0.016</td>
</tr>
<tr>
<td>H9</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>0.011</td>
</tr>
<tr>
<td>H10</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Minor allele frequency 0.47 0.23 0.19 0.09

Base (major/minor allele) G/A G/A T/C T/G
LD between X1 and X2 (D = -0.11) suggesting that opposite alleles tend to lie on the same haplotypes, but strong relative LD when accounting for allele frequencies (D’ = 0.99). LD between X1 and X3 was weak to moderate (D = -0.053, D’ = 0.59) and between X1 and X4 was weak (D = 0.0005, D’ = 0.01). The absolute LD between X2 and X3 was weak (D = -0.04), but the relative LD was strong (D’ = 0.99). Both absolute and relative LD were very weak between X2 and X4 (D = 0.0002 and D’ = 0.004) and between X3 and X4 (D = 0.003 and D’ = 0.035).

As X1 and X2 showed significance for the single locus tests and as they are in strong relative LD with each other, we performed 2-locus tSNP haplotype analyses across these two loci. Only three haplotypes were observed, G-G, G-A and A-G. Table 6.3 shows the association results for these 2-locus tSNP haplotypes for breast cancer risk and age at diagnosis of breast cancer. As would be expected, interesting results were found for age at diagnosis for homozygosity of haplotype A-G (p = 0.028) and carriage of haplotype G-A (p = 0.011), again with the signal coming from those heterozygous for G-A (p = 0.005). However, these results are less significant than those for the individual tSNPs (see above). For breast cancer risk, nominally significant results were observed for homozygosity of haplotype G-G (OR = 0.57, 95% CI: 0.36-0.90, p = 0.014) and carriage of haplotype G-A (OR = 1.30, 95% CI: 1.02-1.67, p = 0.033). It is interesting to note that although not significant after adjusting for multiple testing, these results for breast cancer risk are more significant than the single locus test for X2.

The TDT results for the 2-locus model were nominally significant for over-transmission of the A-G haplotype using the trio TDT statistic (TDT = 4.84, p = 0.039);
Table 6.3. Association of 2-locus XRCC4 tSNP haplotypes with breast cancer risk and for cases, age at diagnosis

<table>
<thead>
<tr>
<th></th>
<th>Breast Cancer Risk</th>
<th>Age at diagnosis</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
<td>95% CI</td>
<td>p</td>
</tr>
<tr>
<td>X1-X2: G-G (freq=0.298)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOM</td>
<td>0.81</td>
<td>(0.63, 1.03)</td>
<td>0.084</td>
</tr>
<tr>
<td>REC</td>
<td>0.61</td>
<td>(0.39, 0.94)</td>
<td><strong>0.033</strong></td>
</tr>
<tr>
<td>HOM vs. other</td>
<td>0.57</td>
<td>(0.36, 0.90)</td>
<td><strong>0.014</strong></td>
</tr>
<tr>
<td>HET vs. other</td>
<td>0.86</td>
<td>(0.67, 1.12)</td>
<td>0.295</td>
</tr>
<tr>
<td>ANOVA</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>X1-X2: G-A (freq=0.235)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOM</td>
<td>1.30</td>
<td>(1.02, 1.67)</td>
<td><strong>0.033</strong></td>
</tr>
<tr>
<td>REC</td>
<td>1.08</td>
<td>(0.63, 1.85)</td>
<td>0.767</td>
</tr>
<tr>
<td>HOM vs. other</td>
<td>1.20</td>
<td>(0.70, 2.05)</td>
<td>0.524</td>
</tr>
<tr>
<td>HET vs. other</td>
<td><strong>1.32</strong></td>
<td>(1.03, 1.70)</td>
<td><strong>0.026</strong></td>
</tr>
<tr>
<td>ANOVA</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>X1-X2: A-G (freq=0.468)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOM</td>
<td>1.05</td>
<td>(0.78, 1.41)</td>
<td>0.781</td>
</tr>
<tr>
<td>REC</td>
<td>1.06</td>
<td>(0.7831, 1.4445)</td>
<td>0.685</td>
</tr>
<tr>
<td>HOM vs. other</td>
<td>1.08</td>
<td>(0.76, 1.55)</td>
<td>0.682</td>
</tr>
<tr>
<td>HET vs. other</td>
<td>1.03</td>
<td>(0.75, 1.41)</td>
<td>0.871</td>
</tr>
<tr>
<td>ANOVA</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

1 Carriage of the haplotype of interest; 2 Homozygous for haplotype of interest compared to all other diplotpe combinations; 3 Homozygous for haplotype; 4 Other diplotpe combinations; 5 Heterozygous for haplotype of interest; 6 ANOVA comparison includes three groups, one for each diplotpe combination for the haplotype of interest (homozygous, heterozygous, and other). Interesting results are shown in bold.
however, the sib-TDT and combined TDT were nonsignificant. All other trio, sib, and combined TDT analyses for the G-G and G-A haplotypes were nonsignificant. The combined trio – sibship quantitative TDT statistics were also not significant for each of these 2-locus haplotypes.

The 2-locus tSNP haplotype results across X1 and X2 indicated that G-G and G-A were the more interesting haplotypes for breast cancer risk, and that G-A and A-G haplotypes were the most interesting for age at diagnosis. Tables 6.4 and 6.5 show association results for carriage of the 4-locus tSNP haplotypes extending from G-A and A-G for age at diagnosis, and G-G and G-A for breast cancer risk, respectively.

For breast cancer age at diagnosis, Table 6.4 considers the six haplotypes that begin with A-G or G-A. It was interesting to observe that only one haplotype from the four possible 4-locus haplotype extensions of A-G showed significant association with age at diagnosis. Carriage of A-G-T-G (H5) resulted in a significantly (p=0.001) later age at diagnosis of breast cancer (mean = 67.17 years) compared to all other diploid combinations of haplotypes (mean = 55.27 years). The other three haplotypes (H1, H6 and H10) indicated no association whatsoever. Similarly, when we considered the 4-locus extensions to the 2-locus haplotype G-A, only one of the two haplotypes indicated association (G-A-T-T, p=0.0085) with an effect towards earlier diagnosis of breast cancer (mean = 54.04 years) compared to all other haplotype combinations (mean = 56.63 years). Both of these results are not only more significant than the single and 2-locus results containing the relevant variants, but the mean diagnosis ages are more extreme. For these age-at-diagnosis results, the extension to 4-locus haplotypes appears to better extract the association evidence.
Table 6.4. Association of selected 4-locus XRCC4 haplotypes and age at diagnosis of breast cancer

<table>
<thead>
<tr>
<th>Haplotype ¹ (code)</th>
<th>Means statistic</th>
<th>p</th>
<th>Mean ± SD age at diagnosis (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-G-T-T (H1)</td>
<td>0.29</td>
<td>0.785</td>
<td>56.00 ± 12.87 (27-86)</td>
</tr>
<tr>
<td>A-G-T-G (H5)</td>
<td>3.61</td>
<td>0.001</td>
<td>67.17 ± 10.94 (49-84)</td>
</tr>
<tr>
<td>A-G-C-T (H6)</td>
<td>0.22</td>
<td>0.837</td>
<td>56.83 ± 15.58 (32-86)</td>
</tr>
<tr>
<td>A-G-C-G (H10)</td>
<td>0.62</td>
<td>0.562</td>
<td>57.14 ± 11.23 (39-81)</td>
</tr>
<tr>
<td>G-A-T-T (H2)</td>
<td>-2.66</td>
<td>0.0085</td>
<td>54.04 ± 11.58 (29-81)</td>
</tr>
<tr>
<td>G-A-T-G (H7)</td>
<td>1.26</td>
<td>0.248</td>
<td>61.43 ± 11.79 (45-74)</td>
</tr>
</tbody>
</table>

¹Carriage of each phased haplotype composed of tSNPs X1-X2-X3-X4 was compared to all other diploid combination of haplotypes.
For breast cancer risk, Table 6.5 considers the six haplotypes that begin G-A or G-G. Both 4-locus haplotypes extending from the 2-locus G-A haplotype showed nominally significant association with breast cancer risk. Carriage of haplotype G-A-T-T (OR=1.29, 95% CI: 1.00-1.66, p=0.048) and carriage of haplotype G-A-T-G (OR=2.23, 95% CI: 1.12-4.46, p=0.02) both retained the association for increased risk, as was seen for the 2-locus haplotype G-A. Similarly, for the four 4-locus haplotypes extending from G-G, two had nominal (G-G-T-T p=0.022 and G-G-C-G p=0.019) and one had significant association with decreased risk (G-G-T-G, p=0.003), consistent with the 2-locus findings for G-G. In contrast to the findings for age at diagnosis, the extension to 4-locus haplotypes does not suggest a single 4-locus haplotype that is superior in extracting the association evidence.

Discussion

Our results suggest that variants in the DNA repair gene XRCC4 may play an important role in the development of breast cancer and in determining the age at diagnosis of breast cancer. Our findings are based on four tSNPs, selected to capture the majority of the underlying variance across the XRCC4 gene, and using a sample of non-BRCA1/2 breast cancer cases from Utah high-risk pedigrees.

For breast cancer risk, we observed that certain haplotypes increased risk while others decreased risk. Haplotypes beginning with G-G showed protection against breast cancer while haplotypes beginning with G-A increased the risk of breast cancer. The increased or decreased risk of cancer was consistent across 2-locus and 4-locus haplotype analyses, with mostly nominal p-values and one significant p-value observed. Although a single, significant 4-locus haplotype was found, the vast majority of all 4-locus
<table>
<thead>
<tr>
<th>Haplotype^1 (code)</th>
<th>OR</th>
<th>95% CI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-A-T-T (H2)</td>
<td>1.29</td>
<td>(1.00, 1.66)</td>
<td>0.048</td>
</tr>
<tr>
<td>G-A-T-G (H7)</td>
<td>2.23</td>
<td>(1.12, 4.46)</td>
<td>0.02</td>
</tr>
<tr>
<td>G-G-T-T (H4)</td>
<td>0.74</td>
<td>(0.58, 0.94)</td>
<td>0.022</td>
</tr>
<tr>
<td>G-G-C-T (H6)</td>
<td>0.93</td>
<td>(0.66, 1.32)</td>
<td>0.891</td>
</tr>
<tr>
<td>G-G-C-G (H8)</td>
<td>0.25</td>
<td>(0.09, 0.69)</td>
<td>0.019</td>
</tr>
<tr>
<td>G-G-T-G (H9)</td>
<td>0.25</td>
<td>(0.11, 0.55)</td>
<td>0.003</td>
</tr>
</tbody>
</table>

^1Carriage of each phased haplotype composed of tSNPs X1-X2-X3-X4 was compared to all other diploid combination of haplotypes.
extensions of G-G and G-A shared the same characteristics of risk as their 2-locus counterparts. In other words, extension to the 4-locus haplotypes did not refine a haplotype that was superior at ‘tagging’ the underlying variant, and thus explaining the association. The fact that the underlying susceptibility variant or variants are tagged just as well by the 2-locus tSNP haplotypes across X1 and X2, indicates that the variant(s) likely would have arisen many generations ago such that they are common and likely in strong LD with X1 and X2. Recombination has allowed for multiple haplotypes containing the risk variant(s).

For our case-only study involving breast cancer age at diagnosis, we observed two very specific 4-locus tSNP haplotypes influencing age at diagnosis; one associated with later diagnosis (A-G-T-G: mean age at diagnosis 67.17 ± 10.14 years) and one associated with earlier diagnosis (G-A-T-T: mean age at diagnosis 54.04 ± 11.58 years). These two 4-locus haplotypes explained all of the association seen for the 2-locus haplotype analyses across X1 and X2. Thus, in contrast to the breast cancer risk analyses where 2-locus haplotypes explained the findings as well as 4-locus haplotypes, here two specific 4-locus haplotypes were superior at ‘tagging’ the underlying variant(s) for age at diagnosis, and thus also better explained the association evidence. The fact that the putative underlying susceptibility variant or variants seem to be tagged better by the 4-locus tSNP haplotypes indicates that the variant(s) likely would have arisen more recently, such that they are rarer and lie on more extended haplotypes.

The tSNPs used for this study were not selected as functional variants, but rather as markers to capture the underlying variation across the \textit{XRCC4} gene. Tagging-SNP rs1478485 lies in the mRNA untranslated region, tSNP rs13180316 and tSNP rs963248
are both intronic, and tSNP rs1056503 results in a synonymous coding change. Although there are no predicted effects for these individual polymorphisms on the protein sequence, they may be involved in the expression or stability of the XRCC4 mRNA, in modification of splicing, and/or in linkage disequilibrium with causal variant(s).

Several previous studies have examined association of a limited number of SNPs in the XRCC4 gene and risk of breast cancer. Fu et al. found a single SNP (rs2075685) in the XRCC4 gene to be significantly associated with breast cancer risk \( (p=0.02) \) in a Taiwanese breast cancer case-control study.\(^\text{15}\) Lee et al. tested the same significant SNP found by Fu et al.\(^\text{15}\) in a Korean hospital-based case control study, but did not find significance.\(^\text{16}\)

To investigate whether the significant SNP in the Fu et al.\(^\text{15}\) study (rs2075685) was in linkage disequilibrium with any of our XRCC4 tSNPs, we selected unrelated parental genotype data from the CEPH Utah families that are part of the HapMap.\(^\text{43}\) Two of our four tSNPs (rs13180316 (X2) and rs963248 (X3)) as well as rs2075685 were available for download. The relative pairwise linkage disequilibrium was moderate between rs2075685 and rs13180316 \( (D' = 0.51) \) and high between rs2075685 and rs963248 \( (D' = 0.81) \) determined using the program EMLD.\(^\text{44}\) The SNP rs2075685 maps outside the XRCC4 gene, but within ~2 kb of the XRCC4 mRNA transcript, near the 5' end of the gene. We also observed significant association for breast cancer risk near the 5' end of the XRCC4 gene through our positive associations with X1 and X2. Hence, it is likely that both our study and the Fu et al. study\(^\text{15}\) are detecting the same XRCC4 variant(s) that predisposes to breast cancer risk, and
interestingly its influence is observed in both a Caucasian and a Taiwanese population, again suggesting that the mutation(s) are more ancient.

To the best of our knowledge, this is the first study to note a significant association between tagging-SNP haplotypes in XRCC4 and age at diagnosis of breast cancer. Confirmation of these results in other populations is necessary. The risk haplotype for later age at diagnosis (and most likely the underlying causal variant) is fairly rare (frequency=0.039 in our population) whereas the opposite haplotype conferring an earlier age at diagnosis was more common (frequency=0.214 in our population) and hence the attributable risk to the breast cancer population could be considerable.

There are limitations of this study. Although the use of heritable breast cancer cases increases the power of association studies to detect low penetrance variants, it is most advantageous to use an independent set of hereditary breast cancer cases and a set of unrelated controls in a study. Hence, our power to detect significant effects for some of the rarer tSNPs may have been limited. Further, our study design was not as large as some studies although our focus was on a specific subset of breast cancer cases that are heritable and also not attributable to mutations in the BRCA1/2 genes. It will, therefore, be of interest to determine whether these results extend to larger cohorts of breast cancer cases, sporadic breast cancer cases, and if the haplotypes we identified modify age at diagnosis of breast cancers attributable to mutations in BRCA1/2. Finally, we did not sequence the XRCC4 gene to determine the tSNPs used for this analysis, rather we tested selected SNPs commercially available at a resolution of 1
SNP/\sim 10 \text{ kb}. Thus some variants may have been missed and these may have stronger associations with breast cancer risk.

In conclusion, the results of this study suggest that variants of the \textit{XRCC4} gene play an important role in both the development of breast cancer and in determining the age at diagnosis of hereditary breast cancer not attributable to \textit{BRCA1/2}. Further studies involving larger cohorts of women and more extensive genotyping across the \textit{XRCC4} gene are required to validate our findings and locate the underlying causal variants.

\textbf{Acknowledgements}

Kristina Allen-Brady is an NLM fellow, supported by NLM grant T15 LM0724. This research was supported by a dissertation research grant from the Susan G. Komen Breast Cancer Foundation for Kristina Allen-Brady (DISS0201521, to NJC) and an NIH NCI grant CA 098364 (to NJC). We appreciate the assistance of Kim Nguyen (Genetic Epidemiology) and Michael Hoffman (Family and Preventive Medicine) for their help in the laboratory. We also thank Helaman Escobar (Director of Sequencing and Genomics) and Michael Klein (Genomics) from the Core Resource Facilities, University of Utah, for use of their equipment and assistance on this project. Data collected for this publication were assisted by the Utah Cancer Registry supported by National Institutes of Health, Contract NO1-PC-35141, Surveillance, Epidemiology and End Results (SEER) Program, with additional support from the Utah Department of Health and the University of Utah. Partial support for all datasets within the Utah Population Database (UPDB) was provided by the University of Utah Huntsman Cancer Institute.
References


34. SNPHAP. [http://www-gene.cimr.cam.ac.uk/clayton/software]

35. Pedigree Analysis Package for Java. [http://hassiedt.genetics.utah.edu/jpap/]


43. HapMap. [http://www.hapmap.org]

44. EMLD. [https://epi.mdanderson.org/~qhuang/Software/pub.htm]


CONCLUSION

In the Introduction of the dissertation, three strategies to assist in identification of additional breast cancer susceptibility loci were proposed. Although these strategies are far from being all-inclusive, accomplishment of these strategies certainly furthers progress in breast cancer research and as expected, leads to additional areas of future research projects. Some of these future projects are listed below.

Our first proposed strategy was to define a novel, informative phenotype that could be used to identify a subset of breast cancer cases that was more likely to be attributable to a single gene, or minimal number of genes, disorder. We showed that the histological subtype lobular breast cancer increases the risk among family members for breast cancer and specifically lobular breast cancer compared to the risk associated with other types of breast cancer, suggesting that lobular cancer is a good phenotype for finding additional predisposition loci. As relatives of lobular breast cancer probands are at higher risk of breast cancer, we recommended that MRI be used to screen these relatives, as it is more sensitive than mammogram to detect breast cancer, and lobular breast cancer is particularly difficult to detect by either physical exam or mammography. Our recommendation was inferred from previous reports focusing mostly on women from BRCA1/2 families and should be studied prospectively to assess whether MRI screening detects breast cancer earlier in these families and is cost effective. Furthermore, future research using lobular breast cancer families to identify
additional breast cancer loci should be pursued. Identification of lobular breast cancer loci was beyond the scope of the dissertation project as it requires such resources as a team of investigators, study coordinators, and laboratory personnel to recruit lobular breast cancer families, genotype these families for lobular candidate genes or genome-wide scans using SNP microarray platforms and rigorous analysis of the data.

The Utah Population Database (UPDB) was utilized to identify the lobular breast cancer phenotype as being more homogenous. The UPDB may also be useful in identifying other homogenous subgroups of breast cancer, such as subgroups defined by other types of cancer occurring in breast cancer families. Additional analyses are planned, but are not part of the dissertation.

The second strategy presented was the need to develop flexible association techniques that could handle case control and pedigree data without restrictions on size or relationship structure, quantitative and qualitative data, haplotype data, and inclusion of covariates. For the dissertation project, the association package PedGenie was expanded, tested, and its functionality illustrated in two manuscripts, one of which has been published. PedGenie is a very useful tool for association analyses of pedigree data as it can handle single locus and haplotype analyses. Although, case-control data are considered more appropriate for association analyses, often pedigree data from previous linkage studies exist, which may have taken years to collect. Tools to perform pedigree-based association analyses are necessary to utilize this wealth of already ascertained data. We plan to expand PedGenie's functionality to assess power and perform gene - gene interaction tests. Gene - gene interaction tests hold promise for
detecting additional breast cancer susceptibility loci as each marker on its own may provide only weak information but in combination may be strong enough to be detected.

The final strategy proposed was multifaceted and provided suggestions to improve the reproducibility of published results and study design. This included such techniques as having an adequate study sample size, performing haplotype analyses using tagging-single nucleotide polymorphisms (tSNPs), and controlling for population stratification through the use of either family data or genomic control. We have incorporated these suggestions in our study of five candidate genes in the DNA repair pathway (ATM, MRE11, XRCC4, NBS1 and RAD50). We have reported linkage disequilibrium structure, haplotype groups, and identified tSNPs. Analysis of each gene separately is completed and hypotheses for interactions are ongoing. As part of the dissertation, a single locus at a time, composite genotype, and haplotype results are reported for XRCC4. Manuscripts including all the genes (ATM, MRE11, NBS1, RAD50 and XRCC4) will be pursued in the future, including haplotype analyses, testing for interactions between the genes, and the inclusion of covariate data.

Although identification of additional breast cancer susceptibility loci is beset with multiple challenges including locus heterogeneity, the possibility of reduced penetrance of multiple genes, inclusion of sporadic cases in family-based studies which complicate the results, high costs to recruit individuals and to perform genome-wide genotype scans on them, and not to mention the medical informatics challenges of storing, retrieving, and analyzing the almost overwhelming amount of data that is generated from genetic studies, the benefits of success will be enormous. Tools can be developed to screen women and identify who is at high-risk for breast cancer. These
high-risk women could be screened more aggressively so that their cancers are
diagnosed earlier, hopefully reducing mortality rates for the tumors that do occur. With
the identification of additional genes involved in breast cancer, therapeutic targets will
be identified and novel treatment strategies can be developed specific to these targets.
The challenges to the study of breast cancer are great, but the potential benefits of
success to all women, their families and society as a whole make the work and
overcoming these challenges worthwhile.
References


