Restriction Enzyme Generated Next-Generation Sequencing Libraries and Genetic Risk Modifiers of BRCA1 Mutation Carriers

Bradley Downs
University of Nebraska Medical Center

Follow this and additional works at: https://digitalcommons.unmc.edu/etd
Part of the Bioinformatics Commons

Recommended Citation
https://digitalcommons.unmc.edu/etd/74

This Dissertation is brought to you for free and open access by the Graduate Studies at DigitalCommons@UNMC. It has been accepted for inclusion in Theses & Dissertations by an authorized administrator of DigitalCommons@UNMC. For more information, please contact digitalcommons@unmc.edu.
RESTRICTION ENZYME GENERATED NEXT-GENERATION SEQUENCING
LIBRARIES AND GENETIC RISK MODIFIERS OF BRCA1 MUTATION CARRIERS

by

Bradley M. Downs

A DISSERTATION

Presented to the Faculty of
the University of Nebraska Graduate College
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Genetics, Cell biology & Anatomy
Graduate Program

Under the Supervision of Professor San Ming Wang

University of Nebraska Medical Center
Omaha, Nebraska

April, 2016

Supervisory Committee:
Karen A. Gould, Ph.D.               Chittibabu Guda, Ph.D.
James D. Eudy, Ph.D.               Andrew T. Dudley, Ph.D.
Acknowledgments

This dissertation wouldn't be possible without the help of many individuals.

I would like to thank Dr. San Ming Wang for his guidance and giving me the opportunity to be a graduate student in his laboratory. Thank you to all past and present members of Dr. Wang’s laboratory.

Special thanks to Divya Bhagirath, Shanshank Sirvastava, Appo Olou, and Nick Griffin for all their help. I would also like to thank Dr. Pudita Dhawa’s lab for helping me with the florescence microscopy experiments, Dr. Henry Lynch’s lab for providing all of the $BRCA1$ mutation carrier DNA samples, Dr. Angelika Noegel’s lab for the Nespin-1 construct and Dr. Andrew Wilde’s for the Anillin construct.

I would also like to give thanks Dr. Robert Norgren, Dr. Shelley Smith, Dr. Lynette Smith, Dr. Hamid Band, Dr. Vimla Band, and my graduate committee members for all the help and mentoring they have given me during the course of my graduate career. Their assistance made my projects more significant and helped me through my Ph.D. training.

Special thanks to Tramy Hoang for the moral and spiritual support she has given me throughout my Ph.D. training. Whether it was editing my take home exams, editing my comprehensive exam or editing my dissertation, Tramy has always been there to help me and for that I am forever grateful.
Next-generation sequencing (NGS) is a high throughput technique used to sequence large amounts of DNA in a short amount of time. However, a limitation to NGS is that the generated data is in a single consensus sequence without distinguishing between variants on homologous chromosomes. Separating or phasing the variants from the maternal and paternal chromosomes can provide information about the genetic origin of disease and information about how DNA nucleotide alterations interact in cis. This dissertation explores a new technical method of using restriction enzymes during NGS library preparation and its ability to increase the amount of phasing information that can be derived from NGS data. This study provides evidence that increasing the fragment size of NGS libraries can increase the amount of variant phasing information derived from NGS data.

BRCA1 is a well-known tumor suppressor that, when mutated, predisposes the mutation carrier to breast cancer. \textit{BRCA1} mutation carriers have a 44-75\% risk of developing breast cancer by age 70. In this study, we used next-generation sequencing data to identify germline genetic variants that modify the risk of breast cancer in \textit{BRCA1} mutation carriers. With the use of both biological and statistical filters, five variants were identified that changed breast cancer risk in \textit{BRCA1} mutation carriers. Furthermore, it was shown that two of the affected genes alter the growth of \textit{BRCA1} mutation breast cell
lines. Perhaps, more importantly, the two variants were shown to alter the function of the affected genes. This is the first study to provide functional evidence on how common genetic variants can modify the risk of breast cancer in *BRCA1* mutation carriers.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................ i

ABSTRACT ......................................................................................................................... ii

TABLE OF CONTENTS ........................................................................................................ iv

LIST OF FIGURES .............................................................................................................. x

LIST OF TABLES ................................................................................................................ xii

LIST OF ABBREVIATIONS ................................................................................................. xiii

INTRODUCTION .................................................................................................................. 1

The human genome .............................................................................................................. 2

The human genome project ............................................................................................... 2

Sanger sequencing .............................................................................................................. 3

Next-generation sequencing ............................................................................................. 3

Illumina Sequencing .......................................................................................................... 4

Computational steps used to align Illumina sequenced reads ............................................. 7

The cancer genome ............................................................................................................ 7

Breast cancer ..................................................................................................................... 9

CHAPTER 1: Restriction enzyme generated next-generation sequencing libraries ....... 11

1.1 Introduction ................................................................................................................ 12

1.1.1 Current limitations of NGS ability to phase the human genome ...................... 12

1.1.2 Phase information derived from NGS ................................................................. 12

1.1.3 Computational and experimental approaches to phase the genome ............ 14
1.1.4 Using restriction enzymes to construct NGS libraries ......................................... 14
1.1.5 Selection of restriction enzyme ........................................................................... 15
1.1.6 Genomic DNA sample selection .......................................................................... 15

1.2 Methods .................................................................................................................... 17
1.2.1 Genomic DNA used to generate restriction enzyme exome libraries .......... 17
1.2.2 Restriction enzyme generated TruSeq DNA sample preparation ................. 17
  1.2.2.1 Restriction Digestion ....................................................................................... 17
  1.2.2.2 Blunting and phosphorylating the gDNA ....................................................... 18
  1.2.2.3 Adenylate of the 3’ end of the blunt DNA fragments ............................... 18
  1.2.2.4 Ligating the Adaptors ....................................................................................... 18
  1.2.2.5 Library Enrichment ......................................................................................... 19
  1.2.2.6 Exome Enrichment ......................................................................................... 19

1.3 Results ..................................................................................................................... 21
1.3.1 Mapping sequenced reads ................................................................................... 21
1.3.2 GATK called variants ........................................................................................ 24
1.3.3 Phasing variants .................................................................................................. 27

1.4 Discussion ............................................................................................................... 27

CHAPTER 2: Identification and characterization of genetic modifiers of breast cancer risk in *BRCA1* mutation carriers .................................................................................................................. 31
2.1 Introduction .............................................................................................................. 32
2.1.2 BRCA1 ................................................................................................................. 32
2.1.3 Genetic modifiers of breast cancer risk ............................................................... 32
2.1.4 Intervention options of BRCA1 mutation carriers .............................................. 33
2.1.5 Benefits of NGS analysis .................................................................................... 33

2.2 Materials and Methods .......................................................................................... 36

2.2.1 Statistical association study .................................................................................. 36
  2.2.1.1 Samples used in the case control study ............................................................... 36
  2.2.1.2 Samples used in the cohort study ..................................................................... 36
  2.2.1.3 Statistical association of genetic modifiers to breast cancer risk ............... 37
  2.2.1.4 Illumina Exome libraries preparation ............................................................... 37
  2.2.1.5 Mapping, variant calling and damaging prediction ........................................ 37
  2.2.1.6 PCR products Sanger sequenced ..................................................................... 38

2.2.2 Functional Study .................................................................................................. 39
  2.2.2.1 Bacterial strains ............................................................................................... 39
  2.2.2.2 Bacterial Cell Growth Conditions .................................................................. 39
  2.2.2.3 Bacterial transformation .................................................................................. 40
    2.2.2.3.1 Electorcompetent transformation ................................................................. 40
    2.2.2.3.2 Chemical competent transformation .......................................................... 40
  2.2.2.4 Mammalian Cell line Growth Conditions ....................................................... 41
  2.2.2.5 shRNA stable cell line production .................................................................. 41
  2.2.2.6 Lentiviral particles production ........................................................................ 41
2.2.2.7 Infecting Target Cells ................................................................. 42
2.2.2.8 DNA Transfection Procedure ..................................................... 42
2.2.2.9 Growth curve ............................................................................. 43
2.2.2.10 Microscopy ............................................................................... 43
2.2.2.11 Western Blot ............................................................................. 43
2.2.2.12 ANLN construct design and verification ..................................... 44
2.2.2.13 SYNE1 construct design and verification .................................... 45

2.3 Results ............................................................................................. 45

2.3.1 Statistical association study ............................................................. 45
  2.3.1.1 \textit{BRCA1} mutated carriers sample selection ............................. 45
  2.3.1.2 NGS mapping, variant calling and variant annotation ................ 46
  2.3.1.3 Case control study to identify candidate genetic modifiers .......... 46
    2.3.1.3.1 Previously identified breast and/or risk modifying variants ...... 46
    2.3.1.3.2 Newly identified breast and/or risk modifying variants .......... 52
  2.3.1.4 Cohort study to calculate the cumulative risk of the candidate genetic modifiers .................................................................................. 53

2.3.2 Functional study ............................................................................. 55
  2.3.2.1 ANLN ......................................................................................... 62
    2.3.2.1.1 Anillin and its role during cytokinesis ................................... 62
    2.3.2.1.2 Anillin functional domains .................................................. 62
    2.3.2.1.3 Anillin role in Cancer ......................................................... 65
2.3.2.1.4 ANLN knockdown ................................................................. 65

2.3.2.1.5 Variant rs3735400 affects the function of Anillin ..................... 67
    2.3.2.1.5.1 Nuclear localization ...................................................... 67
    2.3.2.1.5.2 Knockdown growth rate rescue ....................................... 72

2.3.2.2 SYNE1 ................................................................................. 72
    2.3.2.2.1 SYNE1 and its role in nuclear membrane stability .................. 72
    2.3.2.2.2 SYNE1 and cancer ............................................................. 74
    2.3.2.2.3 Knockdown SYNE1 ......................................................... 75
    2.3.2.2.4 Variant rs2295190 affects the function of Nesprin-1 ............. 75

2.4 Discussion ..................................................................................... 80
    2.4.1 Statistical association that common variants can modify the risk of breast cancer in \textit{BRCA1} mutation carriers ................................................................. 80
    2.4.2 Functional studies with ANLN\textsubscript{rs}3735400 and SYNE1\textsubscript{rs}2295190 ........ 81
        2.4.2.1 ANLN\textsubscript{rs}3735400 ..................................................... 81
        2.4.2.2 SYNE1\textsubscript{rs}2295190 ................................................. 82

DISCUSSION ......................................................................................... 84

Phasing the genome with NGS ............................................................. 85

Future directions for phasing the genome with NGS ................................. 86

Genetic modifiers of breast cancer risk of \textit{BRCA1} mutation carriers ............. 87

Future directions for genetic modifiers of breast cancer risk of \textit{BRCA1} mutation carriers ........................................................................................................ 90
Two *PALB2* germline mutations found in both *BRCA1*+ and *BRCAx* familial breast cancer
LIST OF FIGURES

Figure 1. Illumina based genomic library construction and Illumina HiSeq based NGS ... 5

Figure 2. Exome capture with the Illumina HiSeq based NGS ........................................ 6

Figure 3. Hereditary predispositions of breast cancer ...................................................... 10

Figure 4. Next-generation sequencing read combinations providing variant phasing
information .......................................................................................................................... 13

Figure 5. GATK best practice flow cart ............................................................................. 20

Figure 6. Pooling restriction enzyme libraries increases the percent of exome
nucleotides covered........................................................................................................... 23

Figure 7. Mapped reads generated by sonication and restriction enzymes................. 26

Figure 8. Distribution of variants on the 610k array ......................................................... 35

Figure 9. Probable biological variants previously identified to associate with breast or
ovarian cancer risk ........................................................................................................... 51

Figure 10. Probable biological variants newly identified to associate with breast cancer
risk .................................................................................................................................. 53

Figure 11. SYNE1_rs2295190 cumulative risk of breast cancer ................................. 56

Figure 12. ANKLE1_rs8100241 cumulative risk of breast cancer ............................... 57

Figure 13. ANLN_rs3935400 cumulative risk of breast cancer ...................................... 58

Figure 14. SIPA1L2_rs1547742 cumulative risk of breast cancer .................................. 59

Figure 15. MTHFSD_rs3751800 cumulative risk of breast cancer ............................... 60

Figure 16. Candidate modifiers of breast cancer risk chosen for functional studies ...... 63

Figure 17. Localization of Anillin during the cell cycle in mammals ............................ 64
Figure 18. Secondary structures of Anillin protein in humans........................................66

Figure 19. Growth rate measurement of MCF7 infected shRNA_ANLN cells ............68

Figure 20. N-terminus of Anillin has conserved NLS across different vertebrate species69

Figure 21. Anillin constructs transfected into HeLa cells .................................................70

Figure 22. ANLN_rs3735400 affects the nuclear localization of Anillin ......................71

Figure 23. ANLN_rs3735400 inhibits Anillin ability to rescue the ANLN knockdown.....73

Figure 24. Growth rate measurement of MCF7 infected shRNA_SYNE1 cells ..........76

Figure 25. Secondary structures of Nesprin-1 protein in humans...............................77

Figure 26. C-terminus KASH domain of Nesprin-1 is conserved in different vertebrate

species ..................................................................................................................................78

Figure 27. SYNE1-GFP tagged constructs transfected into HeLa cells......................79
**LIST OF TABLES**

Table 1. Average genomic digested fragment length by restriction enzyme .............. 16

Table 2. Summary of restriction enzyme exome library sequenced with Illumina HiSeq 2000 100bp pair-end reads ................................................................. 22

Table 3. Comparison of GATK called variants from restriction enzyme reads and sonication reads ................................................................. 25

Table 4. Comparison of GATK called variants from ≥2 restriction enzyme library reads and sonication reads ................................................................. 28

Table 5. Phased variants per contig from restriction enzyme and sonication libraries ... 29

Table 6. Summary of published genetic modifiers of breast cancer risk in *BRCA1* mutation carriers................................................................................. 34

Table 7. Breast cancer cases and controls selected for the case control study .......... 47

Table 8. Breast cancer cases and controls sequenced with Illumina HiSeq 2500 150bp pair-end reads ................................................................................. 48

Table 9. Breast cancer cases and controls variants called with GATK and analyzed with PolyPhen2 and SIFT ................................................................. 49

Table 10. Sanger sequenced risk modifying variants Odds Ratio per allele ............... 54
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABD</td>
<td>Actin-binding domain</td>
</tr>
<tr>
<td>ANKLE1</td>
<td>Ankyrin repeat and LEM domain containing 1</td>
</tr>
<tr>
<td>ANLN</td>
<td>Anillin actin binding protein</td>
</tr>
<tr>
<td>APC</td>
<td>Anaphase prompting complex</td>
</tr>
<tr>
<td>BAM</td>
<td>Binary Alignment/Map</td>
</tr>
<tr>
<td>BRCA1</td>
<td>Breast cancer 1, early onset</td>
</tr>
<tr>
<td>BRCA2</td>
<td>Breast cancer 2, early onset</td>
</tr>
<tr>
<td>BRCAx</td>
<td>Non-BRCA1/2 familial breast tumors</td>
</tr>
<tr>
<td>BWA</td>
<td>Burrows-Wheeler Aligner</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>b37</td>
<td>Build 37</td>
</tr>
<tr>
<td>CDH1</td>
<td>Cadherin 1, type 1</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence interval</td>
</tr>
<tr>
<td>C-terminal</td>
<td>Carboxyl terminal</td>
</tr>
<tr>
<td>Dbox</td>
<td>Destruction box</td>
</tr>
<tr>
<td>dbSNP</td>
<td>The single nucleotide polymorphism database</td>
</tr>
<tr>
<td>ddH20</td>
<td>double distilled H20</td>
</tr>
<tr>
<td>mDia2</td>
<td>mammalian homolog of <em>Drosophila</em> diaphanous 2</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTPs</td>
<td>Deoxyribose nucleoside triphosphate</td>
</tr>
<tr>
<td>ER</td>
<td>Estrogen receptor</td>
</tr>
<tr>
<td>F-actin</td>
<td>Filamentous actin</td>
</tr>
<tr>
<td>GATK</td>
<td>Genome Analysis Toolkit</td>
</tr>
<tr>
<td>gDNA</td>
<td>Genomic deoxyribonucleic acid</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-wide association study</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>HER2</td>
<td>Human epidermal growth factor receptor 2</td>
</tr>
<tr>
<td>HFF1</td>
<td>Human foreskin fibroblast 1</td>
</tr>
<tr>
<td>hg19</td>
<td>Human genome reference 19</td>
</tr>
<tr>
<td>HR</td>
<td>Hazard ratio</td>
</tr>
<tr>
<td>IPA</td>
<td>Ingenuity Pathway Analysis</td>
</tr>
<tr>
<td>KASH</td>
<td>Klarsicht-ANC-SYNE-homology</td>
</tr>
<tr>
<td>kb</td>
<td>Kilobase</td>
</tr>
<tr>
<td>LINC</td>
<td>Linker of nucleoskeleton and cytoskeleton</td>
</tr>
<tr>
<td>MAF</td>
<td>Minor allele frequency</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger ribonucleic acid</td>
</tr>
<tr>
<td>MTHFSD</td>
<td>Methenyltetrahydrofolate synthetases domain containing</td>
</tr>
<tr>
<td>NEB</td>
<td>New England Biolabs</td>
</tr>
<tr>
<td>NC-Exon</td>
<td>Noncoding exons</td>
</tr>
<tr>
<td>NHGRI</td>
<td>National Human Genome Research Institute</td>
</tr>
<tr>
<td>NGS</td>
<td>Next-generation sequencing</td>
</tr>
<tr>
<td>NLS</td>
<td>Nuclear localization sequence</td>
</tr>
<tr>
<td>nt</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>N-terminal</td>
<td>Amino terminal</td>
</tr>
<tr>
<td>OR</td>
<td>Odds ratio</td>
</tr>
<tr>
<td>PALB2</td>
<td>Partner and localizer of BRCA2</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PP2</td>
<td>PolyPehn 2</td>
</tr>
<tr>
<td>Pre-mRNA</td>
<td>Precursor messenger ribonucleic acid</td>
</tr>
<tr>
<td>PR</td>
<td>Progesterone receptor</td>
</tr>
<tr>
<td>RhoA</td>
<td>Ras homolog family member A</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>SAM</td>
<td>Sequence Alignment/Map</td>
</tr>
<tr>
<td>SH3</td>
<td>Src-homology-3-binding consensus sequence</td>
</tr>
<tr>
<td>SIFT</td>
<td>Sorting intolerant from tolerant</td>
</tr>
<tr>
<td>SIPA1L2</td>
<td>Single-induced proliferation-associated 1 like 2</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>siRNA</td>
<td>Small interfering ribonucleic acid</td>
</tr>
<tr>
<td>SYNE1</td>
<td>Spectrin repeat containing, nuclear envelope 1</td>
</tr>
<tr>
<td>TCGA</td>
<td>The Cancer Genome Atlas</td>
</tr>
<tr>
<td>TP53</td>
<td>Tumor protein 53</td>
</tr>
<tr>
<td>UTRs</td>
<td>Untranslated regions</td>
</tr>
<tr>
<td>wt</td>
<td>Wild type</td>
</tr>
</tbody>
</table>
INTRODUCTION
The human genome

The genome is comprised of all the genetic material encoded as DNA. For humans (Homo sapiens), the genome is comprised of 23 chromosome pairs (44 autosomes and 2 sex chromosomes) as well as mitochondria DNA. Every cell within the human body with a nucleus has 46 chromosomes, with the exception of germ cells. Germ cells have a haploid genome comprised of 23 chromosomes.

The human genome can be divided into coding and noncoding sequences. Coding DNA sequences occupy ~2% of the genome and can be transcribed into mRNA which is then translated into proteins. Coding regions are called exons while the noncoding regions are called introns, untranslated regions (UTRs), and noncoding exons (NC-Exon). Once a coding region is transcribed into precursor mRNA (pre-mRNA), the primary transcript is converted into mature mRNA by the capping of a 7-methylguanosine residue to the 5' terminal end, polyadenylation of the 3' terminal end, and splicing of the introns.

Noncoding DNA do not encode protein but may contain genes that are transcribed into functional RNA molecules such as ribosomal RNA and microRNAs. Noncoding DNA sequences can also bind transcription factors (ie enhancers, insulators, promoters, etc), which can regulate gene expression and/or genomic stability.

The human Genome project

Prior to the Human Genome Project, very little was known about the contents of the human genome. The Human Genome Project began in 1990, took thirteen years and roughly ~3 billion dollars to complete. The goal for this project was to, for the first time, sequence the entire human genome. The Project sequenced ~3 billion base pairs and ~20,500 genes in the genome (http://www.genome.gov/12011238). The information
gained from this project was largely used to construct the human reference genome and many bioinformatics laboratories still use today.

**Sanger sequencing**

The Human Genome Project relied on the sequencing technique developed by Frederick Sanger in 1977. The Sanger method relies on electrophoretic separation of mixtures of randomly terminated extension products. This type of sequencing uses an automated method of fluorescently labeled terminators, capillary electrophoresis separation, and automated laser signal detection for improved nucleotide sequence detection. The biggest disadvantage of this method is the slow throughput. Despite this, Sanger sequencing utilizes well-defined chemistry, making it the most accurate method for sequencing currently available.

**Next generation sequencing**

In 2004, The National Human Genome Research Institute (NHGRI) initiated a funding program with the aim to reduce cost of genome sequencing. The current high-throughput DNA sequencing technologies frequently used today is next-generation sequencing (NGS) and massive-parallel sequencing. All NGS platforms monitor the sequential addition of nucleotides to immobilize and spatially array DNA templates but differ substantially in how these templates are generated and how they are interrogated to reveal their sequences. The increased throughput of NGS reactions comes at the cost of read length of 30-400bp compared to Sanger sequencing, which can read 0.5-1kb. Because of the short reads, NGS uses resequencing, which compares the density and sequence content of shorter reads to a reference genome to make the genome map. The human reference genomes being used today, hg19 and b37, are largely constructed from the results of the Human Genome Project.
Illumina Sequencing

All NGS requires the starting material to be converted into a library of sequencing reaction templates first (the size of which depends on each sequencing platform’s specifications). Next, platform-specific synthetic DNA adapters are then ligated to the end of the library fragments. Depending on the NGS technology, the NGS library is either sequenced directly or is amplified first, then sequenced.

The only NGS platform used in this dissertation was an Illumina HiSeq based NGS. The Illumina HiSeq platform technology uses genomic DNA hybridization and bridge amplifies on its specific grafted flowcell. Bridge amplification on the specific grafted flowcell is unique to Illumina NGS technology. After the genomic library is annealed and clonally amplified, a fluorescently labeled reversible terminator is imaged as the dNTPs are added (Figure 1).

It is also possible to capture targeted genomic sequences within the NGS library. Arguably, the most widely used targeted sequencing method is exome capturing. These targeted regions are first hybridized by DNA bound biotinylated probes. The biotinylated probes can then be captured with the selected target regions (Figure 2). Sequencing only the exons, ~2% of the genome, saves on cost and increases the rate of data processing.

While NGS has fast throughput, it is difficult to align higher levels of diversity to the reference genome because of the short reads. This issue is combated through the use of longer read lengths or paired-end/mate-pair sequencing. Length selection and the options of paired-end or singled-end sequencing are options on HiSeq sequencing systems. Once the sequencing run is complete, computational tools are needed analyze
the quality of each called nucleotide and to map the sequences to the reference genome.

**Figure 1. Illumina based genomic library construction and Illumina HiSeq based NGS.** Once the genomic DNA is sheared to the correct fragment size, the adapters are attached to the sequenced library. The constructed library is then applied to the flow cell and clutters are generated by bridge amplification. After amplification, a fluorescently labeled reversible terminator is imaged as the dNTPs are added\(^3\).
Figure 2. Exome capture with the Illumina HiSeq based NGS. Once the genomic DNA is processed to an appropriate fragment size, DNA bound biotinylated probes are hybridized to the fragmented DNA. Next, the probes are pulled down and the captured DNA is sequenced on the Illumina HiSeq platform.
Computational steps used to align Illumina sequenced reads

Once the sequencing data from the HiSeq run are collected, the low quality sequences, which could lead to poor mapping and false positive variant calling need to be removed. Quality of sequencing decreases the further the reads are sequenced therefore, Illumina recommends that a percentage of the read ends be trimmed. This percentage is based on the length of read during sequencing. After the low quality sequences are removed, software can align the reads to the reference genome. To increase the accuracy of insertions, deletions, and repetitive regions of the genome, the reads need to be locally realigned. If the library was amplified via PCR before sequencing, the duplicates must be marked to remove errors that might have arisen during the PCR amplification process. The variations between the sequenced library and the reference genome can now be called. The next step is to annotate the called variants. Depending on the type of study and the investigator hypothesis, these annotated variants will be analyzed further in a study specific manner.

The cancer genome

Cancer results when the body is unable to regulate cellular growth. The study of cancer genomes can be divided into two fields of study (i) germline mutations that change the risk of developing cancer and (ii) somatic genomic alterations that contribute to tumorigenesis, invasion, metastasis, and relapse.

Somatic mutations are any alterations of the germline DNA sequence in the genome, which can be separated into passenger and driver mutations. Driver mutations are mutations of the DNA that gives a selective advantage to a cancer cell. Passenger mutations are mutations that have no fitness effect but may be linked to the genome with driver mutations.
Germline mutations are found in every genome within the body and can be passed down in a hereditary fashion. In the general population, familial predispositions for cancer is rare, giving evidence that germline mutations that increase the predisposition for cancer must also be rare. The frequency of minor alleles can be defined by the minor allele frequency (MAF). The MAF refers to the frequency at which the allele occurs in a given population. Alleles with a MAF higher than 1 percent are usually considered common alleles while a MAF lower than 1 percent is considered rare\textsuperscript{5}. While the rarity of the mutation is an important factor when searching for mutations that influence cancer development, the mutation’s effect on the cell is also an important factor to consider when investigating cancer related germline or somatic mutations.

Mutations can be categorized by the effect they have on the protein they affect. Because of the redundancy of codons (multiple codons can translate the same amino acid), silent or synonymous mutations are changes to the DNA nucleotide that do not cause a change in amino acid. Missense or nonsynonymous mutations cause a change in the translated amino acid. Splice mutations affect mRNA processing and can cause deletion of exonic translation. Nonsense or stop gain mutations change the amino acid codon into a stop codon leading to a truncated protein product. The last type of mutation is frame shifts or out of frame insertions or deletions. This type of mutation generally causes a premature stop codon, generating a truncated protein. Nonsense and frame shift mutations generally have the largest effect on the translated protein. Many software packages can predict the effects of missense mutations. However, these mutations are generally considered variants of unknown significance until the mutations undergo rigorous statistical analysis and/or the effects of the mutation are investigated by functional assays.
Breast cancer

Breast cancer is the most frequent cancer diagnosis in women with an estimated 1.38 million new diagnoses worldwide in 2008\textsuperscript{6}. In the United States, the chance of a woman being diagnosed with breast cancer during her lifetime is one in eight\textsuperscript{7}. About 10% of women diagnosed with breast cancer have a hereditary predisposition for breast cancer\textsuperscript{8,9}. BRCA1 and BRCA2 are the most common autodominant predisposition for hereditary breast cancer and also the most well studied, comprising \~20% of all familial breast cancer cases. Interestingly, \~50% of the familial breast cancer cases have an unknown predisposition mutation and are thus labeled as BRCAx\textsuperscript{10} (Figure 3).

This dissertation covers a wide range of topics including analyzing new methods of NGS library construction, investigating potentially new breast cancer predisposition mutations and the identification and characterization of genetic germline modifiers that change the risk of developing breast cancer in BRCA1 mutation carriers.
Figure 3. **Hereditary predispositions of breast cancer.** All breast cancer cases can be separated into sporadic cases and cases with germline predisposition mutations. Mutations in the BRCA genes comprise 20% of all familial breast cancers. The other cases have an unknown genetic predisposition and are thus labeled as BRCAx\textsuperscript{10,11}. 
CHAPTER 1: Restriction enzyme generated next-generation sequencing libraries
1.1 Introduction

1.1.1 Current limitations of NGS ability to phase the human genome

*Homo sapiens* genomes are diploid, receiving one chromosome copy from their mother and one copy form their father. Genome phasing is the process of separating the maternal and paternal chromosomes from a single genome. Information derived from genomic phasing can provide the genetic origin of disease and information about how DNA nucleotide alterations interact *in cis*. Currently, most sequencing techniques generate a single consensus sequence without distinguishing between variants on homologous chromosomes.

Since the conception of next generation sequencing, both the throughput and sequencing read length has increased substantially. However, even with the advances of NGS technology, the biggest disadvantage of NGS system is the relatively short read length of 30-400bp, depending on the NGS platform. Due to the short read length, NGS relies on the method of resequencing to map the reads to the correct position on the reference genome. While this method has been shown to accurately map sequenced reads to a reference genome, it is unable to phase the sequenced chromosomes.

1.1.2 Phase information derived from NGS

Once the genomic libraries are sequenced on the Illumina HiSeq 2000 platform, there are 3 NGS read combinations that can be used to construct contigs and provide variant phasing information; variants can be phased i) in the same read, ii) in different reads, or iii) by paired reads (Figure 4). These 3 NGS read combinations can be integrated to phase multiple variants on a single contig.
Figure 4. Next-generation sequencing read combinations providing variant phasing information. The three next generation sequencing reads that can be used to construct contigs and provide variant phasing information are i) variants phased in the same read, ii) variants phased in different reads, iii) variants phased in pair-end reads. Red lines represent called variants.
1.1.3 Computational and experimental approaches to phase the genome

There are two NGS approaches used for phasing the genome: computational and experimental. Experimental approaches attempt to phase the genome prior to sequencing and rely on the physical separation of the chromosomes. The computational approach attempts to phase the genome post-sequencing. This method uses statistics to estimate the probability that two variants are linked and thus providing phase information. Computational phasing is mostly used to reconstruct haplotypes within a single pedigree. Neither computational nor experimental approaches are commonly used because they are thought to be too experimentally laborious and/or computationally underpowered. In the following project, we propose the use of a new experimental approach that is not experimentally laborious and has the potential to improve the ability of NGS to phase the genome.

1.1.4 Using restriction enzymes to construct NGS libraries

There are two types of experimental approaches that can be used to increase the phasing ability of next-generation sequencing. One can either i) improve NGS platform technology so longer reads can be sequenced or ii) redesign the method of genomic library construction. This study will investigate a new method of genomic library construction.

Currently, the process of making a genomic library is as follows: DNA shearing by sonication, ligation of adapters, selection of exons (if preferred), selection of size, and sequencing of DNA library (Figure 1). We propose that if DNA were specifically cut with restriction enzymes, instead of random shearing, there would be a two-fold effect: i) the library construct size would be more variable which will increase the size of overlapping sequenced data (contigs) and thus increase the number of phased variants and ii) the
number of variant errors due to low quality reads will decrease with the new ability to
filter low quality reads lacking restriction sites.

1.1.5 Selection of restriction enzyme

The Illumina HiSeq 2000 platform has an optimal fragment sequencing length of 150-
200 base pairs. The accuracy of the Illumina system to sequence reads decreases with
fragments sizes larger than 200 bases pairs. Because of this limitation, the restriction
enzymes chosen for genomic DNA digestion must be selected based on average
digestion fragment length. Furthermore, because restriction digest sites are fixed, unlike
random shotgun sequencing, multiple restriction enzyme libraries will need to be pooled
to get adequate genomic coverage. Based on the average genomic digestion fragment
length, 4 restriction enzymes, MluC1, Alu1, HypCH4V, and Fat1 were selected to
generate 4 different restriction enzyme fragmented NGS libraries (Table 1). To
increases the rate of data processing, all of the restriction enzyme NGS libraries
underwent exome capture before being sequenced on the Illumina HiSeq 2000.

1.1.6 Genomic DNA sample selection

Parents and offspring trios are commonly used in phasing experiments due to the ability
to track inherited variants in the offspring. Because we are interested in genome
phasing, we decided to use the genomic DNA from a daughter of a trio (daughter
NA19240, mother NA19238 and father NA19239), which were collected during the 1000 Genomes project. This daughter, NA19240, was chosen because the parents’ and
daughter’s variants have already been identified and are freely available from the
Complete Genomic database. Furthermore, the daughters genomic DNA is
commercially available. The daughters exome has also been captured using the same
exome capture kit and ran on the same NGS platform (Illumina HiSeq 2000) as the
libraries made with the restriction enzymes.
<table>
<thead>
<tr>
<th>Restriction enzyme</th>
<th>MluCl</th>
<th>Alul</th>
<th>HypCH4V</th>
<th>Fatl</th>
<th>Illumina HiSeq 2000 suggested fragment length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognized bases</td>
<td>AATT</td>
<td>TGCA</td>
<td>CATG</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>TTAA</td>
<td>TC</td>
<td>GTAC</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Average digested fragment length</td>
<td>144</td>
<td>241</td>
<td>216</td>
<td>225</td>
<td>150-200</td>
</tr>
</tbody>
</table>

|, cutting site of restriction enzyme
This raw sequenced data is freely available for download from the Complete Genomic database. To compare the two methods of NGS library construction accurately, we ran both the raw data from the restriction enzyme libraries and the raw data from the Complete Genomic database in parallel through a single bioinformatics pipeline.

It was hypothesized that the NGS libraries constructed with fragments generated by restriction enzyme would be more variable in size in comparison to libraries generated from sonication. This variability of fragment size was predicted to increase the contig size and thus increase the number of phased variants. It was also hypothesized that number of variant errors due to low quality reads will decrease in the restriction enzyme generated NGS libraries in comparison to libraries generated by sonication because of the unique ability to filter low quality reads that lacked restriction sites at the end of reads.

1.2 Methods

1.2.1 Genomic DNA used to generate restriction enzyme exome libraries

NA19240 genomic DNA was purchased from Coriel DNA depository.

NA19240 Illumina exonic sequencing data and variant file was downloaded from Complete Genomics database.

1.2.2 Restriction enzyme generated TruSeq DNA sample preparation

1.2.2.1 Restriction Digestion

Each restriction enzyme generated TruSeq DNA sample had the starting concentration of 1.5 µg of genomic DNA. Each restriction enzyme digestion reaction was performed with 1 unit of enzyme per 1µg of DNA. The enzymatic incubation time for each sample was
1 hour. For enzymes MluC1 (NEB), Alu1 (NEB) and HypCH4V (NEB), 5µl of 10X Buffer 4 (NEB), 1.25µl of enzyme (1000u/ml) and 43µl ddH20 was added to 1.25µl of gDNA (1µg/µl) and incubated at 37°C for 1 hour. For enzyme Fat1 (NEB), 5µl of 10X Buffer 2 (NEB), 1.25µl of Enzyme (1000u/ml) and 43µl H20 was added to 1.25µl of genomic DNA (1µg/µl) and incubated at 55°C for 1 hour. Immediately after digestion, all reactions were cleaned using the QIAquick Nucleotide Removal Kit (QIAGEN) and eluted with 50µl ddH20 with an expected 90% retention (~4.5µg).

1.2.2.2 Blunting and phosphorylating the gDNA

To repair the 3’ recessed ends made by the enzymes MluC1 and Fat1 and to phosphorylate the 5’ end, 10µl of phosphorylation reaction buffer 10X (NEB), 5µl of T4 DNA polymerase (3u/µl, NEB), 5µl of T4 polynucleotide kinase (10u/µl, NEB), 4µl of dNTPs (10mM), and 40µl of H20 was added to 50µl of fragmented DNA (0.9µg/ml) and incubated at 20°C for 30 minutes. Immediately after blunting and phosphorylating, all reactions were cleaned using the QIAquick Nucleotide Removal Kit (QIAGEN) and eluted with 32µ ddH20 with expected 80% retention (~3.6µg).

1.2.2.3 Adenylate of the 3’ end of the blunt DNA fragments

12.5µl of A-Tailing Mix (Ilumina) was added to 15µl of repaired DNA (67ng/µl) and incubated for 30 minutes at 37°C.

1.2.2.4 Ligating the Adaptors

2.5µl of DNA Ligase Mix (Ilumina) and 2.5µl of DNA Adapter (Ilumina) was added to the 27.5µl of DNA with A-Tails (above) and incubated for 10 minutes at 30°C. DNA adapter index 4 (Ilumina) was ligated to the library construct with Alu1. DNA adapter index 5 (Ilumina) was ligated to the library constructed with HypCH4V. DNA adapter index 2 (Ilumina) was ligated to the library constructed with Mlu1. DNA adapter index 7 (Ilumina)
was ligated to the library constructed with Fat1. After incubation, 5µl of Stop Ligase Mix (Illumina) was added. The DNA library was then cleaned up twice with AMPure XP Beads and resuspended in 22.5µl of Resuspension Buffer (Illumina).

1.2.2.5 Library Enrichment

5µl of Primer Cocktail (Illumina) and 25µl of PCR Master Mix (Illumina) was then added to 1µg of ligated DNA. The library was then amplified on a thermal cycler with the following conditions: 98˚C for 30 seconds, 10 cycles of (98˚C for 10 seconds, 60˚C for 30 seconds and 72˚C for 30 seconds), 72˚C for 5 minutes and hold at 4˚C. The reaction was then purified with AMPure XP Beads and re-suspended in 32.5µl of Resuspension Buffer (Illumina).

1.2.2.6 Exome Enrichment

Enrichment of the restriction enzyme library exome was performed following the standard Illumina protocol “TruSeq Enrichment Guide”. The TruSeq Exome Enrichment kit targets 62 million bases. The restriction enzyme generated TruSeq exome enriched DNA sample were sequenced at 100 bp pair-end reads in one flow cell in an Illumina sequencer HiSeq 2000. The subsequent sequences were processed by the standard Illumina sequence quality-control pipeline.

Exome sequences were mapped to the human genome reference sequence hg19 by BWA using the default parameters in paired mode. The resulting SAM files were converted to BAM files and duplicates were removed using Picard (http://picard.sourceforge.net). The mapped reads were locally realigned using GATK RealignerTargetCreator. Finally the variants were called with GATK following GATK best practice protocol (Figure 5). To call phased data GATK, ‘ReadBackedPhasing’ was used.
Figure 5. GATK best practice flow chart (http://gatkforums.broadinstitute.org/).
1.3 Results

1.3.1 Mapping sequenced reads

The restriction enzymes (MluC1, Alu1, HypCH4V and Fat1) were chosen based on the average genomic digest fragment length of the restriction enzyme (Table 1). These four restriction enzyme generated exome NGS libraries were pooled and sequenced together and reached 109X coverage and covered 87% of the 62 million TruSeq targeted sequences (Table 2). Next, we analyzed if pooling different combinations of restriction enzyme libraries would increase the percentage of the exome covered (Figure 6). By combining four different restriction enzyme libraries, we are able to nearly double the amount of exome covered in comparison to an exome covered by a single restriction enzyme library (46% to 86%). While the exome NGS library generated by sonication reached 20X coverage, it only covered 76% of the 62 million TruSeq targeted sequences.

Interestingly, ~10% of the reads from the restriction enzyme libraries were unable to be mapped to the human reference genome hg19. This is likely a result of either the library reads length being outside of Illumina sequencing technology specifications or reads being too short to align specifically to the reference genome. Reads that did not include a restriction enzyme sequence at the start of the sequenced read (0.7% of the reads) were removed from the data set.

Next, we compared the average paired end fragment length from the restriction enzyme libraries to the fragment length generation by sonication. While the average read length is comparable between the two libraries (174nt for restriction enzymes, 179nt for sonication) the restriction enzyme libraries standard deviation was much greater than the sonication library (174 ±120nt and 179 ±59nt, respectively).
Table 2. Summary of restriction enzyme exome library sequenced with Illumina HiSeq 2000 100bp pair-end reads

<table>
<thead>
<tr>
<th>Restriction enzyme library</th>
<th>MluC1</th>
<th>Alu1</th>
<th>HypCH4V</th>
<th>Fat1</th>
<th>All four enzymes</th>
<th>1000 Genome (sonication)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reads</td>
<td>35,514,484</td>
<td>54,521,560</td>
<td>44,955,256</td>
<td>49,058,158</td>
<td>184,049,458</td>
<td>32,740,442</td>
</tr>
<tr>
<td>%</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Unmapped reads</td>
<td>4,714,350</td>
<td>6,354,982</td>
<td>4,519,560</td>
<td>4,358,828</td>
<td>19,947,720</td>
<td>188,494</td>
</tr>
<tr>
<td>%</td>
<td>13</td>
<td>12</td>
<td>10</td>
<td>9</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Mapped reads</td>
<td>30,800,134</td>
<td>48,166,578</td>
<td>40,435,696</td>
<td>44,699,330</td>
<td>164,101,738</td>
<td>32,551,948</td>
</tr>
<tr>
<td>%</td>
<td>87</td>
<td>88</td>
<td>90</td>
<td>91</td>
<td>89</td>
<td>99</td>
</tr>
<tr>
<td>Paired reads mapped</td>
<td>320,751</td>
<td>4,116,781</td>
<td>2,726,051</td>
<td>1,205,967</td>
<td>8,369,550</td>
<td>-</td>
</tr>
<tr>
<td>alternatively</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>1</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>Reads without restriction</td>
<td>266,300</td>
<td>35,148</td>
<td>232,294</td>
<td>622,713</td>
<td>1,156,455</td>
<td>-</td>
</tr>
<tr>
<td>sites</td>
<td>.8</td>
<td>.08</td>
<td>.06</td>
<td>1</td>
<td>.7</td>
<td>-</td>
</tr>
<tr>
<td>High quality mapped reads</td>
<td>30,213,083</td>
<td>43,694,649</td>
<td>37,477,351</td>
<td>42,870,650</td>
<td>154,255,733</td>
<td>32,551,948</td>
</tr>
<tr>
<td>%</td>
<td>85</td>
<td>80</td>
<td>83</td>
<td>87</td>
<td>84</td>
<td>100</td>
</tr>
<tr>
<td>High quality mapped reads</td>
<td>30,213,083</td>
<td>43,694,649</td>
<td>37,477,351</td>
<td>42,870,650</td>
<td>154,255,733</td>
<td>32,551,948</td>
</tr>
<tr>
<td>%</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Reads mapped to exome</td>
<td>16,696,935</td>
<td>22,455,881</td>
<td>20,099,998</td>
<td>22,960,533</td>
<td>82,213,347</td>
<td>16,815,776</td>
</tr>
<tr>
<td>%</td>
<td>55</td>
<td>51</td>
<td>54</td>
<td>54</td>
<td>53</td>
<td>52</td>
</tr>
<tr>
<td>nt mapped to exome</td>
<td>1,381,746,304</td>
<td>1,858,148,710</td>
<td>1,665,029,685</td>
<td>1,884,831,045</td>
<td>6,789,755,744</td>
<td>1,221,211,997</td>
</tr>
<tr>
<td>X coverage to exome</td>
<td>22</td>
<td>30</td>
<td>27</td>
<td>30</td>
<td>109</td>
<td>20</td>
</tr>
<tr>
<td>% exome covered</td>
<td>45.14</td>
<td>50.71</td>
<td>45.14</td>
<td>41.98</td>
<td>86.53</td>
<td>76.05</td>
</tr>
</tbody>
</table>
Figure 6. Pooling restriction enzyme libraries increases the percent of exome nucleotides covered. After pooling the four restriction enzyme libraries, the percentage of covered exome doubled from 46% to 86%.
1.3.2 GATK called variants

After the poor quality reads (paired reads that mapped alternatively or reads without restriction enzyme ends) were removed from both the restriction enzyme libraries and sonication library data, the variant caller GATK was used to call variants from both datasets. To compare specificity and sensitivity of the variant calling between the two datasets, the amount of overlapping variants between the two datasets with the published variant list from Complete Genomics was analyzed. The Complete Genomic variant list was compiled from multiple sequencing runs of the NA19240 genome and is considered a highly accurate variant list. The restriction enzyme libraries showed higher sensitivity to call variants (172,571 variants) in comparison to the sonication library (79,330 variants). However, this high sensitivity to call variants in restriction enzyme libraries was accompanied with a low specificity. 49% of the variants from the restriction enzyme libraries overlapped with the Complete Genomic variant list compared to 97% of the sonication library variants overlapped (Table 3). This high error rate might be due to the inability of the restriction enzyme libraries to mark PCR amplified duplicate reads. This is because PCR amplified reads are indistinguishable from multiple reads cut at the same restriction enzyme site (Figure 7).

To increase the specificity of variant calling with the restriction enzyme libraries, only variants that were covered by two different restriction enzyme library reads were called. With this new constraint, the variants called from the sonication library had greater sensitivity and specificity than the restriction enzyme library. 79,330 variants called compared to 47,194 variants called, respectively.
Table 3. Comparison of GATK called variants from restriction enzyme reads and sonication reads

<table>
<thead>
<tr>
<th></th>
<th>Variant Called (%)</th>
<th>Overlapping with Complete Genomics variants (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variants called from both restriction enzyme reads and sonication reads</td>
<td>53,508 (31)</td>
<td>52,999 (99)</td>
</tr>
<tr>
<td>Variants called from restriction enzyme reads and not called from sonication reads</td>
<td>119,063 (69)</td>
<td>30,837 (26)</td>
</tr>
<tr>
<td>Total number of variants from restriction enzyme reads</td>
<td>172,571 (100)</td>
<td>83,886 (49)</td>
</tr>
<tr>
<td>Variants called from both sonication reads and restriction enzyme reads</td>
<td>53,508 (67)</td>
<td>52,999 (99)</td>
</tr>
<tr>
<td>Variants called from sonication reads and not called from restriction enzyme reads</td>
<td>25,822 (33)</td>
<td>24,885 (94)</td>
</tr>
<tr>
<td>Total number of variants from sonication reads</td>
<td>79,330 (100)</td>
<td>77,884 (98)</td>
</tr>
</tbody>
</table>
Figure 7. Mapped reads generated by sonication and restriction enzymes.

A) Fragmenting the genomic DNA by sonication produces randomized fragments.

B) Restriction enzyme generated libraries produces non-randomized fragments that are indistinguishable from PCR amplified reads.
Furthermore, 98% of the sonication library variants overlapped with the Complete Genomic variant list compared to 97% overlap with the restriction enzyme libraries (Table 4).

1.3.3 Phasing variants

To compare the ability of sonication and restriction enzyme libraries to phase variants, GATK ‘ReadBackedPhasing’ was used to call phased variants. To obtain an accurate comparison, only variants that were called by both sonication and restriction enzyme libraries (53,508 variants) were phased. The number of phased contigs were similar between the two techniques (11,936 contigs in restriction enzyme library and 11,494 contigs in sonication library) (Table 5). While the occurrence of contigs with two phased variants are greater in the sonication library, the restriction enzyme libraries constantly had a higher occurrence of contigs with >2 variants phased together than the sonication library.

1.4 Discussion

In this project, I proposed a new method of genomic library construction. It was hypothesized that if random shearing of the genomic DNA was changed to precise cutting of DNA with restriction enzymes, the variant error rate would decrease and the amount of phased variants would increase because the library fragment size would be more variable.

The results from this study showed both the benefits and problems with construction of the genomic NGS library using restriction enzymes. The first problem is that there are 10-fold more unmapped reads than a library constructed by sonication. These unmapped reads are comprised of reads that are too small to be aligned to the reference human genome.
Table 4. Comparison of GATK called variants from ≥2 restriction enzyme library reads and sonication reads

<table>
<thead>
<tr>
<th>Variants called</th>
<th>Variant Called (%)</th>
<th>Overlapping with Complete Genomics variants (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both ≥2 restriction enzyme library reads and sonication reads</td>
<td>30,644 (65)</td>
<td>30,464 (99)</td>
</tr>
<tr>
<td>Only ≥2 restriction enzyme library reads and not called from sonication reads</td>
<td>16,550 (35)</td>
<td>15,462 (93)</td>
</tr>
<tr>
<td>Total number of variants from ≥2 restriction enzyme library reads</td>
<td>47,194 (100)</td>
<td>45,926 (97)</td>
</tr>
</tbody>
</table>

| Both sonication and ≥2 restriction enzyme library reads | 30,644 (39) | 30,464 (99) |
| Only sonication reads and not called from ≥2 restriction enzyme library reads | 48,593 (61) | 47,432 (98) |
| Total number of variants from sonication reads | 79,237 (100) | 77,884 (98) |
Table 5. Phased variants per contig from restriction enzyme and sonication libraries

<table>
<thead>
<tr>
<th>Number of phased variants per contig</th>
<th>Number of contigs</th>
<th>Restriction enzyme library</th>
<th>Sonication library</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5,538</td>
<td>6,503</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2,815</td>
<td>2,593</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1,433</td>
<td>1,114</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>788</td>
<td>554</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>476</td>
<td>272</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>292</td>
<td>161</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>181</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>109</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>97</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>55</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>41</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>26</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>18</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>12</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>12</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>7</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>11</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>12</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total contigs</td>
<td>11,936</td>
<td>11,494</td>
<td></td>
</tr>
</tbody>
</table>
Applying strict read size selection to the methodology of restriction enzyme library construction may help correct this result.

While low quality reads lacking in restriction sequence ends could be filtered out with this new method, the inability to distinguish between PCR duplicate reads and non-PRC duplicate reads with the restriction enzyme libraries lead to a greater variant calling error rate in comparison to the NGS library generated by the sonication of genomic DNA. Because of this limitation, it was decided that only variants that were found in overlapping restriction enzyme libraries should be called. While this filter decreased the variant calling error rate of the restriction enzyme dataset, libraries generated by sonication could still call variants with more specificity and sensitivity in comparison to the restriction enzyme libraries.

Finally, we compared the ability of the two differently generated NGS libraries to call phased variants. While the total number of contigs with phased variant were similar between the restriction enzyme library and sonication library, the number of contigs with >2 variants phased was consistently more in the restriction enzyme library. This observation is most likely due to the increased standard deviation read length in the restriction enzyme library, 174±120 nt) for restriction enzyme and 179±59 nt) for sonication.
CHAPTER 2. Identification and characterization of genetic modifiers of breast cancer risk in \textit{BRCA1} mutation carriers
2.1 Introduction

Breast cancer cases can be separated into two broad categories, sporadic cases and cases with germline mutations. Sporadic breast cancer make up 90% of all breast cancer cases\textsuperscript{11}. The remaining 10% have some type of germline mutations that predisposes the carrier for the development of breast cancer\textsuperscript{14-17}. There are many known germline genes that increase the risk of developing breast cancer\textsuperscript{8,10,18-21}. One of the most common breast cancer germline mutations, which affects about 10% of all germline mutation carriers, are mutations in the gene BRCA1\textsuperscript{10} (Figure 3).

2.1.2 BRCA1

\textit{BRCA1} (breast cancer 1, early onset) is a well-known tumor suppressor implicated in a diverse array of biological processes, including DNA break repair, cell-cycle checkpoint activation, transcription regulation and DNA replication\textsuperscript{22-31}. \textit{BRCA1} is located on chromosome 17 and is comprised of 24 exons spanning 81,188 bp. \textit{BRCA1} mutation carriers from familial breast cancer families have a 44-75% risk of developing breast cancer by age 70\textsuperscript{15}. This wide range of risk is due to different risk modifying factors. These factors include specificity of the \textit{BRCA1} germline mutation, environmental modifiers, and genetic modifiers\textsuperscript{16,17,21,32-38}.

2.1.3 Genetic modifiers of breast cancer risk

Genetic modifiers of cancer risk has been hypothesized for over 20 years. In the mid-1990s, cancer risk modifying alleles were first described in mice cancer models\textsuperscript{39-41}. While there have been many large genome wide association studies (GWAS) that associated germline variants with a change of risk of breast cancer in \textit{BRCA1} germline mutation carriers, there have been no follow up functional studies providing evidence that these variants are actually the risk modifiers\textsuperscript{42-45} (Table 6). There are a number of
reasons why these variants have not been analyzed further. The main reason is the inability to distinguish whether the identified variants are merely linked to the risk modification or if the variant itself is the cause of change of breast cancer risk. Many of the \textit{BRCA1} risk modifier GWAS studies used a 610k array SNP chip to genotype the \textit{BRCA1} mutations carriers, of which 95\% of the variants were from intronic or intergenic regions of the genome (Figure 8). While intronic and intergenic regions are no longer considered “junk DNA”, the functional consequence of variants in these regions is difficult to predict.

\textbf{2.1.4 Intervention options of \textit{BRCA1} mutation carriers}

Currently, there are limited risk assessments and availability in options for the intervention of breast cancer in high-risk women, leading some carriers to undergo an invasive mastectomy. Identifying germline modifiers can both provide better personalized cancer risk assessments to \textit{BRCA1} mutation carriers and can potentially lead to new therapies that may lower the risk for these carriers.

\textbf{2.1.5 Benefits of NGS analysis}

To identify variants that have a functional possibility of modifying the risk of developing breast cancer in \textit{BRCA1} carriers, we used NGS to sequence the exome of 54 \textit{BRCA1} mutation carriers. Using NGS will increase sensitivity for variant detection in comparison to an array SNP chip. This increase in sensitivity will allow the usage of biological filters, which will increase the likelihood of finding functional risk modifying variants. The first biological filter used in this study will be variant function prediction software. All variants that are predicted to be non-damaging to the function of the gene will be filtered out as their influence on the affected gene are less predictable.
<table>
<thead>
<tr>
<th>SNPs</th>
<th>GENE</th>
<th>Molecular consequence</th>
<th>Affected</th>
<th>Unaffected</th>
<th>HR</th>
<th>MAF</th>
<th>P</th>
<th>Ref</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4808611</td>
<td>NR2F6</td>
<td>intronic</td>
<td>4144</td>
<td>3362</td>
<td>1.27</td>
<td>0.102</td>
<td>1.6e-10</td>
<td>42</td>
<td>-</td>
</tr>
<tr>
<td>rs8100241</td>
<td>ANKLE1</td>
<td>exonic (SIFT Damaging)</td>
<td>4151</td>
<td>3368</td>
<td>0.89</td>
<td>0.46</td>
<td>1.6e-10</td>
<td>42</td>
<td>-</td>
</tr>
<tr>
<td>rs2363956</td>
<td>ANKLE1</td>
<td>exonic (SIFT Damaging)</td>
<td>4150</td>
<td>3367</td>
<td>0.84</td>
<td>0.46</td>
<td>2.4e-10</td>
<td>42</td>
<td>-</td>
</tr>
<tr>
<td>rs67397200</td>
<td>ANKLE1,</td>
<td>intergenic</td>
<td>5301</td>
<td>3699</td>
<td>1.16</td>
<td>0.17</td>
<td>4.5e-10</td>
<td>43</td>
<td>-</td>
</tr>
<tr>
<td>rs8170</td>
<td>BABAM1</td>
<td>synonymous</td>
<td>5890</td>
<td>4380</td>
<td>1.23</td>
<td>0.11</td>
<td>2.1e-9</td>
<td>43</td>
<td>-</td>
</tr>
<tr>
<td>rs3745185</td>
<td>BABAM1</td>
<td>intronic</td>
<td>4152</td>
<td>3369</td>
<td>0.86</td>
<td>0.31</td>
<td>7.1e-8</td>
<td>42</td>
<td>-</td>
</tr>
<tr>
<td>rs6138178</td>
<td>SNRPB</td>
<td>intronic</td>
<td>3451</td>
<td>0.78</td>
<td>0.428</td>
<td>0.0015</td>
<td></td>
<td>45</td>
<td>-</td>
</tr>
<tr>
<td>rs6602595</td>
<td>CAMK1D</td>
<td>intronic</td>
<td>3451</td>
<td>1.25</td>
<td>0.481</td>
<td>0.0019</td>
<td></td>
<td>45</td>
<td>-</td>
</tr>
<tr>
<td>rs1048635</td>
<td>TRIM45</td>
<td>UTR3</td>
<td>3451</td>
<td>1.5</td>
<td>0.06</td>
<td>0.0029</td>
<td></td>
<td>45</td>
<td>-</td>
</tr>
<tr>
<td>rs13387042</td>
<td>TNP1,</td>
<td>intergenic</td>
<td>4763</td>
<td>1.18</td>
<td>0.147</td>
<td>0.005</td>
<td></td>
<td>44</td>
<td>-</td>
</tr>
<tr>
<td>rs6590224</td>
<td>KIRREL3</td>
<td>intronic</td>
<td>3451</td>
<td>1.26</td>
<td>0.247</td>
<td>0.0056</td>
<td></td>
<td>45</td>
<td>-</td>
</tr>
</tbody>
</table>

HR, Hazard Ratio  
MAF, Minor allele frequency  
Ref, References
Figure 8. Distribution of variants on the 610k array. A large percentage of SNPs analyzed on the 610k array are in intronic and intragenic regions. Molecular consequence of SNPs in these regions of the genome are difficult to decipher.

* nonsynonymous, stopgain, stoploss and splicing
** synonymous, upstream, downstream, non-coding and UTRs
To further increase the likelihood of selecting a functionally relevant variant, a PubMed literature search will be used to select only variants that occur in genes that have predicted roles in cancer. It is important to note that genes with unknown functions and variants that have no effect on the translated amino acid may, in fact, affect BRCA1 mutation carrier breast cancer risk. However, because this study will analyze the functional consequence of these variants using in vitro assays, strict filters will be used to increase the likelihood of accomplishing this goal.

2.2 Materials and Methods

2.2.1 Statistical association study

2.2.1.1 Samples used in the case control study

The samples used in this study are from Creighton University Breast Cancer Family Registry. Participants provided written informed consent for cancer genetic study. Institutional Review Board of Creighton University and Institutional Review Board University of Nebraska Medical Center approved the study (CU #00-12265, UNMC #718-11-EP).

The cases included 27 BRCA1 mutated carriers that were diagnosed with breast cancer. All breast cancer affected cases were matched with a control breast cancer unaffected individual that shared the same BRCA1 mutation from the same family. If multiple breast cancer affected carriers were available, the individual diagnosed at the earliest age was selected. If multiple unaffected carriers were available, the oldest individual was selected. In total there were 27 cases and 27 controls from 27 families.

2.2.1.2 Samples used in the cohort study

The BRCA1 mutated carrier sample size was increased from 54 BRCA1 mutation carriers in the case control study to 161 BRCA1 mutation carriers in the cohort study.
These 107 new BRCA1 mutation cases were only selected based on the BRCA1 mutation status.

2.2.1.3 Statistical association of genetic modifiers to breast cancer risk

Odds ratio (OR) and their 95% confidence intervals (CI) were estimated using the unconditional maximum likelihood with the R package Epitools software. The variants that passed the biological filters and odds ratio statistics were then re-analyzed in a cohort study using the Cox regression model. The Hazard Ratio (HR), their 95% confidence interval (CI) and Wald’s $P$-value were calculated using the R package Survival software.

2.2.1.4 Illumina Exome libraries preparation

DNA from blood cells was extracted from the selected BRCA1 mutation carriers by Dr. Lynch’s laboratory and was shipped to Dr. San Ming Wang’s lab for analysis. The exome libraries were prepared and captured by using the SureSelectXT2 Target Enrichment System (Agilent Technologies, which has a total amount of targeted sequence of 75 million bases. The libraries were sequenced in an Illumina sequencing platform HiSeq 2500 with 150bp paired-end reads with the goal of 50X coverage, based on nucleotides mapped to the exome.

2.2.1.5 Mapping, variant calling and damaging prediction

After samples were sequenced, the last 25% of each read was trimmed with fastqx-trimmer before being mapped to the human reference genome hg19 by the program BWA. Once mapped, the reads were locally realigned and the duplicates were marked with Picard. The variants were then called with the program GATK_HaplotypeCaller following GATK best practice protocol (Figure 5). All of the variants were analyzed with the program AnnoVar. The computer software PolyPhen2
and Sift was used to determine if the change in amino acid would likely be damaging to the function of the protein.\textsuperscript{52,53}

**2.2.1.6 PCR products Sanger sequenced**

The 5’ and 3’ oligonucleotide primers were manually designed for each DNA sequenced and were ordered from IDT. To genotype the BRCA1 mutation carriers, the following primers were used.

- **SYNE1\_rs2295190**
  
  F 5’- TTGCTTATGACCCGATCCTC-3’
  
  R 5’- GAAGGTGCAGAGGCAAAGAAG-3’

- **ANKLE1\_rs8100241**
  
  F 5’- GAGACGCTGGACTCCATAGC-3’
  
  R 5’- CAGCTCCAGAGACCTCAAACC-3’

- **ANLN\_rs3735400**
  
  F 5’- GCTGAAAGAGAATGGGTTT-3’
  
  R 5’- GCAGATGCTCAGACTCAACTGG-3’

- **SIPA1L2\_rs1547742**
  
  F 5’- AAGAACAGCCAGCCACCTTA-3’
  
  R 5’- CCTCAGTGGTTCTCACCATT-3’

- **MTHFSD\_rs3751800**
  
  F 5’- CAGGGCTCTCTCAGGTCAC-3’
  
  R 5’- GCCTGGAGATGATGGAGAAA-3’
PCR were carried out in a PTC-200 thermal cycler (MJ Research). The following components are added to the final PCR master mix: 15ng of DNA template, 0.1mM dNTP mixture, 1X GoTaq buffer (Promega), 1µM primers (5’ and 3’), and 1u/µl GoTaq (Promega). First, the template DNA was denatured at 95˚C for 7 minutes. Next, 38 cycles of 95˚C for 30sec, 57˚C for 30sec, 72˚C 45sec (depending on expected fragment size, 1kb=45sec). The last PCR cycle was a final extension of 72˚C for 7 minutes. Before the sequences are Sanger sequenced, the PCR products are cleaned with SAP (1u/µg of DNA), 1X SAP buffer, and Exo1 (1u/µg of DNA). This reaction was incubated at 37˚C for 60 minutes and denatured at 95˚C for 20 minutes before the primer is premixed with the sample and sent for Sanger sequencing.

2.2.2 Functional Study

2.2.2.1 Bacterial strains

NEB 10-beta electrocompetent E. coli genotype: Δ(ara-leu)7697 araD139 fhuA ΔlacX74 galK16 galE15 e14- Φ80dlacZΔM15 recA1 relA1 endA1 nupG rpsl (StrR)rph spoT1 Δ(mrr-hsdRMS-mcrBC)

XL10-Gold ultracompetent E. coli genotype: TetR Δ(mcrA)183 Δ(mcrCB-hadSMR-mrr) 173 endA1 supE44 thi-1 recA1 gyrA96 relA1 lac Htee [F’ proAB lacIqZΔM15 Tn10 (TetR) Amy CamR].

2.2.2.2 Bacterial Cell Growth Conditions

LB media was made with 10g NaCl, 10g tryptone, 5g yeast extract (per liter of media) and pH to 7.0 with 5N NaOH. If solid media was required, 20g of agar per liter was added before sterilization by autoclaving. Once the sterilized LB media cooled, depending on desired antibiotic selection, kanamycin (50µg/ml), ampicillin (100µg/ml), or carbenicillin (100µg/ml) was added.
*E. coli* cells inoculated in liquid LB media was grown in a C24 incubator shaker (New Brunswick Scientific) at 37˚C, 200 rpm overnight. *E. coli* cells inoculated on solid LB media was grown at 37˚C overnight.

### 2.2.2.3 Bacterial transformation

#### 2.2.2.3.1 Electrorcompetent transformation

1 µl of DNA was added to *E. coli* beta-10 cells in a chilled 1mm electroporation cuvette (Bioexpress) and mixed by flicking gently. Cells were shocked with the Bio-Rad micropulser at 2.0 kv, 200Ω, and 25µF, 975µl pre-warmed (37˚C) SOC outgrowth medium was transferred to the cuvette and then transferred to a culture tube. The cells were then shaken vigorously (250 rpm) at 37˚C for 1h, then spread onto four pre-warmed LB selection plates and incubated overnight at 37˚C.

#### 2.2.2.3.2 Chemical competent transformation

45µl of XL10-Gold cells was added to a pre-chilled 14-ml polypropylene tube, 2µl of β-mercaptoethanol was added. Cells were incubated on ice and swirled gently every 2 minutes for 10 minutes, then 2µl of DNA was added to the ultracompotent cells. After the cells were incubated on ice for 30 minutes, the cells were heat-pulsed in a 42˚C water bath for 30s. The heat-pulsed cells were then incubated on ice for two minutes before 0.5ml of preheated (42˚C) SOC media was added. These cells were incubated at 37˚C for one hour at 250 rpm before the cells are spread onto four pre-warmed LB selection plates and incubated overnight at 37˚C. Plasmid DNA was purified with the QIAprep Spin Miniprep Kit (QIAGEN).
2.2.2.4 Mammalian Cell line Growth Conditions

HCC1937 cell line was grown at 37˚C, 5% CO₂ atmosphere in Roswell Park Memorial Institute medium (RPMI 1640, ATCC) supplemented with 10% FBS (Gibco), and 1X streptomycin/penicillin (Gibco).

HEK 293T cell line was grown at 37˚C, 5% CO₂ atmosphere in Dulbecco’s Modified Eagle Medium (DMEM, ATCC) supplemented with 10% FBS, and 1X streptomycin/penicillin.

HeLa cell line was grown at 37˚C, 5% CO₂ atmosphere in Dulbecco’s Modified Eagle Medium (DMEM, HyClone) supplemented with 10% FBS, and 1X streptomycin/penicillin.

MCF7 cell line was grown at 37˚C, 5% CO₂ atmosphere in Eagle’s Minimum Essential Medium (EMEM, ATCC) supplemented with 10µg/ml insulin from bovine pancreas, 10% FBS, and 1X streptomycin/penicillin.

2.2.2.5 shRNA stable cell line production

E. coli LX10 gold carrying packaging plasmid psPAX2, envelope plasmid pMD2.G, and scramble shRNA (on the vector backbone pLKO.1) were purchased from Addgene. ANLN shRNA (TRCN0000117257) bacterial glycerol stock were purchased from Sigma-Aldrich.

2.2.2.6 Lentiviral particles production

To start the lentiviral particle packaging, 7×10⁵ HEK-293T cells were plated on a 6 cm tissue culture plate at 37˚C, 5% CO₂ in DMEM +10% FBS without antibiotics overnight. After 12-15 hours, a plasmid cocktail of 1µg pLKO.1, 750ng psPAX2, 250 ng pMD2.G was added to 20µl of OPTI-MEM (Invitrogen). Next 10µl of Lipofectamine 2000 (Invitrogen) was added to 90 µl of OPTI-MEM (Gibco) and incubated at room temperature of 5 minutes. The lipofectamine 2000 master mix was then added to the
plasmid cocktail and incubated for 20-30 minutes at room temperature. This mixture was then added to the cells and incubated at 37°C, 5% CO₂ overnight. In the morning, fresh media with antibiotics was added and the cells were incubated at 37°C, 5% CO₂ for 24 hours. After 24 hours, the media was harvested and 5ml of fresh media containing antibiotics was added to the cells and incubated at 37°C, 5% CO₂ for 24 hours. Again, after 24 hours, the cells were harvested. The harvested media was then filtered through a 0.45μm filter and stored at -20°C. SYNE1 shRNA infection particles (TRCN0000147281) were purchased from Sigma-Aldrich.

2.2.2.7 Infecting Target Cells

Target cells were plated at 50% confluence at 37°C, 5% CO₂ overnight. The next morning, fresh culture media containing 8µg/ml polybrene (Santa Cruz) was added. Next, 0.5ml of lentiviral particles was added to the target cells and incubated at 37°C, 5% CO₂ for 24 hours. The media was then changed after 24 hours. Two days after infection, stable cell lines were selected using 1µg/ml of puromycin (Sigma-Aldrich).

2.2.2.8 DNA Transfection Procedure

Target cells were plated in a 6-well plate in growth medium without antibiotics at 37°C, 5% CO₂ for 24 hours. Once the cells were 90% confluent, 4.0µg of DNA was diluted in 250µl of Opti-MEM and 10µl of Lipofectamine 2000 diluted in 250µl of Opti-MEM. After 5 minutes of incubation at room temperature, the diluted DNA was combined with the diluted Lipofectamine 2000 and incubated for 20 minutes at room temperature. The complexes were added to the cells and medium and mixed by gently rocking the plate back and forth. Cells were grown for 48 hours before imaging or selective medium was added.
2.2.2.9 Growth curve

Stable shRNA cells were seeded at $3.0 \times 10^4$ in 6-well plates with complete EMEM media without antibiotics. The seeded cells were allowed to grow for 24 hours before the first time point was collected. The following time points were collected every 24 hours for five more days. Cell viability was determined with 0.04% trypan blue (Gibco). Cells were counted manually with a hemocytometer (Bright-Line).

2.2.2.10 Microscopy

Two days after transient transfection, cells were transferred and grown on a glass coverslip for 24 hours. Cells were then fixed with 4% paraformaldehyde in PBS. DAPI Fluoromount-G was used to mount and stain the nucleus. These cells were then imaged on a Nikon Eclipse TI-E microscope with a Nikon Digital sight DS-QiMc camera. Images were overlaid in NIS Elements (Nikon) and fluorescence intensity were measured with the software ImageJ.

2.2.2.11 Western Blot

To extract whole cell lysate from living cells, a lysing buffer (5% glycerol, 1.5% SDS) was added directly to the cells. Proteins were then denatured at 95˚C for 10 minutes and protein concentrations were measured using a BCA protein assay Kit (Pierce) on an iMark microplate absorbance reader (Bio Rad).

Proteins were separated by SDS-PAGE and transferred to nitrocellulose membrane (Immobilon-P) for subsequent blotting with antibodies. Both SYNE1 and ANLN proteins were separated by a 9% SDS-PAGE gel. To blot for ANLN, a monoclonal anti-ANLN AMAB90660 antibody produced in mouse was used (Sigma-Aldrich) while SYNE1 was probed using a monoclonal anti-SYNE1 SAB1404967 antibody produced in mouse
(Sigma-Aldrich). To blot Hsc70, a monoclonal anti-Hsc70 ab19136 antibody produced in mouse was used (abcam). The secondary antibody used for all western blot assays was anti-mouse IgG, HRP-linked antibody 7076s (Cell Signaling Technology). All blots were developed on Blue Lite Autorad Film (GeneMate).

2.2.2.12 ANLN construct design and verification

Overexpressing eGFP-anillin vector was a gift from Dr. Andrew Wilde’s lab. The variant was induced with overlapping oligonucleotides with the QuickChange 2 XL Site-Directed Mutagenesis Kit (Agilent Technologies).

Oligonucleotides used to induce the variant in the ANLN_rs3735400 construct:

F 5’- gagaatcttgtacaaacctatGgccatcaaaaaacgctg-3’

R 5’-cagcgttttttgatggcCatggttttgtacaagatttctc-3’.

To verify the induced variant, the vector was Sanger sequenced with primer:

5’-CTGCACCTGAGGAGACACAG-3’.

To generate the eGFP vector without anillin, the following primers were used to amplify only the vector backbone.

F 5’-aatggatccgtttacggagaaactgc -3’

R 5’-ctacaaacctattgaaaagccttaa -3’

After the vector backbone was made, DpnI endonuclease (10 U/µl) was incubated with the PCR product at 37°C for 1 hour to digest the plasmid construct. To add 5’ phosphate to the DNA sequence, T4 polynucleotide kinase (NEB) (10U/µl) with 1x T4 ligase buffer was incubated at 37°C for 1 hour. To ligase the DNA sequence, T4 DNA ligase (NEB) (1U/µl) was incubated with the DNA sequence at room temperature for 2 hours.
2.2.2.13 SYNE1 construct design and verification

EGFP-C2_DN-Nesprin-1 vector was a gift from Dr. Angelika Noegel's lab. The variant was induced with overlapping oligonucleotides with the QuickChange 2 XL Site-Directed Mutagenesis Kit (Agilent Technologies).

Oligonucleotides used to induce variant in the SYNE1_rs2295190 construct:

F 5' - gttggtcagcttcAtgttccgatcctc-3'

R 5' - gaggatccggaacaTgaagctcgacccac-3'

To verify the induced variant, the vector was Sanger sequenced with primer:

5' - TGTGAGTCCCACATCCGAA-3'

2.3 Results

2.3.1 Statistical association study

2.3.1.1 BRCA1 mutated carriers sample selection

To increase the chance of identifying variants that modify the risk of developing breast cancer in BRCA1 mutation carriers, we set strict criteria for sample selection. All of the BRCA1 mutation carriers’ mutations must be listed in breast cancer information core (BIC) database as clinically important and/or large deletions of BRCA1 must have been identified. Because the specificity of the BRCA1 mutation has been shown to modify the risk of breast cancer, the breast cancer affected and unaffected mutation carriers were matched by the specificity of their mutation. Furthermore, because age is a strong risk factor for breast cancer development, the affected and unaffected individuals were also matched by at least 5 years. To increase the likelihood of identifying variants that increase the risk of developing breast cancer, the youngest breast cancer affected
case was selected if multiple cases were available. Likewise, to increase the chance of identifying variants that decreased the risk of developing breast cancer, the oldest breast and ovarian cancer unaffected BRCA1 mutation carrier was selected if multiple cases were available. There were 27 BRCA1 mutated carriers that were diagnosed with breast cancer and 27 BRCA1 mutated carriers unaffected by breast cancer from 27 families that matched all selection criteria (Table 7).

### 2.3.1.2 NGS mapping, variant calling and variant annotation

All 54 genomes were then processed into 54 NGS exome libraries using the SureselectXT2 exome+UTR capture kit. This capture kit was designed to capture 75 million nucleotides per genome. The libraries were sequenced on the Illumina HiSeq 2500 at 150bp pair-end reads with 2.67 libraries seeded per lane. After the 54 samples were sequenced, the average depth of sequenced reads in the exome was 66.5 (66.5X coverage) (Table 8). The 54 samples also had, on average, 97.5 percent of the 75 million nucleotides captured. While GATK called, on average, over 1 million variants per sample, the 54 samples had on average 2,105 variants predicted damaging by PolyPhen2 or SIFT per genome (Table 9).

### 2.3.1.3 Case control study to identify candidate genetic modifiers

#### 2.3.1.3.1 Previously identified breast and/or risk modifying variants

CIMBA and other laboratories have previously published genetic variants that statistically modify the risk of breast and ovarian cancer $^{17,21,37,56,57}$. Most of these studies extracted genomic data through the use of SNP arrays. These arrays are significantly less expensive than NGS. This method of genotyping allowed these laboratories to genotype many thousands of samples per analysis.
Table 7. Breast cancer cases and controls selected for the case control study

<table>
<thead>
<tr>
<th>Samples in families</th>
<th>Ethnicity</th>
<th>BRCA1 Mutation</th>
<th>BIC clinically importance</th>
<th>Sex</th>
<th>BC diag age</th>
<th>Current age of affected</th>
<th>Current age of unaffected</th>
<th>Coefficient of relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Polish</td>
<td>1008 bp del-exon 17</td>
<td>-</td>
<td>F</td>
<td>31</td>
<td>56</td>
<td>57</td>
<td>1.56</td>
</tr>
<tr>
<td>2</td>
<td>Irish</td>
<td>1623del5-ter503</td>
<td>yes</td>
<td>F</td>
<td>35</td>
<td>57</td>
<td>62</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>Ashkenazi Jewish</td>
<td>185delAG-ter39</td>
<td>yes</td>
<td>F</td>
<td>37</td>
<td>64</td>
<td>56</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>Asian</td>
<td>185delAG-ter39</td>
<td>yes</td>
<td>F</td>
<td>38</td>
<td>71</td>
<td>61</td>
<td>0.195</td>
</tr>
<tr>
<td>5</td>
<td>German/Irish/Scottish</td>
<td>185delAG-ter39</td>
<td>yes</td>
<td>F</td>
<td>36</td>
<td>56</td>
<td>57</td>
<td>12.5</td>
</tr>
<tr>
<td>6</td>
<td>English/Irish/Scottish</td>
<td>188del11-ter36</td>
<td>yes</td>
<td>F</td>
<td>31</td>
<td>58</td>
<td>71</td>
<td>6.25</td>
</tr>
<tr>
<td>7</td>
<td>Irish</td>
<td>254-1G-T-exon5</td>
<td>yes</td>
<td>F</td>
<td>22</td>
<td>68</td>
<td>58</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>Dutch/Indian</td>
<td>2841Glu-ter(G-T)</td>
<td>yes</td>
<td>F</td>
<td>27</td>
<td>50</td>
<td>56</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>Indian/Irish/French</td>
<td>300Cys-Gly(T-G)</td>
<td>yes</td>
<td>F</td>
<td>28</td>
<td>70</td>
<td>65</td>
<td>12.5</td>
</tr>
<tr>
<td>10</td>
<td>Scandinavian</td>
<td>332-11T-Gins59-ter75</td>
<td>yes</td>
<td>F</td>
<td>35</td>
<td>75</td>
<td>78</td>
<td>6.25</td>
</tr>
<tr>
<td>11</td>
<td>Scandinavian</td>
<td>5272-2delA-exon 19</td>
<td>yes</td>
<td>F</td>
<td>38</td>
<td>69</td>
<td>64</td>
<td>50</td>
</tr>
<tr>
<td>12</td>
<td>Caucasian</td>
<td>5382insC-ter1829</td>
<td>yes</td>
<td>F</td>
<td>33</td>
<td>72</td>
<td>63</td>
<td>3.125</td>
</tr>
<tr>
<td>13</td>
<td>Russian/Jewish</td>
<td>5382insC-ter1829</td>
<td>yes</td>
<td>F</td>
<td>32</td>
<td>61</td>
<td>62</td>
<td>3.125</td>
</tr>
<tr>
<td>14</td>
<td>Ashkenazi Jewish</td>
<td>5382insC-ter1829</td>
<td>yes</td>
<td>F</td>
<td>44</td>
<td>82</td>
<td>49</td>
<td>50</td>
</tr>
<tr>
<td>15</td>
<td>Unknown</td>
<td>5622Arg-terC-T</td>
<td>yes</td>
<td>F</td>
<td>32</td>
<td>60</td>
<td>61</td>
<td>50</td>
</tr>
<tr>
<td>16</td>
<td>German/English</td>
<td>882Glu-ter(G-T)</td>
<td>yes</td>
<td>F</td>
<td>35</td>
<td>56</td>
<td>54</td>
<td>50</td>
</tr>
<tr>
<td>17</td>
<td>Irish/French</td>
<td>916delITT-ter285</td>
<td>yes</td>
<td>F</td>
<td>24</td>
<td>56</td>
<td>58</td>
<td>3.125</td>
</tr>
<tr>
<td>18</td>
<td>Caucasian</td>
<td>del37kb (exons 1-2)</td>
<td>-</td>
<td>F</td>
<td>36</td>
<td>72</td>
<td>80</td>
<td>12.5</td>
</tr>
<tr>
<td>19</td>
<td>German</td>
<td>Exon 11</td>
<td>yes</td>
<td>F</td>
<td>38</td>
<td>56</td>
<td>69</td>
<td>25</td>
</tr>
<tr>
<td>20</td>
<td>Caucasian</td>
<td>Exon 11</td>
<td>yes</td>
<td>F</td>
<td>59</td>
<td>64</td>
<td>59</td>
<td>12.5</td>
</tr>
<tr>
<td>21</td>
<td>Scandinavian</td>
<td>4184del4(TCAA)</td>
<td>yes</td>
<td>F</td>
<td>29</td>
<td>60</td>
<td>56</td>
<td>12.5</td>
</tr>
<tr>
<td>22</td>
<td>Czech/German/English</td>
<td>Exon 16 C4808G</td>
<td>yes</td>
<td>F</td>
<td>27</td>
<td>64</td>
<td>73</td>
<td>25</td>
</tr>
<tr>
<td>23</td>
<td>German</td>
<td>Glu143ter</td>
<td>yes</td>
<td>F</td>
<td>28</td>
<td>56</td>
<td>53</td>
<td>50</td>
</tr>
<tr>
<td>24</td>
<td>Irish/Dutch</td>
<td>ins6kbEx13-ter1460</td>
<td>yes</td>
<td>F</td>
<td>44</td>
<td>72</td>
<td>69</td>
<td>6.25</td>
</tr>
<tr>
<td>25</td>
<td>Irish/Scottish/English/German</td>
<td>ins6kbEx13-ter1460</td>
<td>yes</td>
<td>F</td>
<td>39</td>
<td>61</td>
<td>58</td>
<td>50</td>
</tr>
<tr>
<td>26</td>
<td>Dutch/German</td>
<td>IVS12-1632del3835</td>
<td>yes</td>
<td>F</td>
<td>29</td>
<td>51</td>
<td>74</td>
<td>1.56</td>
</tr>
<tr>
<td>27</td>
<td>German/Polish</td>
<td>IVS6-2delA</td>
<td>yes</td>
<td>F</td>
<td>34</td>
<td>56</td>
<td>58</td>
<td>50</td>
</tr>
</tbody>
</table>

BIC, breast cancer information core
BC diag, breast cancer diagnosed
Table 8. Breast cancer cases and controls sequenced with Illumina HiSeq 2500 150bp pair-end reads

<table>
<thead>
<tr>
<th>Samples in families</th>
<th>Affected Reads aligned to Exome+UTR</th>
<th>nt mapped to Exome+UTR</th>
<th>% of covered nt in Exome+UTR</th>
<th>Unaffected X coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>86,482,134 56,073,196</td>
<td>7,650,127,187 5,081,972,608</td>
<td>103 68</td>
<td>99.8 99.7</td>
</tr>
<tr>
<td>2</td>
<td>69,068,541 48,956,004</td>
<td>6,394,426,226 4,522,927,280</td>
<td>86 61</td>
<td>99.1 99</td>
</tr>
<tr>
<td>3</td>
<td>47,886,957 77,915,392</td>
<td>4,441,973,829 7,135,423,515</td>
<td>60 96</td>
<td>97.6 99.1</td>
</tr>
<tr>
<td>4</td>
<td>21,985,934 13,626,780</td>
<td>2,074,862,383 1,276,001,973</td>
<td>38 17</td>
<td>93.6 91.8</td>
</tr>
<tr>
<td>5</td>
<td>55,116,882 47,775,867</td>
<td>5,098,710,336 4,436,831,644</td>
<td>68 59</td>
<td>99.2 99</td>
</tr>
<tr>
<td>6</td>
<td>55,166,279 79,944,830</td>
<td>5,066,027,968 7,361,958,066</td>
<td>68 99</td>
<td>99.3 98.6</td>
</tr>
<tr>
<td>7</td>
<td>60,172,743 80,817,890</td>
<td>5,508,038,601 7,437,637,846</td>
<td>74 100</td>
<td>99.3 99.5</td>
</tr>
<tr>
<td>8</td>
<td>37,734,040 59,774,690</td>
<td>3,543,666,305 5,464,037,573</td>
<td>48 73</td>
<td>98.6 99.5</td>
</tr>
<tr>
<td>9</td>
<td>18,011,958 18,187,259</td>
<td>1,614,674,303 1,640,546,264</td>
<td>22 22</td>
<td>99 99</td>
</tr>
<tr>
<td>10</td>
<td>47,773,411 74,349,173</td>
<td>6,689,109,600 7,764,194,016</td>
<td>90 104</td>
<td>99.2 99.5</td>
</tr>
<tr>
<td>11</td>
<td>66,480,847 39,019,760</td>
<td>6,146,505,581 3,644,376,137</td>
<td>82 49</td>
<td>98.9 97.9</td>
</tr>
<tr>
<td>12</td>
<td>35,765,547 54,349,173</td>
<td>3,366,493,766 5,039,766,988</td>
<td>45 68</td>
<td>96.3 98.1</td>
</tr>
<tr>
<td>13</td>
<td>56,547,996 51,214,730</td>
<td>5,267,931,073 4,772,844,570</td>
<td>71 64</td>
<td>98.7 68.7</td>
</tr>
<tr>
<td>14</td>
<td>38,511,142 56,675,359</td>
<td>3,612,334,201 5,341,321,374</td>
<td>48 72</td>
<td>96.7 96.8</td>
</tr>
<tr>
<td>15</td>
<td>54,378,902 48,654,781</td>
<td>4,886,131,853 7,637,212,363</td>
<td>66 102</td>
<td>99.7 98.9</td>
</tr>
<tr>
<td>16</td>
<td>41,335,655 62,937,605</td>
<td>3,856,415,232 5,840,077,590</td>
<td>52 78</td>
<td>98.4 98.8</td>
</tr>
<tr>
<td>17</td>
<td>12,612,972 85,809,640</td>
<td>1,146,953,942 7,930,384,563</td>
<td>15 106</td>
<td>95.6 99</td>
</tr>
<tr>
<td>18</td>
<td>43,517,923 58,476,261</td>
<td>4,032,437,129 5,418,436,409</td>
<td>54 73</td>
<td>97.9 98.6</td>
</tr>
<tr>
<td>19</td>
<td>69,286,743 53,981,810</td>
<td>6,408,976,368 4,982,461,829</td>
<td>86 67</td>
<td>99.2 99.3</td>
</tr>
<tr>
<td>20</td>
<td>63,541,452 46,826,633</td>
<td>5,962,051,678 4,376,285,715</td>
<td>80 59</td>
<td>97.8 97.3</td>
</tr>
<tr>
<td>21</td>
<td>57,691,353 60,787,449</td>
<td>5,322,148,179 5,630,617,835</td>
<td>71 76</td>
<td>98.9 99</td>
</tr>
<tr>
<td>22</td>
<td>42,236,230 70,764,096</td>
<td>3,866,064,306 6,456,266,070</td>
<td>52 87</td>
<td>99.3 99.5</td>
</tr>
<tr>
<td>24</td>
<td>62,454,282 61,252,680</td>
<td>5,694,441,590 5,602,955,013</td>
<td>76 75</td>
<td>99.5 99.4</td>
</tr>
<tr>
<td>25</td>
<td>24,193,171 68,955,803</td>
<td>2,262,053,398 6,365,367,452</td>
<td>30 85</td>
<td>95.6 99.3</td>
</tr>
<tr>
<td>26</td>
<td>19,207,071 48,099,604</td>
<td>1,745,532,463 4,492,557,879</td>
<td>23 60</td>
<td>98.6 98.5</td>
</tr>
<tr>
<td>27</td>
<td>76,433,608 46,069,340</td>
<td>7,086,838,681 4,232,751,636</td>
<td>95 57</td>
<td>99.4 92.5</td>
</tr>
</tbody>
</table>

Average 48,738,379 58,000,669 4,489,162,851 5,343,444,424 61 72 98 97
Table 9. Breast cancer cases and controls variants called with GATK and analyzed with PolyPhen2 and SIFT

<table>
<thead>
<tr>
<th>Samples in families</th>
<th>Total Variants</th>
<th>Total Exonic Variants</th>
<th>Total Predicted Damaging Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Affected</td>
<td>Unaffected</td>
<td>Affected</td>
</tr>
<tr>
<td>1</td>
<td>978,062</td>
<td>1,113,898</td>
<td>19,372</td>
</tr>
<tr>
<td>2</td>
<td>1,636,156</td>
<td>900,886</td>
<td>19,134</td>
</tr>
<tr>
<td>3</td>
<td>1,085,031</td>
<td>1,743,506</td>
<td>18,259</td>
</tr>
<tr>
<td>4</td>
<td>679,774</td>
<td>474,424</td>
<td>15,422</td>
</tr>
<tr>
<td>5</td>
<td>1,082,048</td>
<td>936,887</td>
<td>19,530</td>
</tr>
<tr>
<td>6</td>
<td>1,223,782</td>
<td>1,204,419</td>
<td>18,931</td>
</tr>
<tr>
<td>7</td>
<td>1,319,263</td>
<td>1,225,337</td>
<td>20,681</td>
</tr>
<tr>
<td>8</td>
<td>814,889</td>
<td>1,377,287</td>
<td>19,211</td>
</tr>
<tr>
<td>9</td>
<td>294,220</td>
<td>316,551</td>
<td>18,890</td>
</tr>
<tr>
<td>10</td>
<td>1,263,491</td>
<td>1,424,578</td>
<td>19,513</td>
</tr>
<tr>
<td>11</td>
<td>1,370,974</td>
<td>901,320</td>
<td>19,285</td>
</tr>
<tr>
<td>12</td>
<td>1,029,280</td>
<td>1,586,600</td>
<td>17,270</td>
</tr>
<tr>
<td>13</td>
<td>1,351,599</td>
<td>1,175,098</td>
<td>18,995</td>
</tr>
<tr>
<td>14</td>
<td>1,134,786</td>
<td>1,500,802</td>
<td>17,515</td>
</tr>
<tr>
<td>15</td>
<td>1,013,404</td>
<td>1,254,739</td>
<td>19,677</td>
</tr>
<tr>
<td>16</td>
<td>780,891</td>
<td>1,611,728</td>
<td>18,690</td>
</tr>
<tr>
<td>17</td>
<td>313,035</td>
<td>1,910,332</td>
<td>16,012</td>
</tr>
<tr>
<td>18</td>
<td>1,025,958</td>
<td>1,198,271</td>
<td>18,683</td>
</tr>
<tr>
<td>19</td>
<td>996,549</td>
<td>1,415,127</td>
<td>19,020</td>
</tr>
<tr>
<td>20</td>
<td>1,202,093</td>
<td>838,836</td>
<td>19,131</td>
</tr>
<tr>
<td>21</td>
<td>1,654,218</td>
<td>1,315,488</td>
<td>17,828</td>
</tr>
<tr>
<td>22</td>
<td>1,362,266</td>
<td>1,320,138</td>
<td>19,065</td>
</tr>
<tr>
<td>23</td>
<td>1,046,242</td>
<td>1,497,137</td>
<td>19,259</td>
</tr>
<tr>
<td>24</td>
<td>505,466</td>
<td>734,575</td>
<td>19,269</td>
</tr>
<tr>
<td>25</td>
<td>1,311,678</td>
<td>1,418,653</td>
<td>19,096</td>
</tr>
<tr>
<td>26</td>
<td>424,269</td>
<td>1,155,242</td>
<td>17,294</td>
</tr>
<tr>
<td>27</td>
<td>467,546</td>
<td>1,036,908</td>
<td>18,490</td>
</tr>
<tr>
<td>Average</td>
<td>1,359,857</td>
<td>962,818</td>
<td>19,609</td>
</tr>
</tbody>
</table>
Because we are only analyzing the variants from 54 samples, we must first verify that we have the sensitivity to call variants that are true positives. To do this, we first identified variants previously published to change the risk of breast or ovarian cancer.

We hypothesized that if our sample size was large enough to identify previously identified genetic risk modifiers, we would also be able to identify new genetic risk modifiers.

To identify these previously identified variants, the Odds Ratio was calculated on the variants predicted damaged by either PolyPhen2 or SIFT. Each variant that segregated differently between the breast cancer affected and unaffected with the statistical significance of $P \leq 0.15$ were analyzed further through a search in PubMed. The PubMed search found two variants previously identified to change the risk of breast or ovarian cancer (Figure 9). The variant rs8100241 in ANKLE1 has been identified by CIMBA to decrease the risk of breast cancer in BRCA1 mutation carriers $^{42,43,58,59}$. Similarly, this variant also segregates more in the unaffected samples in the case control study. The variant rs2295190 in SYNE1 has been previously identified to increase the risk of ovarian cancer $^{60}$. Similarly, this variant also segregates more in the breast cancer affected samples than in the case control. These results show that our case control sample size of 54 is large enough to identify previously identified risk modifying variants.
Figure 9. Probable biological variants previously identified to associate with breast or ovarian cancer risk. All variants that segregated with a $p \leq 0.15$ and predicted damaging by PolyPhen2 (PP2) or SIFT were manually searched in PubMed to find published association of the variant with breast or ovarian cancer. The variant rs8100241 in ANKLE1 was previously identified by CIMBA to decrease the risk of breast cancer in *BRCA1* mutation carriers. The variant rs2295190 in SYNE1 has been identified previously to increase the risk of ovarian cancer.
2.3.1.3.2 Newly identified breast and/or risk modifying variants

Next, we attempted to identify new variants that modified the risk of breast cancer in \textit{BRCA1} affected carriers. To decrease the chance of false positive variants, we set both statistical and biological filters as follows:

1. The variants are predicted damaging by both PolyPhen2 and SIFT.
2. The variants segregate with the statistical significances of $P<0.05$.
3. Ingenuity pathway analysis link the affected gene to cancer
4. There is a publication in PubMed that gives evidence that the affected gene is a known or a predicted tumor suppressor/oncogene.
5. The amino acid change is in a known functional domain or odds ratio is $<0.05$ or $>20$ or the affected gene has a protein-protein interaction with a high-medium penetrance breast cancer predisposition gene.

The variant ANLN\_rs3735400, SIPA1L2\_rs1547742 and MTHFSD\_rs3751800 met all five criteria \cite{61-66} (Figure 10). These new and previously identified risk modifiers were then verified with Sanger sequencing and the odds ratios per allele was calculated (Table 10). After Sanger sequencing, all five variants passed all biological and statistical filters.

2.3.1.4 Cohort study to calculate the cumulative risk of the candidate genetic modifiers

I then calculated if these variants changed the cumulative risk of developing breast cancer. To answer this question, the five variants were analyzed further in a more statistically robust cohort study.
Figure 10. Probable biological variants newly identified to associate with breast cancer risk. All variants that segregated with a $P<0.05$ and predicted damaging by PolyPhen2 (PP2) and SIFT were searched in both Ingenuity pathway analysis (IPA) and PubMed for evidence that they affect known/predicted oncogene or tumor suppressors. These variants must also 1) cause an amino acid change in a known functional domain, 2) have an odds ratio $<0.05$ or $>20$, or 3) the affected gene must have a protein-protein interaction with a high-medium penetrance breast cancer predisposition gene. The variants ANLN_rs3735400, SIPA1L2_rs1547742 and MTHFSD_rs3751800 passed through all five filters.
Table 10. Sanger sequenced risk modifying variants Odds Ratio per allele

<table>
<thead>
<tr>
<th>dbSNP</th>
<th>Gene</th>
<th>BRCA1 Affected (%)</th>
<th>BRCA1 Unaffected (%)</th>
<th>MAF*</th>
<th>AA change</th>
<th>Functional Domain</th>
<th>OR</th>
<th>95% CI</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2295190</td>
<td>SYNE1</td>
<td>9 (16.7)</td>
<td>4 (7.4)</td>
<td>17.5</td>
<td>L-M</td>
<td>KASH</td>
<td>2.87</td>
<td>0.760-10.865</td>
<td>0.119</td>
</tr>
<tr>
<td>rs8100241</td>
<td>ANKLE1</td>
<td>17 (31.5)</td>
<td>31 (57.4)</td>
<td>45.5</td>
<td>G-A</td>
<td>-</td>
<td>0.37</td>
<td>0.170-0.808</td>
<td>0.013</td>
</tr>
<tr>
<td>rs3735400</td>
<td>ANLN</td>
<td>2 (3.7)</td>
<td>11 (20.4)</td>
<td>11.3</td>
<td>S-W</td>
<td>NLS</td>
<td>0.15</td>
<td>0.033-0.748</td>
<td>0.020</td>
</tr>
<tr>
<td>rs1547742</td>
<td>SIPA1L2</td>
<td>0 (0)</td>
<td>8 (14.8)</td>
<td>9.7</td>
<td>S-L</td>
<td>-</td>
<td>0.05</td>
<td>0.003-0.931</td>
<td>0.045</td>
</tr>
<tr>
<td>rs3751800</td>
<td>MTHFSD</td>
<td>3 (5.6)</td>
<td>11(20.4)</td>
<td>10.1</td>
<td>R-C</td>
<td>-</td>
<td>0.24</td>
<td>0.063-0.917</td>
<td>0.037</td>
</tr>
</tbody>
</table>

*HapMap-ceu
In this cohort study, ~150 BRCA1 mutation carriers (16 unaffected and 80-91 breast cancer affected), were genotyped by Sanger sequencing and the Cox regression model was used to calculate the cumulative risk. These new BRCA1 mutation carriers were selected based on the BRCA1 mutated status.

After the cumulative risks were calculated, it was shown that all five variants statistically modify the risk of breast cancer in these BRCA1 mutation carriers. SYNE1_rs2295190 was the only variant to increase the risk of developing breast cancer in BRCA1 mutation carriers (HR 1.66, CI 1.05-2.63, \(P=0.031\)) (Figure 11). ANKLE1_rs8100241, which was shown to decrease the risk of developing breast cancer in the CIMBA studies also decrease the risk of breast cancer in BRCA1 mutation carriers (HR 0.64, CI 0.5-0.099, \(P=0.046\)) (Figure 12). The variants in ANLN_rs3735400 (HR 0.56, CI 0.34-0.93, \(P=0.025\)) (Figure 13), SIPA1L2_rs1547742 (HR 0.464, CI 0.28-0.77, \(P=0.003\)) (Figure 14), and MTHFSD_rs3751800 (HR 0.28, CI 0.14-0.58, \(P=0.0006\)) (Figure 15) all decreased the risk of developing breast cancer in BRCA1 mutation carriers.

### 2.3.2 Functional study

As previously stated, candidate variants must be predicted damaging by variant prediction software as well as affect a gene known or predicted to be a tumor suppressor or oncogene. The goal of these functional studies is to show evidence that these predictions were accurate. To show evidence that the affected genes are oncogenes or tumor suppressors, we made knockdown cell lines and measured the cellular growth rate. It is hypothesized that if the affected gene is an oncogene, the knockdown cell line will have a decreased growth rate in comparison to the control cell line. Furthermore, it is also hypothesized that if the affected gene is a tumor suppressor, the knockdown cell line will have an increased growth rate in comparison to the control cell line.
<table>
<thead>
<tr>
<th>Genotype</th>
<th>Unaffected (%)</th>
<th>Breast Cancer (%)</th>
<th>HapMap %</th>
<th>HR</th>
<th>95% CI</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>38 (86.4)</td>
<td>77 (74.8)</td>
<td>70.0</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>6 (13.6)</td>
<td>26 (25.2)</td>
<td>25.0</td>
<td>1.66</td>
<td>1.05-2.63</td>
<td>0.031</td>
</tr>
<tr>
<td>AA</td>
<td>0</td>
<td>2 (1.9)</td>
<td>5.0</td>
<td>2.23</td>
<td>0.51-9.72</td>
<td>0.285</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>105</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 11. SYNE1_rs2295190 cumulative risk of breast cancer
Figure 12. ANKLE1_rs8100241 cumulative risk of breast cancer
<table>
<thead>
<tr>
<th>Genotype</th>
<th>Unaffected (%)</th>
<th>Breast Cancer (%)</th>
<th>HapMap %</th>
<th>HR</th>
<th>95% CI</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>29 (65.9)</td>
<td>93 (83.0)</td>
<td>81.4</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG</td>
<td>15 (34.1)</td>
<td>19 (17.0)</td>
<td>15.3</td>
<td>0.56</td>
<td>0.34-0.93</td>
<td>0.025</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>112</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 13. ANLN_rs3735400 cumulative risk of breast cancer
Figure 14. SIPA1L2_rs1547742 cumulative risk of breast cancer
<table>
<thead>
<tr>
<th>Genotype</th>
<th>Unaffected (%)</th>
<th>Breast Cancer (%)</th>
<th>HapMap %</th>
<th>HR</th>
<th>95% CI</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>29 (65.9)</td>
<td>100 (91.7)</td>
<td>71.2</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>15 (34.1)</td>
<td>8 (7.3)</td>
<td>27.9</td>
<td>0.28</td>
<td>0.14-0.58</td>
<td>0.0006</td>
</tr>
<tr>
<td>TT</td>
<td>0 (0)</td>
<td>1 (.9)</td>
<td>0.9</td>
<td>1.54</td>
<td>0.20-12.01</td>
<td>0.682</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>109</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 15. MTHFSD_rs3751800 cumulative risk of breast cancer
To show evidence that the predicted damaging variant is altering the function of the affected gene, the function of the native allele and candidate modifier variant will be measured in vitro. Because more specific experiments can be designed if a damaging variant is affecting a known functional domain, only candidate modifiers in known functional domains were analyzed. It is hypothesized that if the predicted damaged variant is in a functional domain, this affected domain will be altered functionally.

While there is statistical evidence to show these five variants change the risk of developing breast cancer in BRCA1 mutation carriers, it is unclear whether these variants affect pre- or post-transformed breast cells. Because post-transformed cell lines with known BRCA1 mutations are more established than pre-transformed breast epithelial cell lines, these affected genes would be analyzed in post-transformed breast cancer cell lines. Both HCC1937 and MCF7 have mutations in BRCA1. HCC1937 has a homozygote BRCA1 mutation genotype and is a basal-like ER/PR- normal Her2/neu subtype. MCF7 has a heterozygote null BRCA1 genotype and is an ER/PR+ normal Her2/neu luminal subtype.

Three criteria were set for the selection of candidate modifiers for functional studies.

1. The variants must be found to modify the risk of breast cancer in the cohort study with statistical significant \( P < 0.05 \).
2. The affected gene must be expressed in either MCF7 or HCC1937 cells.
3. The variant must be in a functional domain.

Variants found to modify the risk of breast cancer in the cohort study with statistical significant \( P < 0.05 \) are ANLN_rs3735400, SIPA1L2_rs1547742, MTHFSD_rs3751800, SYNE1_rs2295190 and ANKLE1_rs8100241. To investigate the gene expression in these two cell lines, RNA-seq data from both HCC1937 and MCF7 was downloaded from the cancer genomics hub. The only candidate gene not expressed in either cell line
was ANKLE1. Similarly, it has been previously shown that ANKLE1 is not expressed in breast cell lines. \(^{67}\) Finally, two candidate modifiers, ANLN_rs3735400 and SYNE1_rs2295190 were in known functional domains (Figure 16). ANLN_rs3735400 is in a Nucleus Localization Sequence (NLS) and an mDia2-binding domain while SYNE1_rs2295190 is in a KASH domain important for nuclear membrane localization.\(^{54,68}\)

2.3.2.1 ANLN

2.3.2.1.1 Anillin and its role during cytokinesis

Actin-binding protein Anillin (ANLN) was first identified in Drosophila and plays important roles in cytokinesis.\(^{69}\) Anillin is localized to the nucleus during interphase and relocates to the cortex after the nuclear envelope degrade. It then localizes to the equatorial cortex in anaphase and at the cleavage furrow during telophase.\(^{70}\) The ingression of the cleavage furrow separates daughter cells at the end of mitosis (Figure 17). This furrow ingression is driven by the assembly and contraction of actomyosin filaments, forming a contractile ring. Cytoskeletal protein Anillin has a pivotal role in the organization of the network of cytoskeletal proteins at the cleavage furrow. This network of cytoskeletal proteins acts as a scaffold connecting actomyosin filaments to the plasma membrane.\(^{70-73}\)

2.3.2.1.2 Anillin functional domains

Human Anillin is comprised of four functional domains: Myosin-2-binding domain, Actin-binding domain, RhoA-binding domain and the pleckstrin-homology (PH) domain. Anillin also includes three other secondary structures: destruction box (Dbox), nuclear socialization signal (NLS), and Src-homology-3-binding consensus sequences (SH3).
Figure 16. Candidate modifiers of breast cancer risk chosen for functional studies. Three criteria were set for the selection of candidate modifiers for functional studies. Only SYNE1_rs2295190 and ANLN_rs3735400 were chosen.
The function of Anillin in Drosophila

Function of Anillin-related proteins associating with the growing end. Localize Mid1p in a medial cortical band. Another, as-yet-unidentified protein, Pom1p excludes Mid1p from the non-growing end and cooperates with it to control the division site. This results in the spreading of Mid1p from the medial region to the non-growing end; however, nuclear export per se cannot explain the localization of Mid1p (Echard et al., 2004; Padte et al., 2006). Computer simulations indicate that nuclear export per se cannot explain the localization of Mid1p.

Furthermore, the localization of myosin II and septins is disrupted in Anillin-depleted cells, which causes a series of defects in actomyosin-mediated contractile activity, during both cytokinesis and cellularization (Maddox et al., 2005). RNAi of anillin or ANI-2. ANI-2 localizes to the hermaphrodite gonad and is not required for F-actin accumulation at the furrow but rather for cytokinesis (Echard et al., 2004; Somma et al., 2002). Thus, Anillin is essential for the proper localization of Mid1p to the spindle midzone (D’Avino et al., 2008; Gregory et al., 2008). Finally, although Anillin is not required for RacGAP50C localization to the spindle midzone (D’Avino et al., 2008; Gregory et al., 2008), it does seem to be necessary for RacGAP50C accumulation at the cortex in larval brain cells (Gregory et al., 2008).

Figure 17. Localization of Anillin during the cell cycle in mammals. Anillin protein is depicted in red, chromosomes in black and microtubules, centrosomes, and spindle-pole bodies in blue.

Figure 17. Localization of Anillin during the cell cycle in mammals. Anillin protein is depicted in red, chromosomes in black and microtubules, centrosomes, and spindle-pole bodies in blue.
The destruction box (Dbox) domain of Anillin binds to cadherin 1, type 1 (CDH1) and anaphase prompting complex (APC), which is an ubiquitin protein ligase that ubiquitinates Anillin, promoting its degradation during late mitosis/G1 \(^{74}\). The nuclear socialization signal (NLS) of Anillin binds the protein Important b2, which shuttles Anillin from the cytoplasm to the nucleus during interphase \(^{54}\). The Src-homology-3-binding consensus sequence (SH3) regulates the active state of adaptor proteins and increase the substrate specificity of some tyrosine kinases. The Myosin-2 binding domain of Anillin binds myosin-2, which is responsible for actin-based motility \(^{73}\). The Actin-binding domain of Anillin binds actin, which is a protein that forms microfilaments \(^{75}\). The RhoA domain of Anillin binds the gene Ras Homolog Family Member A (RhoA), which is a member of the Rho family of small GTPases \(^{72}\). The pleckstrin-homolgy (PH) domain of Anillin has been shown to bind septins at the plasma membrane \(^{76}\) (Figure 18).

2.3.2.1.3 Anillin role in Cancer

Anillin has been found to be overexpressed in many different tumors. Furthermore, Anillin expression has also been found to be a prognostic biomarker for cancer \(^{61}\). Other studies have found that high nuclear fraction of ANLN was associated with large tumor size, high histological grade, high proliferation and estrogen receptor (ER) negativity \(^{77}\). Recently it has been shown that stable Anillin knockdown breast cancer cell lines MDA-MB-231 and ZR-75-30 have decreased growth rates. These knockdown cells also have increased percentage of cells in G2/M phase \(^{62}\). This growth rate defect in breast cancer cell lines give evidence that Anillin is not only a cancer biomarker but also an oncogene in breast cancer.

2.3.2.1.4 ANLN knockdown

While both MDA-MB-231 and ZR-75-30 have decreased growth rates after ANLN expression is decreased, these cell lines have normal BRCA1 expression.
Figure 18. Secondary structures of Anillin protein in humans. Variant rs3735400 is located in the N-terminal region of Anillin and causes a change from serine to tryptophan in both the mDia2 binding motif and NLS domain.
To test if breast cancer cell lines with \textit{BRCA1} mutations are affected by decreased expression of ANLN, shRNA specific for ANLN was infected into both MCF7 and HCC1937 cell lines. As stated previously, HCC1937 has a homozygote \textit{BRCA1} mutation genotype and is a basal-like ER/PR- normal Her2/neu subtype cell line while MCF7 has a heterozygote null \textit{BRCA1} genotype and is a luminal subtype cell line. HCC1937 cells infected with shRNA_ANLN are not viable. All infected HCC1937 cells die within two days. This infection was repeated three times with similar results. However, a stable ANLN knockdown in MCF7 showed a decreased growth rate in comparison to the scrambled_shRNA infected MCF7 cells (Figure 19). This growth rate defect is similar to the effects of shRNA_ANLN in both MDA-MB-231 and ZR-75-30 cells. These results show that \textit{BRCA1} mutated cell lines are equally sensitive, if not more sensitive, to the decreased expression of ANLN. Furthermore, this gives evidence that Anillin is a true oncogene.

\textbf{2.3.2.1.5 Variant rs3735400 affects the function of Anillin}

\textbf{2.3.2.1.5.1 Nuclear localization}

The variant rs3735400 is a nonsynonymous mutation that causes a serine to tryptophan mutation of amino acid 65 in Anillin. This variant is localized upstream of the NLS of Anillin (Figure 20). To investigate if the variant rs375400 alters the ability of Anillin to localize to the nucleus, we transfected a N-terminal GFP-tagged Anillin construct into HeLa cells (Figure 21). We found that Anillin with the variant rs3735400 had defected nuclear localization in comparison to the wild type Anillin (Figure 22). While the function of Anillin in the nucleus is unknown, it has been shown that Anillin with a mutated NLS maintain the ability to rescue the bi-nucleated phenotype caused by an Anillin knockdown\textsuperscript{54}.
Figure 19. Growth rate measurement of MCF7 infected shRNA_ANLN cells.

Knockdown of Anillin in MCF7 cells causes the growth rate to decrease in comparison to the shRNA_Scrambled infected MCF7 cells. *P<0.05.
The variant ANLN_rs3735400 affects the amino acid three amino acids upstream from the NLS domain.

Figure 20. N-terminus of Anillin has conserved NLS across different vertebrate species. The variant ANLN_rs3735400 affects the amino acid three amino acids upstream from the NLS domain.
Figure 21. Anillin constructs transfected into HeLa cells. To analyze the effects of ANLN_rs3735400 on Anillin nuclear localization, a full length GFP tagged Anillin_wt and Anillin_(S65W) were transfected into HeLa cells. Both proteins are able to localize to the nucleus.
Figure 22. **ANLN_rs3735400 affects the nuclear localization of Anillin.** Different GFP-Anillin fusion proteins in HeLa cells showing a ratio change of GFP fluorescence between the nuclear and cytoplasmic compartments. **P<0.005.**
2.3.2.1.5.2 Knockdown growth rate rescue

Next, we investigated whether ANLN with the variant rs3735400 can rescue the growth rate phenotype caused by an Anillin knockdown in MCF7 cells. It has been shown that the first 115 amino acids of Anillin is required for the proper localization of mDia2 to the cleavage furrow\(^7\)). Furthermore, truncated Anillin without the first 115 amino acids abolishes the ability of Anillin to rescue the bi-nucleated phenotype caused by an Anillin knockdown\(^78\)). Because rs3735400 affects an amino acid within the first 115 amino acids (ANLN_S65W), it was hypothesized that this variant will disrupt the ability of Anillin to rescue the growth rate defect caused by an infection of shRNA_Anillin in MCF7 cells. Growth rate of infected shRNA_ANLN cells rescued with transfected ANLN_wt was similar to cells with endogenous levels of Anillin protein. However, infected shRNA_ANLN cells transfected with ANLN_(S65W) showed similar growth rates to cells that were infected with shRNA_ANLN and transfected with an empty vector (Figure 23). This result suggests that the variant rs3735400 may alter the function of ANLN and its ability to regulate cellular proliferation.

2.3.2.2 SYNE1

2.3.2.2.1 SYNE1 and its role in nuclear membrane stability

Nesprin-1 (SYNE1) belongs to a family of spectrin-repeat proteins. Alternate initiation and splicing of SYNE1 generate multiple isoforms that vary greatly in size and function. Giant isoform Nesprin-1 is composed of a spectrin-repeat rod domain linked to a C-terminal transmembrane KASH (Klarsicht-ANC-SYNE-homology) domain, which mediates nuclear membrane localization, and N-terminal alpha-actinin-type actin-binding domain (ABD)\(^79\)). Nesprin-1 is an essential component of the nuclear envelope and is part of the LINC (linker of nucleoskeleton and cytoskeleton) complex.
Figure 23. ANLN_rs3735400 inhibits Anillin ability to rescue the ANLN knockdown. The ANLN knockdown could be rescued with ANLN_wt. ANLN_(S65W) protein was unable to rescue the knockdown growth rate defect and had a similar rate to the knockdown ANLN transfected with an empty vector. *$P<0.05$, **$P<0.005$. 
Disruption of the nuclear-cytoskeleton connections affects the stability, size, and shape of the nucleus and alters cell migration and mechanical properties of the cell \(^{80-83}\).

2.3.2.2.2 SYNE1 and cancer

Downregulation of Drop1, an N-terminal isoform of Nesprin-1 has been observed in early tumor stages in a wide range of human carcinomas\(^{84}\). Furthermore, data from TCGA and other publications have shown that SYNE1 is frequently mutated in both breast and colorectal cancer\(^{85}\). It has also been reported that Nesprin-1 has a role in DNA damage response and DNA repair and mismatch repair pathways, which are determinants of genetic instability in cancer. It was observed that human foreskin cells (HFF1) transfected with siRNA_SYNE1 had a larger percentage of gamma-H2AX positive cells than cells with endogenous levels of SYNE1. Gamma-H2AX is a biomarker for DNA double-strand breaks\(^{86}\). These siRNA_SYNE1 HFF1 cells also have defects in size and shape of the nucleus. It is hypothesized that this genomic instability is occurring in these knockdown cells due to a combination of alterations in the DNA damage response and mismatch repair pathways and increased physical stress of the nuclear membrane. This increase in DNA double-strand breaks is especially interesting to this study because a critical function of \(BRCA1\) is to repair DNA double-strand breaks\(^{87}\).

The variant SYNE1_rs2295190 has previously been associated with an increased risk of developing invasive epithelial ovarian cancer\(^{60}\). The result of the cohort study also associated rs2295190 with an increased risk of developing breast cancer in \(BRCA1\) mutation carriers. The mutation frequency in cancer, function in genomic stability, and association of the predicted damaging variant rs2295190 with an increase in breast and ovarian cancer risk gives evidence that SYNE1 is a tumor suppressor.
2.3.2.2.3 Knockdown SYNE1

To investigate the role of SYNE1 in breast cancer, a stable shRNA_SYNE1 MCF7 cell line was made. HCC1937 cells were not infected with SYNE1_shRNA because SYNE1 is not expressed in this cell line. SYNE1 knockdown MCF7 cells have an increased growth rate in comparison to the Scrambled_shRNA infected MCF7 cells (Figure 24). This increased growth rate in MCF7 cells and the increased genomic instability in HFF1 cells gives further functional evidence that SYNE1 is a tumor suppressor.

2.3.2.2.4 Variant rs2295190 affects the function of Nesprin-1

The variant rs2295190 is a nonsynonymous mutation that causes a leucine to methionine alteration of amino acid 8693 in Nesprin-1 (Figure 25). This alteration is in the neck region of the KASH domain, which is the nuclear transmembrane domain of Nesprin-1 (Figure 26). Because this amino acid is very conserved in multiple organisms and located near the nuclear transmembrane domain, we hypothesize that this variant is altering the ability of Nesprin-1 to localize to the nuclear membrane. To test this hypothesis, a N-terminal truncated N-terminal tagged GFP Nesprin-1 construct was used to investigate whether this variant can alter the ability of truncated Nesprin-1 to bind to the nuclear membrane (Figure 27). As previously published, this construct with the native allele binds tightly around the nuclear membrane. However, once the native amino acid was changed from a leucine to a methionine, the GFP-Nesprin-1_(L8693M) was dispersed throughout the cytoplasm and seemed to lose the ability to localize tightly to the nuclear membrane. This localization alteration gives evidence that the variant rs2295190 can alter the function of the tumor suppressor Nesprin-1.
Figure 24. Growth rate measurement of MCF7 infected shRNA_SYNE1 cells. Knockdown of Nesprin-1 in MCF7 cells causes the growth rate to increase in comparison to the shRNA_Scrambled infected MCF7 cells. *$P<0.05$. 
Figure 25. Secondary structures of Nesprin-1 protein in humans. Nesprin-1 binds to the F-Actin and the nuclear membrane. Variant rs2295190 is located in the C-terminal nuclear transmembrane KASH domain of Nesprin-1 and causes a leucine to methionine amino acid change.
Figure 26. **C-terminus KASH domain of Nesprin-1 is conserved in different vertebrate species.** The variant SYNE1_rs2295190 affects an amino acid in the neck region of the KASH domain of Nesprin-1. Mm, *Mus musculus*. Hs, *Homo sapiens.*
Figure 27. SYNE1-GFP tagged constructs transfected into HeLa cells. To analyze the effects of SYNE1_rs2295190 on Nesprin-1 nuclear membrane localization, a truncated GFP-tagged SYNE1_wt and SYNE1_(L8693M) were transfected into HeLa cells. SYNE1_wt localized tightly to the nuclear membrane while SYNE1_(L8693M) was dispersed throughout the cytoplasm and seems to lose the ability to localize tightly to the nuclear membrane. SRs, spectrin repeats. TM, transmembrane domain.
2.4 Discussion

2.4.1 Statistical association that common variants can modify the risk of breast cancer in BRCA1 mutation carriers

Prior to this study, the only evidence that common variants could affect the risk of developing breast cancer in BRCA1 mutation carriers was from GWAS studies. GWAS studies are purely based on statistical analysis and the variants identified are not selected based on biological plausibility. There has been no published studies to functionally link the variants found in these GWAS studies to breast cancer. The ultimate goal of this study was to identify variants that statistically change the risk of developing breast cancer in BRCA1 mutation carriers as well as provide functional evidence linking these variants to breast cancer development.

To identify breast cancer risk variants, we sequenced 27 high risk and 27 low risk BRCA1 mutation carriers with exome next-generation sequencing. We then applied multiple biological and statistical filters to identify variants that are linked both statically and functionally to breast cancer risk.

The statistical association results identified five variants (ANKLE1_rs8100241, ANLN_rs3735400, SIPA1L2_rs1547742, MTHFSD_rs3751800 and SYNE1_rs2295190) that both modify the risk of breast cancer with statistical significant and are predicted to alter genes previously linked to cancer development. A major concern for this study was that the case control sample size would be too small to accurately identify risk modifying variants. Five of the variants identified in the case control study maintained statistical significance when re-examined in the more statistically powerful cohort study. These results provide evidence that the case control study is powerful enough to identify risk modifying variants. Furthermore, the variant ANKLE1_rs8100241 identified in the case control study was previously identified in other larger GWAS studies to modify the risk of...
breast cancer in *BRCA1* mutation carriers. The other four variants found in this study are newly identified *BRCA1* mutation carrier genetic risk modifiers. While all of the variants can be linked to cancer development, two of the five variants were chosen for further investigation with functional studies.

### 2.4.2 Functional studies with ANLN_rs3735400 and SYNE1_rs2295190

There were two aims for these functional studies. The first aim was to provide evidence that the affected gene could alter the growth rate of *BRCA1* mutated breast cancer cells. The second aim was to provide evidence that the identified risk-modifying variants could alter the function of the affected gene. Only two of the five variants, ANLN_rs3735400 and SYNE1_rs2295190, were included in the functional studies because they were the only variants that affected a gene expressed in *BRCA1* mutated breast cancer cells and are located in a known functional domain.

#### 2.4.2.1 ANLN_rs3735400

It was shown that the predicted damaging variant ANLN_rs3735400 decreases the risk of breast cancer in *BRCA1* mutation carriers. Because damaging variants most commonly cause a loss of function genetic phenotype, it would be predicted that ANLN is an oncogene. This is not the first time ANLN was predicted to be an oncogene. ANLN has previously been shown to be overexpressed in different types of cancer and is a biomarker for poor prognosis in breast cancer. Furthermore, this functional study has shown that ANLN expression is linked positively to the proliferation rate of *BRCA1* mutation breast cancer cell lines, providing further evidence that ANLN is an oncogene.

Next, we investigated whether ANLN_rs3735400 could alter the function of Anillin. First, we showed that ANLN_rs3735400 could decrease the ability of Anillin to localize to the nucleus. While the function of ANLN in the nucleus is currently unknown, it has been
hypothesized that Anillin is sequestered into the nucleus during interphase to control the Anillin ability to reorganize the actin filaments in a cell cycle dependent manner. Furthermore, the ability of ANLN to localize to the nucleus has been shown to be independent of its function during cytokinesis.

The ability of ANLN_rs3735400 to rescue the growth rate caused by the shRNA_ANLN knockdown was investigated next. It was shown that while ANLN_wt could fully rescue the phenotype caused by shRNA_ANLN, ANLN_(S65W) was unable to change the growth rate defect caused by shRNA_ANLN in MCF7 cells. This was not suppressing as it has been shown that the N-terminal 115 amino acids of ANLN is critical to maintain Anillin ability to function during cytokinesis. The N-terminal 115 amino acids were found to be critical for the localization and activation of the formin mDia2 to the cleavage furrow. It has been shown that cells with altered mDia2 do not form F-actin required for transition through cytokinesis. Because the variant ANLN_rs373540 is within this N-terminal region, we hypothesize that this variant ANLN_(S65W) is inhibiting ANLN ability to interact with mDia2.

2.4.2.2 SYNE1_rs2295190

It was shown that the predicted damaging variant SYNE1_rs2295190 increases the risk of breast cancer in BRCA1 mutation carriers. Because damaging variants most commonly cause a loss of function genetic phenotype, it would be predicted that SYNE1 is a tumor suppressor. This is not the first time SYNE1 has been predicted to be a tumor suppressor. It has previously been shown that SYNE1_rs2295190 is associated with an increase in risk of ovarian cancer 60. In addition, SYNE1 has been shown to be frequently mutated in different types of cancer and loss of SYNE1 is known to increase genomic instability 68,84. Furthermore, this functional study has shown that decreasing
the expression of SYNE1 increases the proliferation rate of a BRCA1 mutation breast cancer cell line, providing more evidence that SYNE1 is a tumor suppressor.

Next, it was investigated whether SYNE1_rs2295190 could alter the function of Nesprin-1. Giant Nesprin-1 has been shown to link the cytoskeleton to the nucleoskeleton by binding F-actin in both the cytosol and the nuclear membrane. The variant SYNE1_rs2295190 changes a highly conserved amino acid leucine to a methionine in the neck region of the transmembrane KASH domain. This study showed that the variant SYNE1_rs2295190 can alter the ability of Nesprin-1 to bind to the nuclear membrane. It is hypothesized that this alteration of Nesprin-1 localization can affect the nucleoskeleton linkage, which has been shown to affect the function of both DNA repair and mismatch proteins.
DISCUSSION
Phasing the genome with NGS

Information derived from genomic phasing can provide information on the genetic origin of disease and how DNA nucleotide alterations interact *in cis*. However, current next-generation sequencing techniques generate a single consensus sequence without distinguishing between variants on homologous chromosomes. Scientists and engineers are actively advancing genome phasing capabilities by redesigning methods of genomic library construction and improving the NGS platform technology so longer reads can be sequenced.

The current experimental approaches used to phase the genome are too experimentally laborious to be commonly used. The goal of my first project was to analyze a new experimentally non-laborious method of NGS library preparation hypothesized to increase the amount of phasing information derived from NGS data. This new method involved cutting the genomic DNA with restriction enzymes instead of random shearing.

It was shown that NGS libraries processed with restriction enzymes provide more phasing information than libraries made with random shearing. However, there were some technical issues with the reads generated from restriction enzyme libraries. First, libraries generated with a single restriction enzyme is only able to cover 40-50% of the exome. With similar read depths, a NGS library made with random shearing covered about 75% of the exome. However, exome coverage was greatly increased when NGS libraries were generated with a combination of four different restriction enzymes. While this method increases the exome coverage, the cost, time and quantity of genomic DNA required to process four libraries is more than that needed to make a single sonication derived NGS library.

Another issue was that the reads from restriction enzyme libraries were difficult to map to the reference genome. Furthermore, these difficulties might also be the reason why
many variants called from these reads were false variants. These results are likely due to reads that are too small to be aligned accurately to the reference human genome. This technical issue and can be addressed by making changes to the methodology of restriction enzymes NGS library preparation. Another option is to set more strict filters to remove lower quality reads before mapping to the reference genome. However, it may make more fiscal sense to remove these reads before sequencing to save on Illumina sequencing running cost. For now, it seems that the current gain-of-phasing information derived from restriction enzyme NGS libraries does not outweigh the technical issues with this method.

**Future directions for phasing the genome with NGS**

It is proposed that the ideal option to increase the number of phased variants while maintaining a low variant error rate is to modify the sonication method of shearing genomic DNA. By pooling together two libraries, one library with the recommended fragment size and one library with a larger fragment size, the ability to construct larger contigs would increase and the sensitivity and specificity to call variants would be maintained.

The NGS platform technology is currently the limiting factor in our ability to phase the genome with high-throughput sequencing. However, as NGS platform technology improves, the ability to phase the genome will also increase (during my time as a PhD student, Illumina sequence lengths increased from 100bp to 150bp). There are also talks of a fourth generation sequencing platform (Oxford Nanopore) that has no DNA fragment size limitations. If this system works as described, it may be the solution to phasing the genome efficiently. This technology will open up new fields of research that is currently underexplored.
Genetic modifiers of breast cancer risk of BRCA1 mutation carriers

Risk of breast cancer is caused by a combination of environment and genetic modifiers. BRCA1 mutation carriers from familial breast cancer families have a wide range of risk (44-75%) of developing breast cancer by age 70\textsuperscript{15}. There are many known factors that can modify BRCA1 mutation carrier risk.

Currently, there are limited risk assessments and treatment options for the intervention of breast cancer in BRCA1 mutation carriers, leading some carriers to undergo invasive mastectomy. By identifying germline modifiers, physicians can provide better personalize cancer risk assessments to BRCA1 mutation carriers and researchers can potentially design new preventative therapies to lower the risk of cancer in these carriers.

The majority of known modifying factors in BRCA1 mutation carriers are environmental. One reason for this is that the technology to perform GWAS studies has only been available for eleven years. Furthermore, because information about genetic modifiers is from GWAS studies, which can only associate variants with changes of cancer risk, the variants that cause risk of breast cancer to change are currently unknown. To date, there have been no studies to show that genetic variants can functionally changing the risk of breast cancer in BRCA1 mutation carriers. The goal of this study was to identify variants that modify the risk of developing breast cancer in BRCA1 mutation carriers and provide functional evidence linking these variants to breast cancer development.

This study is the first to identify new risk modifying variants (ANLN\textsubscript{rs}3735400, SYNE1\textsubscript{rs}2295190, SIPA1L2\textsubscript{rs}1547742, MTHFSD\textsubscript{rs}3751800, and ANKLE1\textsubscript{rs}8100241) based on statistical and biological evidence. While ANKLE1\textsubscript{rs}8100241 was previously identified by GWAS studies to statistically decrease the risk of breast cancer in BRCA1 mutation carriers, there was no biological evidence
that this variant was functional important. SYNE1_rs2295190 was the only variant found to increase the risk of breast cancer in BRCA1 mutation carriers. The variant SYNE1_rs2295190 has previously been shown to statistically increase the risk of ovarian cancer. Similar to the variant in ANKLE1, SYNE1_rs2295190 was previously identified without biological evidence that this variant was functional important. The three newly identified risk modifying variants ANLN_rs3735400, SIPA1L2_rs1547742, and MTHFSD_rs3751800 were all shown to decrease the risk of breast cancer in BRCA1 mutation carriers.

The variant MTHFSD_rs3751800 was shown to have the greatest modifying effect of breast cancer risk in BRCA1 mutation carriers. BRCA1 mutation carriers with the variant MTHFSD_rs3751800 had a 70% lower risk of developing breast cancer than BRCA1 mutation carriers without this variant. Furthermore, this variant was also found to be the most statistically significant. While this study has shown evidence that a variant in the gene MTHFSD is associated with lower breast cancer risk, there have been no studies to analyze the function of MTHFSD. MTHFSD was previously identified in only one study, during a screen for genes that bind to p53. While there is evidence that MTHFSD and p53 bind, the consequence of MTHFSD binding to p53 remains unknown.

The second goal of this study was to provide functional evidence that common variants can modify the risk of breast cancer in BRCA1 mutation carriers. We decided to focus on the variants ANLN_rs3735400 and SYNE1_rs2295190 because they are both in functional domains and ANLN and SYNE1 are both expressed in cell lines with BRCA1 mutations.

While there is published evidence that ANLN is an oncogene and SYNE1 is a tumor suppressor, the function of these genes have not been investigated in a breast cancer cell line with BRCA1 mutations. As expected, decreasing the expression levels of the
oncogene ANLN caused a decrease in growth rate of BRCA1 mutated breast cancer cells compared to cells with endogenous expression of ANLN. It was also shown that decreasing the expression levels of the tumor suppressor SYNE1 increased the growth rate of BRCA1 mutation breast cancer cells in comparison to cells expressing endogenous levels of SYNE1. These results provide functional evidence that both ANLN and SYNE1 can alter the proliferation rate of human breast cancer cells. Furthermore, there is now functional evidence that ANLN and SYNE1 can regulate the growth rate of BRCA1 mutation breast cancer cells. These results provide evidence that ANLN and SYNE1 are good candidate genes that may be able to modify the risk of breast cancer in BRCA1 mutation carriers.

Next, and perhaps more importantly, we investigated whether these variants could alter the function of the affected gene. We first showed that the variant ANLN_rs3735400 affected the proper nuclear localization of Anillin. While the function of Anillin in the nucleus is unknown, previous studies have shown that the ability of Anillin to localize to the nucleus does not affect the ability of Anillin to rescue the knockdown phenotype. However, because the variant is also in the mDia2-binding domain of Anillin, which is critical for the ability of Anillin to rescue the knockdown phenotype, we investigated whether Anillin with ANLN_rs3735400 could rescue the knockdown phenotype. It was shown that the full length Anillin with the ANLN_rs3735400 variant was unable to rescue the decreased growth rate caused by the shRNA_ANLN in MCF7 cells. It has been shown that the N-terminal 115 amino acids of ANLN is critical to maintain Anillin function during cytokinesis\textsuperscript{78}. Because the variant ANLN_rs3735400 affects the N-terminal 65 amino acid, we hypothesized that this variant ANLN_(S65W) is inhibiting ANLN ability to interact with mDia2.
These results provide evidence that the variant ANLN_rs3735400 can alter the ability of Anillin to drive proliferation. While it remains unknown why Anillin with ANLN_rs3735400 is unable to restore the proliferation rate, we hypothesize that ANLN_rs3735400 is affecting proper Anillin mDia2 binding during cytokinesis. All of the functional results collected from this study provide evidence that *BRCA1* mutation carriers with a mis-functioning oncogene, Anillin, have a decreased risk of breast cancer in comparison to carriers with a fully functional Anillin oncogene.

Next, we tested the ability of variant SYNE1_rs2295190 to disrupt the localization of a truncated Nesprin-1 to the nuclear membrane. The major function of giant Nesprin-1 is to link the cytoskeleton to the nucleoskeleton by binding F-actin in the cytosol with the nuclear membrane. An alteration of Nesprin-1 nuclear membrane localization can affect the nucleoskeleton linkage, which has been shown to affect the function of both DNA repair and mismatch proteins. This disruption of DNA repair and mismatch proteins can increase genomic instability. It was shown that the variant SYNE1_rs2295190 greatly alters the ability of Nesprin-1 to bind to the nuclear membrane. All of the functional results collected from this study give evidence that *BRCA1* mutation carriers with a mutated tumor suppressor, Nesprin-1, have an increased risk of breast cancer in comparison to carriers with a fully functional tumor suppressor.

**Future directions for genetic modifiers of breast cancer risk of *BRCA1* mutation carriers**

While three of the five variants (SIPA1L2_rs1547742, MTHFSD_rs3751800, and ANKLE1_rs8100241) were not functionally studied in this project, there is evidence that these genes are important in cancer development. Because the function of MTHFSD, SIPA1L2 and ANKLE1 in vertebrates are unknown, we would propose to first analyze these genes with a proteomic approach. Once the general function and binding partners...
of these genes are known, one can start designing experiments to test whether these variants alter the function of the affected gene and if these genes are important for cancer development.

The role of these five variants in the normal population should also be investigated. While these alleles may be beneficial in lowering the risk of breast cancer in the \textit{BRCA1} population, they could have deleterious effects in the non-\textit{BRCA1} mutated population. Conversely, these common variants may lower the risk of cancer in the normal population. One would expect that variants that decrease the risk of cancer may be selected for and should be found at a high frequency in the general population. However, this is only true if the variant is not causing a more severe or higher risk state than the benefit of the decreased cancer risk.

It would also be interesting to calculate the frequency of these variants in populations with a high amount of \textit{BRCA1} mutation carriers. If these variants only decrease risk of cancer in \textit{BRCA1} mutation carriers, one would expect that these variants would be more frequent in populations with a higher frequency of \textit{BRCA1} mutation carriers.

Another important goal is to determine how different genetic modifiers affect each other and within the context of known environmental modifiers. These types of studies can be difficult to perform because population size for these studies must increase greatly with the number of variables being accounted for. However, knowledge gained from these studies would greatly improve the ability to accurately calculate risk assessment for \textit{BRCA1} mutation carriers.

More studies should also be performed to investigate how the variant ANLN\_rs3735400 is affecting the function of Anillin. Even though the variant ANLN\_rs3735400 is located in the mDia2 binding domain and a truncated Anillin lacking the mDia2 domain has a similar inability to rescue a phenotype caused by an ANLN knockdown, it is unclear
whether ANLN_rs3735400 affects the ability of Anillin to bind to mDia2. To investigate this, we propose using a protein complex immunoprecipitation pull-down assay. Furthermore, because ANLN_wt has been shown to co-localize with mDia2 at the cleave furrow during cytokinesis, we can also use immunofluorescence to investigate the localization of ANLN_(S65W) and mDia2 during cytokinesis. These experiments would provide evidence on how ANLN_rs3735400 affects the function of Anillin.

Further studies should also be performed to investigate the function of Anillin and Nesprin-1 in non-transformed breast epithelial cells. We propose to use the non-transformed breast epithelial cells MCF10a and MCF10a BRCA1-/ cells. By inducing the variants with the Crisper/Cas9 system, we could change the specific nucleotide while maintaining the endogenous expression of the gene. Results from these experiments will help clarify whether ANLN and SYNE1 are important for tumorigenesis or if they are only important in post-transformed breast cells.

Results from this study would provide insight into whether manipulation of these genes or pathways could be used as cancer therapy and/or a cancer preventive therapy. For example, if these studies provide new evidence that linkage between the cytoskeleton and nucleoskeleton influences breast cancer risk in BRCA1 mutation carriers, one could design preventative therapies to increase this linkage. This type of therapy may be used to decrease the risk of breast cancer in BRCA1 mutation carriers. Furthermore, if this linkage is important for genomic stability in all cell types, this therapy could be used to lower the risk of cancer in non-carriers as well.


Two PALB2 germline mutations found in both BRCA1+ and BRCAx familial breast cancer

Bradley Downs1 · Yeong C. Kim1 · Fengxia Xiao1 · Carrie Snyder2 · Peixian Chen1 · Elizabeth A. Fleissner3 · Dina Becirovic2 · Hongxiu Wen1 · Simon Sherman3 · Kenneth H. Cowan2 · Henry T. Lynch3 · San Ming Wang1

Abstract Partner and localizer of BRCA2 (PALB2), plays an important functional role in DNA damage repair. Recent studies indicate that germline mutations in PALB2 predispose individuals to a high risk of developing familial breast cancer. Therefore, comprehensive identification of PALB2 germline mutations is potentially important for understanding their roles in tumorigenesis and for testing their potential utility as clinical targets. Most of the previous studies of PALB2 have focused on familial breast cancer cases with normal/wild-type BRCA1 and BRCA2 (BRCAx). We hypothesize that PALB2 genetic mutations also exist in individuals with BRCA mutations (BRCAx). To test this hypothesis, PALB2 germline mutations were screened in 107 exome data sets collected from familial breast cancer families who were either BRCA1+ or BRCAx. Two novel heterozygous mutations predicted to alter the function of PALB2 were identified (c.2014G>C, p.E672Q and c.2993G>A, p.G998E). Notably, both of these mutations co-existed in BRCA1+ and BRCAx families. These studies show that mutations in PALB2 can occur independent of the status of BRCA1 mutations, and they highlight the importance to include BRCA1+ families in PALB2 mutation screens.

Keywords BRCA1 · PALB2 · Familial breast cancer · Predisposition

Abbreviations
PALB2 The partner and localizer of BRCA2
BRCA1 Breast cancer 1, early onset
BRCA2 Breast cancer 2, early onset
MAF Minor allele frequency
LOVD Leiden open variation database

Introduction Partner and localizer of BRCA2 (PALB2), plays important roles in double-stranded DNA damage repair through interaction with BRCA1, BRCA2, and RAD51 [1, 2]. Studies by Xia et al. (2007) determined that homozygous mutations in PALB2 cause Fanconi’s anemia [3]. Subsequent studies by Chen et al. (2008) and Janatova et al. (2013) found that women with a clear family history of breast cancer and carried heterozygous mutations in PALB2 had an increased risk of developing familial breast cancer [4, 5]. Further, a large-scale study by Antoniou et al. (2014), which involved 362 members of 154 BRCAx breast cancer families with PALB2 mutations, indicated that the risk carriers of heterozygous PALB2 mutations have for developing breast cancer by age 70 are as high as 35 % [6]. This risk is much greater than previously thought and is similar to that caused...
by \textit{BRCA}2 mutations, which are well known to predispose women to breast and ovarian cancer [7]. Moreover, the higher rate of \textit{PALB}2 mutations implies the high risk of developing breast cancer in the carriers. Thus, \textit{PALB}2 mutations can potentially serve as genetic markers for the clinical diagnosis, treatment, and prognosis of breast cancer. However, before this can occur, it is necessary to determine the full spectrum of \textit{PALB}2 mutations in familial breast cancer. This can be accomplished by screening a large number of families with a history of breast cancer for \textit{PALB}2 mutations [8], similar to the extensive searches for \textit{BRCA}1 and \textit{BRCA}2 mutations (https://research.nhgri.nih.gov/projects/bic/).

It remains undetermined if \textit{PALB}2 mutations are specifically present in familial breast cancer with normal/ wild-type \textit{BRCA}1 and \textit{BRCA}2 (\textit{BRCA}x) or if \textit{PALB}2 mutations universally predispose individuals to familial breast cancer regardless the mutation status of \textit{BRCA}1 and \textit{BRCA}2. This knowledge is expected to help determine if \textit{PALB}2 germline mutations can be used as specific markers to \textit{BRCA}x familial breast cancer or if \textit{PALB}2 germline mutations can be used as general markers for familial breast cancer. Most efforts made toward locating \textit{PALB}2 mutations have screened cases of \textit{BRCA}x familial breast cancer [7, 8]. Our study extends to include breast cancer family members who inherited damage mutations in \textit{BRCA}1 (\textit{BRCA}1+). We analyzed \textit{PALB}2 germline mutations in 107 cases of both \textit{BRCA}1+ and \textit{BRCA}x familial breast cancer and identified two heterozygous mutations predicted to damage the function of \textit{PALB}2 (c.2014G>C, p.E672Q, and c.2993G>A, p.G998E). Most importantly, both mutations were shared between \textit{BRCA}1+ and \textit{BRCA}x familial breast cancer families.

Materials and methods

Informed consent was obtained from all study participants. Institutional Review Boards (IRBs) at both the University of Nebraska Medical Center and Creighton University School of Medicine approved the study.

Exome sequencing and DNA mapping processes applied have been described in previous studies [9–11]. Briefly, samples that were screened for \textit{PALB}2 mutations included both \textit{BRCA}1+ and \textit{BRCA}x carriers (Table 1). Exome libraries were constructed using TruSeq Exome Enrichment Kit (Illumina, Inc., San Diego, CA). A HiSeqTM 2000 sequencer (Illumina) was used to collect exome sequences with paired-end reads (2 x 100). Sequences were then mapped to the human genome reference sequence hg19 using a sequence alignment tool, Bowtie2 [12]. Next, VarScan 2 software was used to call the variants from the mapped sequences [13]. These variants were then annotated using ANNOVAR [14], the \textit{PALB}2 reference sequence NM_024675 [RefSeq: National Center for Biotechnology Information (NCBI)], Single Nucleotide Polymorphism Database 137 (dbSNP137), the 1000 Genomes Project [15], and the Exome Variant Server 6500 (http://evs.gs.washington.edu/EVS/). From the called variants, those originating in the \textit{PALB}2 coding region were selected for further analysis. Variants that caused synonymous changes in \textit{PALB}2 were removed. From the remaining variants, those that caused damaging consequences were further predicted using both sorting intolerant from tolerant (SIFT) [16] and PolyPhen-2 [17] programs, and then validated using Sanger sequencing.

Table 1 Exome-sequenced cases

<table>
<thead>
<tr>
<th>Number of families</th>
<th>Number of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{BRCA}1+</td>
<td>26</td>
</tr>
<tr>
<td>\textit{BRCA}x</td>
<td>52</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>104</td>
</tr>
<tr>
<td>Male</td>
<td>3</td>
</tr>
<tr>
<td>Cancer status</td>
<td></td>
</tr>
<tr>
<td>Breast cancer</td>
<td>77</td>
</tr>
<tr>
<td>No breast cancer</td>
<td>30</td>
</tr>
</tbody>
</table>

Results

To identify the damage mutations in \textit{PALB}2, exome data were mined from 107 cases from 58 different families with familial breast cancer (Table 1). Of these 107 cases, 56 were from 26 families with \textit{BRCA}1+ and 51 were from 32 families with \textit{BRCA}x; 104 cases were women, and three cases were men; 77 cases were diagnosed with breast cancer, and 30 cases were unaffected family members.

Ninety variants across the entire \textit{PALB}2 genomic region were identified from the exome data. Six variants were in coding exons, of which two were synonymous, and four were nonsynonymous. Using Sanger sequencing, the two nonsynonymous mutations were confirmed as c.2014G>C, p.E672Q, and c.2993G>A, p.G998E (Table 2, Fig. 1a). Further, both SIFT and PolyPhen-2 programs predicted each mutation to damage the function of \textit{PALB}2. Each mutation is listed in dbSNP137 at minor allele frequency (MAF) of 0.024 and 0.018, respectively. Notably, neither mutation was reported in the \textit{PALB}2 study by Antoniou et al. [6], but both mutations were listed in the \textit{PALB}2 LOVD database (LOVD v.2.0 Build 36). The mutation c.2014G>C, p.E672Q, was located at the MRG15

\[ \text{Springer} \]
107

interaction domain, and the mutation of c.2993G>A, p.G998E, was located at the BRCA2, RAD51, and POLH interaction domain (Fig. 1b). The mutation of c.2014G>C, p.E672Q, was detected in three families with BRCA1+ (families 1, 2, and 3), and the same mutation was also detected in three families with BRCAx (families 4, 5, 6). Similarly, the mutation c.2993G>A, p.G998E, was detected in two families with BRCA1+ (families 2 & 3) and three families with BRCAx (families 4, 5, 6). Interestingly, the mutation of c.2014G>C, p.E672Q, detected in BRCA1+ family 2 was present in an unaffected family member, suggesting that this mutation had low penetrance in this individual. The two mutations were distributed in eight cases from three BRCA1+ families and six cases from three BRCAx families. The data also indicate that multiple mutations in PALB2 can exist in the same family. Figure 2 shows the pedigree of each BRCA1+ family. Each family contains the following different mutations in BRCA1: ins6kbEx13-ter1460 (family 1), 332-11T-Gins59-ter75 (family 2), and 300Cys-Gly (T-G) (family 3). The distribution in BRCAx families cannot be determined because only the proband DNA samples in each family were available for the study. However, further studies that extend to BRCAx families are expected to help better quantify this distribution.

**Discussion**

Germline mutations in PALB2 have been considered specific to BRCAx familial breast cancer [5, 6, 18]. However, and importantly, our analysis shows that the same germline mutations can be present in both BRCA1+ and BRCAx families. This finding suggests that PALB2 mutations can occur independently of the BRCA status. Further, it suggests that, in addition to BRCAx breast cancer families, BRCA1+ breast cancer families should be included in PALB2 screening. Notably, each mutation identified in this study is listed in the dbSNP, which could suggest that they are normal polymorphisms. However, the two mutations are rare in the normal population, as judged by their low MAF value (0.024 for c.2014G>C, p.E672Q, and 0.018 for c.2993G>A, p.G998E). Moreover, rare variants are known to enrich genetic predispositions for familial diseases [19], as demonstrated by rare variants in XRCC2, FANCC, and BLM that have been found to be genetic predispositions for familial breast cancer [20–24]. Further, six of the 50 PALB2 damage mutations identified in the study by Antoniou et al. (2014) were also rare variants [6]. Our study suggests that an even greater

**Table 2** Two mutations identified in PALB2

<table>
<thead>
<tr>
<th>Base change</th>
<th>Amino acid change</th>
<th>Position (hg19)</th>
<th>BRCA status</th>
<th>dbSNP138</th>
<th>MAF</th>
<th>LOVD</th>
<th>SIFT Score</th>
<th>Prediction Score</th>
<th>PolyPhen-2 Score</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>c.2014G&gt;C</td>
<td>p.E672Q</td>
<td>chr16:23641461</td>
<td>+</td>
<td>rs45532440</td>
<td>0.024</td>
<td>10,092</td>
<td>1</td>
<td>Damage</td>
<td>0.952</td>
<td>Damage</td>
</tr>
<tr>
<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></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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>c.2993G&gt;A</td>
<td>p.G998E</td>
<td>chr16:23634293</td>
<td>+</td>
<td>rs45551636</td>
<td>0.018</td>
<td>10,135</td>
<td>1</td>
<td>Damage</td>
<td>0.996</td>
<td>Damage</td>
</tr>
<tr>
<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></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<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></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1 Mutations in PALB2. a Sanger sequencing validation showing the two mutations was heterozygous mutations in PALB2. b Location of the two mutations of c.2014G>C p.E672Q and c.2993G>A p.G998E in PALB2.
Fig. 2 Pedigree of the three BRCA1+ familial breast cancer families with the two mutations in PALB2. BRCA1 mutations in the three families are family 1 (ins6kbEx13-ter1460), family 2 (332-11T-Gins59-ter75), and family 3 [300Cys-Gly (T-G)]. Exome data from eight members in family 1, two members in family 2, and two members in family 3 were used in the study.
number of PALB2 mutations can exist in familial breast cancer, particularly in families with BRCA1+. Additional screens in more BRCA1+ breast cancer families are needed to confirm this observation as well as to determine if screening for PALB2 is needed regardless of the status of BRCA1, BRCA2, and other genes that predispose individuals to breast cancer. Such a large-scale screen would allow for the entire spectrum of PALB2 mutations to be mapped in order to reveal their roles in familial breast cancer.

Exome data used in this analysis were also used in our previous studies. As such, a logical question arises; why were these mutations in PALB2 not identified in the previous studies? Our analysis of the mapping data shows that most of the PALB2 variants called in current study had lower sequence coverage, which is a known phenomenon for exome sequence data [25]. Similarly, these variants were under the threshold for variant call conditions used in our previous studies. Thus, focusing on PALB2-mapped sequences, lowering the cut-off values for variant calls, and using Sanger sequencing for validation allowed us to identify missed variants. Notably, an extensive amount of exome data has been generated from breast cancer genomic studies [26, 27]. Our study shows that targeted mining of existing exome data plus Sanger validation is a powerful approach to identify the mutations in specific genes. In summary, our study shows that mutations in PALB2 can occur independent of the status of BRCA1 mutations.

Acknowledgments

The authors acknowledge the Breast Cancer Family Registry (BCFR), developed and maintained by the University of Nebraska Medical Center (UNMC) Biomedical Informatics Core at the Fred & Pamela Buffett Cancer Center. The study was supported by a pilot grant from the UNMC Fred & Pamela Buffett Cancer Center, an NIH Grant 1R21CA183008 (SMW), an overseas study scholarship (sponsored by China Scholarship Council (F(X)), the Charles F. and Mary C. Heider Chair in Cancer Research (JL), and revenue from Nebraska cigarette taxes awarded to Creighton University by the Nebraska Department of Health and Human Services. The funding bodies play no roles in design, collection, analysis, and interpretation of data. The authors kindly thank Melody A. Montgomery at the University of Nebraska Medical Center (UNMC) Research Editorial Office for the professional editing of this manuscript.

Conflict of interests

The authors declare that they have no competing interests.

References


