8-16-2013 12:00 AM

Detection, prioritization and analysis of variants of unknown significance in familial breast cancer genes

Eddie A. Dovigi
The University of Western Ontario

Supervisor
Dr. Peter Rogan
The University of Western Ontario

Graduate Program in Biochemistry

A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

DETECTION, PRIORITIZATION AND ANALYSIS OF VARIANTS OF UNKNOWN SIGNIFICANCE IN FAMILIAL BREAST CANCER GENES

Thesis format: Monograph

by

Edwin Dovigi

Graduate Program in Biochemistry

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

Currently, Molecular Diagnostics Laboratories in Ontario sequence coding and adjacent intronic regions in *BRCA1* and *BRCA2* in patients with a family history of breast cancer. At LHSC it is estimated that ~15% of patients have *BRCA1* or *BRCA2* variants of clinical significance, and ~15-20% patients have variants of unknown clinical significance (VUS), while the remaining patients have variants of no clinical significance, making patient prognosis difficult to ascertain. To elucidate VUS and improve deleterious variant detection, my study has three aims, 1) assess the effects of VUS on splicing using bioinformatics and transfection assays; 2) investigate the limitations of *BRCA1* and *BRCA2* routine sequencing in deleterious variant detection and expand deleterious variant detection by sequencing seven breast cancer associated genes in 21 familial breast cancer patients and 3) prioritize detected variants *in silico* for effects on: splicing, transcription factor binding, mRNA structure, miRNA binding and amino acids.

Keywords:

Familial breast cancer, variants of unknown significance, splicing, *BRCA1*, *BRCA2*, molecular diagnostics, next-generation sequencing, variant prioritization

Acknowledgements

I would like to thank my supervisor Dr. Peter Rogan as well as my advisory committee, Drs. Jim Koropatnick, Joan Knoll and Peter Ainsworth, as well as the members of the Rogan/Knoll lab, especially Eliseos Mucaki who designed the target-enrichment arrays and who I worked with during DNA library preparation and sequencing. I would also like to thank the Molecular Diagnostics Laboratory at the London Health Sciences Centre (senior technologist: Alan Stuart) for providing DNA samples.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ASSA: Automated Splice Site Analysis
ASSEDA: Automate Splice Site and Exon Definition Analysis
ATM: ataxia telangiectasia mutated
BAM: Binary sequence alignment map
BCLAF: Bcl2 associated transcription factor 1
BIC: Breast Cancer Information Core database
BCLC: Breast Cancer Linkage Consortium
BLAST: Basic local alignment serach tool
BLAT: BLAST-like alignment tool
BRCA1: early onset breast cancer gene 1
BRCA2: early onset breast cancer gene 2
CDH1: cadherin 1
CHEK2: checkpoint kinase 2
dsDNA: Double stranded DNA
EEO: Electroendosmosis
ENCODE: Encyclopedia of DNA Elements
ENIGMA: Evidence-based Network for the Interpretation of Germline Mutant Alleles
ESE: Exonic splicing enhancer
ESS: Exonic splicing silencer
GAIIe: Genome Analyzer IIe
GATK: Genome Analysis Toolkit
gDNA: Genomic DNA
HBOC: Hereditary breast ovarian cancer
HGVS: Human Genome Variation Society
HMEC: Human Mammary Epithelial Cell line
hnRNPA/B: Heteronuclear ribonuclear protein A/B
IDT: Integrated DNA Technologies
IGV: Integrative genome viewer
IM7: 1-Methyl-7-Nitro-Isatoic Anhydride
INSERM: Institute National de la Santé et de la Recherche Médicale
ISE: Intronic splicing enhancer
LHSC: London Health Sciences Centre
LOVD: Leiden Open Variation Database
MRN: A protein complex consisting of Mre11–Rad50–Nbs1
NCBI: National Center for Biotechnology Information
PALB2: Partner and localizer of BRCA2
RBPBS: RNA-binding protein binding site
RBPDB: RNA-binding protein database
RefSeq: Referense Sequence
Ri: Information content
RR: Relative Risk
SAM: Sequence alignment map
SF2/ASF: Splicing factor 2/Alternative splicing factor
SHAPE: Selective 2'-hydroxyl acylation analyzed by primer extension

SHARCNET: Shared Hierarchical Academic Research Computing Network
SNRPA: small nuclear ribonuclear protein A
SR: Serine-arginine rich
TERT: Telomerase reverse transcriptase
TFBS: Transcription factor binding site
TP53: Tumor protein 53
UCSC: University of California Santa Cruz
VUS: variant of unknown significance
ZOOPS: Zero or one bipartite site occurrence per sequence
ΔRi: Change in information content

Chapter 1 Introduction

1.1 Hereditary breast cancer and high penetrant genes *BRCA1* and *BRCA2*

The GLOBOCAN project, which is part of the International Agency for Cancer Research at the World Health Organization, estimates that over 1.3 million new cases of breast cancer were diagnosed worldwide in 2008, as well as 458,000 breast cancer related deaths, making breast cancer the most prevalent malignant cancer in women (Ferlay J. *et al.* 2008) [1]. The American Cancer Society describes several risk factors associated with the development of breast cancer including: gender, environmental factors, age, family history of breast cancer and ethnicity. Many research studies have been performed to determine the effects of these risk factors on breast cancer etiology. Current estimates indicate that 5-10% of breast cancer cases are due to hereditary factors, with the remaining cases being due to sporadic breast cancer as a result of risk factors listed previously [2].

In searching for breast cancer associated genes, Miki *et al.* (1994) conducted a series of mapping experiments on 65 expressed sequences within a 600kb region on chromosome 17q which was previously linked to early onset breast cancer [3,4,5]. After extensive experimentation, Miki *et al.* (1994) identified the ~100kb breast cancer 1 early onset gene (*BRCA1*). The following year, Wooster *et al.* (1995) identified a second breast cancer associated gene, the breast cancer 2 early onset gene (*BRCA2*) on chromosome 13q by using genetic recombination information from a cohort of families with early-onset breast

cancer [6]. Together, these two genes make up the majority of high penetrant breast cancer associated genes [7]. Numerous linkage and mutation analysis studies have shown that harboring a deleterious mutation in *BRCA1* or *BRCA2* can significantly increase the risk of acquiring breast cancer [7,8].

1.2 Current molecular diagnostic sequencing and breast cancer variant databases

The Molecular Diagnostics Laboratory at the London Health Sciences Centre (LHSC) sequences coding and adjacent intronic regions of *BRCA1* and *BRCA2* (-20 to +10bp around exons) in patients with a strong family history of breast cancer as defined on page two of the, "Ontario Cancer Genetic Testing Program" requisition form (http://www.lhsc.on.ca/lab/molegen/brca_req.pdf). Coding and adjacent intronic sequencing of *BRCA1* and *BRCA2* revealed 158 novel variants, with patients having between 0 to 10 or more variants detected. Examining these variants in the Breast Cancer Information Core (BIC) database (an online database compiling mutations and polymorphisms in breast cancer associated genes with the aim of facilitating VUS characterization) [9], revealed that 107 are listed as of unknown clinical significance or are not reported, 40 are listed as of no clinical significance, and 11 are listed as clinically significant. Currently, at LHSC it is estimated that ~15% of patients have *BRCA1* or *BRCA2* variants of clinical significance, and ~15-20% patients have variants of unknown clinical significance (VUS), while the remaining patients have variants of no unknown clinical significance. To aid the clinical understanding of VUS, variant data should be converted to a standardized, consistent nomenclature for submission to a database

accessible to many researchers and clinicians who can perform functional and segregation analyses to ascertain the clinical significance of VUS.

Genetic variants can be expressed in various nomenclatures. Variant nomenclatures are standardized descriptions of sequence variants that include: the coordinate of the variant (genomic, transcript etc.), the reference nucleotide and the sequence variation. Different variant cataloging databases can use different nomenclatures to describe sequence variants. The Human Genome Variation Society (HGVS) [10] collects genomic variant information and associated clinical findings, and aims to foster the discovery and characterization of genomic variants and their phenotypic association. HGVS and BIC use different variant nomenclatures. In BIC nomenclature (http://research.nhgri.nih.gov/projects/bic/index.shtml), variant position numbering begins at the first transcribed nucleotide, whereas in HGVS nomenclature, variant position numbering begins at the first translated nucleotide (i.e. the "A" of the ATG start codon is position +1). All nomenclatures require a DNA reference sequence from which to call genetic variants. The National Center for Biotechnology Information curates a Reference Sequence Database (*RefSeq*) that houses human genomic, mRNA transcript and protein sequence data (as well as data for other organisms) that is used as a reference sequence in a variety of nomenclatures including HGVS and BIC.

In the case of breast cancer related genes, after variants are detected and described in a particular nomenclature, researchers consult databases such as BIC to see if the variant has been previously reported and if its phenotypic effects are known. In a population

based study conducted by Borg *et al.* (2010) on young women with contralateral (n=705) or unilateral (n=1398) breast cancer, 470 novel sequence variants were detected, 113 of which are deleterious mutations, 373 of which are VUS. The Evidence-based Network for the Interpretation of Germline Mutant Alleles consortium (ENIGMA; http://enigmaconsortium.org/)[11] was created for researchers to deposit *BRCA1* and *BRCA2* VUS's to facilitate collaboration and variant classification.

BRCA1 and *BRCA2* VUS detected by the Molecular Diagnostics Laboratory at LHSC were catalogued in BIC and HGVS nomenclatures, as well as a clinical format different from HGVS that describes the amino acid position and change as well as the nucleotide change. However, before variants can be submitted to the ENIGMA consortium, they need to be described in a consistent, standardized nomenclature, such as HGVS. To facilitate VUS classification, VUS need to be submitted to the ENIGMA consortium to make data accessible to researchers that use functional and segregation analyses to investigate pathological effects of genetic variants of unknown significance. However, unclassified VUS in coding and adjacent intronic regions (-20bp to +10bp around exons) of *BRCA1* and *BRCA2* may not account for all disease-causing variants in these two genes.

1.3 mRNA splicing and variants in non-coding regions affecting disease etiology

It has been estimated that 15-50% of genetic diseases are due to mRNA splicing abnormalities depending on the gene [12]. mRNA splicing is a critical step in human gene expression that involves many proteins working in concert to identify the exons to

include in mature mRNA transcript and removing intervening introns. Canonical splicing sequence motifs located at intron-exon junctions, called the splicing acceptor site (3' end of intron) and donor site (5' end of intron), as well as the branch point, which is necessary for mRNA lariat formation, are required for accurate splicing. However in humans, these sequence elements are not sufficient to recruit the spliceosome to all exons. Computational analysis of human mRNA transcripts conducted by Lim and Burge (2001) concluded that the information encoded within the canonical acceptor, donor and branch point splice sites is not sufficient for the spliceosome to recognize short introns for splicing in the human genome [13]. Yet, the splicing machinery is very accurate in identifying exons separated by large introns containing many "decoy" canonical splice sites, termed cryptic sites, which resemble functional acceptor and donor sites located at intron-exon boundaries, but are not used during splicing [14]. Additional *cis* elements called splicing regulatory elements can enhance or silence exon recognition by the spliceosome, thus influencing the final processed transcript [14]. Exonic splicing enhancer (ESE) motifs are sequences located in exons that recruit splicing regulatory proteins and serve as a scaffold for protein-protein interactions with components of the spliceosome, facilitating spliceosome assembly and promoting exon recognition [14]. These splicing enhancer proteins are from the serine-arginine (SR) rich family, and include: SRp40, SRp55, SC35, and splicing factor 2/alternative splicing factor (SF2/ASF), each of which has an N-terminal RNA-recognition domain and a C-terminal serine-arginine (SR) rich domain, which is responsible for protein-protein interactions with SR domains of other proteins, including core spliceosome components [15]. Each splicing enhancer binds unique mRNA sequences, for instance the consensus SRp40 binding sequence contains

an inverted repeat (GAGCAGTCGGCTC) with the potential to form a step-loop structure, whereas high affinity ASF/SF2 and SC35 binding sites do not display any predicted sequence structure [16]. The SF2/ASF splicing enhancer protein has been shown to interact with U1 and U2AF snRNPs, and facilitate acceptor and donor splice site recognition by the spliceosome, thus promoting exon recognition [17]. The SRp40 protein when phosphorylated, has been shown to localize to ESE elements adjacent to acceptor and donor sites, and facilitate acceptor and donor site recognition [18].

Alternatively, exonic splicing silencer (ESS) motifs recruit proteins that antagonize splicing, such as the heteronuclear ribonucleotide proteins A/B (hnRNPA/B), which have been shown to interfere with both core spliceosome binding to splice sites, as well as splicing enhancer protein binding [19]. Heteronuclear ribonucleotide proteins H and F have been shown to bind intronic splicing enhancer (ISE) motifs and promote intron splicing [20]. However, when these same proteins are recruited to an exon, splicing is inhibited [20]. Therefore splicing regulatory elements exist in both introns and exons and have been shown to modulate both alternative and constitutive splicing by recruiting splicing enhancer or silencer proteins to affect spliceosome assembly at canonical acceptor and donor splice sites [21]. Due to the significant role of *cis* splicing elements in mRNA splicing, genetic variants in the canonical spliceosome binding acceptor and donor sites, as well as the splicing branch point and splicing regulatory sites (such as: SRp40, ASF/SF2, SC35, SRp55 and hnRNPH) can result in disease [22,23].

Some variants, called "inactivating variants" can completely abolish a splice site, and result in no spliceosome/splicing regulatory protein binding. Inactivating a natural donor or acceptor site results in the failure of the spliceosome to bind, and can result in exon skipping, or exon elongation/truncation if a cryptic splice site is present in close proximity to the inactivated site. Variants can also weaken splice sites without fully inactivating them, and are called, "leaky variants", which result in a mixture of wild type and aberrant mRNA transcripts. Finally, variants can strengthen a cryptic splice site in an exon or intron, which can be used in place of a natural splice site depending on its proximity to the natural site and if it the spliceosome or splicing regulatory protein binds it with greater affinity than the natural site. Cryptic splice sites are found throughout genes and can result in aberrant splicing if activated by a mutation.

Disease causing DNA variants have been demonstrated to occur outside of exonic and adjacent intronic regions [24,25]. A study by Anczukow *et al.* (2012) found a *BRCA2* T>G mutation in nine families with breast cancer that is located 594 nucleotides into intron 12 and strengthens a cryptic donor site, resulting in the creation of a 95-nucleotide pseudo-exon that is included in the final mRNA transcript, which alters the reading frame [24]. Accurate splicing in this case was restored through antisense oligonucleotides suppressing cryptic exon recognition. This variant would have gone undetected in routine sequencing of coding and adjacent intronic regions of *BRCA1* and *BRCA2*. Other regions not sequenced during routine sequencing that may be influential in disease etiology include adjacent intergenic regions, where variants within transcription factor binding sites could result in aberrant transcriptional regulation. A study conducted by

Horn *et al.* (2013) [25] revealed that a mutation linked to familial melanoma, which is found in the promoter region of a gene encoding a telomerase subunit (*TERT*), created a transcription factor-binding site resulting in 2.2 fold increased gene expression in various cell lines through a luciferase reporter gene assay. Therefore sequencing non-coding regions in *BRCA1* and *BRCA2* is necessary to increase deleterious variant detection in patients with familial breast cancer.

1.4 Breast cancer as a polygenic disease

Sequencing *BRCA1* and *BRCA2* in their entirety alone, may not be sufficient to detect all deleterious variants involved in familial breast cancer etiology. Of familial breast cancer cases, ~13% of individuals with familial breast cancer carry identified pathogenic *BRCA1* or *BRCA2* variants,[27,28,7,29] suggesting that mutations in other genes are influential in breast cancer etiology. A polygenic model of breast cancer as a complex disease may explain why pathogenic high-penetrant *BRCA1* and *BRCA2* mutations comprise only approximately 13% of all familial breast cancer cases [7]. Additionally, low to moderate penetrant alleles with a range of relative risks have been identified in individuals with breast cancer through association studies, as linkage studies have been successful mostly in cases of families with rarer high-penetrant monogenic mutations [30]. High penetrant deleterious variants in tumor suppressor encoding gene *TP53* have also been identified in patients with hereditary breast cancer [31].

Thompson *et al.* (2005) estimated that mutations in the ataxia telangiectasia mutated gene (*ATM*) can confer a breast cancer relative risk of approximately 2.23 (95% C.I. 1.16-4.28)

[32]. Several additional studies have linked low to moderate penetrant genes to breast cancer etiology, including: *CHEK2* (odds ratio of breast cancer when harboring a deleterious *CHEK2* variant is calculated to be 5.18 [33]), *CDH1* (estimated 39-60% elevated risk of lobular breast cancer in individuals with a germline mutation [34,35]) and *PALB2* (partner and localizer of *BRCA2;* estimated 2.3 fold elevated risk of breast cancer in germline mutation carriers who do not harbor a *BRCA1* or *BRCA2* variant) [36,37,29]. Excluding *CDH1*, all of these gene products are involved in the double strand DNA (dsDNA) break repair pathway. Figure 1 depicts breast cancer associated gene product involvement in dsDNA break repair. The serine-threonine kinase encoded by *ATM* undergoes autophosphorylation in response to DNA double stranded breaks, and goes on to mediate downstream phosphorylation events involved in double strand break repair, including the phosphorylation of cell cycle checkpoint kinase encoded by *CHEK2* (Chk2). Phosphorylated Chk2 phosphorylates p53 (*TP53*), which mediates cell-cycle arrest. BRCA1, BRCA2 and PALB2 proteins are part of a multi-subunit complex that localizes to DNA double stranded breaks resulting in strand invasion and homologous mediated repair. Therefore, deleterious mutations in the genes that encode the proteins involved in double stranded DNA break repair can abrogate this process, and result in genome instability and cancer. In addition to genes involved in the DNA double-strand break repair pathway, the E-cadherin encoding gene (*CDH1*) which is responsible for cell-cell adhesion, and has been shown to be associated with lobular breast cancer when mutated. It has also been shown that loss of E-cadherin expression eliminates adherens junction formation, important in cell mobility and proliferation, and is associated with the transition from adenoma to carcinoma, and metastatic capacity [38]. Restoring E-cadherin

expression in cancer cell lines has been shown to restore adherens junction formation and exert tumor suppressive effects including reduced proliferation and motility [39].

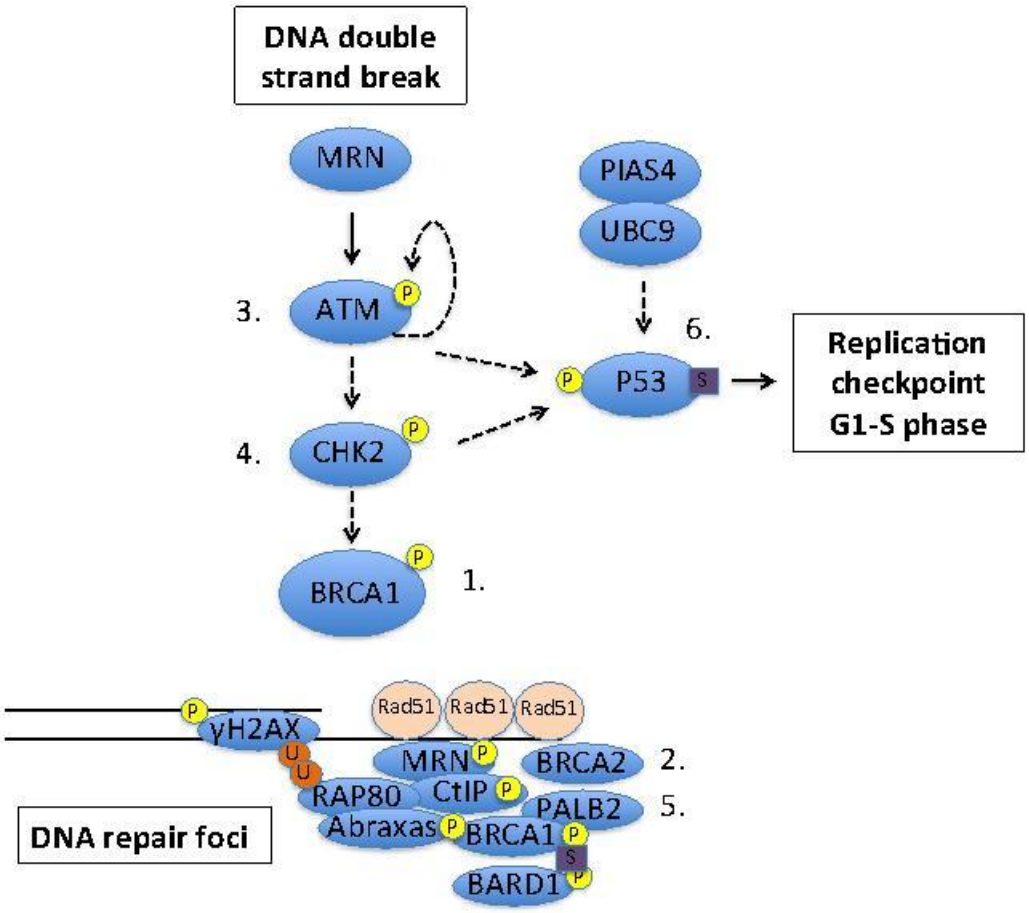

Figure 1. (Modified from Shuen A.Y. and Foulkes W.D. 2011[29]) Double stranded DNA break repair pathway. High penetrant breast cancer associated gene products (1 and 2), as well as low to moderate penetrant breast cancer associated gene products mentioned previously (3-6) are numbered above. Dashed arrows indicate phosphorylation or sumoylation events. The genes encoding the numbered gene products above each account for a percentage of familial breast cancer when mutated, *BRCA1* (5-10%), *BRCA2* (5-10%), *CHEK2* (2%), *ATM* (2%), *PALB2* (0.4%), *CDH1* (0.1%), TP53 (0.1%) [7]. This figure was modified from Shuen and Foulkes (2011) [29] to include DNA repair foci and initial steps in DNA double strand break repair pathway. Intermediate steps in DNA repair foci assembly have been omitted in this figure.

1.5 DNA targeted enrichment for downstream variant detection and DNA sequencing on the Illumina Genome Analyzer

Various methods have been employed to capture specific DNA regions from a genome for sequencing and variant detection, including microarray-based *in situ* hybridization [40,41] and solution-based hybridization [42]. There are several advantages of using solution-based methods for DNA enrichment. One is the kinetic favorability of using excess probe to drive hybridization to sheared DNA compared to *in situ* hybridization, which is limited by the number of probes on the array, and thus requires a larger amount of precious DNA sample to achieve efficient hybridization. Therefore, solution-based hybridization can be more favorable because it reduces the amount of DNA sample required to achieve efficient enrichment. In addition, synthesizing probes in excess through *in vitro* enzymatic reactions enables researchers to perform multiple hybridizations compared to on-array hybridization, in which the efficiency and consistency of multiple hybridizations is limited by repeated microarray stripping. For these reasons, solution-based capture is a preferential means of DNA targeted enrichment for downstream sequencing. A workflow for solution-based library DNA enrichment by Gnirke *et al.* (2009) [42] details the enzymatic steps involved in genomic DNA library and RNA probe preparation (Figure 2).

After DNA enrichment, samples are paired-end sequenced on the Illumina Genome Analyzer II*e*. Paired-end sequencing is when 36-150 nucleotides on both ends of sheared DNA fragments are sequenced, enabling for more stringent sequence alignment to the reference genome while increasing the number of times each individual nucleotide is sequenced, termed base coverage, thus improving variant detection accuracy. In addition to sequence alignment and base coverage, another aspect of variant detection accuracy is base quality. The Illumina platform in this study uses a phred base quality score metric, which uses a logarithmic equation to estimate the probability of an incorrect base call during sequencing. For example, a phred base quality score of 30 is associated with a probability of an incorrect base call of 1 in 1000, [43] and can be used as a threshold for accurate base quality on the Illumina platform as described in their technical notes (http://res.illumina.com/documents/products/technotes/technote_q-scores.pdf). On the Illumina platform, fluorescently labeled nucleotides containing a reversible blocking group are incorporated onto the sample DNA fragment one at a time, with the blocking group serving as a polymerization terminator. After incorporation, individual fluorophores corresponding to the A/C/G/T incorporated base are excited using a laser and imaged. Single nucleotide incorporation prevents base calling errors resulting from difficult to sequence regions, such as GC rich and repetitive regions, improving base calling.

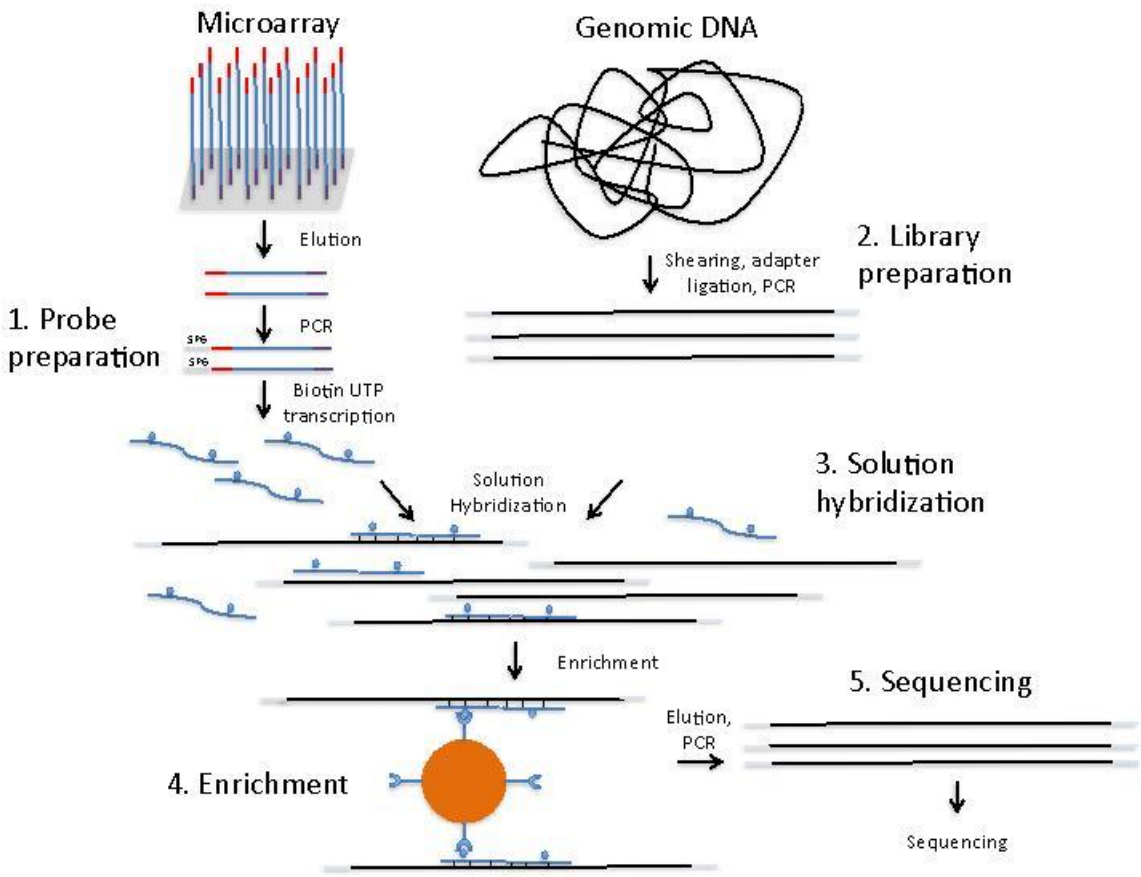

Figure 2. (Gnirke *et al.* 2009) [42] Library preparation workflow for Illumina sequencing. RNA probes (wavy lines) are transcribed in excess *in vitro*, incorporating biotinylated UTP, from PCR amplified oligonucleotides cleaved from a microarray. Probes are hybridized in solution to sheared, adapter ligated genomic DNA and subsequently eluted using streptavidin-coated magnetic beads. Eluted DNA is PCR amplified and sequenced on an Illumina Genome Analyzer instrument.

Detected genomic variants can impact a variety of processes during the course of gene expression that ultimately affect gene product function. These include the disruption of: transcription factor binding sites (TFBS), mRNA splicing, miRNA binding, 5' and 3' UTR mRNA structure and protein binding sites, amino acids etc. One way to examine the putative effects of genetic variants on protein binding sites *in silico* is information theory.

1.6 Prioritization of detected variants on disease etiology using information theory

Information theory indicates the number of choices needed to select a particular variable from two or more alternatives [44]. This can be applied to genetic information. For nucleotides, two questions need to be asked to determine whether a nucleotide is A, G, C or T (i.e. Is it a purine? Is it adenine?). Expressing the number of binary choices needed to determine nucleotide identity (i.e. "information") in "bits" results in 2 bits of information needed to specify a nucleic acid base.

Information theory has been applied to examining protein binding site sequence conservation, which is a more useful tool for investigating protein binding sites than consensus sequences as proteins can still bind non-consensus sequences *in vivo* [44]. The information content (R_i) of a nucleotide at a defined position in a protein binding sequence can be determined from a set of aligned, experimentally validated protein binding sites. The information content of a nucleotide at a particular position in an aligned set of protein binding sequences is inversely related to the amount of base uncertainty at the particular position, which is associated with the frequency of each base

found at that position in the set of aligned sequences. A logarithmic equation is used to calculate the base uncertainty, which can then be converted to information content [44]. The information content of an entire protein binding site can be calculated by summing the information contents of the individual positions comprising that binding site [44]. With regard to disease etiology, the information content of a mutated protein binding sequence can be compared to the information content of the set of aligned wild-type sequences to determine the likelihood of a particular mutation affecting protein binding. The resulting change in information content (ΔRi) of a protein binding sequence is related to fold change of protein binding affinity according to $2^{\text{delta } Ri}$. [45] Therefore a 1 bit difference in information content of a protein binding site with and without a mutation represents a two-fold change in protein binding affinity. Information theory has been used to prioritize variants found in protein binding sites, such as the acceptor and donor splice sites, for further wet-lab studies. Mucaki *et al.* (2011) compared information theory predictions of *BRCA1/2* variants in non-canonical splice sites with published mRNA expression data and found a total of 57 of 62 variants predicted to affect splicing using information theory were concordant with mRNA expression patterns [46]. A high success rate has been achieved in predicting putative splicing effects of variants in other genes as well, including *MLH1* and *MLH2* (16/17 affecting canonical splice sites validated in Tournier *et al.* study) [47].

1.7 Prioritized variant validation studies

After variants are detected, prioritized *in silico* for effects putative effects on WT gene expression using various bioinformatics tools discussed in Chapter 2.6, and Sanger

sequencing validated, functional assays need to be conducted to validate effects predicted using bioinformatics. Variants predicted to alter transcription factor binding sites for example can be validated through reporter gene assays. The familial TFBS mutation in *TERT* described in Chapter 1.3 discovered by Horn *et al.* (2013) [25] through targeted enrichment-based sequencing was examined using luciferase reporter gene constructs, one containing the promoter encoding the mutated TFBS, the other with the wild-type binding site, which is fused to a luciferase gene and transfected into a mammalian cell line. Luciferase intensity of mutant condition is then measured using a luminometer and compared to wild-type control.

Variants predicted to alter splicing can be examined through RNA expression analysis, either using tissue-specific patient RNA, or when that is unavailable, an *ex vivo* mini-gene assay. In a mini-gene assay, a putative splicing mutation is cloned into a plasmid harboring a "mini-gene", a human gene fragment under CMV promoter control containing 2-3 exons and intervening introns, as well as two restriction enzyme sites where the splicing variant of interest and adjacent regions is cloned in. The recombinant plasmid is then transiently transfected into a mammalian cell line along with the wild-type control, mRNA is then extracted, reverse transcribed and PCR amplified using primers specific to the mini-gene cDNA to examine the effects of the variant on mRNA splicing through gel electrophoresis and later sequencing. Two separate mini-gene plasmids containing ampicillin and neomycin resistance genes, pCAS2 and pcDNA-Dup, have been described by Tournier *et al.* (2008) [47] and are designed to examine effects of variants affecting canonical acceptor and donor splice sites, and splicing regulatory

elements respectively. The cloning site in pCAS2 is within the intron separating *SERPING1/CINH* exons 1-2 of the mini-gene, where the sample exon containing the putative splicing-affecting variant of interest as well as ~150bp of flanking intron is ligated in. The cloning site of pcDNA-Dup is within a 47bp cassette exon, which is located within an intron separating two exons derived from human β-globin gene, where a ~30bp region containing the mutated regulatory splice site is ligated in. Due to the weak 3' splice site of the cassette exon, the spliceosome will not recognize the cassette exon without cloning in a functional exonic splicing enhancer (ESE) element, and will therefore be skipped [47]. A schematic of pcDNA-Dup and pCAS2 mini-gene vectors is shown in Figure 3.

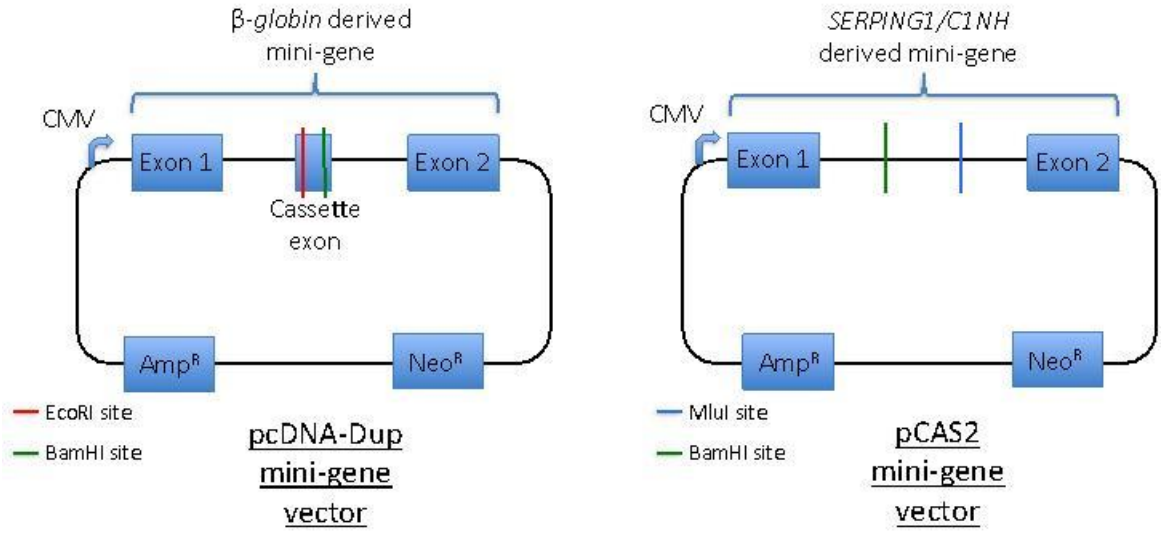

Figure 3. (Adapted from Tournier *et al.* 2008) [47] A schematic of pcDNA-Dup and pCAS2 mini-gene vectors. Both vectors harbor ampicillin and neomycin resistance genes. Colored vertical lines indicate restriction enzyme sites used during recombinant plasmid construction. In pcDNA-Dup schematic, EcoRI restriction site is on the 5' end of the cassette exon (appears on left in image) and BamHI restriction site is on 3' end of cassette exon. For pCAS2 schematic, BamHI restriction site appears at 5' end of intron (left-most vertical line) and MluI restriction site appears at 3' end of intron (right-most vertical line). A CMV promoter regulates expression of mini-genes.

Variants predicted to alter mRNA structure can be investigated through SHAPE analysis, which stands for "Selective 2'-hydroxyl acylation analyzed by primer extension" [48]. SHAPE examines mRNA flexibility at one-base resolution by chemically modifying 2'OH groups on ribonucleotides with a reagent such as 1-methyl-7-nitroisatoic anhydride (1M7). Flexible, single stranded ribonucleotides have greater 2'OH nucleophilic reactivity for this reagent and are preferentially modified compared to nucleotides in structure. Chemically modified, 5' labeled sample mRNA is then digested using RNase R, which is capable of digesting highly structured RNA in the 3'>5' direction, and recognizes chemically modified 2'OH groups as stops. Digested mRNA encoding a mutation in one condition and the wild-type nucleotide in another condition, can be resolved using denaturing PAGE. Position-specificity is conferred by running conditions alongside a "sequencing ladder", which is generated by covalent modification of single stranded guanosines in 5' labeled mRNA transcript using kethoxal reagent. The transcript is subsequently digested using RNase R, which stops at modified guanosines.

1.8 Project Aims

Currently, Clinical Molecular Genetics Laboratories in Ontario sequence coding and adjacent intronic regions (-20bp and +10bp around exons) of *BRCA1* and *BRCA2* in patients with a strong family history of breast cancer. However, as mentioned in section 1.3, variants can occur within introns and other non-coding regions (such as transcription factor binding sites upstream of genes) and cause disease. Such variants would go undetected by current routine diagnostic screening. Approximately 10-20% of familial

breast cancer is due to pathogenic variants in *BRCA1* and *BRCA2* [7]. Variants in other low to moderate penetrant genes that are not routinely sequenced have also been shown to contribute to hereditary breast cancer etiology [7,29].

Based on the potential limitations of routine diagnostic sequencing of breast cancer patient DNA in identifying deleterious variants, there are three aims to this study. The first aim is to 1a) convert all *BRCA1* and *BRCA2* variants previously detected at LHSC to a standardized nomenclature and submit VUS to the ENIGMA consortium; and 1b) assess the effects of variants on splicing using bioinformatics tools and validate predictions using a mini-gene assay as described in Chapter 1.7. The second aim is to 2a) investigate the limitations routine sequencing of coding and adjacent intronic regions of *BRCA1* and *BRCA2* in deleterious variant detection. This will be achieved by comparing the number of deleterious *BRCA1* and *BRCA2* variants detected by LHSC to linkage and mutation studies using a power analysis. And 2b) improve variant detection by sequencing coding, non-coding and adjacent intergenic regions ($\pm$ 10kbp of gene) of seven breast cancer associated genes, including: *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, *PALB2* and *TP53*. The third aim of this study is to prioritize variants detected in sequencing study based on potential disruption of important mechanisms required for expression of these genes. We will predict variants that affect: splicing, TFBS, 5' and 3' UTR mRNA structure and RNA binding protein binding sites (RBPBS), amino acids, and miRNA binding sites using *in silico* bioinformatics tools to reduce the number of clinically non-significant variants and select only those potentially influential in breast cancer etiology.

It is hypothesized that 1) examining only coding and adjacent intronic regions (-20 to +10 bp around exons) of *BRCA1* and *BRCA2* limits deleterious variant detection in patients with a family history of breast cancer. Multiple gene sequencing will reveal variants potentially influential in breast cancer etiology, and 2) variants of unknown significance predicted to lead to splicing aberrations *in silico* will be validated through *ex vivo* transient transfection studies.

1.9 Roles and responsibilities

Coding variants of unknown significance in the LHSC breast cancer patient cohort were identified by the Molecular Diagnostics team (senior technologist: Alan Stuart). Variants were converted to HGVS nomenclature by myself in collaboration with Alan Stuart, Dr. Peter Ainsworth and Dr. Peter Rogan and published to the ENIGMA consortium by Dr. Peter Rogan and Dr. Peter Ainsworth. VUS were prioritized *in silico* for effects on splicing and analyzed *ex vivo* through transient transfection and RT-PCR by myself.

The capture array used in this study to target seven breast cancer associated genes was designed by lab technician Eliseos Mucaki, and was synthesized by both Eliseos Mucaki and myself. Purified patient genomic DNA was obtained from the Molecular Diagnostics Laboratory. DNA library preparations and DNA sequencing were performed by both Eliseos Mucaki and myself for the first two sequencing runs, and by Eliseos Mucaki for the final run. Variants were called and analyzed *in silico* for effects on: splicing using software developed by computer science student Ben Shirley by myself and Ben Shirley, transcription factor binding sites using a program called "Mutation Analysis" developed

by computer science student Coby Viner and prioritized by myself and Coby Viner, and 5’ and 3’ UTR structure using SNPfold by Eliseos Mucaki, and RNA-binding protein binding sites by Eliseos Mucaki, miRNA binding and amino acid sequence using publically available software as described in Chapter 2.6 by myself.

Chapter 2. Analysis of *BRCA1* and *BRCA2* variants of unknown significance on mRNA splicing and breast cancer patient DNA sequencing methodology

2.1 *BRCA1* and *BRCA2* variant nomenclature conversion

Previously, all intronic *BRCA1* and *BRCA2* variants identified at LHSC were collated in BIC nomenclature [9] in Intervening Sequence (IVS) format, which uses an intron numbering system to identify the intron containing the genetic variant (for example, a T>C variant one nucleotide into intron 1 is described as IVS1+1T>C). Exonic variants were collated in HGVS nomenclature [10], and a clinical format describing the amino acid and nucleotide change. To create one file describing all variants in HGVS and BIC nomenclature for submission to the Evidence-based Network for the Interpretation of Germline Mutant Alleles consortium (ENIGMA; http://enigmaconsortium.org/), all variants were converted to HGVS cDNA nomenclature. HGVS cDNA nomenclature requires: the gene name harboring the variant, the HGVS cDNA coordinate of the variant, and the HGVS amino acid coordinate and effect when applicable. Conversion of variants to HGVS nomenclature used *RefSeq* database [49] reference sequences for BRCA1 transcript (NM_007294.3) and BRCA2 transcript (NM_000059.3) approved by the ENIGMA consortium and Leiden Open Variation Database (LOVD) [50].

To convert intronic variant coordinates in BIC nomenclature to HGVS nomenclature, a previously gathered list matching cDNA coordinates of the first and last nucleotide positions of each exon in *BRCA1* and *BRCA2* to the IVS numbering system, was used. After conversion, the reference base of all variants in HGVS format was compared to the

RefSeq database reference base to ensure no discrepancies emerged during coordinate conversion.

Exonic variants in clinical format were converted to HGVS nomenclature by linking the amino acid change reported by the clinic to the corresponding HGVS nucleotide position through the BIC database, which expresses *BRCA1* and *BRCA2* variants, as well as amino acid changes, in both BIC and HGVS format.

All variants were then submitted to the ENIGMA consortium. For future *in silico* analyses, an HGVS position conversion website https://mutalyzer.nl/positionConverter [51] was used to convert all variants to hg19 genomic coordinates, which is the most recent sequence assembly of the human genome at the time of this study [52]. Differences between *BRCA1* and *BRCA2* hg18 and hg19 genome builds include chromosomal coordinates. Total number of exons, and amino acids encoded in the coding region remains the same between builds.

2.1.1 *BRCA1* and *BRCA2* variant prioritization for effects on splicing

All *BRCA1* and *BRCA2* variants were run through the Automated Splice Site Analysis (ASSA) Server [53] (splice.uwo.ca) to predict putative effects on splicing. First, all variants were converted from HGVS to hg15 nomenclature for input into the ASSA server using a list previously compiled translating cDNA coordinates to corresponding hg15 genomic coordinates. A 300bp window was used when examining effects of variant on splice site information content to look for cryptic splice sites with greater information contents than

the affected site that may be selected by the spliceosome or splicing enhancer/repressor proteins in response to a mutation. Cryptic splice sites located outside of a 300bp window are less likely to bind the spliceosome with a natural splice site as described in more detail in Chapter 2.6. All variants were examined for information content changes in both canonical and regulatory splice sites including: acceptor, donor, branch point, SRp40, SC35, hnRNPH, SF2/ASF and SRP55. Output was filtered to include only variants resulting in splice site abolishing mutations, leaky mutations, or cryptic site mutations with information content changes of $\geq$1 bit, corresponding to a two fold change in protein binding affinity [42], *In vitro* mRNA splicing assays and gene expression analyses have determined that a two-fold change in binding affinity is a threshold for detecting splicing and gene expression changes [46,54]. Results were further filtered based on: 1) the amount of information content change (ΔR_i), 2) the number of splice sites affected by the mutation, 3) and the presence and proximity of a cryptic site of greater information content than an affected natural site. After prioritization, a literature search of all filtered variants expressed in HGVS, BIC and hg19 nomenclatures was performed using: Google Scholar™, PubMed, Web of Science Database (Thomson Reuters), Quertle™ and Google™ to ensure prioritized variants were classified as variants of unknown significance or had no published findings.

After information theory based prioritization, prioritized variants were examined using the Human Splicing Finder version 2.4.1 [55], an online bioinformatics tool that predicts the effects of mutations on canonical and regulatory splicing elements using position weight matrices generated from human genetic and transcript data from Ensembl

(www.ensembl.org). The strength of a putative acceptor/donor site is defined as the sum of each nucleotide's frequency at positions within a site plus a constant used for normalization [55]. The mean score of acceptor and donor splice sites are 86.81 (SD=6.33) and 87.53 (SD=8.34) respectively, strong sites are defined as having score >80, with weaker sites having scores between 70-80 [55]. Prioritized variants affecting the information content of an ESE element were also examined using the ESEfinder tool (http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home), which is an online tool that uses position weight matrices of ESE motifs generated from functional experimental data to predict the presence of ESE elements within a submitted sequence [56]. The Human Splicing Finder was chosen as an additional prediction tool because the algorithms used to predict putative variant effects on splicing use experiment-derived functional canonical and regulatory splice sites from Ensembl. A comparison between the Human Splicing Finder, MaxEntScan (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html) [57] and information theory was conducted by Mucaki *et al.* (2011) for *BRCA1* and *BRCA2* variants with published mRNA expression data [46]. Of 36 variants examined, 22 were concordant for all bioinformatics methods as well as 26 of 26 variants reported to have no effect on splicing [46]. Of discordant mutations, nine were correctly predicted by information analysis when at least one other prediction tool did not. *BRCA1* and *BRCA2* cryptic splice sites not predicted by information theory shown to be influential in splicing have been predicted by Human Splicing Finder and not MaxEntScan [23] and by both Human Splicing Finder and MaxEntScan [58]. In addition, putative splicing variants can be analyzed from a genomic context using the Human Splicing Finder but not MaxEntScan

or other splicing prediction tools such as the Spliceman webserver (http://fairbrother.biomed.brown.edu/spliceman/html/tools.html) [59], which have only a user sequence submission option, thus requiring a criteria for standardization of sequence input length to ensure all sites relevant in splicing are considered in predictions. It is for these reasons that the Human Splicing Finder was chosen as an additional prediction tool for putative effects of *BRCA1* and *BRCA2* variants on splicing.

A VUS was also selected for downstream *ex vivo* analysis from the BIC database, as described by Mucaki *et al.* (2011) [46], using the same prioritization criteria listed above. A workflow of variant prioritization for splicing, as well as VUS submission to ENIGMA consortium is displayed in Figure 4.

Putative canonical leaky splicing variants were also identified using the ASSA server. Leaky splicing variants are defined as variants that reduce the information content of the natural splice site, but have a final information content of >1.6 bits, which has been used as a threshold for minimum functional splice site information content based on analysis of minimally functional splice sites [53]. Patient family history of hereditary breast and/or ovarian cancer (HBOC), which is defined as a predisposition to developing breast and/or ovarian cancer due to germline genetic mutations in *BRCA1* or *BRCA2* [60], was obtained for patients with putative leaky splicing variants to examine disease penetrance. Average number of 1st, 2nd and 3rd degree relatives with breast cancer for patients with each variant was calculated. Also, variant rsID, as well as alternate allele population frequencies were

recorded as found in dbSNP. The clinical significance of the variant as reported in BIC was also recorded.

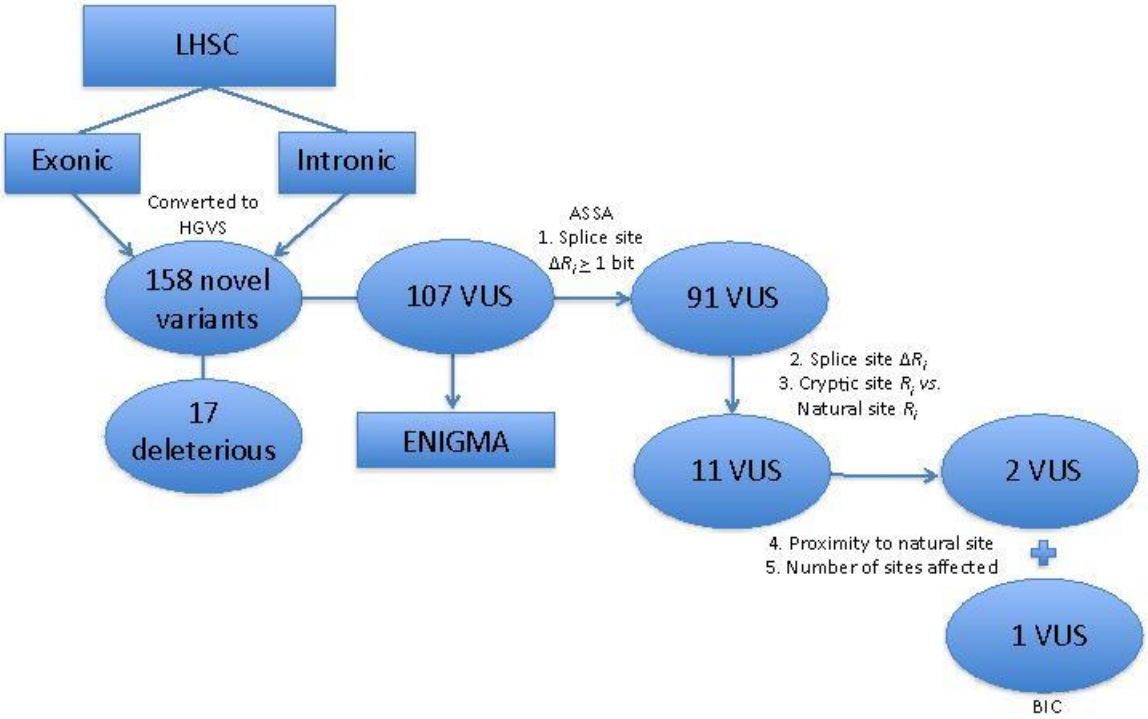

Figure 4. **Workflow of *BRCA1* and *BRCA2* variant *in silico* prioritization for effects on splicing.** Two files containing *BRCA1* and *BRCA2* variants were obtained from LHSC, one describing intronic variants in IVS format (BIC nomenclature) and one describing exonic variants (HGVS nomenclature and clinical format, as described in Chapter 1.2). All variants were converted to HGVS nomenclature. VUS were submitted to the ENIGMA consortium and prioritized *in silico* for effects on splicing using the ASSA server for downstream *ex vivo* transfection analysis.

2.1.2 Preparation of mini-gene construct for investigation of putative splicing variant, *BRCA2*:c.7319A>G

Patient genomic DNA (gDNA) with the *BRCA2*:c.7319A>G variant (rs4986860, dbSNP alternate allele population frequency 0.007, listed in BIC as unknown clinical significance) was purified from a patient blood sample at LHSC using MagNA Pure Compact Nucleic Acid Isolation kit I (Roche, catalog#: 03 730 964 001). This variant was chosen because it is predicted to create a cryptic SRp40 site in close proximity to a cryptic donor site in exon 14 of *BRCA2*. A more detailed explanation is provided in Chapter 3.1. The following steps were undertaken to prepare mini-gene construct: i) PCR amplification of variant containing region in patient and HapMap gDNA; ii) restriction enzyme digestion of PCR amplified DNA products; iii) ligation of digested PCR product into digested, phosphatase-treated mini-gene plasmid, iv) transformation of recombinant mini-gene vector into DH5-α *Escherichia Coli* cells, v) PCR amplification of plasmid insert from *E. coli* colony, followed by plasmid purification and sequencing validation of insert.

i) PCR amplification of variant and Wild-type (WT) region. Primers were designed to amplify exon 14 of *BRCA2,* which contains c.7319A>G, as well as 146nt of 5' flanking intron, and 235nt of 3' flanking intron (total amplicon size: 809 nucleotides) using Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/), an online tool for primer design [61]. To do this, *BRCA2* exon 14 and 250 nt of flanking intron was inputted into Primer3 using hg19 sequence build on the UCSC genome browser [62]. The SantaLucia salt adjusted melting temperature parameter, which refers to a publication by SantLucia *et al.* (1998) [63], which

accounts for the presence of divalent cations during primer melting temperature calculation, was used as recommended by Primer3 [64]. Also, the recommended SantaLucia Table of Thermodynamic parameters was used during primer design, which incorporates primer duplex initiation parameters based on terminal G-C and A-T base pairs in the standard entropy calculation. Further parameters include: a minimum forward and reverse primer melting temperature difference of 2°C, primer length between 20-30nt (27nt optimal), and GC content between 20-80%. A primer pair amplifying a 809bp fragment (*BRCA2*:g.32,928,852-32,929,660) fitting all parameters was chosen. The DNA amplicon was then submitted as a "user sequence" to the ASSA server to ensure the information content changes observed when examining the variant in the genomic context was the same. The 5' end of forward primer was designed to include a BamHI restriction enzyme binding site, and the 5' end of the reverse primer was designed to include an MluI binding site for subsequent digestion, and ligation into mini-gene plasmid. Complementarity to *BRCA2* begins at the first nucleotide after the restriction sites. Primer sequences shown below were ordered from Integrated DNA Technologies (IDT):

Forward: 5' GACCGGATCCGTTTTGGAATGGCAACCATGGTGAATA 3'
(BamHI site) Tm=56.7°C, Length: 37nt

Reverse: 5' GACCACGCGTGTAACAGCAGTAATAGATTGTGTGTCT 3'
(MluI site) Tm=58.2°C, Length: 37nt

In addition to patient genomic DNA, HapMap (GM18859) genomic DNA homozygous for the reference allele at *BRCA2:*c.7319, as determined using the Coriell SNP Browser (*Coriell Institute for Medical Research*) was used to act as a WT control. First, patient and HapMap GM18859 genomic DNA (80ng) were PCR amplified using Platinum *Pfx* DNA polymerase (Invitrogen catalog#: 11708-013) according to the manufacturer's instructions and primers shown above at an annealing temperature range of 47-58°C for 27 cycles. PCR products were spin-column purified (Qiagen catalog#: 28104) and gel visualized on a 1% agarose gel; a band was observed at expected size of 809 base pairs at an annealing temperature of 53°C.

ii) Restriction enzyme digestion of PCR products and mini-gene plasmids. PCR products were quantified using Nanodrop 2000 spectrophotometer (Thermo Scientific), and 333ng of patient and HapMap PCR products were double digested using: 25 units each of BamHI and MluI restriction enzymes (New England Biosciences catalog#: R0136S and R0198S respectively), 5µL NEBuffer 3 and 5µg Bovine Serum Albumin in a 50µL reaction volume at 37°C for three hours. Products were spin column purified according to manufacturer's specifications (Qiagen catalog#: 28104).

pcDNA-Dup and pCAS2 mini-gene plasmids were obtained from Dr. Mario Tosi at *Institute National de la Santé et de la Recherche Médicale* (INSERM) Unité 614 Rouen, France. For ligation of PCR products into mini-gene plasmid, pCAS2 (4.18µg) was digested using 25 units each of BamHI and MluI restriction enzymes, 5µL NEBuffer 3 and 5µg Bovine Serum Albumin in a 50µL reaction volume at 37°C for three hours.

After digestion, pCAS2 was treated with 5 units of Antarctic Phosphatase (New England Biosciences catalog#: M0289S) according to manufacturer's instructions, to prevent vector self-ligation.

iii) Ligation of digested PCR products into mini-gene plasmid. 0.023pmol of pCAS2 was used as a starting point for ligation reactions of various insert:vector molar ratios. To calculate the mass of 0.023pmol of pCAS2, the size of digested pCAS2 (7556bp) was multiplied by the average molar mass of one DNA base pair (~660 g/mol/bp), which results in $5.0x10^6$ g/mol of pCAS2. Multiplying the molar mass of pCAS2 by 0.025pmol, results in 115ng. Using the same formula, the molar mass of PCR product insert was calculated to be $5.339x10^5$ g/mol.

Digested patient and HapMap PCR products were ligated into digested, phosphatase-treated pCAS2 at 0:1, 0.5:1, 1:1, 3:1 and 5:1 PCR-product:pCAS2 molar ratios, using 0ng, 12.2ng, 12.2ng, 36.6ng and 55ng of PCR product respectively, and 115ng, 330ng, 115ng, 115ng, and 101ng of pCAS2 vector respectively. Digested patient and HapMap DNA were combined with digested, phosphatase-treated pCAS2 in separate reactions, and were ligated using 400 units of T4 DNA ligase and 1/10 volume of T4 DNA ligase buffer. Samples were then incubated at 16°C for 16.5 hours, followed by 15 minute 65°C enzyme heat inactivation. Ligation products were visualized on a 1% agarose gel.

iv) Recombinant mini-gene plasmid transformation into DH5-α *Escherichia Coli* cells. Approximately 58ng of ligation products of all molar ratio conditions were

transformed into 200µL of competent DH5-α *Escherichia Coli* cells by first combining DNA and cells, and resting them on ice for 30 minutes. Tubes were transferred to a circulating water bath at 42°C for 90 seconds, and placed back on ice for 2 minutes. Next, 800µL of Super Optimal Broth was added to each sample (2g tryptone, 0.5g yeast extract and 0.05g NaCl was added to 95mL of deionized water and mixed to dissolve. Next, 1mL of 250mM KCl was added, solution was brought up to 100mL using deionized water and autoclaved, when cooled, 0.5mL of 0.22µm filter sterilized 2M $MgCl_2$ was added). Samples were incubated at 37°C on a shaker at 225 cycles/minute for 40 minutes. After transformation, 10µL, 15µL, 100µL and 150µL of cells and media were plated onto separate ampicillin-containing Luria broth agar plates (1.25g tryptone, 0.625g yeast extract, 1.25g NaCl, 1.875g agar, add deionized water to 125mL, boil, let cool, add 12.5mg ampicillin, mix and pour ~3mm thin layer into sterile 100x15mm petri dishes), and incubated overnight at 37°C.

To confirm the success of transformation of insert-containing plasmid as well as determine the optimal ligation ratio condition, PCR amplification of mini-gene insert was performed the next day on four or five colonies from each ligation condition plate depending on colony density. Autoclaved toothpicks were used to gently touch individual colonies, which were then suspended into separate 0.2mL Eppendorf tubes containing 5µL of nuclease free water. To reduce the risk of inadvertently sampling multiple colonies, which would result in template heterogeneity, only isolated, easily accessible colonies were chosen from each plate. The fraction of isolated colonies varied based on colony density. The fraction of isolated colonies in the 10µL, 1:1 ligation ratio condition

corresponds to 4 of 12 colonies, whereas in the 100μL, 1:1 ligation ratio condition, only 1 of 310 colonies was chosen. Therefore, five colonies in total were examined from the 1:1 ligation ratio condition, as well as five from the 3:1 ligation ratio condition, and four from each 0.5:1 and 5:1 based on sampling only isolated colonies. Fifteen microliters of *Taq* Polymerase Master mix prepared according to manufacturer's instructions (Invitrogen catalog#: 10342-053) containing primers specific to insert, shown previously, was added to each tube. Tubes were incubated at 94°C for four minutes to lyse cells, which was followed by 27 cycles of amplification. Amplification products were spin column purified (Qiagen catalog#: 28104) and run on a 1% agarose gel. Two colonies from 0.5:1 ligation ratio plates showing amplification at the desired size of 809bp were transformed into DH5-α cells and grown in 25mL of Super Optimal Broth overnight at 37°C on a shaker at 225 cycles/minute. Plasmids were then purified using a PureYield Midi prep Kit (Promega, catalog # A2492) and sent to the London Regional Genomics Facility sequencing facility (*London Regional Genomics Centre*, London ON) using insert specific forward primer to confirm genotype of patient and HapMap DNA at the *BRCA2*:c.7319 locus.

2.1.3 Preparation of mini-gene construct for investigation of putative splicing variant, *BRCA1*:c.548-16G>A

The *BRCA1*:c.548-16G>A variant (rs80358171, no alternate allele population frequency data, listed in BIC as unknown clinical significance), is predicted to create a cryptic acceptor site (ΔR_i = -2.8>5.4 bits) 14 nucleotides upstream of the exon 8 while decreasing the information content of the natural acceptor site (ΔR_i = 0.5>0.2 bits). Further explanation is provided in Chapter 3.1. To examine this variant, a recombinant mini-

gene construct was created using the same workflow described in the previous section. This variant was found in Mucaki *et al.* (2011) and is not part of the London Health Sciences Centre patient cohort. Since patient DNA was unavailable, the *BRCA1*:c.548-16G>A mutation was introduced using the primers used during PCR amplification. Therefore, two forward primers, one encoding the mutation of interest, and one encoding the WT nucleotide, were designed to amplify HapMap genomic DNA (GM18860) to create mutant and WT products respectively.

i) PCR amplification of variant and WT

For PCR amplification, primers were designed using Primer3 under the same parameters and constraints listed previously, to amplify exon 8 of *BRCA1* as well as 44nt of 5' flanking intron and 210nt of 3' flanking intron (total amplicon size 300bp). The 5' end of forward primers were designed to include a BamHI recognition site, and the 3' end of the reverse primer was designed to include an MluI recognition site for subsequent digestion, and ligation into digested pCAS2. Complementarity to *BRCA1* begins at first nucleotide downstream of restriction sites. Primer sequences shown below were ordered from IDT:

Mutant forward (mutant nucleotide in bold):

5' GACCGGATCCAAGAAAACTTTTATTGATTTATTTTTG**A**GGGGAAATT 3'

(BamHI) Tm=55.8°C, Length: 47nt

Wild-type forward (wild-type nucleotide in bold):

5’ GACCGGATCCAAGAAAACTTTTATTGATTTATTTTTG**G**GGGGAAATT 3’

(BamHI) Tm=57.4°C, Length: 47nt

Reverse:

5’ GACCACGCGTGGTATGCTTAGTACCCGGATGATGA 3’

(MluI) Tm=57.7°C, Length: 35nt

HapMap genomic DNA (100ng) was PCR amplified in two conditions, one using WT forward primer and the other using mutant forward primer. Amplification was performed using Platinum *Pfx* DNA Polymerase (Invitrogen catalog#: 11708-013) according to the manufacturer’s instructions, with primer annealing temperatures ranging from 57-67°C, and in the presence of 5% v/v DMSO in the WT condition to prevent primer secondary structure. Amplification products were spin column purified (Qiagen catalog#: 28104) and visualized on a 2.5% agarose gel, an amplicon was observed at the expected size of 300bp at an annealing temperature of 61.1°C for the WT condition and 58.6°C for the mutant condition.

ii) Restriction enzyme digestion of PCR products and mini-gene plasmids

PCR products were quantified using a Nanodrop 2000 spectrophotometer, and 333ng of WT and mutant amplification products were double digested using: 25 units each of BamHI and MluI restriction enzymes, 5µL NEBuffer3, and 5µg Bovine Serum Albumin in a 50µL reaction volume at 37°C for three hours.

iii) Ligation of digested PCR products into mini-gene plasmid

Digested PCR products were then ligated into digested, phosphatase-treated pCAS2 vector using T4 DNA ligase in 0:1, 1:1, 1.5:1 and 3:1 insert:pCAS2 molar ratios, using 0ng, 20.9ng, 31.35ng and 62.7ng of digested PCR products respectively, as well as 115ng of digested, phosphatase-treated pCAS2 in each condition.

Digested PCR product insert and digested, phosphatase-treated pCAS2 were ligated using 400units of T4 DNA ligase and 1/10 volume of T4 DNA ligase buffer according to manufacturer's instructions (New England Biosciences catalog#: M0202S) at 16°C for 16.5 hours, followed by 15 minute 65°C enzyme heat inactivation.

iv) Recombinant mini-gene plasmid transformation into DH5-α *Escherichia Coli* cells

Wild type and mutant recombinant vectors were transformed into competent DH5-α cells, plated onto ampicillin-containing Luria broth agar plates, and incubated overnight at 37°C as described in the previous section. To confirm the success of transformation of insert-containing plasmid as well as determine the optimal ligation ratio condition, PCR amplification of mini-gene insert was performed the next day on three colonies from each ligation condition plate, using the same autoclaved toothpick, lysis and PCR method described previously. Amplification products were spin column purified (Qiagen catalog#: 28104) and run on a 2.5% agarose gel. Colonies from (1:1 ligation ratio for WT, 1:1.5 ligation ratio for mutant) plates showing amplification at the desired size of

300 bp were transformed into DH5-α cells and grown in 25mL of Super Optimal Broth overnight at 37°C and 225 cycles/min. Plasmids were then purified using a PureYield Midi Prep Kit (Promega, catalog#: A2492) and sent to the London Regional Genomics Facility for Sanger sequencing using standard T7Pro forward primer to confirm WT and mutant conditions had the proper genotype.

2.1.4 Preparation of mini-gene construct for investigation of putative splicing variant, *BRCA1*:c.288C>T

The *BRCA1*:c.288C>T variant is predicted to abolish ESE elements:
SRp40 (ΔR_i = 2.6>-2.9 bits) and SF2/ASF (ΔR_i = 6.6>0.9 bits), both located in exon 5, 14 nucleotides upstream of the natural donor site. Abolishing splicing regulatory binding sites in close proximity to the natural donor site is predicted to lead to reduced donor site recognition by the spliceosome and lead to aberrant splicing. A more detailed explanation of why this variant was chosen for splice site analysis is provided in Chapter 3.1. To investigate this variant, the variant and 17 nucleotides upstream and downstream was inputted into the ASSA server as a user defined sequence to confirm that information theory predictions are the same as when examining the variant from the genomic context. To examine the effects of *BRCA1*:c.288C>T on splicing, two complementary 55-mer DNA oligonucleotides encoding the putative mutated SRp40 and SF2/ASF site were designed, as well as a pair of oligonucleotides encoding the WT nucleotide. Oligonucleotides were designed to encode a EcoRI and BamHI restriction site for subsequent digestion and ligation into pcDNA-Dup mini-gene vector. Undigested pcDNA-Dup not containing recombinant insert was used as a control for cassette exon exclusion in downstream transfection and RT-PCR experiments. Oligonucleotide

sequences shown below were ordered from IDT, regions complementary to *BRCA1* are shown below in brackets:

Mutant Forward (mutation in bold):

5’GACCGAATTC(TGTGCTTTTCAGCTTGA**T**ACAGGTTTGGAGTGTAA)GGATCC GACC 3’

(5’ EcoRI and 3’ BamHI) Tm=72.1°C, Length: 55nt

Reverse complement:

5’GGTCGGATCC(TTACACTCCAAACCTGT**A**TCAAGCTGAAAAGCACA)GAATT CGGTC 3’

Tm=72.1°C, Length: 55nt

Wild-type Forward (WT nucleotide in bold):

5’GACCGAATTC(TGTGCTTTTCAGCTTGA**C**ACAGGTTTGGAGTGTAA)GGATC CGACC 3’

(5’ EcoRI and 3’ BamHI) Tm=72.8°C, Length: 55nt

Reverse complement:

5’GGTCGGATCC(TTACACTCCAAACCTGT**G**TCAAGCTGAAAAGCACA)GAATT CGGTC 3’

Tm=72.8°C, Length 55nt

Next, WT and mutant forward and reverse strands (2µg each) were annealed separately in two tubes by first heating at 95°C for 6 minutes, and ramping down 1°C per minute until samples reached 26°C. After annealing, 1µg of double-stranded product was digested using: EcoRI and BamHI restriction enzymes (New England Biosciences catalog #: R0101S and R0136S respectively), 1.5µL $MgCl_2$, 5µL NEBuffer 3 and 5µg Bovine Serum Albumin in 50µL reaction volume at 37°C for three hours. Next, 5µg of pcDNA-Dup was sequentially digested with EcoRI and BamHI restriction enzymes and treated with Antarctic Phosphatase according to manufacturer's instructions. Digested insert and vector were then ligated using T4 DNA ligase as described previously in 0:1, 5:1 and 10:1 insert:vector molar ratios. Ligation products were then transformed into competent DH5-α cells and plated onto ampicillin containing plates as previously described. Colonies were then selected from each ligation condition using autoclaved toothpicks for heat lysis and PCR amplification of insert as described previously. Amplification products were run out on a 4% gel and bands at the expected size of 97bp were observed. Colonies were then transformed into DH5-α cells and grown in 25mL of Super Optimal Broth overnight at 37°C. Plasmids were then purified using a PureYield Midi prep Kit (Promega, catalog # A2492) and sent to the London Regional Genomics Facility for Sanger sequencing using standard T7Pro forward primer to ensure the insert was present.

2.1.5 Transfection of recombinant mini-gene vectors in MDA-MB-231 breast cancer cell line

Different mRNA splice isoforms transcribed from the same gene exist in different tissue types [65]. Tissue-specific expression of RNA binding proteins involved in splicing promotes alternative splicing across tissues. To mimic endogenous levels of proteins

involved in splicing in breast tissue, recombinant mini-gene vectors were transfected into breast cancer cell line MDA-MB-231. This cell line has been used to transfect vectors derived from the same backbone using the same transfection reagent (FuGene6) in a previous study conducted by Morelli *et al.* (2003) [66]. One 1mL vial of MDA-MB-231 cells was removed from a liquid nitrogen tank and immediately placed in 37°C water bath. When thawed, vial was wiped with 70% ethanol and placed in cell culture hood sterilized with 10% bleach and 70% ethanol. Cells were placed in 15mL Falcon tube, and 10mL of pre-warmed Dulbecco's Modified Eagle Medium (Sigma-Aldrich product#: RNBC3976), containing 10% Fetal Bovine Serum (Fisher product#: SH3007003), 1% penicillin and 1% streptomycin (Invitrogen product#: 15140122) was added and mixed. Cells were pelleted at 1000xg for 9 minutes. Supernatant was discarded, and cells were resuspended in 7mL of complete media. Cells were then transferred to T25 flask (BDBiosciences catalog#: 353108) and placed in an incubator at 37°C, 5% CO_2. When cells reached 80% confluency, cells were trypsinized, plated onto two T75 flasks (BDBiosciences catalog#: 353136) and returned to the incubator. When cells reached 80% confluency, they were trypsinized again and split into four T75 flasks and returned to incubator. When flasks reached 80% confluency, cells were trypsinized and counted using a hemacytometer. Next, eight six-well plates (VWR product#: 82050-846) were seeded with $3x10^5$ cells/well in complete media to total 3mL of cells and media per well as recommended by Fugene6 transfection reagent protocol (Promega, catalog#: E2691). Cells were returned to incubator and retrieved the following day for transfection when they reached ~80% confluency.

Cells were transfected with eight plasmids, one plasmid per plate, these include: empty pcDNA-Dup (containing no insert), empty pCAS2 (containing no insert), pCAS2 mini-gene containing *BRCA1*:c.548-16G WT and *BRCA1*:c.548-16G>A mutant, pCAS2 mini-gene containing *BRCA2*:c.7319A WT and *BRCA2*:c.7319A>G mutant, and pcDNA-Dup mini-gene containing *BRCA1*:c.288C WT and *BRCA1*:c.288C>T mutant. A total of 3µg of plasmid DNA was transfected in each well using FuGene6 transfection reagent. This was calculated by multiplying the amount of DNA suggested for one well of a 96 well plate (0.1µg) by the relative growth area multiplication factor for a 6-well plate (30X) provided as described in the FuGene6 transfection reagent protocol.

Plasmids were transfected in three ratios of FuGene6 transfection reagent to plasmid DNA: 1.5:1, 3:1 and 6:1 (µL:µg). First, transfection reagent was incubated for 5 minutes at room temperature with a volume of pre-warmed serum free media necessary to bring the final volume up to the recommended 150 uL/well.

Next, 3µg of plasmid DNA was added to transfection reagent/media mixture. Tubes were mixed and incubated for 15 minutes at room temperature. Next, 150µL of mixture was added to each well corresponding to plasmid and reagent:plasmid ratio, with one plasmid per plate, and two wells per ratio condition. Plates were returned to incubator. To select for transfected cells harboring plasmid, the backbone of which encodes a neomycin resistance gene, twenty-four hours later, 1.5mg of filter sterilized 200µg/µL G418 (VWR, product#: E859-1G) was added to each well of each plate, to a final concentration of 500ug/mL, as used by Morelli *et al.* [66]. After adding G418, cells were returned to

incubator. Forty-eight hours later, 25µg of 5ug/µL filter-sterilized puromycin (Sigma catalog#: P8833-10MG) was added to three of six wells of each plate (final concentration per well: 8.33µg), one for each reagent:plasmid ratio, as recommended by Tournier *et al.* (2008) [47] to inhibit nonsense mediated decay and thus allow for extraction of RNA transcripts encoding premature stop codons. Next, 5.5 hours after puromycin was added, cells were washed with 1XPBS and trypsinized. Cells for each condition were pelleted and divided into two tubes, one for RNA extraction, one for DNA extraction.

Total RNA was extracted using Trizol LS (Invitrogen catalog#: 10296-028) according to the manufacturer's instructions. RNA was resuspended in 48µL of nuclease free water and 2µL of RNaseIN (Promega catalog#: N2611), and split into two 25µL aliquots. Samples were ethanol precipitated twice to remove residual Trizol reagent contamination, and achieve 260/280 and 260/230 absorbance ratios of between 1.9-2.1 and 1.7-2.3 respectively, as determined on Nanodrop spectrophotometer.

Plasmid DNA was extracted from cells using Trizol reagent, and was subsequently purified using a Qiagen spin mini-prep kit according to manufacturers instructions, as has been performed by Ziegler *et al.* in 2004 to extract low-molecular weight circular DNA from mammalian cells [67].

2.1.6 Reverse transcription of extracted mRNA

RNA was isolated from both WT and mutant plasmid conditions corresponding to: *BRCA2*:c.7319A>G, *BRCA1*:c.548-16G>A and *BRCA1*:c.288C>T variants at 1.5:1 and

3:1 transfection ratios. Next, 10µL of DNase I solution (stock DNase I solution made by mixing: 10µL of 100mM $MgCl_2$, 0.5µL of 1.0 mg/mL DNase I, 0.5µL of RNaseIN, and 39µL of TE buffer) was added to 10µL of RNA (1.1µg total) and mixed. Solution was incubated at 37°C for 15 minutes before adding 1µL of 0.05M EDTA. Solution was mixed and heat inactivated at 65°C for 20 minutes. Solutions were placed on ice.

RNA was reverse transcribed (RT) using Superscript II Reverse Transcriptase (Invitrogen: 18064-022) using 250pg of random hexamers according to the manufactures instructions. To validate success of RNA extraction and reverse transcription, an aliquot of the RT reaction (1µL) was PCR amplified using *Taq* DNA Polymerase according to manufacturer's instructions in addition to HapMap GM19200 RNA to act as a positive control. cDNA derived from endogenously expressed VSP39 mRNA, was PCR amplified using primers designed using Primer3 using parameters described in 2.1.2 to amplify a 103bp cDNA fragment at an annealing temperature of 59°C. VSP39 was chosen as a positive control for RNA extraction and RT-PCR because previous microarray expression data revealed that splicing index and intensity levels of this exon of VSP39 had the lowest variability across 176 HapMap samples examined. Low levels of expression and splicing variability across samples supports use of VSP39 as a stable gene to act as a positive control for RNA extraction and RT-PCR. Primer sequences are shown below:

Forward: 5' CTCAGTTCCATCAAAACTCGGTTT 3' Tm= 55.6°C, Length: 24nt

Reverse: 5' GAAGCCAAGACACACAGTGCTG 3' Tm= 58.7°C, Length: 22nt

Next, cDNA derived from endogenously expressed *BRCA2* mRNA was PCR amplified at 52°C using primers designed using Primer3 using parameters described in 2.1.2 to bind the 3' end of exon 11 (forward primer, chr13:g.32,152,254-32,915,276), and across the exon11-exon12 junction (reverse primer), yielding an expected amplicon size of 96bp. This PCR was conducted as an additional control to ensure that the RNA extraction and reverse transcription steps were successful, and that mRNA splicing of endogenously expressed genes is efficient in MDA-MB-231 cell line. Primer sequences are shown below:

Forward: 5' TACATGTCCCGAAAATGAGGAAA 3' Tm=54.3°C, Length: 23nt

Reverse: 5' TTGATTGAGGGTTCTCCCACTAAG 3' Tm=56.3°C, Length: 24nt

BRCA1:c.548-16G and *BRCA2*:c.7319A WT cDNA was PCR amplified for 30 cycles using primers specific to the first and last exon of pCAS2 mini-gene (Table 1, primer identifier numbers 5 and 6) as recommended by INSERM at an annealing temperature of 57°C. PCR amplification was performed with and without 1M betaine. An additional PCR reaction was performed at a lower annealing temperature of 55°C. Products were electrophoresed on a 2% agarose gel.

cDNA corresponding to pcDNA-Dup mini-gene vectors was PCR amplified using Taq polymerase according to manufacturer's instructions and primers designed to bind the T7 Promoter (forward) and exon 2 of pcDNA-Dup mini-gene (reverse) and an annealing

temperature of 57°C as recommended by INSERM, for 30 cycles. The sequence of the forward primer is presented in Table 1 (identifier 1) and the sequence of the reverse primer is presented below. PCR products were electrophoresed and resolved on a 2.5% agarose gel.

Reverse primer: 5' GGACTCAAAGAACCTCTGGG 3' Tm= 60.5°C, Length: 20nt

As a control experiment to determine the effectiveness of DNase I digestion, varying concentrations of stock pcDNA-Dup plasmid DNA were digested with DNase I using the same procedure listed previously except using plasmid template amounts of 440ng, 220ng, 110ng and 20ng. DNase I digested plasmid was then PCR amplified for 35 cycles using primers specific to pcDNA-Dup mini-gene listed above. pcDNA-Dup plasmid, DNase I treated pcDNA-Dup plasmid and mini-gene PCR amplified DNase I treated plasmid was electrophoresed on a 1.5% agarose gel. DNase I digested plasmid was then PCR amplified a second time for 25 cycles using the same mini-gene specific primers.

Next, primers were designed to span exon-exon junctions, using Primer3 under standard parameters described in 2.1.2, to selectively amplify cDNA corresponding to spliced RNA. Each cDNA product was PCR amplified in two conditions, one using primers designed to amplify WT mini-gene cDNA, the other using primers designed to amplify mutant mini-gene cDNA. Mini-gene cDNA (1μL of RT reaction) was PCR amplified using *Taq* polymerase according to manufacturer's instructions under a range of annealing temperatures (52-56°C) for pcDNA-Dup cDNA (includes: empty pcDNA-Dup,

BRCA1:c.288C WT, *BRCA1*:c.288C>T Mutant), and pCAS2 cDNA (includes: empty pCAS2, *BRCA1*:c.548-16G WT, *BRCA1*:c.548-16G>A mutant, *BRCA2*:c.7319A WT, and *BRCA2*:c.7319A>G Mutant). PCR products were resolved on 2% agarose gel using electrophoresis. Primers used during PCR are shown in Table 1.

Table 1. Primers for mini-gene cDNA PCR amplification following reverse transcription

Plasmid condition	Primer sequence (5'>3')	Binding site	Amplicon size (bp)	Tm (°C)	Length (nt)	Identifier
pcDNA-Dup no cassette exon Forward	TAATACGACTCACTATAGGG	Upstream of pcDNA-Dup mini-gene	209	54.3	20	1
pcDNA-Dup no cassette exon Reverse	AGCCTGCCCAGGGCCTCA	Across first and second exon of pcDNA-Dup mini-gene	209	62.9	18	2
pcDNA-Dup cassette exon Forward	GGCAGGCTGAATTCTGTGCTT	Across first and cassette exon of pcDNA-Dup mini-gene	63	61.2	21	3
pcDNA-Dup cassette exon Reverse	AGCCTGCCGGATCCTTACAC	Across cassette and third exon of mini-gene	63	62.5	20	4
pCAS2 *BRCA1*:c.548-16 no middle exon Forward	TGACGTCGCCGCCCATCAC	First exon in pCAS2 mini-gene	280	57	19	5
pCAS2 *BRCA1*:c.548-16 no middle exon Reverse	ATTGGTTGTTGAGTTGGTTGTC	Last exon in pCAS2 mini-gene	280	57	22	6
pCAS2 *BRCA1*:c.548-16 middle exon Forward	GGCTGGGGATCTGATTCTTC	Across first and middle exon of pCAS2 mini-	220	60.5	20	7

		gene				
pCAS2 *BRCA1*:c.548-16 middle exon Reverse	GACAACCAACTCAACAACCAAT	Last exon in pCAS2 mini-gene	220	57	22	8
pCAS2 *BRCA2*:c.7319 no middle exon Forward	TGACGTCGCCGCCCATCAC	First exon of pCAS2 mini-gene	280 (no middle exon), 667 (truncated middle exon)	57	19	9
pCAS2 *BRCA2*:c.7319 no middle exon Reverse	ATTGGTTGTTGAGTTGGTTGTC	Last exon of pCAS2 mini-gene	280 (no middle exon), 667 (truncated middle exon)	57	22	10
pCAS2 *BRCA2*:c.7319 middle exon Forward	TGCTGGCTGGGCACAACTAA	First exon and middle exon of pCAS2 mini-gene	610	60.5	20	11
pCAS2 *BRCA2*:c.7319 middle exon Reverse	ATTGGTTGTTGAGTTGGTTGTC	Last exon of pCAS2 mini-gene	610	57	22	12

Wild-type and mutant *BRCA2*:c.7319A>G were PCR amplified using primers specific to WT and mutant mini-gene transcript displayed in Table 1 (primer identifiers 11 and 12 for WT mini-gene transcript and 9 and 10 for mutant mini-gene transcript). cDNA was PCR amplified using *Taq* polymerase according to manufacturer's instructions according to the thermocycler program: 95°C (3 min), 10 cycles of: 94°C (45 seconds), 67°C (20 seconds) ramping down -1°C per cycle, 72°C (25 seconds); this was followed by 30 cycles of: 94°C (45 seconds), 57°C (20 seconds), 72°C (25 seconds), and a final elongation step at 72°C (1 minute). cDNA was also amplified using final annealing temperatures of 53°C, 54°C for WT primers and 55°C and 56°C for mutant primers. All products were electrophoresed on a 2% agarose gel.

pcDNA-Dup derived plasmid cDNA corresponding to *BRCA1*:c.288C WT, *BRCA1*:c.288C>T mutant and empty pcDNA-Dup were PCR amplified at an annealing temperature of 56°C under the following conditions using *Taq* polymerase: 95°C (3 min), 10 cycles of: 94°C (45 seconds), 66°C (20 seconds) ramping down -1°C per cycle, 72°C (25 seconds); this was followed by 30 cycles of: 94°C (45 seconds), 56°C (20 seconds), 72°C (25 seconds), and a final elongation step at 72°C (1 minute). Products were electrophoresed on a 2% agarose gel.

cDNA corresponding to WT *BRCA2*:c.7319A and *BRCA1*:c.548-16G and 3:1 transfection ratio conditions were PCR amplified using Taq Polymerase according to manufacturer's instructions using primers designed to amplify: pCAS2 mini-gene containing middle exon and no middle exon (primer identifiers: 7 and 8; 5 and 6

respectively, at an annealing temperature of 52°C) for *BRCA1*:c.548-16G>A, and pCAS2 mini-gene containing middle exon and no/aberrant middle exon (primer identifiers: 11 and 12; 9 and 10 respectively, at an annealing temperature of 52°C). PCR products were electrophoresed on a 2% agarose gel.

2.2 Estimation of number of breast cancer families with *BRCA1* and *BRCA2* variants using a Power Analysis

Breast cancer patients at LHSC are categorized into a series of groups defined by the incidence and age-of-onset of breast, ovarian and/or other HBOC in first, second or third degree relatives. The Molecular Diagnostics Laboratory at LHSC has provided the number of patients in each group. The estimated proportion of families with *BRCA1* and *BRCA2* mutations was determined through linkage frequency published by Ford *et al.* (1998) [28]. In the Ford *et al.* (1998) study, genetic information was obtained for 237 families with breast and/or ovarian cancer from the Breast Cancer Linkage Consortium (BCLC), which is a consortium that harbors genetic information for over 700 families with hereditary breast cancer and aims to evaluate the cancer risks conferred by *BRCA1* and *BRCA2*. Genetic markers on 17q21 flanking *BRCA1*, including D17S579 (Hall et al. 1992) and D17S250 (centromeric to *BRCA1*) were used to determine linkage frequency of *BRCA1* to breast and/or ovarian cancer. Genetic markers flanking *BRCA2*, D13S260 (centromeric to *BRCA2*) and D13S267 were used to determine linkage of breast and/or ovarian cancer to *BRCA2*. Additional linkage and mutation studies were used to estimate the frequency of *BRCA1* and *BRCA2* mutations in patients with bilateral breast cancer [68] and breast-ovarian [69] cancer. The total number of deleterious *BRCA1* and *BRCA2* mutations detected in each group of breast cancer patients was obtained from the

molecular diagnostics lab. Dividing this by the total number of patients in each group gives the fraction of patients with detected deleterious variants. Subtracting the fraction of detected *BRCA1* and *BRCA2* variants by the fraction of *BRCA1* and *BRCA2* variants predicted by linkage and mutation analyses, gives the estimated fraction of undetected variants by patient group. Multiplying the estimated fraction of undetected variants by the total number of patients in each group gives the predicted number of patients with undetected variants in *BRCA1* and *BRCA2*. An online power analysis calculator (Soper D.S. 2010; http://www.danielsoper.com/statcalc3/calc.aspx?id=1) was used to determine the minimal sample size in each group needed to detect a pathogenic variant in *BRCA1* and *BRCA2,* using the fraction of undetected variants as effect size, for a statistical power of 0.8 and probability cutoff of <0.05.

2.3 Capture Array design and synthesis and library preparation methodology

Patient genomic DNA (gDNA) was purified by the Molecular Diagnostics Laboratory at LHSC from patient blood samples using MagNA Pure Compact Nucleic Acid Isolation kit I (Roche, catalog#: 03 730 964 001). Genomic DNA from twenty-one breast cancer patients with a family history of breast cancer was obtained from the Molecular Diagnostics Laboratory at LHSC. Patient gDNA was prepared for sequencing on the Illumina Genome Analyzer IIe using a genomic DNA solution capture procedure described by Gnirke *et al.* (2009) [42]. In the protocol outlined by Gnirke *et al.,* oligonucleotide probes are designed to hybridize to regions of interest on sheared genomic DNA, and are synthesized on a microarray chip. Probes are then cleaved from the array, and transcribed *in vitro* to biotinylated RNA, using biotinylated uracil as

described in section 2.4 (see Figure 2). These probes are then hybridized in solution to sheared, adapter ligated DNA and subsequently captured on streptavidin-coated magnetic beads. After enrichment, samples are PCR amplified and paired-end sequenced on the Illumina GAII*e*. Several modifications of the protocol outlined by Gnirke *et al.* (2009) were introduced in this study, including: 1) 24,000 custom designed oligonucleotide probes targeting 36-70mer single-copy and divergent repetitive sequences (both forward and reverse strands) were synthesized on two separate microarrays (versus 10,000 200mer probe sequences from Agilent Technologies), 2) cleaved probes were PCR amplified using a forward primer encoding an SP6 promoter for downstream *in vitro* transcription (versus T7 Promoter), 3) enzymatic library DNA preparation in the presence of Agencourt magnetic beads in final sequencing run, 3) 36bp (first two sequencing runs) and 50bp (final sequencing run) paired-end sequencing runs were conducted on the Illumina Genome Analyzer II on enriched patient DNA (versus 76bp paired-end sequencing).

2.4 Probe design and synthesis

Twenty-four thousand 36-70mer oligonucleotide probes were designed for independent hybridizations to the forward and reverse strands of seven breast cancer-associated gene sequences. These included exonic, intronic, and adjacent intergenic regions separately for targeted enrichment and downstream sequencing. Separating hybridizations into two reactions, one with forward strand specific probes, the other with reverse strand specific probes, ensures that complementary forward and reverse strand probes do not anneal to each other during incubation. These seven genes include: *ATM*, *BRCA1*, *BRCA2*, *CDH1*,

CHEK2, *PALB2* and *TP53*, and were chosen based on hereditary breast cancer linkage studies described in Chapter 1.4 [29,32,34–36]. Only these seven genes with published causative alleles involved in breast cancer etiology were chosen for probe design in this study.

Within these seven genes, probes were designed to hybridize to intronic, exonic and 5' and 3' untranslated regions (UTR), in addition to 10kb upstream and downstream of each gene. Only *ab initio* identified, single-copy sequences were chosen for probe design, as described by Dorman *et al.* (2013) [70], as well as regions within repetitive elements lacking non-divergent repeats (<30% divergence). We used *ab initio* single copy sequences for probe design in order to increase the coverage of unique regions for capture and sequencing, including regions containing divergent repetitive elements, which under the stringent conditions used, should hybridize only to these targets.

Eligible 36-70mer oligonucleotide probes within *ab initio* single-copy and divergent repeat regions, were chosen using microarray design software PICKY 2.2. Settings used for probe design include: 30-70% GC content, Tm=65°C$\pm$2°C, a 20nt maximum overlap between probes, and a maximum of five probes per 100nt interval, all other settings were set to default. Probes were designed for both forward and reverse strands to be synthesized on two separate microarrays to prevent annealing to each other. Using the program, "Amalgamated-Post-Picky-Program", sequences were analyzed for overlap and cross-hybridization to other regions of genome (hg19) and transcriptome using BLAT [62]. Any gaps in probe coverage within exons greater than five nucleotides were filled

manually by designing unique probes to span these regions with similar predicted melting temperature. To examine the effects of RNA probe secondary structure, a program called "UNAFOLD" was used to calculate the Gibbs free energy of probes at hybridization temperature. Those with values less than -10 kcal/mol were redesigned to minimize secondary structures.

Probes were then synthesized on two cleavable 12K CustomArray chips, one corresponding to probes on the forward strand, the other to probes on the reverse strand, using a Combimatrix CustomArray Synthesizer. To PCR amplify probes after cleavage from the array, primer-binding sequences taken from Gnirke *et al.* 2009 were added to the 5' and 3' ends of 36-70mer probe sequences, as shown below, where $N_{36\text{-}70}$ corresponds to the target region probe:

5' ATCGCACCAGCGTGT-$N_{36\text{-}70}$-CACTGCGGCTCCTCA 3'

This Excel spreadsheet was copied and pasted into Notepad and saved on a computer containing a program called ProbeWeaver. This file was then opened in ProbeWeaver and saved as a chip design file.

To synthesize arrays, a pre-capping protocol designed to expose the chip to capping reagents CapA (Sigma, product#: L040250-1) and CapB (Sigma, product#: L050250-01) 15 minutes prior to synthesis was performed to reduce the number of initiation sites and oligonucleotide density, reducing the load on the membrane. Next, the chip design file

was loaded, and reagent volumes were checked, the acetonitrile keg was full and the waste container was empty. A fresh solution of reagents needed for electrochemical oligonucleotide synthesis on the CustomArray microarray synthesizer, termed E-chem solution, was prepared prior to synthesis by adding 700mL of acetonitrile (Sigma product#: 271004-20L-P2) and 200mL of methanol (Fisher product#: AC14280) to a large jar containing magnetic stir bar, followed by 1.08g of benzoquinone (Fluka Analytical, product#: 12309-100G), 110.11g of hydroquinone (Sigma, product#: H17902-2KG), 15.16g of tetra-ethylammonium p-toluenesulfonate (Sigma, product#: 134473-500G), and 0.59mL of 2,6-lutidine (Sigma, product#: 336106-100ML). Solution was mixed for ten minutes and connected to Combimatrix CustomArray Synthesizer. Two 12K cleavable CustomArray chips, one for forward strand specific probe synthesis, one for reverse strand specific probe synthesis, were loaded into two slots on the Combimatrix CustomArray Synthesizer. Next, the microarray chip was tested for connections and the chip design protocol was run.

After synthesis, oligonucleotide probes were cleaved from the arrays using 14.5M ammonium hydroxide (Carolina, item# 844010) to cleave sulfonyl-amidite bond linking oligonucleotide to array, leaving 3' phosphate on oligonucleotides. First, the newly synthesized arrays were placed in CombiMatrix CustomArray Stripping Clamp manifolds and sealed. Next, 500μL of 14.5M ammonium hydroxide was added to each stripping clamp. The arrays were then incubated in the manifold at 65˚C for four hours. The liquid was removed from the manifolds and placed in two 1.5mL Eppendorf tubes, which were then placed into a SpeedVac RC110A for one hour at 65˚C. The resulting pellet

was resuspended in 100μL of 1X TE buffer and purified using a MicroSpin column (Roche catalog#: 11814419001). Forward and reverse oligonucleotide probes (1μL of final reaction) were then PCR amplified separately using Platinum *Pfx* DNA Polymerase (Invitrogen, catalog#: 11708-013) and primers complimentary to primer binding sites on probes [42] according to the program: 95°C (4 minutes), 25 cycles of: 98° (20 seconds), 55°C (15 seconds), 72°C (15 seconds), and a final elongation step of 72°C (5 minutes), followed by 4°C (hold). Nested PCR was then performed on resultant product (3μL of 25μL previous reaction) using a forward primer encoding a 5' SP6 promoter site, retrieved from SP6 KapaBiosystems kit, using Kapa HiFi DNA polymerase ReadyMix (KapaBiosystems Code: KK2101) for 25 cycles under the same program described above. PCR product was purified on MinElute column and eluted in 30μL of nuclease free water.

Next, SP6 promoter-containing probes (1μg from column purified DNA brought up to 12μL using nuclease free water) were transcribed *in vitro* using MAXIscript SP6 *in vitro* transcription kit according to manufacturer's instructions (Ambion cat. no. AM1308), while incorporating biotin-labeled UTP (Roche catalog#: 11388908910) in a 1:4 biotinylated:non-biotinylated ratio. Nascent RNA mixture was incubated with TURBO DNase I (1μL) provided in MAXIscript SP6 *in vitro* transcription kit at 37°C for 15 minutes. DNase I digestion was stopped by adding 0.5M EDTA (1μL). Products were ethanol precipitated and resuspended in 30μL of nuclease free water before adding 2μL of RNaseIN (Promega catalog#: N2611). RNA bait pool was stored at -80°C.

2.5 Library preparation methodology

Six patient gDNA samples and one HapMap gDNA sample (GM19145) were prepared for the first sequencing run on the Illumina Genome Analyzer IIe. Genomic DNA samples (5µg) were suspended in 100µL of nuclease free water and sonicated using Covaris Ultrasonicator Model S220 at the London Regional Genomics Facility under the following settings: Time: 120 seconds, Duty Cycle: 20%, Intensity: 4, Cycles per burst: 200.

After sonication, samples were removed from the instrument and purified on a QiaQuick spin column (Qiagen, product#: 28104), eluting in 50µL of nuclease free water. Samples were imaged on a 3% agarose gel to ensure shearing, which is indicated by a smear of DNA fragments. Using Kapa Library Preparation kit (KapaBiosystems KK8201), sheared DNA was end repaired according to manufacturer's instructions to produce blunt ends. Samples were immediately purified using MinElute PCR purification kit (Qiagen product#: 28004) and eluted in 31µL of provided elution buffer. Next, addition of a 3' adenine nucleotide ("A-tailing") was performed on end-repaired DNA using Kapa Library Preparation kit; samples were purified using MinElute PCR purification kit.

Single-stranded Illumina Paired End (PE) Adapters, as well as their complements, were ordered from IDT. Sequences are shown below:

i) PE Adapter 1: 5' GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 3'

ii) PE Adapter 1 complement:

5' CTCGGCATTCCTGCTGAACCGCTCTTCCGATC 3'

iii) PE Adapter 2: 5’ ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3’

iv) PE Adapter 2 complement:

5’ AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT 3’

Adapter sequences obtained from IDT were resuspended in nuclease free water to 200μM. Complementary single-stranded adapters were annealed by mixing equal volumes of 200μM of each strand, i) and ii), and iii) and iv) separately, and incubated on a thermocycler at 98°C while ramping down 0.1°C per second, until samples reach 4°C. Products were diluted to 30μM using nuclease free water. Nascent 30uM double-stranded Illumina PE adapters (5μL) were ligated to sheared, A-tailed patient DNA using Kapa Library Preparation kit, and incubating at 20°C for 15 minutes. Samples were purified using a MinElute PCR purification kit. Following adapter ligation, samples were resolved on 2% low electroendoosmosis (EEO) agarose gel, and DNA within 200-300bp was gel extracted using Qiagen Gel Extraction kit. Extracted DNA was PCR amplified under the following parameters using Kapa Library Preparation kit: 98°C (45 seconds), followed by 15 cycles of 98°C (15 seconds), 65°C (30 seconds), 72°C (30 seconds), and finally one 72°C (60 seconds) step and holding at 4°C.

Adapter ligated, sheared patient and GM19145 samples were then hybridized to RNA probes using the using a series of solutions made with the following reagents as described in CustomArray targeted sequencing kit version 2.2: Neutralization buffer (1M Tris-HCl, pH 7.5), Wash buffer 1 (2mL 1X Saline Sodium Citrate buffer, 4mL 0.1% Sodium Dodecyl Sulfate, 34mL nuclease free water), Wash buffer 2 (0.1X Saline Sodium Citrate

buffer, 4mL 0.1% Sodium Dodecyl Sulfate (SDS), 34mL nuclease free water), Bead Binding buffer (2.3376g of 1M NaCl, 400µL of 10mM Tris HCl, 80µL of 1µM EDTA and nuclease free water up to 40mL), Denhardt's solution, and 20X Saline-sodium phosphate EDTA (SSPE) buffer.

Adapter-ligated, sheared patient and HapMap DNA solution (500ng in varying volumes depending on patient sample) was concentrated to 3.4µL using a SpeedVac RC110A. Next, Library Mix was prepared by adding 2.4µL of 1µg/µL human COT-1 DNA, 2.5µL of 1µg/µL salmon sperm DNA and 0.6µL of 25µM Paired-end PCR primers (each) to 3.4µL of patient and HapMap DNA in a thermocycler tube. A Hybridization buffer mix was prepared by combining: 25µL of 20X SSPE buffer, 1µL of 0.5M EDTA, 10µL of 50X Denhardt's solution, and 13µL of 1% SDS. Oligo Bait Library mix was prepared by mixing 5µL of 100ng/µL of biotinylated RNA bait and 1µL of RNaseIN (Promega catalog#: N2611). Seven times the volume of solutions was made above to accommodate 7 samples. To hybridize patient and HapMap DNA to biotinylated RNA bait, Library DNA mix was incubated at 95°C for 5 minutes to denature DNA. Next, samples were cooled to 65°C. When thermocycler reached 65°C, Hybridization buffer mix was added to fresh 0.2mL Eppendorf tubes placed in thermocycler. Library and Hybridization buffer mixes were incubated at 65°C for 3 minutes, when Oligo Bait mix was added to clean 0.2mL Eppendorf tubes and placed in thermocycler. After 2 minutes of incubation, Library mix and Hybridization buffer mix were added to tubes containing Oligo Bait mix, and solution was mixed by pipetting. Solution was hybridized at 65°C for 66 hours.

To capture DNA hybridized to biotinylated RNA bait, 100µL of MyOne Streptavidin coated magnetic Dynabeads T1 (Invitrogen #656-01) was placed into each of six 1.5mL Eppendorf tubes, which were washed with Bead binding buffer three times, and then suspended in 400µL of Bead binding buffer. This solution was partitioned equally into six additional clean 1.5mL tubes, for a total volume of 200µL of beads and Bead binding buffer in each tube. The Hybridization/DNA mix was added to tubes containing beads, and tubes were placed onto a nutator at room temperature for 30 minutes to promote binding of DNA:RNA hybrid to magnetic beads. Beads were then pelleted on magnetic separator (Ambion product#: AM10026) and supernatant was removed and replaced with 1mL of Wash buffer 1. Beads were briefly vortexed and incubated for 15 minutes at room temperature. Beads were separated again, and supernatant was removed and replaced with 1mL of Wash buffer 2 that was pre-warmed to 65°C. Beads were incubated for 10 minutes and briefly vortexed. Washing with Wash buffer 2 was repeated two more times. Beads were then mixed with 100µL of Elution Buffer (0.1M NaOH) and incubated for 10 minutes at room temperature. Beads were separated and supernatant containing DNA was transferred to six tubes containing 140µL of Neutralization buffer. Samples were purified using MinElute PCR purification kit, and eluted in supplied Elution Buffer.

Purified, captured DNA was then PCR amplified using Kapa HiFi DNA Polymerase (KapaBiosystems Code: KK2101) for 27 cycles. Quantitative PCR was then performed on amplified DNA to determine volume needed for concentration of 10nM required for Illumina GAIIe sequencer. This was performed using Kapa qPCR kit according to the

manufacturers instructions (KapaBiosystems Code: KK4600). Next, 5μL of 10mM captured DNA samples were sequenced on the Illumina GAIIe through 36bp, paired-end reads in addition to Illumina PhiX control v3. The PhiX control v3 is an adapter ligated DNA library generated from PhiX-174 bacteriophage genome that is used to assess cluster generation, sequencing and alignment (Illumina Technical Note: Sequencing) [71].

Seven patient genomic DNA samples were prepared for a second sequencing run according to the same protocol with a few alterations. One alteration includes the post-capture PCR amplification step, which was reduced from 27 to 24 cycles to reduce PCR amplification bias, and maximize the number of unique DNA fragments to improve sequencing coverage and accurate variant detection. Also, magnetic bead separation was performed using six-channel DynaMag Spin Magnet (Invitrogen product#: 12320D), thus reducing sample preparation time. After library preparation, samples were sequenced on the Illumina GAIIe through 36 base pair, paired-end reads in addition to PhiX control.

Finally, eight patient genomic DNA samples were prepared for the final sequencing run according to the same protocol performed during the second library preparation with one major alteration. All pre-hybridization enzymatic reactions, including: end-repair, A-tailing, and adapter ligation, were prepared in the presence of Agencourt AMPure XP beads (Beckman Coulter, item#: A63880). Agencourt AMPure beads are carboxyl-coated paramagnetic beads that remove DNA from solution using a mixture of 20% polyethyleneglycol (PEG) and 2.5M NaCl, which forces DNA molecules to bind to carboxyl group on beads, which can then be separated on a magnet [72]. Supernatant can

then be removed and replaced with the reagents of a subsequent enzymatic reaction after a brief 70% ethanol wash. Use of Agencourt AMPure beads eliminates sample loss arising from gel extraction, MinElute spin-column purification and sample transfer, resulting in fewer PCR amplification cycles needed to achieve sufficient DNA quantities for sequencing, thus reducing PCR amplification bias, generating more unique reads and higher sequence coverage. In addition to using AMPure beads during the final library preparation, samples were paired-end sequenced using two 50 base reads, instead of 36 as used in the previous two sequencing runs. Finally, Illumina GAII Sequencing Control Software (SCS) and Real Time Analysis software was upgraded to version 2.10 (from 2.9) and 1.13 respectively. This update resulted in improved image analysis, producing more high-quality data, which impacts overall base coverage and quality scores.

2.6 Sequencing data analysis

All raw base intensity data generated during sequencing was converted to DNA sequence and aligned to the human genome using CASAVA v.1.7 (first two sequencing runs) and v.1.8 (final sequencing run) software (Illumina). In version 1.7, from Intensity files produced during sequencing, base calls are generated using a program called "Bustard". From base calls, paired-end reads are aligned to the hg19 reference genome using a program called "Gerald", which is a collection of Perl scripts and C++ executables that use an alignment algorithm called, Efficient large-scale alignment of nucleotide databases version 2 (ELANDv2), to perform multiseed and gapped alignment of paired-end reads to the reference sequence in parallel (Illumina CASAVA1.7 user guide, pg 43). Where a "seed" is defined as a substring of bases within a read that is extracted and aligned to the

reference sequence. A gapped alignment is when an alignment program maps a read containing a gap within its sequence compared to the reference sequence to the reference sequence and can allow the detection of indels. Aligned sequences are converted to Sequence Alignment Map (SAM) file format in CASAVA v.1.7 using a Perl script called "export2sam", which takes read 1 and read 2 export files generated by Gerald as input and converts them to a text-based sequence alignment map (SAM) format. SAM files were then used to generate variant calls. The final sequencing run of eight patients used an updated version of CASAVA, version 1.8. In this version, several program directories were changed, in addition to the final file output, which is in Binary Alignment Map (BAM) format, as opposed to SAM format, which is a compressed, indexed binary version of SAM file. In the final sequencing run, sequences were aligned only to chromosomes containing the genes of interest on the hg19 version of the human genome, as opposed to the entire genome, to reduce computation time.

SAM and BAM files generated by CASAVA were then further processed on the Shared Hierarchical Academic Research Computing Network (SHARCNET) for variant calling using "Picard" (http://picard.sourceforge.net), a program containing a set of Java-based command-line utilities that process BAM files in preparation for variant calling by removing PCR duplicates, identifying paired-end mate pairs and realigning reads around indels, "Samtools" (http://samtools.sourceforge.net/) [73], a program containing utilities designed to index, sort and merge SAM files, and the "Genome Analysis Toolkit" (GATK) [74], a software package developed at the Broad institute for genotyping and sequencing variant detection. All SAM files generated from CASAVA v.1.7 were

converted to BAM files using the samtools "view" command in SHARCNET. From BAM files, the MarkDuplicates.jar file from Picard (http://picard.sourceforge.net/) was used to flag PCR duplicates. All flagged duplicates were disregarded during downstream variant calling. After duplicate marking, read groups specifying the sequencing date, lane and platform were added to each BAM file using the AddOrReplaceReadGroups.jar file from Picard. BAM files were then indexed using samtools index command. Indexed BAM files were then realigned around indels using the RealignerTargetCreator.jar and FixMatePair.jar files from Picard. Finally variants and indels were called using the UnifiedGenotyper tool in GATK with the GATK recommended standard call confidence value of 30 (a minimum phred base quality score to call variants), and files were exported in vcfv4.1 format. After all variants were called, only variants within target regions were examined further. Reference and alternate allele coverage values were extracted from the column 10 of the final output (corresponding to the AD value in vcf format, "DepthPerAlleleBySample").

To determine the location of the variant (i.e. intronic, exonic, upstream of gene, downstream of gene, or 5' and 3' UTR), the final variant output file was formatted for input into an online database called SNPNexus [75], which a SNP Annotation tool that determines the location of a genetic variant (i.e. exon, intron, UTR etc. in hg19 genome nomenclature), amino acid changes (when applicable) and any SNP's corresponding to that variant (dbSNP and HapMap). Variants can also be examined for putative effects on protein structure and miRNA binding sites using SNPNexus which accesses publically available mutation prediction tools such as PolyPhen [76], PolyPhen 2

(http://genetics.bwh.harvard.edu/pph2/index.shtml), and SIFT [77], which predicts the effects of amino acid substitution on protein structure and function, and miRBase [78] and TargetScanHuman[79], which predict putative miRNA binding sites by examining the presence of conserved 7-8mer sites in human UTRs matching miRNA seed region. After inputting variants and selecting individual prediction tools, such as miRBase, the output file will contain predicted effects of variants on miRNA binding. All variants within exonic regions were catalogued along with corresponding amino acid changes, and whether variant corresponds to a previously described SNP.

Reference and alternate allele coverage values were used to filter out likely false positive variant calls as a result of sequencing artifacts. Although variants with a low fraction of reads calling the variant allele may be true heterozygous variants, all variants need to fall within a confidence interval for reliable variant detection. To do this, an online binomial probability calculator was used to determine the minimum number of reads calling the alternate allele to be considered a true variant within a 95% confidence interval for variants of varying total coverage (http://stattrek.com/online-calculator/binomial.aspx).

First, all variants within each gene were examined *in silico* for putative effects on splicing using the, "Shannon Human Splicing Pipeline" [80], a high-throughput software that analyzes the effects of variants detected in large sequencing experiments on canonical splice sites using information theory, followed by the Automated Splice Site and Exon Definition Analysis (ASSEDA) server (http://ossify.sg.csd.uwo.ca), which predicts the effects of variants on splice site information content and the resulting mRNA

isoform using information theory and the exon definition model. The ASSEDA server predicts aberrant mRNA isoforms as a result of splice site inactivation or cryptic splice site activation through changes in the total amount of information content ($R_{i\,\mathrm{TOTAL}}$) of each splice site contributing to exon recognition by the spliceosome. Exons with the greatest total information content have the highest abundance within mRNA isoforms. When predicting pairs of acceptor and donor splice sites defining an exon, in addition to the sum of information content of the constitutive splice sites, a logarithmic gap surprisal function is used in the calculation of the total information content of spliced exon. The gap surprisal function is based on the observation by Robberson *et al.* (1990) [81] that long exons, although present in the genome (n=1115 known exons >1000bp), are recognized by the spliceosome more inefficiently than shorter exons [82]. Studies using transcriptome-wide distribution of exon lengths have determined the most common internal exon length (96bp), which is used to normalize the gap surprisal penalty when calculating total exon information content [82]. Putative exons with lengths deviating from the normalized exon length are penalized, and the total information content of the putative exon goes down. The exons with greater $R_{i\,\mathrm{TOTAL}}$ values are predicted to have a more predominant presence in final mRNA transcripts.

Variants were formatted to accommodate the Shannon pipeline's specifications, which includes four tab-delineated columns in a text file: chromosome (i.e. chr17), identifier, hg19 variant coordinate, and nucleotide variant (i.e. C/T). A file containing all single nucleotide variants was uploaded to the Shannon pipeline plug-in on the CLC-Bio

Genomics Workbench to analyze effects of variants on natural and cryptic donor and acceptor splice sites using the hg19 human genome as a reference.

The resultant information was then further prioritized for effects on splicing based on the following criteria: 1) variant results in ≥1 bit information content decrease in a natural splice site (corresponds to a two-fold decrease in spliceosome binding affinity) [45]. *In vitro* mRNA splicing assays, and gene expression analyses have found that a two-fold change in binding affinity is a threshold for detecting splicing and gene expression changes [46,54]. 2) Information content of affected cryptic splice site must be greater than nearest proximal natural site, and must reside within 300nt of the natural site. This is based on the observation by Robberson *et al.* (1990) that stable exons will be formed only if 3' and 5' splice sites are found in the correct orientation and are usually within 300nt of each other [81]. Therefore, although some variants >300bp from the natural splice site can produce cryptic sites with greater information content than the natural splice site, the change in exon length produced if the site were chosen would result in a total information content less than that of the natural exon due to the gap surprisal function.

Variants fitting these criteria were organized into three categories: putative leaky splicing mutations, inactivating mutations and cryptic mutations. Prioritized variants in hg19 nomenclature were examined on the ASSEDA server to predict mutant mRNA isoforms.

Variants in all probe-designed regions were examined for effects on sixty transcription factor binding sites (TFBS) using a program called, "Mutation Analysis", which

calculates changes in information content of TFBS as a result of variants, and generates an output file containing all changes TFBS changes in information content as a result of mutations. Transcription factor ChIP-seq data was downloaded from the Encyclopedia of DNA Elements (ENCODE) consortium, and filtered with ENCODE H3K27 acetylation and DNase I hypersensitivity tracks to select for only DNA regions accessible to transcription factors for binding. Transcription factors identified through ENCODE ChIP-seq studies that map to regions within the promoters of the seven genes in this study, as determined by turning on ENCODE transcription factor binding track on the UCSC Genome Browser were included in the analysis. Epigenetically-filtered ChIP-seq data was aligned using a program called Bipad [83], which performs multiple local alignments by entropy minimization to predict sequence elements within unaligned sequence data. Common sequences recognized by each individual transcription factor within a set of defined lowest entropy sequences identified using Bipad are used to build transcription factor binding site models. Sequence models are constrained by specifying that each sequence contain either one or zero binding sites per submitted sequence (ZOOPS). The probe-design regions are filtered using ENCODE H3K27 acetylation and DNase I HS in HMEC cell line (human mammary epithelial tissue) and resultant regions are scanned for putative transcription factor binding sites generated from aligned sequence models. Changes in transcription factor binding site information content due to sequence variants, is calculated using Bipad-generated position weight matrices in a program called “Mutation Analysis”.

The following criteria is used to prioritize variants for effects on transcription factor binding sites from the Mutation Analysis output data: 1) variants must occur within 10kb of transcription start site, because in addition to the core promoter located within 200bp of the transcription start site (TSS), DNA sequence clusters that insulate or enhance transcriptional activation can be located up to 100kb upstream of the TSS in humans, such as the transcriptional enhancer motif regulating *Igf2* expression in humans and mice [84,85]. Examining transcription factor ChIP-seq data from ENCODE on the UCSC genome browser reveals clusters of transcription factor binding sites located several kilobases upstream of breast cancer associated genes examined in this study, as well as within coding sequences such as *CDH1*. Therefore probing regions up to 10kb outside of the TSS is necessary for detection of potential functionally relevant TFBS altering mutations associated with breast cancer etiology. 2) Variants must occur in accessible chromatin regions in mammary tissue as determined through HMEC (mammary epithelial cell line) epigenetic data from ENCODE. 3) Variants must affect TFBS's with information content greater than one standard deviation (SD) below the mean information content of each particular TFBS. The mean TFBS information content filter prioritizes variants affecting only putative TFBS that bind transcription factors with moderate to high affinity (with respect to other TFBS of the same transcription factor). TFBS with information content less than 1 SD below the mean information content are less likely to be functionally relevant due to weak interaction with transcription factor proteins. 4) Finally, variants must result in a $\geq$1 bit information content decrease of TFBS, which corresponds to a 2-fold protein binding affinity change. Although it has been shown that the loss of transcription factor binding as a result of a mutation is proportional to the

impact of its match with the TFBS position weight matrix [86], variants that significantly alter a TFBS position weight matrix score (i.e. information content change), can still result in TF binding, as in the case of human CFTF binding [67], which is thought to be due to heterologous protein-protein interactions. Therefore although a mutation can significantly alter TFBS information content, it may still be active *in vivo* through protein-protein interactions with other TF's [67,87], thus limiting prioritization to mutations resulting in >1 bit decrease of TFBS information content may detect some functionally relevant mutations, but not all due to regulation mediated by protein-protein interactions with other TF's. Filtered variants were further prioritized based on degree of information content change, proximity to transcription start site and the information content of the TFBS prior to mutation.

Variants found within the 5' and 3' UTRs of the seven genes were examined *in silico* for potential effects on mRNA structure using an online tool called "SNPfold" [88]. SNPfold compares the base-pairing probability of each base in a mutated RNA sequence to the wild-type RNA sequence in a matrix called the partition function. Each base in the sequence is then ranked according to the degree to which it likely affects the RNA structure based on base-pairing probabilities. Then, sorted p-values are calculated for the set of input variants by dividing the base rank by the total number of bases in the sequence. Complete 5' and 3' UTR sequences of each gene containing one or more variants identified in breast cancer patients were inputted SNPfold including all variants encoded within their sequences. Variants with p-values <0.05 (in a few instances <0.10) were prioritized as variants most likely to affect the stability of the RNA structure of the

sequenced gene. Variants meeting p-value cut-offs were submitted to Alain Laederach to confirm that UTR's harboring variants should be validated by SHAPE analysis.

Variants predicted to alter RNA structure were then assessed for effects on RNA binding protein-binding sites (RBPBS) using RBPBS position weight matrices obtained from the RNA binding protein database (RBPDB) [89]. Position weight matrices were entered into the Shannon pipeline, and differences in information content of RBPBS as a result of putative RNA structure-altering UTR variants were generated. Variants predicted to result in information content changes of $\geq$2 bits in RBPBS's with initial information contents of 3 bits or greater were prioritized. The number of RNA-binding protein binding sites used in developing position weight matrices and models are relatively small, resulting in a higher standard deviation in information content. Therefore a higher threshold of information content change in RBPBS with high average information content was used as a prioritization criteria to increase variant prioritization stringency and reduce false positives.

Variants in 3' UTRs were inputted into SNPNexus while selecting for predicted effects on miRNA binding through TargetScanHuman [79] and miRBase databases [78]. Since SNPNexus does not report the location of miRNA binding sites and only if a variant is predicted to affect a binding site, the location of miRNA binding sites was determined through the miRBase Sequence Database by turning on snoRNA/miRNA tracks on the UCSC genome browser [78] and examining the 3' UTR of genes harboring variants. Searching TargetHumanScan for predicted miRNA binding sites in these genes was also

performed by inputting the gene into TargetHumanScan and mapping predicting sites to the UCSC Genome Browser. To support SNPNexus results, variant location was compared to putative miRNA binding sites in 3' UTRs of genes harboring variant.

Primers were designed to amplify regions containing prioritized variants using Primer3 for amplification of patient gDNA for variant Sanger sequencing validation. Prioritized variants from patients were amplified using Platinum *Pfx* DNA Polymerase and Kapa HiFi DNA Polymerase according to manufacturer's instructions. Products were purified using Qiagen MinElute spin columns, and sent to the London Regional Genomics Facility for Sanger sequencing.

Chapter 3 Results

3.1 Results of LHSC patient *BRCA1* and *BRCA2* variant nomenclature change and prioritization on splicing

After all patient *BRCA1* and *BRCA2* variants identified by the Molecular Diagnostics Laboratory at LHSC were converted to HGVS cDNA nomenclature, a single file describing all variants was created. A total of 158 novel variants are described (see Table 3), with 89 variants in *BRCA2* and 69 variants in *BRCA1*. A breakdown of variant location, as well as amino acid alterations is summarized in Table 2.

Table 2. Summary of *BRCA1* and *BRCA2* variants in LHSC patient cohort

Gene	No. variants in exon	No. variants in intron	No. variants in UTR	No. variants causing missense changes	No. truncating variants	No. synonymous variants	No. indels
BRCA1	45	24	0	30	1	11	5
BRCA2	75	13	1	46	2	24	4

After examining 158 variants in the BIC database, it was found that 107 are listed as of unknown clinical significance or are not reported, 40 are listed as of no clinical significance, and 11 are listed as clinically significant. All variants were converted to hg15 reference genome coordinates and analyzed with the ASSA server to examine putative effects on splicing. A total of 30 of 158 variants are predicted to result in splice site inactivation, 38 variants are predicted to strengthen cryptic splice sites, and 7, variants are predicted to weaken constitutive splice sites (resulting in leaky splicing). A literature search of all variants expressed in HGVS (cDNA and hg19 reference genome)

and BIC genomic nomenclatures as defined in Chapter 1.2, was performed using: Google Scholar™, PubMed, Web of Science Database (Thomson Reuters), Quertle™ and Google™. A table detailing all variants detected by routine sequencing converted to HGVS nomenclature for submission to the ENIGMA consortium is displayed in Table 3.

Table 3. *BRCA1* and *BRCA2* variants detected by routine sequencing at the molecular diagnostics lab.

Gene	Variant (HGVS)	Amino acid	Number of patients
BRCA1	c.42C>T	p.Val14Val	1
BRCA1	c.75C>T	p.Pro25Pro	2
BRCA1	c.80+77A>G		1
BRCA1	c.81-11delT		1
BRCA1	c.81-65G>C		1
BRCA1	c.81-6T>C		3
BRCA1	c.135-1G>T*		3
BRCA1	c.140G>A	p.Cys47Tyr	3
BRCA1	c.181T>G*	p.Cys61Gly	21
BRCA1	c.212+3A>G		5
BRCA1	c.213-11T>G*		2
BRCA1	c.288C>T	p.Asp96Asp	1
BRCA1	c.301+7G>A		1
BRCA1	c.441+36_49del		1
BRCA1	c.441+36CTT>TTC		2
BRCA1	c.442-34T>C		2
BRCA1	c.455T>C	p.Leu152Pro	3
BRCA1	c.547+8T>G		1
BRCA1	c.548-57delT		114
BRCA1	c.591C>T	p.Cys197Cys	3
BRCA1	c.594-24A>C		4
BRCA1	c.736T>G	p.Leu246Val	1
BRCA1	c.802A>G	p.Asn268Asp	1
BRCA1	c.981A>G	p.Thr327Thr	2
BRCA1	c.1067A>G	p.Gln356Arg	148
BRCA1	c.1308T>C	p.Pro336Pro	1
BRCA1	c.1487G>A	p.Arg496His	4
BRCA1	c.2050C>T	p.Pro684Ser	2
BRCA1	c.2077G>A	p.Asp693Asn	335
BRCA1	c.2082C>T	p.Ser694Ser	889
BRCA1	c.2311T>C	p.Leu771Leu	23
BRCA1	c.2315T>C	p.Val772Ala	1

BRCA1	c.2412G>C	p.Gln804His	2
BRCA1	c.2477C>A	p.Thr826Lys	1
BRCA1	c.2521C>T	p.Arg841Trp	8
BRCA1	c.2596C>T	p.Arg866Cys	1
BRCA1	c.2612C>T	p.Pro871Leu	1115
BRCA1	c.3113A>G	p.Glu1038Gly	548
BRCA1	c.3119G>A	p.Ser1040Asn	31
BRCA1	c.3296C>T	p.Pro1099Leu	4
BRCA1	c.3302G>A	p.Ser1101Asn	1
BRCA1	c.3418A>G	p.Ser1140Gly	3
BRCA1	c.3454G>A	p.Asp1152Asn	1
BRCA1	c.3548A>G	p.Lys1183Arg	525
BRCA1	c.4039A>G	p.Arg1347Gly	18
BRCA1	c.4132G>A	p.Val1378Ile	1
BRCA1	c.4308T>C	p.Ser1436Ser	555
BRCA1	c.4327C>T*	p.Arg1443X	2
BRCA1	c.4357+17A>G		1
BRCA1	c.4357+6T>C		2
BRCA1	c.4362A>G	p.Val1454Val	2
BRCA1	c.4484+61G>T		1
BRCA1	c.4535G>T	p.Ser1512Ile	12
BRCA1	c.4600G>A	p.Val1534Met	1
BRCA1	c.4654T>C	p.Tyr1552His	1
BRCA1	c.4750G>T*	p.Ala1584Ser	1
BRCA1	c.4812A>G	p.Gln1604Gln	1
BRCA1	c.4816A>G	p.Lys1606Glu	1
BRCA1	c.4837A>G	p.Ser1613Gly	596
BRCA1	c.4956G>A	p.Met1652Ile	25
BRCA1	c.4986+6T>G*		4
BRCA1	c.4987-92A>G		22
BRCA1	c.5074+3A>G*		4
BRCA1	c.5075-53 C>T		1
BRCA1	c.5083del19*		1
BRCA1	c.5152+66G>A		1
BRCA1	c.5277+48_5277+59Du		1
BRCA1	c.5333-36del510*		1
BRCA1	c.5406A>C	p.Thr1802Thr	1

BRCA2	c.-26G>A		10
BRCA2	c.3G>A*	p.Met1Ile	6
BRCA2	c.67+62T>G		2
BRCA2	c.68-7T>A		3
BRCA2	c.125A>G	p.Tyr42Cys	9
BRCA2	c.198A>G	p.Gln66Gln	2
BRCA2	c.223G>C	p.Ala75Pro	3
BRCA2	c.241T>A	p.Phe81Ile	2
BRCA2	c.387T>C	p.Asp129Asp	1
BRCA2	c.426-2A>G*		4
BRCA2	c.475+1G>T*		2
BRCA2	c.506A>G	p.Lys169Arg	2
BRCA2	c.865A>C	p.Asn289His	106
BRCA2	c.865A>G	p.Asn289Asp	1
BRCA2	c.978C>A	p.Ser326Arg	4
BRCA2	c.1151C>T	p.Ser384Phe	3
BRCA2	c.1365A>G	p.Ser455Ser	77
BRCA2	c.1395A>C	p.Val465Val	1
BRCA2	c.1514T>C	p.Ile505Thr	1
BRCA2	c.1599T>C	p.Thr533Thr	4
BRCA2	c.1698C>G	p.Thr566Thr	2
BRCA2	c.1786G>C	p.Asp596His	3
BRCA2	c.1792A>G	p.Thr598Ala	5
BRCA2	c.1909+12delT		10
BRCA2	c.1938C>T	p.Ser646Ser	3
BRCA2	c.1964C>G	p.Pro655Arg	2
BRCA2	c.2229T>C	p.His743His	76
BRCA2	c.2803G>A	p.Asp935Asn	5
BRCA2	c.2883G>A	p.Gln961Gln	9
BRCA2	c.2971A>G	p.Asn991Asp	75
BRCA2	c.3218A>G	p.Gln1073Arg	1
BRCA2	c.3264T>C	p.Pro1088Pro	2
BRCA2	c.3396A>G	p.Lys1132Lys	516
BRCA2	c.3762G>T	p.Glu1254Asp	1
BRCA2	c.3807T>C	p.Val1269Val	392
BRCA2	c.4008_4009insCATC*	p.Asp1337Ile*fs*X2	1
BRCA2	c.4046T>C	p.Ile1349Thr	1

BRCA2	c.4068G>A	p.Leu1356Leu	12
BRCA2	c.4258G>T	p.Asp1420Tyr	30
BRCA2	c.4308T>C	p.Ile1436Ile	4
BRCA2	c.4319A>G	p.Lys1440Arg	5
BRCA2	c.4530C>T	p.Pro1510Pro	3
BRCA2	c.4563A>G	p.Leu1521Leu	33
BRCA2	c.4585G>A	p.Gly1529Arg	1
BRCA2	c.4656T>C	p.Gly1552Gly	1
BRCA2	c.4672A>T	p.Ser1558Cys	2
BRCA2	c.4686A>G	p.Gln1562Gln	1
BRCA2	c.5199C>T	p.Ser1733Ser	18
BRCA2	c.5312G>A	p.Gly1771Asp	1
BRCA2	c.5418A>G	p.Glu1806Glu	1
BRCA2	c.5455C>T	p.Pro1819Ser	1
BRCA2	c.5569G>T*	p.Glu1857X	1
BRCA2	c.5744C>T	p.Thr1915Met	149
BRCA2	c.5901G>A	p.Lys1967Lys	2
BRCA2	c.6100C>T	p.Arg2034Cys	36
BRCA2	c.6317T>C	p.Leu2106Pro	2
BRCA2	c.6322C>T	p.Arg2108Cys	1
BRCA2	c.6323G>A	p.Arg2108His	6
BRCA2	c.6757C>G	p.Leu2253Val	1
BRCA2	c.6842-68T>C		1
BRCA2	c.6938-78A>T		1
BRCA2	c.7007G>A*	p.Arg2336His	6
BRCA2	c.7008-62A>G		3
BRCA2	c.7017G>C	p.Lys2339Asn	1
BRCA2	c.7242A>G	p.Ser2414Ser	419
BRCA2	c.7319A>G	p.His2440Arg	1
BRCA2	c.7435G>T	p.Asp2479Tyr	8
BRCA2	c.7469T>C	p.Ile2490Thr	2
BRCA2	c.7544C>T	p.Thr2515Ile	4
BRCA2	c.7806-14T>C		9
BRCA2	c.7992T>A	p.Ile2664Ile	4
BRCA2	c.8149G>T	p.Ala2717Ser	2
BRCA2	c.8182G>A	p.Val2728Ile	7
BRCA2	c.8567A>C	p.Glu2856Ala	6

BRCA2	c.8573A>G	p.Gln2858Arg	1
BRCA2	c.8755-66T>C		1
BRCA2	c.8795A>C	p.His2932Pro	2
BRCA2	c.8904delC*	p.Val2969Cys*fs*X7	1
BRCA2	c.8918G>A	p.Arg2973His	3
BRCA2	c.8953+98T>C		1
BRCA2	c.9088A>C	p.Thr3030Pro	1
BRCA2	c.9257-16T>C		7
BRCA2	c.9292T>C	p.Tyr3098His	1
BRCA2	c.9501+3A>T*		1
BRCA2	c.9730G>A	p.Val3244Ile	2
BRCA2	c.9976A>T	p.Lys3326X	77
BRCA2	c.10095delCinsGAATTATATCT	p.Ser3366AsnfsX4	2
BRCA2	c.10110G>A	p.Arg3370Arg	4
BRCA2	c.10234A>G	p.Ile3412Val	10

* Variants previously determined to be clinically significant, and not submitted to ENIGMA consortium

Putative splicing variants were prioritized for transfection analyses based on the following criteria: 1) the amount of information content change in the affected canonical or regulatory splice site (ΔR_i), 2) the proximity of the affected regulatory splice site to the natural canonical splice site if applicable. A study conducted by Lastella *et al.* (2004) examined 16 mutations in putative SF2/ASF motifs in exon 12 of *hMLH1* predicted using the ESEfinder. After conducting a mini-gene transfection study to examine the effects of variants in putative SF2/ASF sites, Lastella *et al.* (2004) determined that only 3 of 16 lead to aberrant splicing, all three of which affected only SF2/ASF sites located within 40 nucleotides of the acceptor and donor site [90]. As SR proteins binding ESE elements have been shown to facilitate spliceosome assembly at canonical splice sites through SR domain-mediated protein-protein interactions with the spliceosome components, the findings published by Lastella *et al.* (2004) should not be surprising [15]. Therefore

proximity to the canonical splice site is a criterion that was included when examining variants in putative ESE elements. 3) The number of canonical and regulatory splice sites affected by the mutation, and 4) the presence and proximity of a cryptic site of greater information content than an affected natural site. Based on these criteria, two mutations, *BRCA1*:c.288C>T and *BRCA2*:c.7319A>G, shown in Table 4 were selected for further analysis, both of which may affect regulatory splice sites in close proximity to canonical splice sites. Further explanation of the reasons why these variants were prioritized, as well as the putative mRNA splicing effects of these variants is described in the following paragraph.

Table 4. Prioritized putative splicing variants in LHSC patient cohort

Mutation	Location	Splice site(s) effect	R_i initial (bits)	R_i final (bits)
***BRCA1*:c.288C>T**	12 nucleotides upstream of natural donor site	Abolished SRp40 and SF2/ASF	2.6 and 6.6	-2.9 and 0.9
***BRCA2*:c.7319A>G**	117 nucleotides upstream of natural donor site	Creation of SRp40	0.7	4.6

The *BRCA1*:c.288C>T variant is located 12bp upstream of the natural donor site in exon 5 of *BRCA1* (see Figure 5). The variant is predicted to abolish putative exonic splicing enhancer SRp40 (R_i = 2.6>-2.9 bits) and SF2/ASF (R_i = 6.6>0.9 bits) binding sites, both located 14 nucleotides upstream of the natural donor site. As described in Chapter 1.3, SRp40 and SF2/ASF proteins facilitate 5' and 3' splice site recognition. Therefore abolishing splicing regulatory binding sites in close proximity to the natural donor site is

predicted to lead to reduced donor site recognition by the spliceosome, and result in exon 5 skipping, as no cryptic donor sites of greater information content than the natural donor site are found within 200 nucleotides of the natural donor. Exon 5 skipping would reduce the mRNA transcript by 84 nucleotides, and result in a frameshift mutation. The Human Splicing Finder version 2.4.1 [55] predicts an abolished SRp40 site and reduced SF2/ASF site as a result of this mutation. However, the ESEfinder version 3.0 predicts an SRp40 site at the position of the mutation, but does not predict an SF2/ASF site.

The *BRCA2*:c.7319A>G variant is located 117 nucleotides upstream of the natural donor site of exon 14 and is predicted to result in the creation of a cryptic SRp40 site (R_i = 0.7>4.6 bits) located 120 nucleotides upstream of the natural donor site of exon 14 (hg19 genomic coordinate, chr13:g.32,929,306; see Figure 5). Examining the variant region on the ASSA server uncovered a cryptic donor site (R_i=2.3 bits) located 18 nucleotides downstream of the strengthened cryptic SRp40 site (hg19 genomic coordinate, chr13:g.32,929,324). No ESE elements of greater information content are found within 40 nucleotides of the natural donor site (R_i=2.4 bits). Therefore this variant creates a putative cryptic SRp40 ESE site in close proximity to a cryptic donor site of comparable information content to the natural donor site with no nearby ESE elements of information content greater than or equal to the cryptic SRp40 site produced by the variant. If the cryptic donor site is chosen, exon 14 could be truncated from 427 nucleotides to 326 nucleotides, and thus result in a frameshift. The Human Splicing Finder predicts creation of a cryptic SF2/ASF site four nucleotides upstream of the variant (final score=76.70), but does not predict the creation of a cryptic SRp40 site.

The *BRCA1*:c.548-16G>A variant described by Mucaki *et al.* (2011) [46] from the BIC database deposited by Myriad Genetics Inc. is located 16 nucleotides upstream of exon 8 in *BRCA1* (see Figure 5). ASSA server analysis predicted natural acceptor splice site weakening (R_i=0.5>0.2 bits), and the creation of a cryptic acceptor site 14 nucleotides upstream of the first nucleotide in exon 8 (R_i =-2.8>5.4 bits). If the cryptic acceptor site is activated, it is predicted to increase the length of exon 8 from 46 to 60 nucleotides and result in a frameshift mutation. These results are consistent with results obtained from the Human Splicing Finder, which predicts the creation of a strong cryptic acceptor site (initial score=54.62, final weight score=83.57) that if used would increase exon length by 14 nucleotides. A schematic of the effects and location of prioritized variants on canonical and regulatory splice site information content is shown in Figure 5.

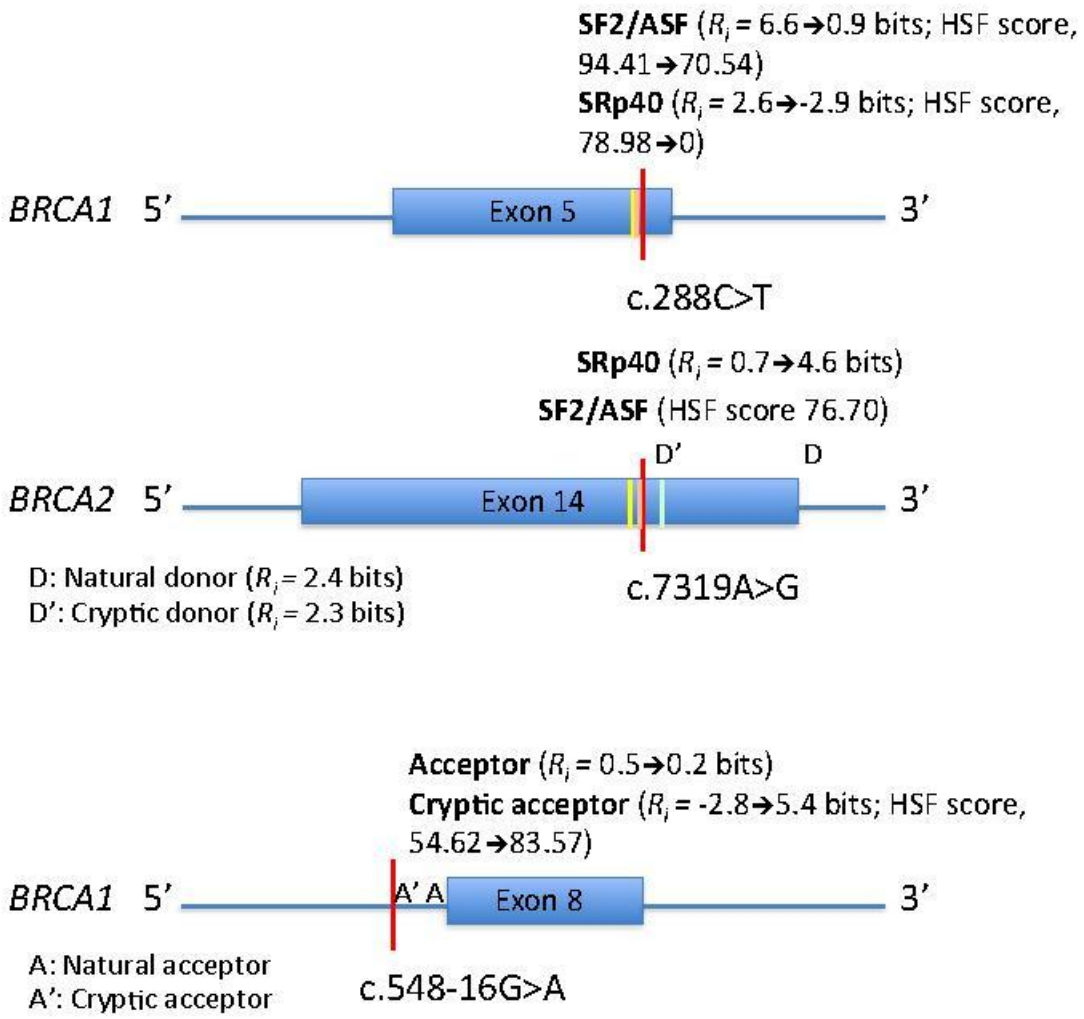

Figure 5. **Schematic of effects and location of prioritized variants on canonical and regulatory splice site information content**. Variants are marked by a long red, vertical line. In the case of the *BRCA2*:c.7319A>G variant, cryptic donor site location is marked by a green line below D'. SRp40 sites are marked in orange (vertical line most proximal to variant in both cases), and SF2/ASF sites are marked in yellow (vertical line distal to variant in both cases).

In addition to the variants prioritized for *ex vivo* transfection analysis, putative leaky splicing variants were also identified using the ASSA server, and are displayed in Table 5. Variant rsID, alternate allele population frequency as well as clinical significance (BIC) is also reported in the table. A total of 15 patients harbor the putative leaky splicing variants. Pedigree information was obtained from LHSC to examine patient family history of HBOC. The average number of 1^{st}, 2^{nd} and 3^{rd} degree relatives with HBOC for patients with each putative splicing variant is displayed in Table 5.

Table 5. Breast cancer family history of patients with putative leaky splicing variants LHSC

Gene	Variant (HGVS)	R_i initial (bits)	R_i final (bits)	ΔR_i	Type	BIC clinically significant (yes/no/ unknown)	rsID (dbSNP)	Alternate allele population frequency (%)	No. patients with variant	Average no. 1st degree relatives with HBOC	Average no. 2nd degree relatives with HBOC	Average no. 3rd degree relatives with HBOC
BRCA2	c.68-7T>A	5.25	3.15	-2.1	Acceptor	unknown	rs81002830	0.274	2	1.5	3.5	0
BRCA2	c.9501+3A>T	9.04	4.55	-4.49	Donor	unknown	rs61757642	-	4	1.25	0.75	1
BRCA1	c.4986+6T>G	3.58	2.25	-1.33	Donor	yes	rs80358086	-	1	2	1	0
BRCA1	c.4357+6T>C	8.28	6.56	-1.72	Donor	unknown	rs80358143	-	1	1	1	2
BRCA2	c.7007G>A	5.51	2.50	-3.01	Donor	yes	rs28897743	1.091	4	0.75	0.75	1
BRCA1	c.5406A>C	6.05	4.15	-1.9	Donor	-	-	-	1	1	0	0
BRCA1	c.591C>T	10.4	8.8	-1.6	Donor	unknown	rs1799965	0.074	1	1	0	2

3.2 Estimation of number of breast cancer families with *BRCA1* and *BRCA2* variants using a Power Analysis

Previous literature has suggested a higher prevalence of deleterious *BRCA1* and *BRCA2* variants in patients with family history of breast cancer compared to the LHSC patient cohort (57-100% vs. 3.6-35.7%) [27]. After publishing and prioritizing *BRCA1* and *BRCA2* variants detected through routine sequencing on splicing *in silico*, a power analysis was conducted to estimate the number of patients with undetected deleterious *BRCA1* and *BRCA2* mutations with 80% power and $p \leq 0.05$. The number of deleterious variants detected in patients with varying family history of breast cancer as described in column 2 of Table 6 was compared with published *BRCA1* and *BRCA2* linkage frequency in patients with familial breast cancer [26], as described in Chapter 2.2. Proband pedigree information was only available for patients within certain family history groups (Groups: 5-10). Results indicate that the fraction of undetected *BRCA1* and *BRCA2* pathogenic variants across all groups ranges from 0.364 to 0.78, and the predicted minimum sample size needed to detect a *BRCA1* and *BRCA2* variant across all groups is 13 (Groups 5 and 10). Therefore to improve deleterious variant detection, whole gene sequencing should be employed to probe regions unexamined during routine sequencing.

Table 6. Estimation of number of breast cancer families with *BRCA1* and *BRCA2* variants using a Power Analysis

Group	Definition	No. Eligible Patients	Estimation of fraction of families with *BRCA1* and *BRCA2* variants	Fraction of Detected *BRCA1* and *BRCA2* variants	Fraction of undetected variants	Estimated no. patients with undetected mutations	Minimum sample size for *BRCA1* and *BRCA2* variant detection*
5	Breast cancer under 60, and a 1st or 2nd degree relative with ovarian or male breast cancer	65	0.91	0.13	0.78	51	13
6	Breast and ovarian cancer in the same individual, or bilateral breast cancer with first case before 50	37	0.57	0.206	0.364	14	24
7	Two cases of breast cancer, both before 50, in 1st or 2nd degree relatives	93	0.7173	0.133	0.5843	55	16
8	Two cases of ovarian cancer, any age, if 1st and 2nd degree relatives	4	1	0.357	0.643	3	15
10	Three or more cases of breast or ovarian cancer at any age on the same side of the family	213	0.784	0.036	0.748	159	13
Total no. family pedigrees		412					

* Using "fraction undetected" (Column 6) as an effect size

3.3 Mini-gene transfection assays of prioritized putative *BRCA1* and *BRCA2* splicing variants

After prioritizing variants of unknown significance for effects on splicing using information theory-based analysis and the Human Splicing Finder, a recombinant mini-gene transfection assay was designed as described in Chapter 2.1.2-2.1.6 to examine the effects of *BRCA1*:c.288C>T, *BRCA2*:c.7319A>G and *BRCA1*:c.548-16G>A variants on mRNA splicing. Variants producing splicing changes predicted in Chapter 3.1 (Figure 5) will have the following corresponding effects on this mini-gene transfection assay: 1) *BRCA1*:c.288C>T would result in pcDNA-Dup mini-gene cassette exon skipping, whereas WT *BRCA1*:c.288C would result in cassette-exon inclusion, 2) *BRCA2*:c.7319A>G would result in pCAS2 mini-gene middle exon truncation from 427bp to 326bp, whereas mini-gene middle exon containing WT *BRCA2*c.7319A would remain 427bp. Finally, 3) *BRCA1:*c.548-16G>A would result in pCAS2 mini-gene middle exon elongation from 46bp to 60bp, whereas mini-gene middle exon containing WT *BRCA1*:c.548-16G would remain 46bp. After transfection of recombinant mini-gene containing plasmids into MDA-MB-231 cells, total RNA was extracted, treated with DNase I and reverse transcribed. Expected amplicon sizes of all resultant RT-PCR products are given when each experiment is described.

cDNA products corresponding to WT *BRCA1*:c.548-16G>A and *BRCA2*:c.7319A>G were PCR amplified for 30 cycles using primers specific to the first and last exon of pCAS2 mini-gene (Table 1, primer identifier numbers 5 and 6) as recommended by INSERM. PCR products were electrophoresed on 2% agarose gel. Expected amplicon

size of *BRCA1*:c.548-16G WT mini-gene cDNA is 280bp and expected amplicon size of *BRCA2*:c.7319A WT mini-gene cDNA is 663bp. Non-specific, low molecular weight fragments of ~100bp were observed in all conditions corresponding to both *BRCA1*:c.548-16G and *BRCA2*:c.7319A WT cDNA amplified both with and without 1M betaine at 57°C and 55°C annealing temperatures. DNA fragments corresponding to mini-gene specific cDNA were not observed.

pcDNA-Dup cDNA (empty pcDNA-Dup, WT *BRCA1*:c.288C>T, mutant *BRCA1*:c.288C>T) was PCR amplified using primers specific to mini-gene insert as given by collaborators at INSERM and described in Chapter 2.1.6. PCR products were electrophoresed on a 2.5% agarose gel, shown in Figure 6. The expected amplicon size of pcDNA-Dup mini-gene containing cassette exon is 304bp, and the expected amplicon size of mini-gene not containing cassette exon is 246bp. DNA fragments greater than 500bp were observed in all conditions. Faint DNA fragments at sizes between 300bp and 400bp are also observed in all conditions. This may be an artifact corresponding to single stranded DNA derived from non-duplex mini-gene amplicon. Investigating the identity of fragments >500bp revealed that the expected amplicon sizes of WT, mutant and empty pcDNA-Dup containing both exons and introns are 567bp, 567bp and 559bp respectively, which are approximately the sizes of fragments observed in Figure 6. One explanation for the detection of fragments containing both intronic and exonic sequences is PCR amplification of cDNA derived from unspliced heteronuclear RNA transcribed from pcDNA-Dup mini-gene. A more detailed explanation of this is discussed in Chapter 4.2. Another possible explanation that was investigated is plasmid DNA contamination.

To investigate the role of plasmid DNA contamination and DNase I digestion efficiency in amplification of mini-gene cDNA containing introns (Figure 6), varying concentrations of stock pcDNA-Dup plasmid DNA were digested with DNase I using the same procedure used on RNA samples, as described in Chapter 2.1.6. Products of DNase I digestion were PCR amplified for 35 cycles using the same primers used to amplify pcDNA-Dup mini-gene cDNA. If residual plasmid DNA remained after treatment with DNase I, the expected amplicon size of PCR product is 559bp. If DNase I digestion ran to completion, then no template would be present for subsequent PCR reaction, and no fragments would be visualized after gel electrophoresis. pcDNA-Dup plasmid DNA, DNase I treated pcDNA-Dup plasmid, and PCR amplification products of DNase I treated pcDNA-Dup plasmid were electrophoresed on a 1.5% agarose gel. DNA fragments corresponding to plasmid DNA disappeared in all plasmid template concentration conditions after DNase I digestion (see Figure 7, Lane 2, 5, 8 and 11). However when DNase I treated plasmids were PCR amplified using mini-gene specific primers, faint fragments corresponding to the size of pcDNA-Dup mini-gene amplicon, as observed previously in Figure 10, were observed in all conditions (see Figure 7, Lane 3, 6, 9 and 12), indicating residual plasmid DNA remained after DNase I digestion, which was subsequently PCR amplified and detected using gel electrophoresis.

A second PCR amplification on DNase I digested pcDNA-Dup plasmid using the same mini-gene specific primers was performed for 25 cycles, products were electrophoresed on a 2.5% gel alongside products from 35 cycles amplification (see Figure 8). This

shows that mini-gene specific PCR amplicon derived from DNase I digested pcDNA-Dup plasmid was detectable at fewer PCR amplification cycles than performed on cDNA derived from RNA extracted from MDA-MB-231 cells. This indicates that it may be possible that PCR amplification of mini-gene derived from plasmid DNA contamination account for fragments observed above 500bp in Figure 10. However, fragments corresponding to PCR amplified, DNase I treated pcDNA-Dup in Figure 7 and 8 are very faint after 35 and 25 cycles of amplification compared to plasmid DNA loaded in known concentrations shown in Figure 7. Therefore, although plasmid DNA may account for residual amounts of fragments observed in Figure 6, the identity of these fragments are likely due to RT-PCR of unspliced hnRNA template.

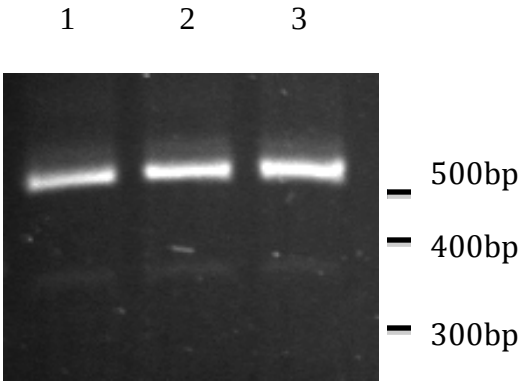

Figure 6. **RT-PCR of DNase I treated pcDNA-Dup mini-gene RNA extracted from transfected MDA-MB-231 cells using primers recommended by INSERM**. Numbers above all gel images presented in this section indicate lane number, and numbers beside images indicate DNA molecular weight marker. Lane 1 (empty pcDNA-Dup, expected amplicon size without cassette exon: 246bp), Lane 2 (WT *BRCA1*:c.288C>T, expected amplicon size with cassette exon: 304 bp), Lane 3 (mutant *BRCA1*:c.288C>T, expected amplicon size without cassette exon: 246 bp). Fragments above 500 bp were observed in all conditions. Faint DNA fragments at sizes between 300bp and 400bp are also observed in all conditions. This may be an artifact corresponding to single stranded DNA derived from non-duplex mini-gene amplicon. The expected amplicon sizes of WT, mutant and empty pcDNA-Dup mini-gene containing both exons and introns are 567bp, 567bp and 559bp respectively.

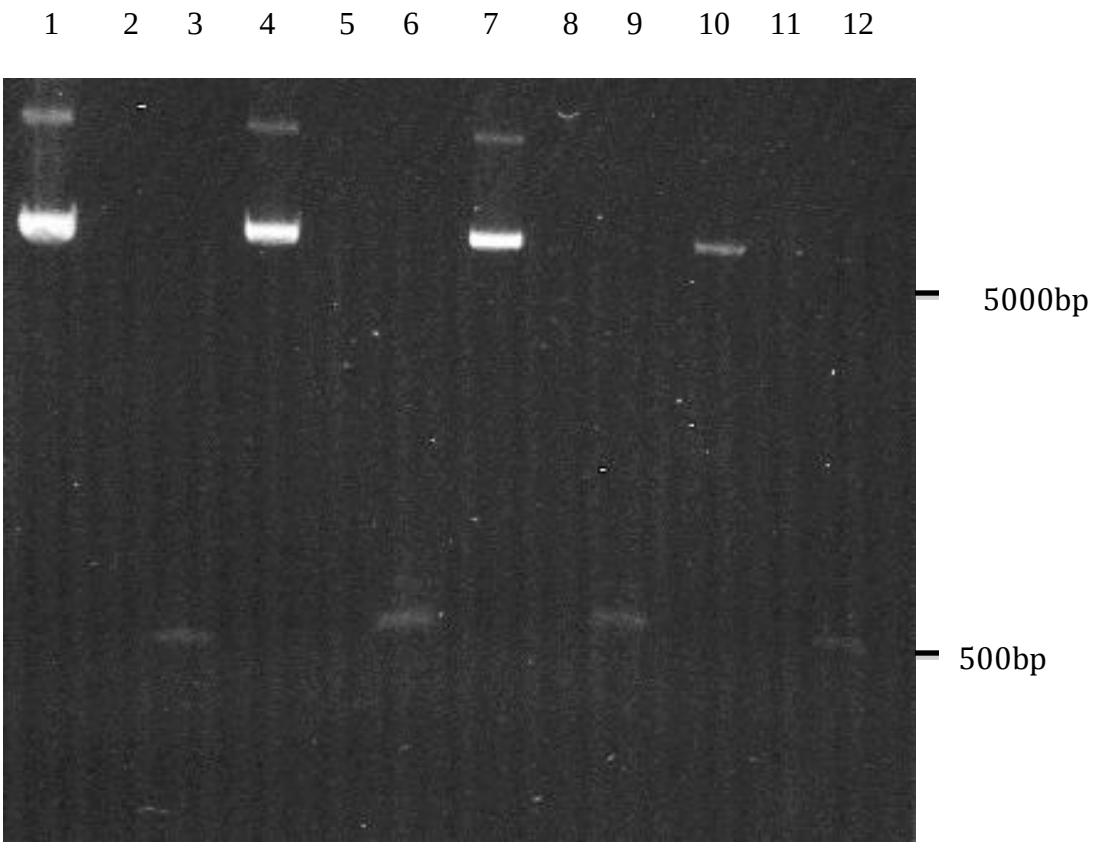

Figure 7. DNase I digestion of stock pcDNA-Dup plasmid and subsequent PCR amplification. pcDNA-Dup plasmid of varying concentrations was electrophoresed on a 1.5% gel (Lane 1: 440ng, Lane 4: 220ng, Lane 7: 110ng, Lane 10: 20ng). The same concentration of pcDNA-Dup plasmid was treated with DNase I as described in Chapter 2.1.6 (Lane 2: 440ng, Lane 5: 220ng, Lane 8: 110ng, Lane 11: 20ng). DNase I digested plasmid DNA of each concentration was PCR amplified using primers specific to mini-gene insert for 35 cycles (Lane 3: 440ng, Lane 6: 220ng, Lane 9: 110ng, Lane 12: 20ng), fragments corresponding to the expected size of mini-gene amplicon were observed in all conditions (559 bp).

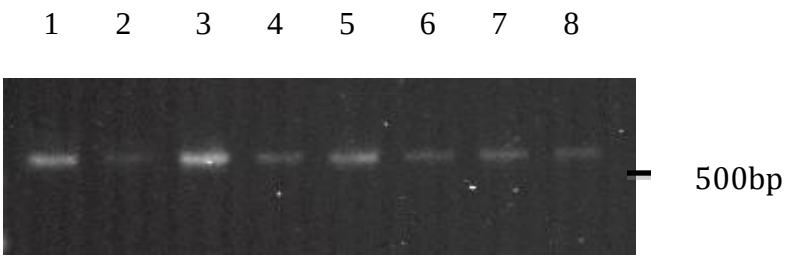

Figure 8. 35 and 25 cycle PCR amplification of pcDNA-Dup mini-gene following DNase I digestion. Varying concentrations of pcDNA-Dup plasmid were digested with DNase I and PCR amplified for 35 and 25 cycles using primers specific to mini-gene. Fragments corresponding to the expected size of mini-gene amplicon (559 bp) were observed in all conditions in both 35 cycle and 25 cycle PCR amplification (Lane 1: 440ng template, 35 cycles, Lane 2: 440ng template, 25 cycles, Lane 3: 220ng template, 35 cycles, Lane 4: 220ng template, 25 cycles, Lane 5: 110ng template, 35 cycles, Lane 6: 110ng template, 25 cycles, Lane 7: 20ng template, 35 cycles, Lane 8: 20ng template, 25 cycles).

Following the PCR amplification reactions described previously, primers were designed to span exon-exon junctions to selectively amplify spliced mini-gene cDNA. PCR amplification of WT and mutant mini-gene transcripts corresponding to *BRCA2*:c.7319A>G was conducted as described in Chapter 2.1.6. Expected amplicon sizes for *BRCA2*:c.7319A>G pCAS2 mini-gene cDNA: containing middle exon is 610bp, and with hypothesized truncated middle exon is 667 bp. Products were resolved on a 2% agarose gel. Faint low molecular weight DNA fragments were observed in *BRCA2*:c.7319A>G WT and mutant conditions amplified using primers designed to amplify mini-gene with no/aberrant middle exon (primer identifiers 9 and 10 in Table 1). Mini-gene specific DNA fragments were not observed in any conditions. No fragments were observed in both WT and mutant conditions amplified using primers specific to mini-gene with middle exon (primer identifiers 11 and 12 in Table 1). Following this, PCR amplification was conducted at lower annealing temperatures of 53°C and 54°C for primers 11 and 12 and 55°C and 56°C for primers 9 and 10. Products were resolved on a 2% gel. Low molecular weight DNA fragments of the same size were observed in all conditions. However, mini-gene specific DNA fragments were not observed in any conditions.

Next, primers specific to amplify pcDNA-Dup containing cassette exon (primer identifier numbers 3 and 4 in Table 1) and not containing cassette exon (primer identifier numbers 1 and 2 in Table 1) were used to PCR amplify pcDNA-Dup mini-gene cDNA. Products were resolved on a 2.5% agarose gel, as shown in Figure 9. cDNA fragments appearing at sizes corresponding to both cassette exon inclusion and exclusion in both

BRCA1:c.288C>T WT and mutant conditions were observed after resolving PCR products on a 2% agarose gel, as shown in Figure 9. Therefore mini-gene transcripts containing both WT and mutant *BRCA1*:c.288C>T appear to undergo leaky splicing in MDA-MB-231 breast cancer cell line.

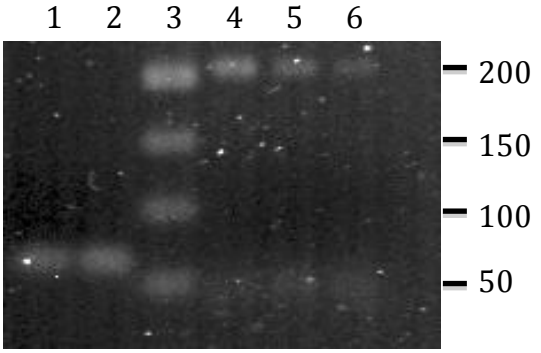

Figure 9. PCR amplification of empty pcDNA-Dup, *BRCA1*:c.288C>T WT and mutant cDNA. *BRCA1*:c.288C WT and *BRCA1*:c.288C>T Mutant mini-gene cDNA was PCR amplified using primers specific to amplify mini-gene containing cassette exon (Lane 1 and 2 respectively). Empty pcDNA-Dup, *BRCA1*:c.288C WT and *BRCA1*:c.288C>T Mutant mini-gene cDNA was also PCR amplified using primers specific to mini-gene not containing cassette exon (Lane 4, 5 and 6 respectively). Fragments appear at expected size of 63bp in WT and mutant condition (Lane 1 and 2 respectively). Fragments also appear at expected size of 209bp in WT, mutant and empty pcDNA-Dup conditions (Lanes 4, 5 and 6). A 50bp ladder (Lane 3) was used to approximate fragment sizes. PCR and gel electrophoresis indicates both WT and mutant conditions show partial cassette exon skipping.

Cassette exon exclusion of WT pcDNA-Dup cDNA (Figure 9) and initial RT-PCR fragments corresponding to mini-gene template containing intronic and exonic sequences (Figure 6), suggest splicing of transfected pcDNA-Dup plasmid is not efficient (see Chapter 4.2 for discussion). Therefore the hypothesis that *BRCA1*:c.288C>T leads to cassette exon skipping as a result of abolishing SRp40 and SF2/ASF ESE elements can neither be supported nor rejected based on these results.

cDNA corresponding to WT *BRCA2*:c.7319A>G and *BRCA1*:c.548-16G>A at 3:1 transfection ratios were PCR amplified as described in Chapter 2.1.6. Products were electrophoresed on a 2% agarose gel. Expected amplicon size for *BRCA1*:c.548-16G>A mini-gene cDNA containing middle exon is 220bp, not containing middle exon is 280bp. Expected amplicon size for *BRCA2*:c.7319A>G cDNA containing middle exon is 610bp, not containing middle exon is 280bp, and containing truncated middle exon is 667bp. Faint low molecular weight fragments were observed in all conditions. Fragments corresponding to mini-gene specific cDNA were not observed in any conditions.

Before the preceding RT-PCR experiments were conducted, DNase I treated reverse transcribed RNA extracted from all 1.5:1 transfection ratio WT and mutant conditions (in the presence of puromycin) as well as control HapMap cDNA from a previous study, were PCR amplified using primers specific to cDNA derived from stable gene *VSP39* to ensure RNA extraction and reverse transcription procedures were successful. Results are shown in Figure 10. DNA fragments corresponding to expected amplicon sizes of 103bp were observed in all conditions, Lane 7 corresponding to *BRCA1*:c.548-16G>A mutant

condition intensity is very faint. An additional control will be required to verify that RNA extraction and RT-PCR was successful in this condition. This is described in the following paragraph. Another PCR reaction designed to amplify VSP39 cDNA under the same conditions was conducted on 1.5:1 transfection ratio pcDNA-Dup plasmid conditions (with puromycin) including empty pcDNA-Dup, *BRCA1*:c.288C WT and *BRCA1*:c.288C>T mutant because amplification data for *BRCA1*:c.288C WT condition is not presented in Figure 10 (see Figure 11). Presence of fragments at expected sizes indicate that VSP39 is spliced in MDA-MB-231 cells and RNA extraction and RT-PCR protocols were successful.

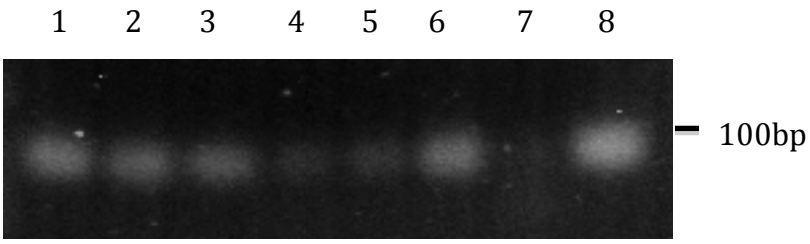

Figure 10. VSP39 cDNA amplification of 1.5:1 transfection ratio samples and control HapMap (GM19200) cDNA. PCR products of cDNA corresponding to the following transfection conditions: empty pCAS2 (Lane 1), *BRCA2*:c.7319A WT (Lane 2), *BRCA2*:c.7319A>G Mutant (Lane 3), empty pcDNA-Dup (Lane 4), *BRCA1*:c.288C>T Mutant (Lane 5), *BRCA1*:c.548-16G WT (Lane 6), *BRCA1*:c.548-16G>A Mutant (Lane 7), GM19200 (Lane 8) was electrophoresed on a 2% agarose gel. Fragments corresponding to expected amplicon sizes of 103bp were observed in all conditions, Lane 7 corresponding to *BRCA1*:c.548-16G>A Mutant condition intensity is faint..

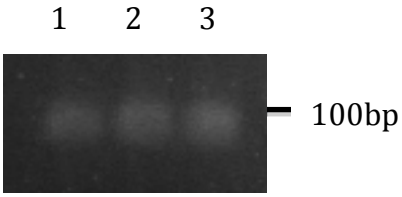

Figure 11. VSP39 cDNA amplification of 1.5:1 transfection ratio pcDNA-Dup samples. PCR products of cDNA corresponding to empty pcDNA-Dup (Lane 1), *BRCA1*:c.288C WT (Lane 2) and *BRCA1*:c.288C>T Mutant (Lane 3) was electrophoresed on a 2% agarose gel. Fragments corresponding to expected amplicon sizes of 103bp were observed in all conditions.

cDNA from an additional endogenously expressed gene, *BRCA2*, was PCR amplified to confirm the success of RNA extraction and RT-PCR. DNase I treated reverse transcribed RNA extracted from all 1.5:1 transfection ratio WT and mutant conditions (in the presence of puromycin) was PCR amplified using primers spanning the BRCA2 exon11-12 junction as described in Chapter 2.1.6. The expected amplicon size for these products is 96bp. PCR products were electrophoresed on a 2% agarose gel, shown in Figure 12. Fragments of the expected size are seen in all conditions, indicating splicing of BRCA2 mRNA in MDA-MB-231 cells, and successful RNA extraction and reverse-transcription in each condition.

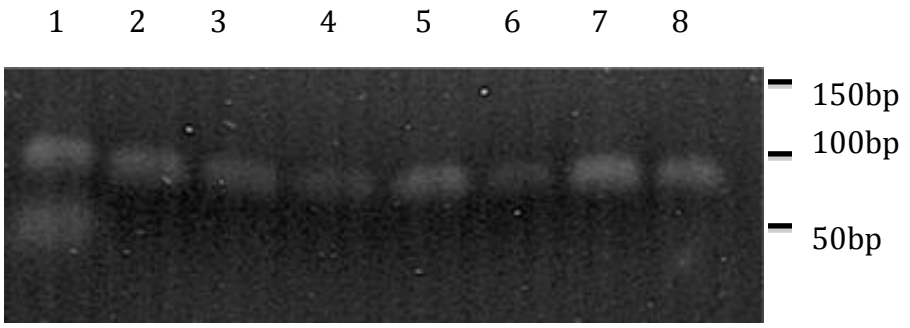

Figure 12. BRCA2 cDNA amplification of 1.5:1 transfection ratio (with puromycin). PCR products of cDNA corresponding to the following transfection conditions: empty pcDNA-Dup (Lane 1), empty pCAS2 (Lane 2), *BRCA1*:c.288C WT (Lane 3), *BRCA1*:c.288C>T Mutant (Lane 4), *BRCA1*:c.548-16G WT (Lane 5), *BRCA1*:c.548-16G>A Mutant (Lane 6), *BRCA2*:c.7319A WT (Lane 7), *BRCA2*:c.7319A>G Mutant (Lane 8) was electrophoresed on a 2% agarose gel. A 50bp DNA ladder was loaded into Lane 2, 50bp and 100bp markers are annotated. DNA fragments of varying intensities are observed at expected size of 96 bp. A second, non-specific PCR product above the 50bp band on the ladder is observed in the empty pcDNA-Dup condition.

3.4 Capture Array design and synthesis

Twenty-four thousand, 36-70mer oligonucleotide probes were designed to capture seven breast cancer associated genes and adjacent intergenic regions for downstream sequencing. Tiling-array tracks displaying designed probes targeting regions of interest are displayed in Figure 13. Gaps in coverage are non-single copy regions, as determined using the *ab initio* method described in Chapter 2.4, and include interspersed repetitive regions such as: short interspersed nuclear elements (SINE), long interspersed nuclear elements (LINE), microsatellites, long terminal repeats (LTR) etc. Table 7 displays the percentage of each gene covered by probes.

Table 7. Percent probe coverage across each gene

Gene	Percent coverage across gene and 10kb intergenic region (%)	Percent exon coverage (%)	Percent UTR coverage (%)	Percent intronic coverage (%)	Percent intergenic coverage (%)
ATM	60.7	97.1	73.9	56.7	64.4
BRCA1	47.5	96.1	72.1	27.6	50.9
BRCA2	66.6	97.4	79.1	59.6	75.4
CDH1	47.3	99.3	61.8	47.9	35.6
CHEK2	32.6	71.8	29.7	31.8	62.3
PALB2	38.4	93.9	86.9	39.1	26.2
TP53	49.0	91.9	78.7	37.2	53.0

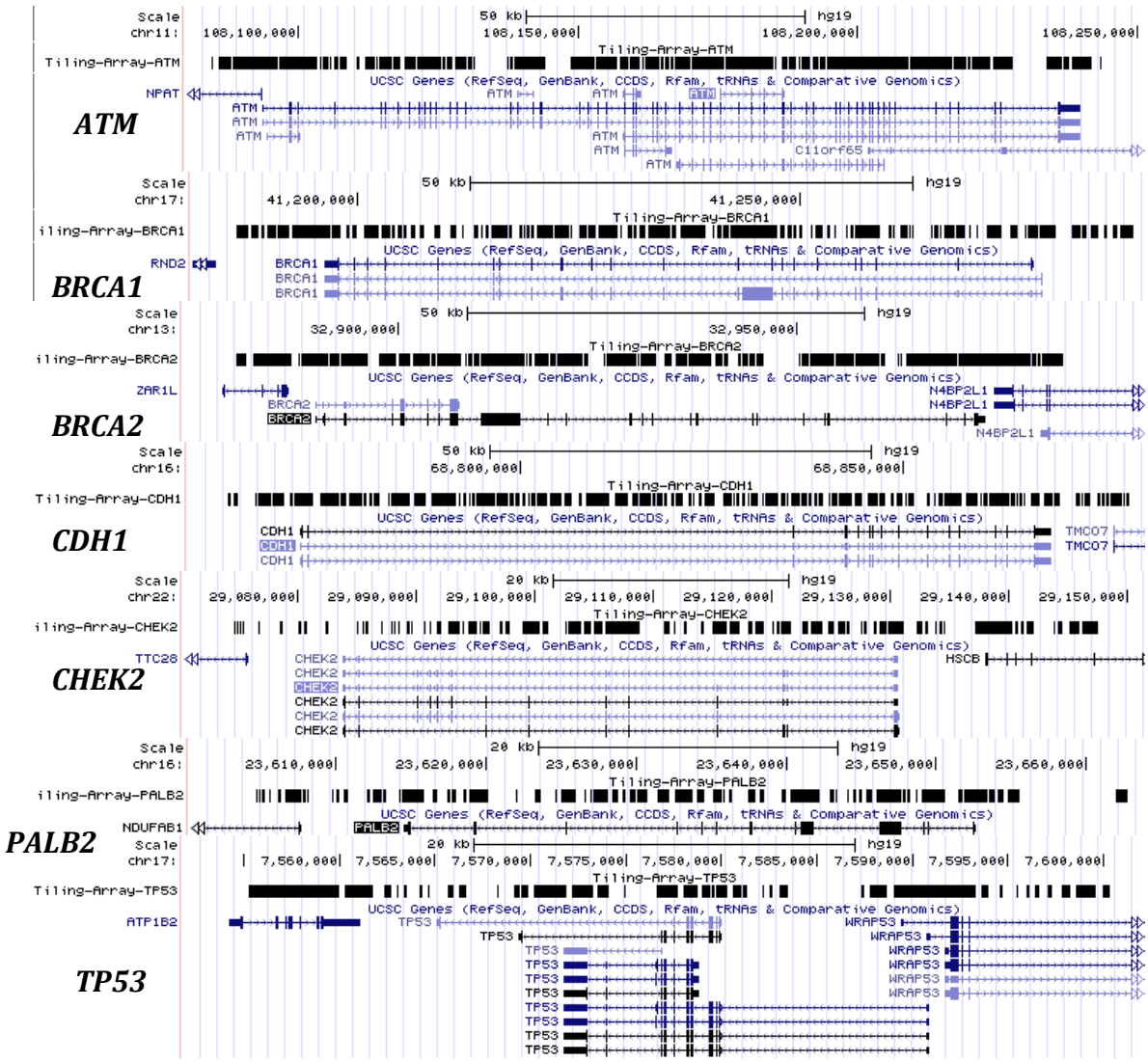

Figure 13. Target-enrichment probe tiling arrays. BED tracks showing probes designed to target seven breast cancer associated genes and 10kb of adjacent intergenic regions (shown in black bars above gene). Gaps in coverage are non-single copy regions.

A total of 7869 variants were called in all 21 sequenced patients using Picard and GATK after using binomial calculator to determine putative false positive variants as described in Chapter 2.6. When accounting for common variants detected in multiple patients, a total of 2871 unique variants were detected, 1504 of which are previously identified SNPs (dbSNP and HapMap). Variants were found throughout each gene and adjacent intergenic regions ($\pm$10 kbp), as shown in Table 8.

Table 8. Variant location and average base coverage

Gene	Upstream of Gene	5' UTR	Intron	Exon	3' UTR	Downstream of Gene	Avg. base coverage per variant
ATM	106	12	1299	26	38	67	32.8
BRCA1	131	1	749	68	19	124	32.6
BRCA2	202	6	958	83	19	135	35.5
CDH1	284	0	1554	13	18	89	36.2
CHEK2	87	0	762	71	0	56	23.5
PALB2	36	0	177	7	0	58	25.4
TP53	90	1	174	25	14	350	42.2

Each of three separate sequencing runs achieved higher variant coverage than the run previous, with the first sequencing run attaining an average of 4.46x variant coverage, the second, 15.3x variant coverage and third, 47.93x variant base coverage. After all variants were identified and filtered to exclude putative false positives, coding and adjacent intronic SNP's detected by the Molecular Diagnostics Laboratory were examined in our targeted sequencing data to determine the number of corroborated SNPs. The results are shown in Table 9. In the first sequencing run, 1/22 *BRCA1* and *BRCA2* SNPs identified by the Molecular Diagnostics Laboratory were detected. The reason for low concordancy is inadequate base coverage. Although coverage improved in the second sequencing run,

still only 15/49 SNPs were validated. However, sequencing DNA samples prepared using Agencourt magnetic beads in the final run, as described in Chapter 2.5, validated 81/81 SNPs, thus highlighting the significance sample loss resulting from gel extraction, column purification and sample transfer on sequencing coverage, and reliable variant detection. This increase in coverage can also be attributed to longer sequencing reads (from 36>50bp paired-end reads) and an Illumina imaging software upgrade.

A second BED track displaying captured and sequenced regions was created is displayed below each probe design tiling-array track in Figure 14. Sequencing reads aligned to greater than 99% of regions containing probes. However, sequenced regions aligned to non-target areas as well, including repetitive regions in the targeted genes due to capture of sequences spanning both single copy regions hybridizing to probe, and repetitive regions that are paired-end sequenced and aligned to both on-target single-copy and off-target repetitive regions. In addition, some reads aligned to regions outside of the targeted areas. One potential cause of this is what has been termed, "daisy-chaining", and is the result of repetitive regions on a DNA fragment encoding a probe-specific single copy sequence binding to a complementary repetitive sequence on a DNA fragment that is off-target [91]. Although COT-1 DNA has been used during hybridization to compete with off-target repetitive sequences for binding on-target DNA fragments harboring repeats, off-target repetitive regions binding to on-target DNA cannot be prevented with 100% efficiency. Additionally, DNA fragments containing segments of segmental duplicons mapping to both on-target and off-target regions may bind probes with high enough affinity to be captured and sequenced. Also, post-hybridization wash

steps are designed to be stringent enough to prevent washing away DNA fragments bound to RNA probes with high affinity, but not so stringent as to wash away on-target DNA fragments binding with moderate affinity. Therefore, it is possible that residual amounts of off-target, non-specifically bound DNA is retained during wash steps.

Table 9. Validation of *BRCA1* and *BRCA2* SNPs detected by Molecular Diagnostics Laboratory.

***BRCA1* region**	**Intron 8**	**Exon 11**	**Exon 13**	**Exon 16**	**Intron 16**	**Intron 17**	**Exon 18**	**Intron 18**		
Variants validated	0/2	38/54	7/10	7/11	0/4	0/1	1/1	1/1		
***BRCA2* region**	**5' UTR**	**Exon 3**	**Intron 4**	**Exon 10**	**Intron 11**	**Exon 11**	**Exon 14**	**Intron 14**	**Intron 16**	**Exon 22**
Variants validated	0/1	1/1	0/1	11/16	0/1	20/32	2/2	0/1	8/13	1/1

*First sequencing run: 1/22 *BRCA1* and *BRCA2* SNPs previously detected by the Molecular Diagnostics Laboratory were validated; second sequencing run: 15/49 SNPs validated; third sequencing run: 81/81 SNPs validated.

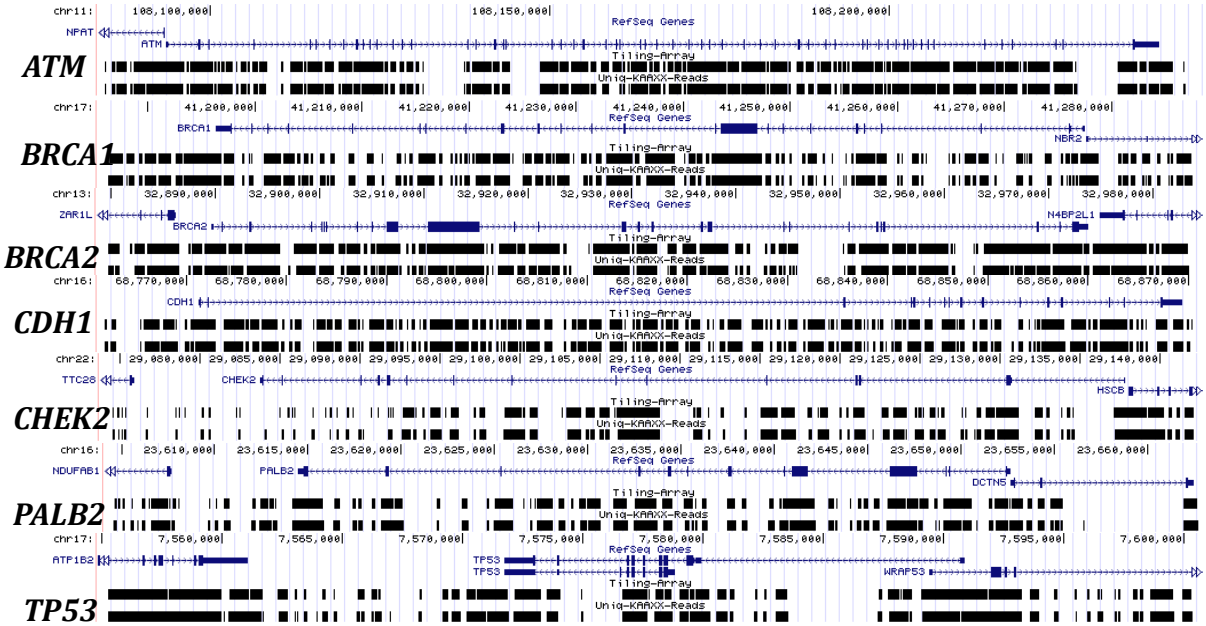

Figure 14. Target enrichment probe tiling array and captured sequence tracks. Tracks of seven breast cancer associated genes (top track), tiling-array probe regions (second track), and captured and sequenced DNA (third track).

Identified variants were analyzed for effects on: splicing, TFBS, and 5' and 3' UTR structure and RBPBS, amino acids and miRNA binding sites. All variants were inputted into the Shannon pipeline to examine putative effects on canonical splice sites. Of 7869 variants submitted, 2052 variants lead to changes in information content of natural and cryptic donor and acceptor sites. These variants were prioritized for effects on splicing based on the criteria described in Chapter 2.2.4. The majority of variants (1898) altered the information content of cryptic intronic splice sites. Of these variants, 449 had a final information content less than that of the nearest natural splice site and are unlikely to bind the spliceosome with greater affinity than the natural site and were thus disregarded. However, 61 variants resulted in an information content greater than the most proximal natural site. Fourteen of these variants were within 300bp of the natural splice site and

were prioritized. Of 2052 variants, 138 resulted in changes in information content of exonic cryptic splice sites, however, all occurred >300bp from the natural splice site, and none resulted in the creation of a cryptic site of greater information content than the natural splice site and so, were not prioritized. Finally, 15 variants were found to alter natural splice site information content, 11 of which strengthened natural splice sites by 1.12, 1.33 and 1.72 bits and were not prioritized. Three of these variants resulted in reduced splice site information content, and one variant abolished a natural acceptor site. Table 10 displays the final list of variants prioritized for splicing effects, as well as the site affected by the variant, the initial and final information content of the splice site affected, the SNP associated with the variant (if applicable), base coverage associated with the variant and the information content of the nearest natural splice site in the case of variants creating cryptic splice sites.

Table 10. Prioritized putative splicing variants in 21 breast cancer patients.

Gene	Variant (HGVS)	R_i initial (bits)	R_i final (bits)	ΔR_i	Type	R_i of nearest natural site	rsID	Alt. Allele population frequency (dbSNP)	Coverage[1]
Inactivating Natural Splice Site Variants (R_i final <1.6 bits)									
PALB2	c.2997-1G>A	7.41	-3.47	-10.88	Acceptor	-	-	-	6
Leaky Natural Splice Site Variants (R_i final ≥1.6 bits)									
ATM	c.3403-12T>A	9.81	7.21	-2.6	Acceptor	-	-	-	4
BRCA2	c.9502-909A>T [2]	19.87	12.17	-7.7	Acceptor	-	-	-	5
BRCA1	c.4358-2801T>A[3]	10.17	8.32	-1.85	Acceptor	-	-	-	20
Cryptic Splice Site Strengthening Variants (R_i final equivalent or greater than natural site R_i)									
ATM	c.6199-61C>G	-11.2	0.5	11.7	Acceptor	-1.14	-	-	6
BRCA2	c.197T>C *	0.63	1.75	1.12	Acceptor	1.72	rs4942486	0.468	15, 6, 6, 48, 105, 89, 28, 64, 78
PALB2	c.3351-49T>A	5.76	13.46	7.7	Acceptor	9.03	-	-	4
BRCA1	c.548-24A>T	0.14	2.32	2.18	Acceptor	0.49	-	-	4
CHEK2	c.-6-98C>T	9.41	10.5	1.09	Acceptor	1.76	rs73881129	No data	25, 81

* Sanger sequencing validated

[1]Variants with more than one coverage value are found in multiple patients

[2] EST BE869603

[3] 26nt upstream of acceptor in BRCA1 uc002ict.3 transcript

All identified variants were then examined for effects on TFBS's *in silico* as described in Chapter 2.2.4. After all changes in information content were calculated using the "Mutation Analysis" program, variants were filtered to include only those affecting particular TFBS's with information contents greater than one standard deviation above the mean information content for that TFBS. The total number of putative TFBS altering variants was reduced to 3108. Next, variants were filtered using HMEC mammary epithelial cell line DNase I HS HotSpot data from ENCODE; putative TFBS variants were reduced from 3108 to 228. Variants were then filtered for presence within 10kb of transcription start site (TSS), which reduced the number of putative TFBS variants from 228 to 48. Variants were then prioritized based on change in information content ($\geq$1.0 bits) and proximity to TSS. A final list of variants prioritized for effects on TFBS, including the change in information content, rsID's associated with variant, variant coverage, and proximity to TSS, are shown in Table 11. Transcription factor binding sites affected by the variants presented in Table 11 have been shown to be influential in breast cancer etiology.

Table 11. Prioritized variants altering TFBS information content

Gene	Variant (HGVS)	Transcription Factor	R_i initial (bits)	R_i final (bits)	ΔR_i	Distance from TSS[1]	rsID	Alternate allele population frequency (dbSNP)	Coverage[3]
CHEK2	c.-346T>C	HSF1	9.5	-1.6	-11.1	271	rs200681712	No data	8, 69, 206
CDH1	c.48+76C>T	HSF1	9.5	-0.9	-10.4	*251*	-	-	15
CHEK2	c.-346T>C	BCLAF	10.0	0.4	-9.6	282	rs200681712	No data	8, 69, 206
ATM	c.-1520G>A	GRP20	6.3	-3.3	-9.6	1129	rs190491586	No data	4
CDH1	c.163+1880G>T	TCF7L2	10.3	1.2	-9.1	*2999*	-	-	5
CDH1	c.163+1757A>T	YY1	9.7	8.4	-1.3	*2880*	-	-	6
ATM	c.-365C>G	ETS1[2]	5.7	-3	-8.7	*22*	-	-	29
CDH1	c.48+136C>G	EBF1[2]	6.9	0.8	-6.0	*305*	rs36037011	No data	18, 34
CDH1	c.163+1771A>G	GATA3[2]	7.4	5.6	-1.8	*2897*	-	-	20
BRCA1	c.-20G>T	TCF7L2[2]	6.7	5	-1.7	*203*	-	-	5

[1]Italicized Distances are downstream of TSS

[2]TFBS initial R_i is below Mean R_i

[3]Variants with more than one coverage value listed were found in multiple patients

Variants within 5' and 3' UTRs were then examined *in silico* for effects on mRNA structure using SNPfold. Altering UTR mRNA structure can affect translation initiation and influence disease [92]. Both UTR regions have been submitted to Alain Laederach's lab at UNC Chapel Hill for SHAPE analysis as described in Chapter 1.7, to validate SNPfold predictions on RNA structure alterations. Both variants predicted to alter structure were subsequently examined for effects on RBPBS's. Variants are shown in Table 12. RNA-binding protein binding sites altered include an abolished SNRPA binding site (the U1 snRNP of spliceosome), as a result of the 5' UTR variant, and a strengthened cryptic SF2/ASF binding site (3' UTR variant).

Variants in 3' UTRs were inputted into SNPNexus while selecting for predicted effects on miRNA binding. No variants were predicted to affect miRNA binding. Next, predicted miRNA binding sites in *BRCA1* and *BRCA2* were examined using TargetScanHuman [79], and mapped to *BRCA1* and *BRCA2* 3' UTR regions on the UCSC genome browser. No miRNA's were predicted to bind to *BRCA2* 3' UTR, whereas 5 putative miRNA binding sites were found on *BRCA1* 3' UTR, including: miR-7/7ab (chr17:41196419-41196425), miR-132/212 (chr17:41196447-41196453), miR-205 (chr17:41196487-41196493), miR-125a-3p (chr17:41197454-41197461), and miR-218 (chr17:41197477-41197483). miRNA binding sites were also examined using the miRBase through the UCSC Genome browser snoRNA/miRNA gene prediction track. No variants were found within predicted miRNA binding sites identified through either TargetScan or miRBase.

Table 12. Variants predicted to affect UTR mRNA structure and RNA-binding protein binding sites

Gene	Variant (HGVS)	UTR	rsID	Alternate allele population frequency (dbSNP)	SNPfold rank	p-value	RBPBS	R_i initial (bits)	R_i final (bits)	ΔR_i	Coverage[1]
BRCA2	c.-52A>G *	5'	rs206118	0.156	2/900	0.0022	SNRPA	8.2	4.4	-3.8	17, 107, 54
BRCA2	c.*369A>G *	3'	rs7334543	0.231	241/4500	0.0535	SF2/ASF	0.68	4.8	4.2	30, 37, 30, 2

*Sanger sequencing validated

[1]Variants with more than one coverage value indicate the variant was found in multiple patient samples, each with a unique coverage value

A total of 293 exonic variants were detected in all patients, which were subsequently sorted based on SNP alternate allele population frequency and variant coverage. This was done with the assumption that variants with lower alternate allele population frequencies have a greater likelihood of influencing disease etiology than those with high alternate allele population frequency. Results of sorted variants are shown in Table 13. Variants presented in Table 13 were not predicted to result in splicing or TFBS changes. One variant has been reported previously by the Molecular Diagnostics Laboratory (*BRCA1*:c.5136G>A) to result in a protein truncation, and was included in this study as a positive control. Another variant resulting in a missense change (*PALB2*:c.2993G>A) predicted to affect protein function by PolyPhen and SIFT as has been reported by Hellebrand H. *et al.* (2011) and found to be prevalent in both patients with breast cancer and healthy controls in three independent studies, and is not likely to be influential in breast cancer etiology [93,94]. An additional putative variant leading to a premature stop codon was identified in *ATM* (c.7051G>T), however this variant would need to be Sanger sequencing validated.

Table 13. Exonic variants filtered based on coverage, amino acid and SNP alternate allele population frequency

Gene	Variant (HGVS)	Amino acid	SNP	Alternate allele population frequency (dbSNP)	Coverage
ATM	c.735C>T	V>V	rs3218674	0.007	41
PALB2	c.2993G>A [2]	G>E	rs45551636	0.011	27
PALB2	c.3300T>G	T>T	rs45516100	0.018	26
BRCA1	c.5136G>A [1]	W>Stop	rs80357418	No data	16
TP53	c.699C>A	H>Q	-	-	15
TP53	c.691A>C	T>P	-	-	15
BRCA2	c.1814T>A	I>K	rs80358468	No data	13
ATM	c.7051G>T	E>Stop	-	-	12
ATM	c.146C>G	S>C	rs1800054	0.006	11
CHEK2	c.1298A>G	Y>C	-	-	11

[1]Also detected by Molecular Diagnostics Laboratory at LHSC
[2] VUS reported by Hellebrand H. *et al.* (2011), found in both breast cancer patient and control samples. PolyPhen, SIFT and MutationTaster predict variant affects protein function.

After putative splicing, TFBS, and 5' and 3' UTR affecting variants were PCR amplified and sent for Sanger sequencing validation, three were confirmed and include putative UTR affecting variants: *BRCA2*:c.-52A>G, *BRCA2*:c.*369A>G, as well as putative splicing variant: *BRCA2*:c.197T>C. Putative effects of these variants on disease are discussed in Chapter 4.4.

Chapter 4 Discussion

4.1 Project summary

Familial breast cancer is a complex disease with multiple genes associated with its etiology [7,29]. High penetrant breast cancer associated genes *BRCA1* and *BRCA2* were first discovered through familial linkage analyses [3,6] and account for approximately 13% of total familial breast cancer cases [7]. Subsequent association studies have identified other low to moderate penetrant genes associated with breast cancer, including: *ATM*, *CDH1*, *CHEK2*, *PALB2*, and an additional high penetrant gene *TP53*, the gene products of which are involved in the double strand DNA break repair pathway and cell-to-cell adhesion (*CDH1*) [29].

Currently, Molecular Diagnostics Laboratories in Ontario sequence coding and adjacent intronic regions (-20 to +10 bp around exons) of the high penetrant breast cancer associated genes *BRCA1* and *BRCA2* in individuals with a strong family history of breast cancer to detect deleterious variants. A power analysis examining *BRCA1* and *BRCA2* variant prevalence in individuals with family history of breast cancer was conducted to estimate the minimum number of individuals needed for whole gene sequencing to identify pathological *BRCA1* and *BRCA2* variants. This revealed that many individuals with a family history of breast cancer are predicted to have undetected pathogenic variants in *BRCA1* and *BRCA2* (Table 6). Therefore, regions outside of exons and adjacent introns (-20bp to +10bp) may harbor deleterious mutations. Since variants in non-coding regions can affect: splicing, transcription factor binding, mRNA structure and

RBPBS, miRNA binding etc., sequencing whole genes is necessary to achieve more sensitive deleterious variant detection. In addition to non-coding regions in *BRCA1* and *BRCA2*, variants may be present in other genes associated with familial breast cancer. To better detect deleterious variants in individuals with a family history of breast cancer, multiple low to high penetrant genes previously determined to be associated with breast cancer etiology need to be sequenced in their entirety, including adjacent intergenic regions.

Based on the limitations of routine diagnostic sequencing of breast cancer patients in identifying deleterious variants, there are three aims to this study. The first is to publish VUS previously detected by the Molecular Diagnostics Laboratory at LHSC to the ENIGMA consortium and to predict the effects of variants on splicing *in silico* and validate predictions using an *ex vivo* mini-gene assay. The second aim is to investigate the deleterious variant detection limits of the molecular diagnostics through a power analysis, and to improve variant detection by sequencing coding, non-coding and adjacent intergenic regions ($\pm$10 kbp) of seven breast cancer associated genes, including: *ATM, BRCA1, BRCA2, CDH1, CHEK2, PALB2* and *TP53*. The final aim of this study is to prioritize variants detected in this sequencing study for effects on: splicing, transcription factor binding sites (TFBS), 5' and 3' UTR mRNA structure and RNA binding protein binding sites (RBPBS), and miRNA binding using *in silico* bioinformatics tools to reduce the number of clinically non-significant variants and select only those potentially influential in breast cancer etiology for future analyses to determine variant pathogenicity.

4.2 Aim 1: Variants of unknown significance publication and splicing analysis

To facilitate the elucidation of variants of unknown significance detected by the molecular diagnostics lab, all variants were converted to HGVS nomenclature and submitted to the ENIGMA consortium. Next, variants were prioritized *in silico* for effects on splicing. Of 158 variants, two were prioritized for effects on splicing using the Automated Splice Site Analysis server and the Human Splicing Finder as described in Chapter 2.1.1, these variants include *BRCA1*:c.288C>T, predicted *in silico* to abolish splicing regulatory SRp40 and SF2/ASF binding sites 14 nucleotides upstream of natural donor site, and *BRCA2*:c.7319A>G, predicted *in silico* to create a cryptic SRp40 site 120 nucleotides upstream of natural donor site, and in close proximity to a cryptic donor site of comparable information content to the natural donor (2.3 bits and 2.4 bits respectively). Both variants, as well as a variant described by Mucaki *et al.* (2011) [46], *BRCA1*:c.548-16G>A, predicted *in silico* to create a cryptic acceptor site 14 nucleotides upstream of natural acceptor site, were examined for effects on splicing using a mini-gene assay.

Putative leaky splicing variants were also identified within the LHSC patient cohort and are displayed in Table 5. To examine disease penetrance, pedigree information for all probands harboring a putative leaky splicing variant was obtained from LHSC. The average number of 1st, 2nd and 3rd degree relatives with HBOC is reported in Table 5. At least one patient with each variant has a 1st degree relative with breast and/or ovarian cancer. Based on information theory predictions and disease penetrance of 15 patients

with putative leaky splicing variants, each of these variants is likely to result in deleterious *BRCA1/BRCA2* levels as a result of aberrant splicing, which is likely to be influential in hereditary breast and/or ovarian cancer predisposition.

The *BRCA1*:c.4986+6T>G and *BRCA2*:c.7007G>A variants are listed as clinically significant in the BIC database, which supports their role in breast cancer etiology. Investigating putative leaky splicing variants of unknown clinical significance (BIC) in the literature revealed that *BRCA2*:c.68-7T>A, *BRCA2*:c.9501+3A>T and *BRCA1*:c.591C>T lead to aberrant splicing [95–97]. RT-PCR analysis of patient blood samples and lymphoblastoid cell lines by Muller *et al.* (2011) revealed that patients with the *BRCA2*:c.68-7T>A have 22% of total BRCA2 mRNA missing exon 3, compared to 6% in WT controls [95]. However, they concluded that further co-segregation and case/control mutation frequency analyses are needed to characterize this variant as pathogenic, and recommend that the variant remain unclassified. The *BRCA2*:c.9501+3A>T variant has been characterized as deleterious by Borg *et al.* (2010) due to segregation data and cDNA confirming *BRCA2* exon 25 skipping, resulting in a frameshift and a pre-mature stop codon [97]. Finally, the *BRCA1*:c.591C>T variant has been shown by Dosil *et al.* (2010) to decrease the efficiency of intron 9 donor site recognition through RT-PCR of RNA extracted from patient peripheral blood lymphocytes. However, they characterize this variant as non-pathogenic because although they demonstrate that recognition of the intron 9 donor site is reduced, further analyses of patient-specific BRCA1 transcript revealed in-frame skipping of both exon 9 and exon 10, which is a naturally occurring, non-pathogenic splice isoform of *BRCA1* [96].

Therefore based on patient family history of HBOC and the clinical significance of these variants as reported both in the literature and BIC database, the leaky splicing variants identified in the LHSC cohort that are likely to be influential in breast cancer etiology include: *BRCA2*:c.9501+3A>T, *BRCA1*:c.4986+6T>G and *BRCA2*:c.7007G>A, whereas *BRCA1*:c.591C>T is likely to be non-pathogenic. And although patients with *BRCA2*:c.68-7T>A have a strong family history of HBOC as indicated in Table 5, and it has been shown that this variant promotes aberrant splicing, further case/control mutation frequency and segregation studies need to be conducted before this variant is labeled pathogenic, as recommended by Muller *et al.* (2011).

Recombinant plasmids harboring WT and mutant mini-gene constructs corresponding to *BRCA1*:c.288C>T (pcDNA-Dup plasmid), *BRCA2*:c.7319A>G (pCAS2 plasmid) and *BRCA1*:c.548-16G>A (pCAS2 plasmid) were constructed and transfected into MDA-MB-231 cell line to examine effects of variants on splicing (Chapter 2.1.2-2.1.6). After transfection of all mini-gene plasmids, RNA was extracted, digested with DNase I and reverse transcribed. First, RT-PCR products were PCR amplified to amplify cDNA from endogenously expressed *VSP39* and *BRCA2*. Amplicons of expected sizes (103bp for VSP39 and 96bp for BRCA2 respectively) were observed in all plasmid conditions (Figures 10, 11 and 12). This indicates that these transcripts are correctly spliced in MDA-MB-231 cells, and RNA extraction and reverse transcription experiments were successful.

Next, PCR amplification of cDNA corresponding to pcDNA-Dup transfection conditions yielded fragments each of approximately the same size at just greater than 500bp (Figure 6). The expected amplicon size of pcDNA-Dup mini-gene containing both introns and exons is 567bp, 567bp and 559bp for WT, mutant and empty pcDNA-Dup conditions respectively. One explanation of this result is that these fragments are derived from RT-PCR of unspliced, heteronuclear pcDNA-Dup mini-gene RNA. After investigating the role of plasmid DNA contamination in producing fragments shown in Figure 6, faint fragments corresponding to PCR amplification of residual amounts of pcDNA-Dup plasmid remaining after DNase I digestion were observed. However, levels of amplicon produced appear to be very low compared to fragments corresponding to pcDNA-Dup plasmid of known concentrations loaded into the gel in Figure 7. Therefore, to account for the high intensity fragments observed in Figure 6, the identity of these fragments is likely due to RT-PCR of unspliced hnRNA derived from pcDNA-Dup mini-gene rather than residual plasmid DNA.

To examine spliced RNA, primers spanning mini-gene exon junctions were designed for PCR amplification as described in Chapter 2.1.6. PCR amplification of cDNA derived from pcDNA-Dup using primers specific to mini-gene not containing cassette exon, followed by 2% agarose gel electrophoresis, revealed fragments corresponding to the expected amplicon size of 209bp in WT *BRCA1*:c.288C, mutant *BRCA1*:c.288C>T, and empty pcDNA-Dup control conditions (Figure 9, Lane 5, 6 and 4 respectively). Also, PCR amplification of mini-gene containing cassette exon revealed fragments at the expected size of 63bp in WT and *BRCA1*:c.288C>T mutant conditions (Figure 9, Lane 1

and 2 respectively). Therefore, based on this data, both WT and mutant *BRCA1*:c.288C>T show leaky splicing of the cassette exon. However, the detection of DNA fragment corresponding to cassette exon exclusion in the WT condition using exon-exon spanning primers (Figure 9, Lane 5), and the detection of fragments corresponding to mini-gene constructs containing intronic sequences using a T7Pro primer and a primer binding the last exon of the mini-gene (Figure 6) suggest that pcDNA-Dup mini-gene splicing has poor efficiency and fidelity. The choice of cell line used in this experiment may be a factor affecting transfected plasmid splicing.

The MDA-MB-231 breast cancer cell line was chosen for this study based on a study conducted by Morelli *et al.* (2003) which demonstrated transfection of plasmid derived from pcDNA3.1 backbone into MDA-MB-231 cells using FuGene6 transfection reagent [66]. However, this group stably transfected cells with plasmid, in contrast to this study in which transient transfection was performed to replicate the mini-gene assay performed by Tournier *et al.* (2008) [47]. mRNA splicing from a gene transcribed in a genomic context is influenced by its epigenetic environment, which consists of histone modification proteins that have been demonstrated to associate with splicing factors and influence splicing [98]. Therefore although pcDNA-Dup vector was transfected into MDA-MB-231 cell line as supported by the identification of spliced pcDNA-Dup mini-gene cDNA (Figure 9), the fidelity and efficiency of splicing of pcDNA-Dup mini-gene as a result of transient transfection into MDA-MB-231 cell line may be a cause of PCR amplification of unspliced hnRNA (Figure 6) and mini-gene not containing cassette exon in the WT condition (Figure 9, Lane 5). To remediate this, WT pcDNA-Dup mini-gene can be

transfected into a variety of cell lines, and the splicing fidelity and efficiency can be examined through RNA extraction and RT-PCR to identify a cell line in which splicing is efficient and accurate. This could then be followed by examining plasmids harboring a mini-gene containing putative splicing mutations. In conclusion, the hypothesis that *BRCA1*:c.288C>T affects splicing by abolishing SF2/ASF and SRp40 sites can neither be supported nor rejected based on this data as splicing fidelity and efficiency of transiently transfected pcDNA-Dup in MDA-MB-231 cell line does not appear to be sufficient.

PCR amplification of cDNA derived from limited amounts of spliced mini-gene RNA compared to RNA derived from a genomic context, may skew splice isoform levels with respect to endogenous BRCA1 mRNA containing c.288C>T. Therefore although leaky mini-gene cassette exon splicing was observed after RT-PCR of WT and mutant RNA using primers spanning exon-exon junctions, if the pool of spliced RNA from which these results were derived is small, then the accuracy of this assay as a representation of endogenous mRNA splice isoforms encoding *BRCA1*:c.288C>T transcribed from a genomic context may not be high enough for the reliable interpretation of this variant.

The following explanation explains the RT-PCR results of the mini-gene containing *BRCA1*:c.288C>T variant (Figure 9) if they are in fact consistent with what is shown in Figure 9 independent of inefficient pcDNA-Dup mini-gene splicing. Tournier *et al.* (2008) explain that due to a weak 3' splice site, the cassette exon is not recognized by the spliceosome and is therefore excluded in the final processed mRNA transcript unless an exonic splicing enhancer element (ESE) is inserted within its sequence, and only a

cassette exon harboring a functional ESE will be included in the final transcript [47]. Since both WT and mutant conditions show cassette exon inclusion, a functional ESE was cloned into cassette exon. This ESE could not have been abolished by the *BRCA1*:c.288C>T variant, because cassette exon appears in both the WT and mutant condition. Examining the 35bp sequence cloned into cassette exon on the ASSEDA server as a user defined sequence shows that additional putative ESE elements unaffected by the *BRCA1*:c.288C>T variant are found within this insert, including SRp40 and SRp55 binding sites of information content greater than abolished SRp40 site, but lower than abolished SF2/ASF binding site. Therefore it may be possible that one of these ESE elements is facilitating cassette exon recognition independent of abolished SF2/ASF and SRp40 sites. However, since DNA fragments consistent with cassette exon exclusion were observed in both WT and mutant conditions (Figure 9, Lane 5 and 6 respectively), the ESE element responsible for exon inclusion shown in Lane 1 and 2 may not promote constitutive exon recognition, and may require additional ESE elements outside of the 35bp region examined in this mini-gene study. However, this is speculation. This could be investigated by examining the effects of *BRCA1*:c.288C>T on whole exon 5 in pCAS2 mini-gene vector, and looking for constitutive exon inclusion after transient transfection, RNA extraction and RT-PCR. Constitutive exon inclusion would indicate sequence motifs outside of the 35bp region surrounding the variant examined in this study could facilitate exon recognition by the spliceosome. Based on RT-PCR analysis, the *BRCA1*:c.288C>T variant predicted using information theory to abolish an SF2/ASF site (R_i= 6.6>0.9 bits) and SRp40 site (R_i= 2.6>-2.9 bits) does not appear to affect cassette exon recognition by the spliceosome in pcDNA-Dup mini-gene vector. It may be

possible that the putative abolished SF2/ASF site within the cassette exon is not functional and does not exist, as the ESEfinder, but not Human Splicing Finder and Information theory has indicated (see section 3.1).

Next, PCR amplification cDNA derived from pCAS2 transfection conditions revealed no mini-gene specific PCR products corresponding to WT and mutant *BRCA1*:c.548-16G>A and *BRCA2*:c.7319A>G. Preliminary control experiments investigating the effectiveness of RNA extraction and RT-PCR determined that these experiments were successful (Figure 10, 11, 12), therefore RNA extraction and RT-PCR experiments were not the cause of no findings. Furthermore, mini-gene cDNA was PCR amplified under a range of annealing temperatures and conditions as described in Chapter 2.1.6. Therefore plasmid transfection and mini-gene expression may be the cause of lack of results.

First looking at mini-gene expression, all mini-genes in this study are under the control of a CMV promoter, the most commonly used promoter in mammalian expression vectors [99]. Chen *et al.* (2011) studied the effects of c-Myc-GFP fusion protein, encoded on a vector controlled by CMV promoter that was transiently transfected into MDA-MB-231 cell line, and determined *BRCA1* promoter is controlled by c-Myc, and *BRCA1* promoter activity increases in proportion to amount of vector transfected [100]. Also, in the case of pcDNA-Dup derived plasmids, spliced mini-gene RNA was detected (Figure 9), therefore the CMV promoter is active in MDA-MB-231 cells. Therefore, it is very unlikely that low transcriptional expression of mini-gene plasmid in MDA-MB-231 cell line is due to the choice of promoter. Also, since cells were treated with puromycin prior to

harvesting, mini-gene transcripts would not have been degraded due to non-sense mediated decay. Therefore low expression of mini-gene due to promoter choice, or degradation of expressed mini-gene due to nonsense-mediated decay are not likely causes of lack of fragments corresponding to mini-gene cDNA. Therefore a more likely cause of lack of mini-gene specific RNA is suboptimal transfection.

Although multiple transfection reagent:DNA ratios were performed as recommended in the FuGene6 manual, altering the total amount of DNA transfected into each well was not optimized. A total of 3μg of all plasmid DNA (derived from both pcDNA-Dup and pCAS2 mini-genes) was transfected into each well containing MDA-MB-231 cells, as recommended by the FuGene6 transfection reagent protocol described in Chapter 2. It may be possible that size differences between recombinant pcDNA-Dup vectors (3725bp with insert) and pCAS2 vectors (7838bp for *BRCA1*:c.548-16G>A vector, and 8350bp for *BRCA2*:c.7319A>G vector) result in differences in transfection efficiency, and could account for the lower transfection efficiency of pCAS2 derived plasmids, since the number of plasmids transfected is less than half of pcDNA-Dup. To remediate this, a gradient of vector DNA (μg) could be transfected into MDA-MB-231 cells to account for the larger size of pCAS2 compared to pcDNA-Dup.

One other potential source of reduced transfection efficiency is passage number. All vectors were transfected into cells at the same time, requiring a large number of cells to accommodate transfection ratio, and $\pm$ puromycin conditions. To lower passage number and increase transfection efficiency, only a WT condition would first be transfected to

ensure adequate expression and correct splicing. Once the ideal transfection conditions are identified (i.e. plasmid transfection, expression and correct splicing), subsequent experiments examining mutant mini-gene sequences could be conducted. Therefore, optimizing plasmid DNA amount (μg) during transfection by transfecting a gradient of plasmid DNA, while examining only plasmids harboring WT mini-genes, would reduce the total number of cells needed for transfection and thus passage number, and could increase transfection efficiency and promote mini-gene expression to yield more RNA for downstream analyses.

Finally, although a selection agent (G418) was used during cell culture, as described in Chapter 2.1.5, to a final concentration of 500ug/mL, as used by Morelli *et al.* (2003) [57] when culturing MDA-MB-231 cells after transfection with pcDNA3.1 backbone plasmid containing neomycin resistance gene, it may be possible that cells weren't exposed to drug long enough to kill all cells not harboring plasmid. The Morelli *et al.* (2003) study examined stable gene transfection and exposed cells to G418 for a longer period of time than in this study. MDA-MB-231 cells were cultured for three days at 500ug/mL of G418, followed by 2mg/ml once plasmids were incorporated into the genome, whereas in this study, plasmids were transiently transfected to replicate the Tournier *et al.* (2008) study [47], and cultured for 48 hours with 500ug/mL G418. To remediate this, a killing curve experiment designed to determine optimal G418 concentration and duration of exposure to kill MDA-MB-231 cells not harboring neomycin-resistance gene containing plasmid and prevent harvesting cells not containing mini-gene plasmid could be performed. Therefore it may be possible that suboptimal transfection conditions

(potentially remediated through transfecting cells of lower passage number under DNA concentration gradient) coupled with insufficient exposure to G418 could lead to lack of pCAS2 mini-gene RNA.

4.3 Aim 2: Current molecular diagnostics variant detection limitations and expanded sequencing study

A power analysis using *BRCA1* and *BRCA2* mutation positivity information from the Molecular Diagnostics Laboratory indicated that only a fraction of deleterious variants are detected when compared to linkage analyses as described in Chapter 2.3 (3.6-35.7% vs. 57-100%). A minimum patient sample size of 13 across all groups defined in Table 6, was calculated with 80% power and $p \leq 0.05$, to be necessary to detect a patient with a pathogenic *BRCA1* or *BRCA2* mutation. As previous studies have demonstrated that pathogenic variants can lie in regions outside of exons and immediate introns [24,25], non-coding regions may harbor deleterious mutations in *BRCA1* and *BRCA2* that are going undetected in routine sequencing. After coding, non-coding and adjacent intergenic regions in seven breast cancer associated genes were sequenced in 21 breast cancer patients, 97/152 *BRCA1* and *BRCA2* SNPs previously detected by the Molecular Diagnostics Laboratory were detected in this sequencing study. The final sequencing run achieved an average variant coverage of 47.93x and validated 81/81 *BRCA1* and *BRCA2* SNP's previously detected in coding and adjacent intronic regions by the molecular diagnostics lab, compared to the first sequencing run which achieved only a 4.46x average variant coverage and a *BRCA1* and *BRCA2* SNP validation fraction of 1/22. This can be attributed to the protocol alterations described in Chapter 2.2.3, which include the use of Agencourt magnetic beads during library preparation, 50bp paired-end reads

during sequencing, and an Illumina software upgrade. This increase in coverage allows for sequencing more samples in a single sequencing run in the future, thus increasing throughput.

4.4 Aim 3: Variant prioritization

After *in silico* prioritization of 2871 unique variants detected in all genes over all sequencing runs, the number of variants potentially influencing breast cancer etiology as predicted using bioinformatics tools was reduced to 19, thus greatly reducing the number of putative deleterious variants to examine through future wet-lab functional analyses. Of prioritized variants, three were Sanger sequencing validated, none of which were previously reported in patients by the molecular diagnostics lab. Two variants were found in the *BRCA2* 5' and 3' UTR (c.-52A>G and c.*369A>G respectively) and were predicted by SNPfold and information theory to significantly alter mRNA structure and RNA-binding protein binding sites respectively, including abolishing an SNRPA binding site (5' UTR variant) and creating a cryptic SF2/ASF binding site (3' UTR variant) (Table 12). Abolishing an SNRPA binding site could affect splicing between the 5' UTR and first exon of *BRCA2* since this variant is located 12bp upstream of a natural donor site and the U1 snRNP (SNRPA) is known to bind the donor site. The inability of the spliceosome to recognize the donor site of the 5' UTR could prevent splicing of the 5' UTR and exon 1 intervening intron and prevent 5' UTR from being included in the final transcript, thus affecting mRNA stability, and translation initiation. However, since this variant is reported in dbSNP to have an alternate allele population frequency of 0.156, it is unlikely that it alone is a causative variant in breast cancer etiology due to its high

prevalence in the population. However obtaining genotype data for this SNP in a greater disease patient sample size in the future would allow one to construct an OR value for this SNP when comparing it to healthy control SNP information from dbSNP. This is also true of the *BRCA1* 3' UTR, c.*369A>G (rs7334543; alternate allele population frequency of 0.231). The UTRs in each patient DNA sample harboring these variants were PCR amplified and sent to a collaborator at UNC (Alain Laederach) for SHAPE analysis to validate SNPfold predictions on mRNA structure.

At the time of writing this thesis, *in vitro* analyses of 85 human RNA-binding proteins (RBP) were conducted by Ray *et al.* (2013) to examine sequence recognition specificity [101]. RNA-binding protein sequence motif conservation was examined in three mRNA regulatory regions, including: a) 5' and b) 3' UTRs, and c) alternatively spliced exons and flanking introns. Approximately two thirds of RBPs analyzed (104 of 165) showed a high degree of conservation in at least one of these three regulatory regions. The majority of motifs in most RBP families are found in multiple regulatory regions (38 out of 49 RBP families), suggesting RNA-binding protein multifunctionality. Of all RBP isomers analyzed, the sequence recognition motifs of 20 of them were scanned in the UTR regions in this thesis for information content changes as a result of a genetic variant. Of the 20 RBPBS examined in this thesis also reported in Ray *et al.* (2013), three are associated with stabilizing 3' UTR mRNA (PUM2, EIF4B, ELAVL1) and three promote 3' UTR mRNA instability (SRSF1/2/9) according to *in vitro* RBP binding experiments, and mRNA stability experiments correlating RBP and target gene expression conducted on a set of 34 human tissues and cell lines (Ray *et al.,* 2013) [101].

Also, one putative splicing variant located 66 nucleotides upstream of an intron-exon junction in *BRCA2* (c.197T>C) not previously detected by the Molecular Diagnostics Laboratory in these patients was identified and validated by Sanger sequencing. This variant is predicted to result in strengthening a cryptic acceptor site (R_i = 0.63> 1.75 bits) to approximately the same information content as the corresponding natural site (1.72 bits). If the spliceosome binds this site, exon could be extended by 66 nucleotides upstream. However, this variant is a SNP (rs4942486) with high alternate allele population frequency (0.468) and therefore likely not influential in disease etiology.

Two putative heat shock factor 1 (HSF1) TFBS inactivating variants, one in *CDH1* (c.48+76C>T), the other in *CHEK2* (c.-346T>C), both within 300bp of the transcription start site were prioritized (see Table 11). HSF1 function has been shown to support aberrant proliferation and survival of a variety of human cell lines, including both breast tumor and mammary epithelial lines [102]. Dai *et al.* found that small hairpin RNA mediated knockdown of HSF1 in human mammary epithelial cells had only a minimal impact on cell viability (~90% viable cells remaining after 4 days), whereas the number of viable tumorigenic mammary epithelial cells subjected to the same shRNA-mediated HSF1 knockdown greatly decreased in comparison (<20% viable cells remaining after 4 days) [69]. This group concluded that while HSF1 can enhance survival and longevity under most conditions, tumor cell proliferation and survival is dependent on HSF1 activity. Therefore in the context of this study, inactivating mutations in HSF1 binding

sites within the *CDH1* and *CHEK2* promoter may be more influential in women with breast cancer versus women without breast cancer, and may linked with better prognosis.

Another putative TFBS inactivating variant was found in a BCLAF binding site in *CHEK2* (c.-346T>C). The Bcl-associated factor 1 transcription factor (BCLAF) has been shown to be involved in tumor suppression. A study conducted by Kasof *et al.* (1999) demonstrated that overexpression of BCLAF in HeLa cells suppresses transcription induced apoptosis [103]. A Northern blot analysis in several human tumor cell lines indicated that the chromosomal region encoding BCLAF is deleted in some tumor types. Therefore it is conceivable that abolishing a strong BCLAF binding site in *CHEK2*, which encodes a kinase involved in the apoptosis pathway, may promote oncogenesis. Another TFBS with a large information content change as a result of a variant in *CDH1* was TCF7L2 (c.163+1880G>T) Δ R_i 10.3>1.2 bits). An association study by Connor *et al.* (2012) has shown that several SNPs in the TCF7L2 gene are linked to increased breast cancer risk [104], although the underlying mechanisms associated with breast cancer etiology are unclear. ChIP-seq analysis performed by Frietze *et al.* (2012), has shown it to be expressed in the breast cancer cell line MCF7, and bind to enhancer regions of many genes [105]. Therefore although SNPs in TCF7L2 have been associated with breast cancer, and ChIP-seq analyses have demonstrated that it is expressed in breast cancer cell lines and binds to enhancer regions, it is difficult to predict what the effects of an abolished TCF7L2 site would have on breast cancer etiology. Variants within these TFBS were not Sanger sequencing validated.

Finally one protein truncation mutation was identified in *BRCA1* previously by the Molecular Diagnostics Laboratory (c.5136G>A, p.W1712Stop), and used as a positive control, was confirmed in this study. Additionally, a second putative protein truncation mutation in *ATM* was identified in this study (c.7051G>T), although its coverage value is 12 and needs to be Sanger sequencing validated.

Therefore, after examining prioritized, Sanger sequencing validated variants, it is possible that two are influential in disease etiology, both residing in *BRCA2*, one in the 5' UTR, the other in the 3'UTR. However, *in silico* predictions need to be validated through SHAPE analysis and ultimately an additional functional analysis relating structural changes to *BRCA2* transcript stability and RNA-binding protein binding. Increasing sample throughput as mentioned earlier would allow for organizing patient samples according to family history as outlined in Table 6, as well as increasing the number of patient DNA sequenced in a single sequencing run, thus bringing the number of sequenced patients in each group to the threshold predicted using the power analysis needed to detect a deleterious variant in *BRCA1* and *BRCA2*.

4.5 Project limitations

PCR amplification of unsliced hnRNA transcribed from pcDNA-Dup mini-gene (Figure 6) and WT pcDNA-Dup mini-gene not containing cassette exon suggests improper splicing of transiently transfected pcDNA-Dup plasmid in MDA-MB-231 cell line. Examining RNA from WT pcDNA-Dup plasmid transfected into multiple cell lines may identify a cell line in which mini-gene splicing is correct and efficient, which can then be

used to examine plasmids containing mutant mini-gene sequences. Suboptimal DNA concentration, passage number and exposure to G418 selection agent during transfection of pCAS2 recombinant plasmids may account for low transfection efficiency, which resulted in no fragments corresponding to mini-gene cDNA. Determining optimal G418 selection agent by performing a "killing curve", while reducing passage number by examining mutations individually and transfecting using a gradient of DNA may identify optimal transfection conditions to ascertain the effects of *BRCA1*:c.548-16G>A and *BRCA2*:c.7319A>G variants on mini-gene splicing.

For the patient breast cancer associated gene sequencing study, of the 19 prioritized variants, only three were Sanger sequencing validated. This can be attributed to two factors, low coverage and read alignment error. Meynert A.M. *et al.* (2013) [106] examined minimum read depths required for single nucleotide variant detection in DNA enriched through a variety of commercial capture arrays such as Agilent and Nimblegen, which varied from 19.7$\pm$0.9 fold coverage to 45.6$\pm$4.2 fold coverage for 95% sensitivity for heterozygous mutations. Eleven of nineteen prioritized variants in this study had a minimum read depth of <10x, all of which were not Sanger sequencing validated.

In contrast, some high coverage variants were also not Sanger sequencing validated, including the putative TFBS variant *CHEK2*:c.-346T>C, which had 69x coverage in one patient and 206x coverage in another. This can be attributed to a read alignment error. Limitations in DNA shearing can result in DNA fragments containing both unique sequences hybridizing to probe, and low complexity regions at the ends of the fragment.

RNA probes will hybridize to single-copy region on DNA fragment containing low-complexity regions on the fragment ends, which are subsequently enriched and paired-end sequenced, leading to misalignment to the genome and resulting in false positive variant calling. Unique regions within segmental duplicons can also be captured and misaligned to the genome, resulting in false positive variant calling. It was discovered that multiple patients carried several of the same "novel" exonic mutations in *CHEK2*, after investigating further, it was discovered that these mutations lie in a pseudogene spreading across exons 10-14 within a segmental duplicon that maps to chromosome 15. A total of twelve segmental duplicons were found in a ~21kb region in *CHEK2* (chr22:g.29,071,654-29,092,380). Therefore, false positive variant calls within this region may be attributed to probe hybridizing to off-target chromosomes, which is subsequently sequenced and incorrectly aligned to *CHEK2*. Therefore it is critical to validate variants called within segmental duplicon regions through Sanger sequencing to ensure they are not false positives due to alignment errors.

Finally, a variant with moderate coverage, the putative TFBS variant, *ATM*:c.-365C>G (29x coverage, 13 reads calling alternate allele) was not Sanger sequencing validated. Examining read information on the Integrated Genomics Viewer (IGV; Robinson J.T. *et al.* 2011) revealed that the 13 reads calling the alternate allele had low phred base quality scores (all <2) upstream of the variant, which itself had a phred base quality score of 213. Therefore a false positive variant call was generated through overall poor read quality, despite the variant base itself passing the stand_call_conf filter threshold of 30 during GATK variant calling.

No putative deleterious variants were prioritized in low to moderate penetrant genes. This may be attributed to two causes, including small sample size and sample selection. The criteria for molecular diagnostic sequencing of high penetrant breast cancer associated gene *BRCA1* and *BRCA2* is a strong family history of breast as defined on page two of the, "Ontario Cancer Genetic Testing Program" requisition form (http://www.lhsc.on.ca/lab/molegen/brca_req.pdf). Therefore sequencing low to moderate penetrant breast cancer associated genes in a cohort of patients with a high incidence of breast cancer may skew results in favor of variant detection in the high penetrant breast cancer associated genes *BRCA1* and *BRCA2*. To detect variants in low to moderate penetrant genes, it may be necessary to identify women with lower incidence of breast cancer family history than that described in the "Ontario Cancer Genetic testing Program requisition form". Alternatively, there may be other high penetrant breast cancer associated variants lying outside of the high penetrant breast cancer associated genes *BRCA1* and *BRCA2* that are in regions not examined in this study that contribute to breast cancer risk. Lalloo F. and Evans D.G. (2012) estimate that ~61% of the familial component of breast cancer is linked to unidentified regions in the genome which could harbor high penetrant variants not probed for in this experiment [7].

4.5 Conclusion and implications

The limitations of routine diagnostic sequencing in detecting pathogenic DNA variants in individuals with or at risk of breast cancer warrants deeper probing into non-coding and intergenic regions in *BRCA1* and *BRCA2* as well as other genes associated with familial

breast cancer, such as *ATM, CDH1, CHEK2, PALB2* and *TP53*. Library DNA samples prepared with Agencourt Magnetic beads, show less sample loss from column and gel purification steps as well as sample transferring, and result in increased coverage and ultimately, more reliable variant calling. Since the vast majority of true variants detected in large sequencing studies are polymorphisms or variants of unknown significance [97,11], it is necessary to first prioritize variants *in silico* for effects on cellular processes integral in healthy gene expression such as: splicing, transcription factor binding, 5' and 3' UTR structure and mRNA binding protein binding, amino acid changes and miRNA binding. All *in silico* prioritized variants need to be examined for read alignment errors prior to Sanger sequencing to reduce the incidence of false positive variant calls as discussed in Chapter 4.4. After Sanger validation, prioritized variants need to be examined in the literature to ascertain clinical significance, if unknown, it is necessary to perform functional analyses to validate the effects of the variant on the process in question through assays such as those described in Chapter 1.7 before the variant can be classified as deleterious.

This project provides a workflow for variant detection in several breast cancer associated genes of varying penetrance, as well as an *in silico* framework that prioritizes variants predicted to affect a variety of cellular processes involved in WT gene expression for downstream function analyses. Sanger validated variants likely most influential in disease etiology, undetected by the molecular diagnostics lab, have been submitted to a collaborator at UNC (Dr. Alain Laederach) for functional analysis to validate *in silico* prioritization. Overall, this work can improve variant detection in individuals with or at

risk of breast cancer in which routine sequencing revealed no deleterious findings and impact patient genetic counseling and quality of care.

References

1. GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10. at <http://globocan.iarc.fr/references.htm>

2. De Grève, J., Sermijn, E., De Brakeleer, S., Ren, Z. & Teugels, E. Hereditary breast cancer: from bench to bedside. *Current opinion in oncology* **20,** 605–13 (2008).

3. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science (New York, N.Y.)* **266,** 66–71 (1994).

4. Neuhausen, S. L. *et al.* A P1-based physical map of the region from D17S776 to D17S78 containing the breast cancer susceptibility gene BRCA1. *Human Molecular Genetics* **3,** 1919–1926 (1994).

5. Hall, J. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250,** 1684–1689 (1990).

6. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378,** 789–92 (1995).

7. Lalloo, F. & Evans, D. G. Familial breast cancer. *Clinical genetics* **82,** 105–14 (2012).

8. Antoniou, A. *et al.* Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *American journal of human genetics* **72,** 1117–30 (2003).

9. Szabo, C., Masiello, A., Ryan, J. F. & Brody, L. C. The breast cancer information core: database design, structure, and scope. *Human mutation* **16,** 123–31 (2000).

10. Horaitis, O., Talbot, C. C., Phommarinh, M., Phillips, K. M. & Cotton, R. G. H. A database of locus-specific databases. *Nature genetics* **39,** 425 (2007).

11. Spurdle, A. B. *et al.* ENIGMA--evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Human mutation* **33,** 2–7 (2012).

12. Krawczak, M., Reiss, J. & Cooper, D. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: Causes and consequences. *Human Genetics* **90,** (1992).

13. Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 11193–8 (2001).

14. Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA (New York, N.Y.)* **14,** 802–13 (2008).

15. Graveley, B. R. Sorting out the complexity of SR protein functions. *RNA (New York, N.Y.)* **6,** 1197–211 (2000).

16. Tacke, R., Chen, Y. & Manley, J. L. Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer. *Proceedings of the National Academy of Sciences of the United States of America* **94,** 1148–53 (1997).

17. Wu, J. Y. & Maniatis, T. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* **75,** 1061–70 (1993).

18. Caputi, M., Freund, M., Kammler, S., Asang, C. & Schaal, H. A bidirectional SF2/ASF- and SRp40-dependent splicing enhancer regulates human immunodeficiency virus type 1 rev, env, vpu, and nef gene expression. *Journal of virology* **78,** 6517–26 (2004).

19. Martinez-Contreras, R. *et al.* hnRNP proteins and splicing control. *Advances in experimental medicine and biology* **623,** 123–47 (2007).

20. Wang, Y., Ma, M., Xiao, X. & Wang, Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nature structural & molecular biology* **19,** 1044–52 (2012).

21. Fu, X. D. The superfamily of arginine/serine-rich splicing factors. *RNA (New York, N.Y.)* **1,** 663–80 (1995).

22. Buratti, E. *et al.* hnRNP H binding at the 5' splice site correlates with the pathological effect of two intronic mutations in the NF-1 and TSHbeta genes. *Nucleic acids research* **32,** 4224–36 (2004).

23. Chen, X. *et al.* Intronic alterations in BRCA1 and BRCA2: effect on mRNA splicing fidelity and expression. *Human mutation* **27,** 427–35 (2006).

24. Anczuków, O. *et al.* BRCA2 deep intronic mutation causing activation of a cryptic exon: opening toward a new preventive therapeutic strategy. *Clinical cancer research : an official journal of the American Association for Cancer Research* **18,** 4903–9 (2012).

25. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science (New York, N.Y.)* **339,** 959–61 (2013).

26. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science (New York, N.Y.)* **339,** 959–61 (2013).

27. Peto, J. *et al.* Prevalence of BRCA1 and BRCA2 Gene Mutations in Patients With Early-Onset Breast Cancer. *JNCI Journal of the National Cancer Institute* **91,** 943–949 (1999).

28. Ford, D. *et al.* Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *American journal of human genetics* **62,** 676–89 (1998).

29. Shuen, A. Y. & Foulkes, W. D. Inherited mutations in breast cancer genes--risk and response. *Journal of mammary gland biology and neoplasia* **16,** 3–15 (2011).

30. Pharoah, P. D. P., Dunning, A. M., Ponder, B. A. J. & Easton, D. F. Association studies for finding cancer-susceptibility genetic variants. *Nature reviews. Cancer* **4,** 850–60 (2004).

31. Gasco, M., Shami, S. & Crook, T. The p53 pathway in breast cancer. *Breast Cancer Res* **4,** 70–76 (2002).

32. Thompson, D. *et al.* Cancer risks and mortality in heterozygous ATM mutation carriers. *Journal of the National Cancer Institute* **97,** 813–22 (2005).

33. Desrichard, A., Bidet, Y., Uhrhammer, N. & Bignon, Y.-J. CHEK2 contribution to hereditary breast cancer in non-BRCA families. *Breast cancer research : BCR* **13,** R119 (2011).

34. Masciari, S. *et al.* Germline E-cadherin mutations in familial lobular breast cancer. *Journal of medical genetics* **44,** 726–31 (2007).

35. Kaurah, P. *et al.* Founder and recurrent CDH1 mutations in families with hereditary diffuse gastric cancer. *JAMA : the journal of the American Medical Association* **297,** 2360–72 (2007).

36. Vahteristo, P. *et al.* A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer. *American journal of human genetics* **71,** 432–8 (2002).

37. Einarsdóttir, K. *et al.* Comprehensive analysis of the ATM, CHEK2 and ERBB2 genes in relation to breast tumour characteristics and survival: a population-based case-control and follow-up study. *Breast cancer research : BCR* **8,** R67 (2006).

38. Perl, A. K., Wilgenbus, P., Dahl, U., Semb, H. & Christofori, G. A causal role for E-cadherin in the transition from adenoma to carcinoma. *Nature* **392,** 190–3 (1998).

39. Vleminckx, K., Vakaet, L., Mareel, M., Fiers, W. & van Roy, F. Genetic manipulation of E-cadherin expression by epithelial tumor cells reveals an invasion suppressor role. *Cell* **66,** 107–19 (1991).

40. Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nature methods* **4,** 903–5 (2007).

41. Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nature genetics* **39,** 1522–7 (2007).

42. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature biotechnology* **27,** 182–9 (2009).

43. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research* **8,** 175–185 (1998).

44. Stephens, R. M. & Schneider, T. D. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *Journal of Molecular Biology* **228,** 1124–1136 (1992).

45. Schneider, T. D. Information content of individual genetic sequences. *Journal of theoretical biology* **189,** 427–41 (1997).

46. Mucaki, E. J., Ainsworth, P. & Rogan, P. K. Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants. *Human mutation* **32,** 735–42 (2011).

47. Tournier, I. *et al.* A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. *Human mutation* **29,** 1412–24 (2008).

48. Steen, K.-A., Siegfried, N. A. & Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by protection from exoribonuclease (RNase-detected SHAPE) for direct analysis of covalent adducts and of nucleotide flexibility in RNA. *Nature protocols* **6,** 1683–94 (2011).

49. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **33,** D501–4 (2005).

50. Fokkema, I. F. A. C., den Dunnen, J. T. & Taschner, P. E. M. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. *Human mutation* **26,** 63–8 (2005).

51. Wildeman, M., van Ophuizen, E., den Dunnen, J. T. & Taschner, P. E. M. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Human mutation* **29,** 6–13 (2008).

52. Fujita, P. A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic acids research* **39,** D876–82 (2011).

53. Nalla, V. K. & Rogan, P. K. Automated splicing mutation analysis by information theory. *Human mutation* **25,** 334–42 (2005).

54. Lin, L. *et al.* Using high-density exon arrays to profile gene expression in closely related species. *Nucleic acids research* **37,** e90 (2009).

55. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research* **37,** e67 (2009).

56. Smith, P. J. *et al.* An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Human molecular genetics* **15,** 2490–508 (2006).

57. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology : a journal of computational molecular cell biology* **11,** 377–94 (2004).

58. Gutiérrez-Enríquez, S., Coderch, V., Masas, M., Balmaña, J. & Diez, O. The variants BRCA1 IVS6-1G>A and BRCA2 IVS15+1G>A lead to aberrant splicing of the transcripts. *Breast cancer research and treatment* **117,** 461–5 (2009).

59. Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J. & Fairbrother, W. G. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 11093–8 (2011).

60. Hereditary Breast and Ovarian Cancer | Cancer.Net. at <http://www.cancer.net/cancer-types/hereditary-breast-and-ovarian-cancer>

61. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods in molecular biology (Clifton, N.J.)* **132,** 365–86 (2000).

62. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Research* **12,** 996–1006 (2002).

63. SantaLucia, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences* **95,** 1460–1465 (1998).

64. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics (Oxford, England)* **23,** 1289–91 (2007).

65. Zhang, C. *et al.* Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes & development* **22,** 2550–63 (2008).

66. Morelli, C., Garofalo, C., Bartucci, M. & Surmacz, E. Estrogen receptor-alpha regulates the degradation of insulin receptor substrates 1 and 2 in breast cancer cells. *Oncogene* **22,** 4007–16 (2003).

67. Ziegler, K., Bui, T., Frisque, R. J., Grandinetti, A. & Nerurkar, V. R. A rapid in vitro polyomavirus DNA replication assay. *Journal of Virological Methods* **122,** 123–127 (2004).

68. Rogozińska-Szczepka, J. *et al.* BRCA1 and BRCA2 mutations as prognostic factors in bilateral breast cancer patients. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO* **15,** 1373–6 (2004).

69. Shih, H. A. *et al.* BRCA1 and BRCA2 mutations in breast cancer families with multiple primary cancers. *Clinical cancer research : an official journal of the American Association for Cancer Research* **6,** 4259–64 (2000).

70. Dorman, S. N., Shirley, B. C., Knoll, J. H. M. & Rogan, P. K. Expanding probe repertoire and improving reproducibility in human genomic hybridization. *Nucleic acids research* **41,** e81 (2013).

71. Using a PhiX Control for HiSeq Sequencing Runs - technote_phixcontrolv3.pdf. at <http://res.illumina.com/documents/products/technotes/technote_phixcontrolv3.pdf>

72. Lennon, N. J. *et al.* A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome biology* **11,** R15 (2010).

73. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078–9 (2009).

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

75. Chelala, C., Khan, A. & Lemoine, N. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* **25,** 655–661 (2009).

76. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7,** 248–9 (2010).

77. Sim, N.-L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research* **40,** W452–7 (2012).

78. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research* **39,** D152–7 (2011).

79. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).

80. Shirley, B. C. *et al.* Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences. *Genomics, proteomics & bioinformatics* **11,** 77–85 (2013).

81. Robberson, B. L., Cote, G. J. & Berget, S. M. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Molecular and cellular biology* **10,** 84–94 (1990).

82. Mucaki, E. J., Shirley, B. C. & Rogan, P. K. Prediction of mutant mRNA splice isoforms by information theory-based exon definition. *Human mutation* **34,** 557–65 (2013).

83. Bi, C. & Rogan, P. K. BIPAD: a web server for modeling bipartite sequence elements. *BMC bioinformatics* **7,** 76 (2006).

84. Webber, A. L., Ingram, R. S., Levorse, J. M. & Tilghman, S. M. Location of enhancers is essential for the imprinting of H19 and Igf2 genes. *Nature* **391,** 711–5 (1998).

85. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424,** 147–51 (2003).

86. Spivakov, M. *et al.* Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome biology* **13,** R49 (2012).

87. Buchler, N. E. & Cross, F. R. Protein sequestration generates a flexible ultrasensitive response in a genetic network. *Molecular systems biology* **5,** 272 (2009).

88. Halvorsen, M., Martin, J. S., Broadaway, S. & Laederach, A. Disease-associated mutations that alter the RNA structural ensemble. *PLoS genetics* **6,** e1001074 (2010).

89. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic acids research* **39,** D301–8 (2011).

90. Lastella, P. *et al.* Site directed mutagenesis of hMLH1 exonic splicing enhancers does not correlate with splicing disruption. *Journal of medical genetics* **41,** e72 (2004).

91. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nature methods* **7,** 111–8 (2010).

92. Pickering, B. M. & Willis, A. E. The implications of structured 5' untranslated regions on translation and disease. *Seminars in cell & developmental biology* **16,** 39–47 (2005).

93. Rahman, N. *et al.* PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nature genetics* **39,** 165–7 (2007).

94. Hellebrand, H. *et al.* Germline mutations in the PALB2 gene are population specific and occur with low frequencies in familial breast cancer. *Human mutation* **32,** E2176–88 (2011).

95. Muller, D. *et al.* An entire exon 3 germ-line rearrangement in the BRCA2 gene: pathogenic relevance of exon 3 deletion in breast cancer predisposition. *BMC medical genetics* **12,** 121 (2011).

96. Dosil, V. *et al.* Alternative splicing and molecular characterization of splice site variants: BRCA1 c.591C>T as a case study. *Clinical chemistry* **56,** 53–61 (2010).

97. Borg, A. *et al.* Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the WECARE study. *Human mutation* **31,** E1200–40 (2010).

98. Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R. & Misteli, T. Epigenetics in alternative pre-mRNA splicing. *Cell* **144,** 16–26 (2011).

99. Xia, W. *et al.* High levels of protein expression using different mammalian CMV promoters in several cell lines. *Protein expression and purification* **45,** 115–24 (2006).

100. Chen, Y. *et al.* c-Myc activates BRCA1 gene expression through distal promoter elements in breast cancer cells. *BMC cancer* **11,** 246 (2011).

101. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499,** 172–7 (2013).

102. Dai, C., Whitesell, L., Rogers, A. B. & Lindquist, S. Heat shock factor 1 is a powerful multifaceted modifier of carcinogenesis. *Cell* **130,** 1005–18 (2007).

103. Kasof, G. M., Goyal, L. & White, E. Btf, a novel death-promoting transcriptional repressor that interacts with Bcl-2-related proteins. *Molecular and cellular biology* **19,** 4390–404 (1999).

104. Connor, A. E. *et al.* Associations between TCF7L2 polymorphisms and risk of breast cancer among Hispanic and non-Hispanic white women: the Breast Cancer Health Disparities Study. *Breast cancer research and treatment* **136,** 593–602 (2012).

105. Frietze, S. *et al.* Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. *Genome biology* **13,** R52 (2012).

106. Meynert, A. M., Bicknell, L. S., Hurles, M. E., Jackson, A. P. & Taylor, M. S. Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics* **14,** 195 (2013).

Curriculum Vitae

Name:	Edwin Dovigi
Post-secondary Education and Degrees:	Western University London, Ontario, Canada 2007-2011 Honors Bachelor of Medical Sciences
	Western University London, Ontario, Canada 2011-present Master of Science (Biochemistry)
Honours and Awards:	Continuing Admissions Scholarship 2007-2009
	Translational Breast Cancer Studentship Award 2011-2012

Publications and Presentations:

London Health Research Day (poster presentation). "Functional analysis of variants of unknown significance and discovery of potential splicing mutations in inherited breast cancer". E. Dovigi, A. Stuart, E. Mucaki, N. Bryans, C. Viner, P. Ainsworth, P.K. Rogan (2013)

Oncology Research and Education Day (poster presentation). "Functional analysis of variants of unknown significance and discovery of potential splicing mutations in inherited breast cancer" E. Dovigi, A. Stuart, E. Mucaki, N. Bryans, C. Viner, P. Ainsworth, P.K. Rogan (2013)

American Society of Human Genetics (poster presentation). "Strategy for identification, prediction and prioritization of variants of uncertain significance in heritable breast cancer" P.K. Rogan, E.J. Mucaki, A. Stuart, N. Bryans, E. Dovigi, B.C. Shirley, C. Viner, J.H. Knoll, P. Ainsworth (2012).

International Congress of Genetics/HGM (poster presentation). "STRATEGY FOR IDENTIFICATION, PREDICTION, AND PRIORITIZATION OF NON-CODING VARIANTS OF UNCERTAIN SIGNIFICANCE IN HERITABLE BREAST CANCER." P. Rogan*, E. Mucaki, A. Stuart, E. Dovigi, C. Viner, B. Shirley, J. Knoll, P. Ainsworth (2013).

Human Genome Variation Society Meeting (oral presentation). "Strategy for identification, prediction and prioritization of non-coding variants of uncertain

significance in heritable breast cancer" P. Rogan, E. Mucaki, A. Stuart, E. Dovigi, C. Viner, B. Shirley, J.H. Knoll, P. Ainsworth (2013)

European Society of Human Genetics (poster presentation). "Prediction and prioritization of non-coding variant of unknown significance in heritable breast cancer" P.K. Rogan, E.J. Mucaki, E. Dovigi, C. Viner, B.C. Shirley, S. Dorman, A. Stuart, J. Knoll, P. Ainsworth (2013).