Search for influential genomic changes in breast cancer families

Anna Marzellíusardóttir

Thesis for the degree of Master of Science
University of Iceland
Faculty of Medicine
School of Health Science

HÁSKÓLI ÍSLANDS

Leit að áhrifabreytingum í erfðaefni fjölskyldna með háa tíðni brjóstakrabbameins

Anna Marzellíusardóttir

Ritgerð til meistaragráðu í Líf- og læknavísindum

Umsjónarkennari: Rósa Björk Barkardóttir

Meistaranámsnefnd: Rósa Björk Barkardóttir, Aðalgeir Arason, Laufey Ámundadóttir

Læknadeild

Heilbrigðisvísindasvið Háskóla Íslands

Október 2015

Search for influential genomic changes in breast cancer families

Anna Marzellíusardóttir

Thesis for the degree of Master of Science

Supervisor: Rósa Björk Barkardóttir

Masters committee: Rósa Björk Barkardóttir, Aðalgeir Arason, Laufey Ámundadóttir

Faculty of Medicine

School of Health Sciences

October 2015

Prentun: Háskólaprent
Reykjavík, Ísland 2015

Ágrip

Brjóstakrabbamein (BK) er algengasta dánarorsök vegna krabbameins meðal kvenna og annað algengasta krabbamein í heiminum. Áætlað er að um 5-10% BK tilfella séu ættgeng með sterk áhrif erfðaþátta. Þó tekist hafi að útskýra stóran hluta slíkra tilfella með stökkbreytingum í BK genum líkt og *BRCA1* og *BRCA2* er rúmlega helmingur BK fjölskyldna ekki skýrður af þekktum erfðaþáttum. Í þessu verkefni voru gögn úr háhraðaraðgreiningu á öllu erfðamengi (e. WGS: Whole Genome Sequencing) þrettán einstaklinga úr fjórum íslenskum BK fjölskyldum skoðuð með það markmið að bera kennsl á stökkbreytingar sem auka hættu á BK. Í verkefninu voru tvær aðferðir notaðar. Í fyrri hluta verkefnisins voru gögn úr einni fjölskyldu greind og var markmiðið að finna stökkbreytingu sem bæri með sér mikla áhættuaukningu og gæti skýrt háa tíðni BK í fjölskyldunni. Í seinni hluta verkefnisins var háhraðaraðgreiningargögnum úr þremur fjölskyldum til viðbótar bætt við. Þessar þrjár fjölskyldur höfðu áður verið rannsakaðar án þess að niðurstöður bentu til þess að í þeim fyndust stökkbreytingar tengdar mikilli áhættuaukningu. Af þeim sökum var markmið seinni hluta verkefnisins að finna stökkbreytingar sem hver fyrir sig bera með sér meðal áhættuaukningu á BK, en samanlögð áhrif stökkbreytinganna gætu valdið þeirri miklu áhættuaukningu sem sjá má í fjölskyldunum.

Við úrvinnslu gagnanna var notast við forritið Ingenuity$^{®}$ Variant Analysis^{TM} (IVA) þar sem gögnin voru síuð með tilliti til gæða raðgreiningarinnar, tíðni breytinga, hversu skaðlegar þær voru taldar, hvernig þær erfðust og líffræðilegs samhengis. Í fyrri hluta verkefnisins voru teknar fyrir breytingar sem spáð var að leiddu til taps á virkni gens eða hefðu eyðileggjandi áhrif á bindingu miRNA sameindar við gen og breyttu þannig tjáningu þess. Tuttugu og fjórar slíkar breytingar fundust og eftir staðfestingarferli og frekara mat á breytingunum var ákveðið að skima fyrir fimm þeirra í hópi óvalinna BK sjúklinga og í viðmiðunarhópi. Engin breytinganna reyndist vera í marktækt hærri tíðni í hópi óvalinna BK sjúklinga. Þrátt fyrir það er ekki hægt að útiloka að ein þeirra, *TRMT44* c.1928-2_1929delAGAG, sé tengd áhættuaukningu á BK. Í seinni hluta verkefnisins voru teknar fyrir breytingar í próteinkóðandi svæðum (útröðum) DNA-viðgerðargena sem metnar voru skaðlegar af spáforritum sem meta líkleg áhrif stökkbreytinga á próteinafurð. Skimað var fyrir 13 slíkum breytingum í hópi óvalinna BK sjúklinga. Breytingarnar voru allar með þekkta tíðni milli 1-5% í AGES-Reykjavík sýnahópi Hjartaverndar. Niðurstöður skimunarinnar bentu ekki til þess að nein breytinganna væri í marktækt hærri tíðni í hópi BK sjúklinga en í viðmiðunarhópi.

Út frá niðurstöðunum má álykta að há tíðni BK í fjölskyldunni sem skoðuð var í fyrri hluta verkefnisins skýrist trúlega ekki af stakri stökkbreytingu sem ber með sér mikla áhættuaukningu. Líklegra er að um fjölgena erfðir sé að ræða, þar sem sameiginleg áhrif tveggja eða fleiri breytinga leiða til mikillar áhættuaukningar. *TRMT44* c.1928-2_1929delAGAG gæti verið ein slík breyting. Enn fremur má álykta að þær skorður sem settar voru seinni hluta verkefnisins með því að nýta tíðniupplýsingar frá Hjartavernd (sem aðeins eru til staðar fyrir þekktar einsbasabreytingar í útröðum gena) og spáforrit sem meta aðeins áhrif skiptibreytinga (e. missense variants) hafi mögulega komið í veg fyrir að borin væru kennsl á áhættuaukandi breytingar. Næstu skref fela í sér skoðun á áhugaverðum breytingum sem ekki eru til tíðniupplýsingar fyrir hjá Hjartavernd eða eru ekki metnar af spáforritunum.

Abstract

Breast cancer (BC) is the leading cause of cancer deaths in women and the second most common cancer in the world. An estimated 5-10% of BC has an autosomal dominant inheritance pattern and is considered hereditary BC (HBC). Although a considerable proportion of HBC families can be explained by mutations in BC susceptibility genes such as *BRCA1* and *BRCA2*, over half of HBC remains unexplained. In this project, WGS (Whole Genome Sequencing) data from four Icelandic HBC families were analyzed with the aim of identifying mutations which confer an increased risk of BC. Two different approaches were applied. In the first part of the project, WGS data from one family were analyzed with the aim of identifying a high-risk mutation explaining the BC clustering within the family. In the second part of the project, WGS data from three additional families were added to the analysis. These three families had been analyzed in previous studies without preliminary data revealing evidence of a high-risk mutation, and therefore the aim was to identify moderate-risk variants that contribute to the increase in BC risk in the families in a polygenic fashion.

Initial analysis of the WGS data was performed using Ingenuity® Variant Analysis™ (IVA), in which identified variants were filtered according to quality, frequency, predicted deleteriousness, mode of inheritance and biological context. In each part of the project, the filtering process in IVA was adjusted according to the aim. In the first part of the project, the focus was on identifying variants predicted to cause loss-of-function (LOF) of a gene or to disrupt the binding of a miRNA to the 3' UTR of a gene and thereby altering its expression. Twenty-four such variants were identified and following a validation process and further assessment of the variants, five were screened for in a series of unselected BC cases and controls. This revealed that none of the variants confer a significantly increased risk of BC, although it cannot be ruled out that one of them, *TRMT44* c.1928-2_1929delAGAG, is a very rare moderate-risk variant. In the second part of the project, variants located in the coding regions of DNA repair genes, predicted deleterious by online function-prediction tools, were the focus of the study. This analysis resulted in thirteen variants, with control frequencies between 1-5% in the Icelandic Heart Association's (IHA) AGES-Reykjavik cohort, being screened for in a group of BC cases. The frequencies of the variants were compared to their control frequencies from AGES-Reykjavik, revealing that none of them are likely to increase the risk of BC.

From the results of the first part of this project, it can be concluded that the familial aggregation of BC in the family is probably not explained by a single high-risk mutation but rather by two or more variants that each has low to moderate effects on BC risk. *TRMT44* c.1928-2_1929delAGAG is possibly one such variant. Furthermore, it can be concluded that the limits placed on the study by using control frequencies from the IHA, which are only available for recorded exonic SNPs, and using online prediction tools, which only assess missense variants, possibly prevented us from identifying predisposing variants. Next immediate steps include studying candidate variants for which control frequency was not available or were not assessed by function-prediction tools.

Acknowledgements

First of all, I would like to thank my supervisors **Rósa Björk Barkardóttir**, **Aðalgeir Arason** and **Laufey Ámundadóttir** for introducing me to the very exciting field of breast cancer genomics by giving me the opportunity to participate in this project. Your excellent supervision and great support is highly appreciated.

Thank you to my wonderful coworkers at the Laboratory of Cell Biology **Guðrún Jóhannesdóttir**, **Inga Reynisdóttir**, **Edda Sigríður Freysteinsdóttir** and **Eydís Þórunn Guðmundsdóttir** for always being helpful and encouraging. You truly are a pleasure to work and spend time with. Special thanks must go to Guðrún and Inga, for providing outstanding guidance and for always being supportive and willing to help.

I would also like to extend gratitude to **Vilmundur Guðnason**, **Guðný Eiríksdóttir** and **Albert Vernon Smith** at the Icelandic Heart Association for providing control samples and frequencies, to **Óskar Örn Hálfdánarson** for sharing his knowledge and helping me navigate when I was taking my first steps working on this project and to coworkers at the Department of Pathology, especially **Bjarni A. Agnarsson** and **Sigrún Kristjánsdóttir,** for their help.

Last but not least, I want to thank my family and friends for their invaluable support in everything I do and my boyfriend Einar for being an endless source of encouragement and interest.

This project was supported by the Research fund of Landspitali University Hospital, The Nordic Cancer Union and the Icelandic Association “Walking for Breast Cancer Research”.

Contents

List of figures

List of tables

Abbreviations

53BP1: p53 Binding Protein 1

aa: amino acid

ACMG: American College of Medical Genetics and Genomics

AGES: Age, Gene/Environment Susceptibility study

AP: Alkaline Phosphatase

ASPSCR1: Alveolar Soft Part Sarcoma Chromosome Region, Candidate 1

AT: Ataxia Telangiectasia

ATM: Ataxia Telangiectasia Mutated

BARD1: BRCA1-Associated RING Domain Protein 1

BC: Breast Cancer

BCAC: The Breast Cancer Association Consortium

BRCA1: Breast Cancer 1, Early Onset Gene

BRCA2: Breast Cancer 2, Early Onset Gene

BRCT: BRCA1 C-Terminal

BRIP1: BRCA1 Interacting Helicase 1

CDH1: Cadherin-1

CDK7: Cyclin-Dependent Kinase 7

cDNA: Complementary DNA

CG: Complete Genomics

CHEK2: Checkpoint-Kinase 2

CIMBA: Consortium of Investigators of Modifiers of *BRCA1/2*

cM: Centimorgan

CNV: Copy Number Variation

Condel: Consensus Deleteriousness Score

COPZ2: Coatomer Protein Complex, Subunit Zeta 2

CQ: Call Quality

CRC: Colorectal Cancer

CS: Cowden-Syndrome

CSF: Cytostatic Factor

DCAF7: DDB1 and CUL4 Associated Factor 7

DDR: DNA Damage Response

DNA: Deoxyribonucleic Acid

dNTP: Deoxyribonucleotide

DSB: Double Strand Break

EDTA: Ethylenediaminetetraacetic Acid

EMBL-EBI: European Molecular Biology Laboratory – European Bioinformatics Institute

ER: Endoplastic Reticulum

ER: Estrogen Receptor

ERCC5: DNA Repair Protein Complementing XP-G cells

ERCC6: DNA Excision Repair Protein ERCC-6

ESP: Exome Sequencing Project

EtOH: Ethanol

EXO1: Exonuclease 1

FANCD1: Fanconi Anemia Type D1

FANCJ: Fanconi Anemia Type J

FANCM: Fanconi Anemia Type M

FATHMM: Functional Analysis Through Hidden Markov Models

GERP: Genomic Evolutionary Rate Profiling

GID8: Glucose-induced Degradation Protein 8 Homolog

GLUT4: Glucose Transporter Type 4

GWAS: Genome-Wide Association Studies

GWL: Genome-Wide Linkage

HBC: Hereditary Breast Cancer

HBOC: Hereditary Breast- and Ovarian Cancer

HDGC: Hereditary Diffuse Gastric Cancer

HGMD: Human Gene Mutation Database

HPA: The Human Protein Atlas

HR: Homologous Recombination

HWE: Hardy-Weinberg Equilibrium

ICR: Icelandic Cancer Registry

IDC-NOS: Invasive Ductal Carcinoma, Not Otherwise Specified

IHA: Icelandic Heart Association

ILC: Invasive Lobular Carcinoma

INMT: Indolethylamine N-Methyltransferase

IVA: Ingenuity® Variant Analysis™

LFS: Li-Fraumeni Syndrome

LIG1: DNA Ligase 1

LOF: Loss-Of-Function

LOH: Loss-Of-Heterozygosity

LSH: Landspitali University Hospital

MAF: Minor Allele Frequency

MAF: Mutation Analysis Facility

MAPK1IP1L: Mitogen-Activated Protein Kinase 1 Interacting Protein 1-Like

MDC1: Mediator of DNA Damage Checkpoint Protein 1

miRNA: MicroRNA

MMR: Mismatch Repair

MPS: Massively Parallel Sequencing

MRE11: Double-Strand Break Repair Protein MRE11

mRNA: Messenger RNA

MTMR3: Myotubularin Related Protein 3

MYT1: Myelin Transcription Factor 1

NBS1: Nibrin

NCOR2: Nuclear Receptor Corepressor 2

NGS: Next Generation Sequencing

NHEJ: Non-Homologous End Joining

NHERF-2: Na(+)/H(+) Exchange Regulatory Cofactor NHE-RF2

NHLBI: NIH (National Institutes of Health) Heart, Lung and Blood Institute

NSCLC: Non Small Cell Lung Cancer

OB: Oligonucleotide-Binding

OC: Ovarian Cancer

OR: Odds Ratio

PALB2: Partner And Localizer Of *BRCA2*

PARP2: Poly [ADP-ribose] Polymerase 2

PCR: Polymerase Chain Reaction

PDGFR: Platelet Derived Growth Factor

PEPS: Partial Epilepsy with Pericentral Spikes

PHB2: Prohibitin-2

Phred: Phil's Read Editor

PI3: Phosphoinositide 3-Kinase

PIF1: ATP-Dependent DNA Helicase PIF1

PJS: Peutz-Jeghers Syndrome

POLG2: DNA Polymerase Subunit Gamma-2, Mitochondrial

PolyPhen-2: Polymorphism Phenotyping v2

PPP6R1: Protein Phosphatase 6, Regulatory Subunit 1

PR: Progesterone Receptor

PROVEAN: Protein Variation Effect Analyzer

PRS: Polygenic Risk Scores

PTEN: Phosphatase and Tensin Homolog

RAD50: DNA Repair Protein RAD50

RAD51: DNA Repair Protein RAD51

RD: Read Depth

RECQL4: ATP-Dependent DNA Helicase Q4

REFINE: Risk Evaluation For INfarct Estimates Reykjavik Study

RNA: Ribonucleic Acid

RPA: Replication Protein A

RR: Relative Risk

SIFT: Sorting Intolerant From Tolerant

SLC9A3R2: Solute Carrier Family 9, Subfamily A (NHE3, Cation Proton Antiporter 3), Member 3 Regulator 2

SNP: Single Nucleotide Polymorphism

ssDNA: Single-Stranded DNA

STK11: Serine/Threonine Kinase 11

SUPT16H: FACT Complex Subunit SPT16

TBE: Tris-Borate-EDTA

TFE3: Transcription Factor E3

TICRR: Treslin

TNBC: Triple Negative Breast Cancer

TP53: Tumor Protein P53

Tris: Tris(hydroxymethyl)aminomethane

TRMT44: tRNA Methyltransferase 44 Homolog

TROAP: Trophinin Associated Protein

UCSC: University of California Santa Cruz

WES: Whole Exome Sequencing

WGS: Whole Genome Sequencing

WHO: World Health Organization

ZNF488: Zinc Finger Protein 488

ZNF534: Zinc Finger Protein 534

ZNF703: Zinc Finger Protein 703

1 Introduction

1.1 Breast cancer

Breast cancer (BC) is the second most common cancer in the world and the 5th leading cause of cancer deaths with an estimated 1.67 million diagnoses and 522,000 deaths in 2012. BC is both the most common cancer in women and the leading cause of cancer deaths, accounting for a total of 25.2% of new cancer cases and 14.7% of cancer related deaths in 2012. Rates of BC incidence vary greatly between different regions of the world, ranging from 27 per 100,000 in Eastern Asia and Middle Africa and up to 92 and 96 per 100,000 in Northern America and Western Europe, respectively. Although incidence rates vary greatly around the world, mortality rates are quite similar. In 2012, the mortality rate was 14.9 per 100,000 in more developed regions and 11.5 in less developed regions (figure 1). This can be attributed to earlier detection and better survival in developed regions (1). In Iceland, BC is also the most common cancer diagnosed in women, accounting for 29.7% of all cancers diagnosed between 2008 and 2012. The BC incidence rates have been steadily increasing since the Icelandic Cancer Registry (ICR) opened in the 1950s, from an average of 38.5 per 100,000 in the period of 1958-1962 up to 93 per 100,000 in 2008-2012. Although there has been a large increase in 5 year relative survival rates, from 60% in 1959-1968 to 90% in 1999-2008, this has not been enough to counter the growing incidence rates as mortality rates have also increased, albeit slightly, from 14.8 deaths per 100,000 in 1958-1962 to 16.4 per 100,000 in 2008-2012 (2).

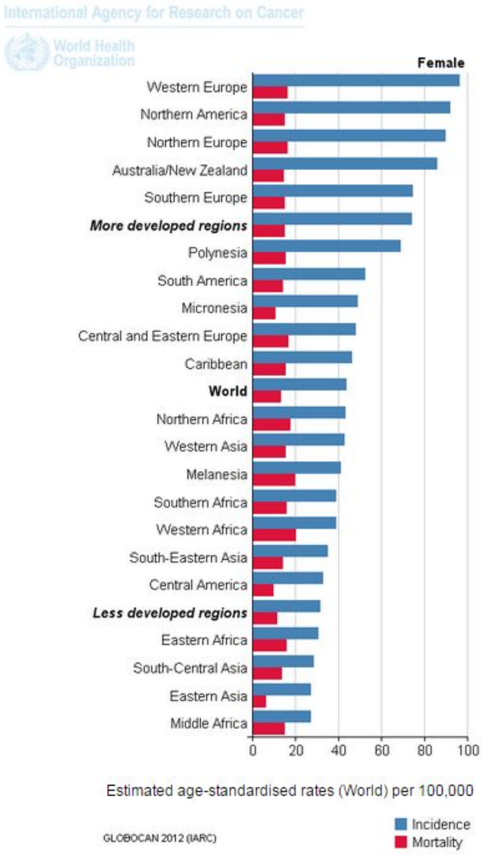


Figure 1: The estimated age-standardised rates of incidence (blue) and mortality (red) per 100,000 women in different regions of the world in 2012 (adapted from the WHO: International Agency for Research on Cancer, http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx).

Breast cancer is a disease characterized by malignant tumor growth in the glandular tissue of the breast. It is a heterogeneous disease and can be classified in a few different ways. First of all, BC tumors are classified by histological appearance. Until recently, this classification was mainly based on the origin of the tumor within the breast. The majority of invasive BC could be classified as either invasive lobular carcinoma (ILC) if the origin was in the lobules of the breast or as invasive ductal carcinoma, not otherwise specified (IDC-NOS), if the tumor had no specific differentiating features. In the latest edition of the WHO Classification of Tumors of the Breast, the name of IDC-NOS has been changed to invasive breast carcinoma of no special type (NST) as it is not certain that these tumors originate in the ducts (3). Invasive breast carcinoma (NST) is by far the most common type of breast cancer, followed by ILC as the second most common and the most common "special" histological subtype (4). Second, BC tumors are clinically classified by their receptor status into three groups; estrogen receptor (ER) positive, HER2 amplified and triple negative breast cancer (TNBC). TNBC tumors do not express ER or progesterone receptor (PR) and also lack HER2 amplification (5). Third, BC tumors can be further classified by their molecular subtype. In 2000, Perou *et.al.* described four molecular subtypes of breast cancer based on a cDNA microarray analysis; luminal, HER2 amplified, basal-like and normal-breast-like (6). In the following year, the same group published another study in which they found that the luminal subgroup could be separated into at least two groups with different gene expression profiles and prognosis (7). The tumors of the luminal subgroups are ER positive and can therefore be treated with endocrine therapy. The difference between the two luminal groups is mainly that the luminal A subgroup exhibits low expression of proliferative markers and very good prognosis while the luminal B group has higher proliferation rates and worse prognosis. In spite of usually being an aggressive form of cancer, the HER2 amplified subgroup is a clinical success since HER2 can be effectively targeted therapeutically (5). The basal-like subtype is also an aggressive form of cancer. Basal-like tumors are generally TNBC and therefore do not respond to targeted therapies (5, 8). The normal-breast-like group is still poorly characterized, and could possibly be an artifact of high levels of normal tissue contamination. A few other molecular subtypes have been identified, including molecular apocrine tumors, the interferon subtype and the claudin-low subgroup. The clinical significance of these subgroups has yet to be determined (8).

Not all individuals are at equal risk of being diagnosed with BC. First of all, men are about 100 times less likely to develop BC than women. Other risk factors that are beyond our control include age and ethnicity. When it comes to age, only 1 in 8 invasive BC cases are diagnosed in women under the age of 45 while 2 of 3 invasive breast cancers are found in women over the age of 55. Regarding ethnicity, white women are more likely to be diagnosed with breast cancer than women of other ethnic groups, but African-American women are more likely to die from the disease (9). Women subjected to longer exposure to endogenous sex hormones related to menstrual duration are also at higher risk of being diagnosed with BC, and early age of menarche increases risk. BC risk is also increased by exogenous exposure to sex hormones through hormonal replacement therapy or use of contraceptives. Other lifestyle factors such as not having children (nulliparity) or having them later in life, low physical activity, obesity, increased alcohol intake and smoking can also increase BC risk (9-11). One of the most important risk factors, however, is family history of the disease (11).

1.2 Familial breast cancer

Epidemiological studies have revealed that women who have a first-degree relative diagnosed with BC are about twice as likely to get the disease as compared to the general population (12). The risk increases with increasing number of relatives with BC (11, 13). Although it could be argued that a shared environment and a similar lifestyle influence these numbers, twin studies suggest that around 27% of BC are caused by hereditary factors (14). BC cases without a family history of the disease are called sporadic, and account for 70-75% of all BC. Although sporadic BC cases are mainly caused by non-hereditary factors, polygenic inheritance (where many genetic variants are involved, each with a small effect on BC risk) likely has a role as well (15-17). BCs caused by inherited germline mutations are described as being either familial or hereditary and account for 15-20% and 5-10% of BC cases, respectively. A family with more cases of BC than expected by chance (generally two or more first- or second-degree affected relatives) without a specific pattern of inheritance, is classified as familial BC. Hereditary BC (HBC), on the other hand, has an autosomal dominant inheritance pattern indicating a strong inherited component (figure 2) (18).

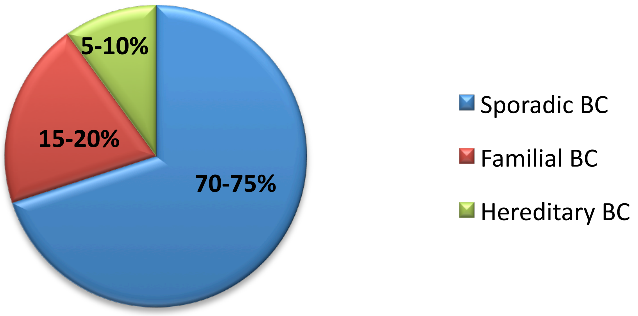


Figure 2: A pie chart showing the distribution of BC cases into groups based on their heredity. Sporadic BC cases are thought to be mainly caused by non-hereditary factors. Cases with two or more first- or second-degree relatives affected are considered familial and when BC is inherited in an autosomal dominant pattern it is classified as hereditary BC (HBC) (14-16).

For over three decades, scientists have been working tirelessly trying to identify the genomic variants underlying genetic predisposition to BC. Substantial progress has been made so far with the discoveries of a number of susceptibility variants. These variants can be divided into the following three groups depending on their frequency in the general population and their level of risk: 1) rare and highly penetrant mutations, 2) rare variants with moderate penetrance, and 3) common variants with low penetrance (10-13). The terms describing the penetrance of pathogenic variants are well defined. Highly-penetrant variants are associated with a relative risk (RR) of 5 or higher, moderately penetrant variants with a RR of 1.5-5 and variants with low penetrance with a RR of under 1.5 (13, 18). While the terminology regarding the prevalence of variants is flexible, variants with moderate to high penetrance generally have a frequency of under 1% in the general population and are considered rare, while most known variants with low penetrance have a frequency of over 5% and are considered common (figure 3) (10, 13). Due to their different levels of risk, variants in each of the three groups contribute to BC susceptibility in a different manner. While single pathogenic mutations in genes with high penetrance can be enough to cause BC, variants with moderate and low effects on risk are more likely to contribute to BC susceptibility in a polygenic manner where many individual variants come together within a family and their combined effect explains the substantially elevated risk (10, 12, 18, 19). Variants of lower penetrance are also likely to act synergistically with environmental factors and lifestyle (18, 19).

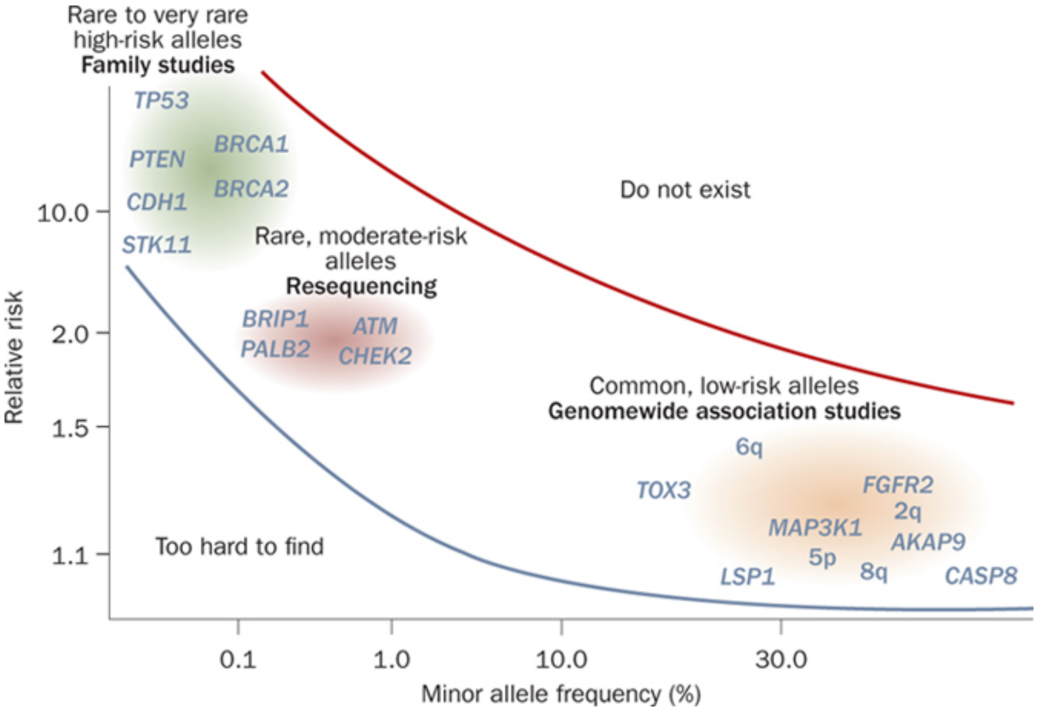


Figure 3: Breast cancer susceptibility genes and loci. Known high-risk BC susceptibility genes are highlighted in the green area and confirmed moderate-risk genes are highlighted in the red area. These are all rare variants with a frequency under 1% in the general population. Well-established low-risk genes and loci are highlighted in the orange area. Approximately 100 low-risk loci, detailed further in the "Common low-penetrance variants" chapter on pages 29-30, have been identified through genome-wide association studies (GWAS). The methods that were used to identify the genes within each group of variants are written in bold text. Variants below the blue line probably exist, but very big sample sizes are needed to confirm them. Variants above the red line are not likely to exist (adapted from Harris & McCormick, 2010 (20)).

1.3 Rare mutations with high penetrance

1.3.1 *BRCA1* and *BRCA2*

The first gene to be linked to hereditary BC was breast cancer 1, early onset gene (*BRCA1*). *BRCA1* was first linked to chromosome 17q in 1990 and was cloned in 1994 (21). Pathogenic mutations in *BRCA1* generally confer a 60-85% lifetime risk of BC (11). A genome-wide association study (GWAS) published in 2013 concluded that although the average BC risk for *BRCA1* mutation carriers is around 65%, the risk can go as low as 28-50% for the 5% of carriers with the lowest risk and as high as 81-100% for the 5% of carriers with the highest risk (22). *BRCA1* is a very large gene containing 24 exons and mutations occur throughout the coding sequence of the gene. Hundreds of pathogenic mutations have been described, most of them small insertions or deletions (indels) that shift the reading frame (frameshift mutations) causing either the translation of a truncated protein or nonsense-mediated decay. The BRCA1 protein has 1863 amino acid residues and plays an important role as a "gatekeeper of genomic integrity", carrying out roles in various cellular processes including DNA repair, checkpoint control and spindle regulation (23, 24). BCs caused by *BRCA1* mutations tend to have different clinical features than those of non-*BRCA1* mutated BC cases. First of all, over 90% of BCs in *BRCA1* mutation carriers are ER-negative and *BRCA1* associated tumors are more likely to lack HER2 amplification than sporadic tumors (11, 13). Second, *BRCA1* associated BCs have an immunohistological profile that resembles sporadic basal carcinomas. Although these characteristics would imply a worse prognosis, studies regarding the prognosis of BC patients with *BRCA1* mutations are conflicting and a recently published meta-analysis found that current evidence does not support worse BC survival of *BRCA1* mutation carriers (11, 25). Third, women with *BRCA1* mutations are not only at high risk of getting a primary BC but they also have a 64% risk of contralateral BC before they reach 70 years of age (11).

In 1994, Wooster *et.al*. mapped a second BC susceptibility gene to chromosome 13q (26). In the following year, the same group managed to identify the associated gene, breast cancer 2, early onset gene (*BRCA2)* (27). Pathogenic mutations in *BRCA2* confer a 40-85% lifetime risk of BC and there is more variability of the risk associated with mutations in *BRCA2* than with mutations in *BRCA1*, which suggests a more modifiable gene. *BRCA2* is an even larger gene than *BRCA1*, containing 27 exons, and like in *BRCA1*, mutations are found throughout the gene with most of the pathogenic ones being frameshift mutations. The BRCA2 protein has 3418 residues, almost twice as many as BRCA1, and is known to play an important role in DNA repair (11, 24). Clinically relevant features associated with *BRCA2* BCs are not as obvious as with *BRCA1* tumors. Although lobular carcinoma is more common in *BRCA2* associated tumors than in *BRCA1* tumors and these tumors are more often ER+ than sporadic cases, prognosis for *BRCA2* tumors is similar to sporadic BCs and they generally have similar features (11).

Early work on BRCA research found that cells deficient of BRCA1 or BRCA2 accumulate genetic abnormalities, which signify defects in DNA damage response (DDR). Further research has implicated both proteins in a DNA repair mechanism called homologous DNA recombination (HR) (24, 28). HR is an error-free DDR mechanism that mammalian cells have evolved to repair double strand DNA breaks (DSBs). When DSBs happen as a by-product of replication in the S/G2-phase of the cell-cycle, the undamaged sister chromatid is situated nearby and can be used as a template to guide high-fidelity repair of the broken chromatid. BRCA1 is important for the initiation of HR. BRCA1 C-terminal (BRCT) domains recruit it to damage sites, where it displaces the HR-suppressing factor p53 binding protein 1 (53BP1) from broken ends, triggering end resection. BRCA2 acts in the following steps of HR, working more directly in the resolution of lesions. After end resection, the exposed single-stranded (ss) DNA is coated with a protein called replication protein A (RPA), which is subsequently displaced by BRCA2, allowing loading of the recombination enzyme RAD51 to the ssDNA. The RAD51-ssDNA filament, stabilized by BRCA2, mediates synapsis with the sister chromatid to initiate strand exchange allowing for the damaged chromatid to be repaired. In the absence of BRCA1 or BRCA2, replication-associated DSBs are repaired by error-prone mechanisms such as non-homologous end joining (NHEJ), causing genomic instability and an increased mutation load that can drive malignant transformation (24, 28). Although BRCA1 and BRCA2 are inherited in an autosomal dominant manner, they act as recessive cancer genes, usually through loss-of-heterozygosity (LOH) where the normal allele is lost in the tumor (10, 12). There is however emerging evidence suggesting that heterozygous mutations in

BRCA2 might suffice to drive carcinogenesis in some tissues (24, 28). A simplified view of the function of BRCA1 and BRCA2 in HR (and other cellular processes) as well as a schematic representation of their functional domains can be seen in figure 4.

Although pathogenic mutations in *BRCA1* and *BRCA2* have the greatest effect on BC risk, they also confer an increased risk of other cancers. Mutations in both genes significantly increase the risk of ovarian cancer (OC). For *BRCA1* mutation carriers the lifetime risk of OC is 40-60% and for carriers of *BRCA2* mutations the lifetime risk of OC is up to 30% (11). Defective *BRCA1* also increases the risk of prostate cancer and fallopian tube cancer (19, 29) and has been linked to pancreatic cancer and male breast cancer (29). Pathogenic mutations in *BRCA2* increase risk of prostate cancer, male breast cancer and pancreatic cancers and have been linked to biliary cancers and melanoma (13, 19, 29). Biallelic mutations in *BRCA2* cause Fanconi anemia type D1 (FANCD1), a condition characterized by developmental anomalies and substantially increased risk of childhood cancers. Biallelic mutations in *BRCA1* have rarely been described and are probably embryonic lethal in most cases (11, 12, 19). However, a recent publication reports the presence of biallelic *BRCA1* mutations in a woman that presented with symptoms consistent with a Fanconi anemia-like disorder, suggesting *BRCA1* is a new Fanconi anemia subtype (type S) (30).

Pathogenic mutations in the BRCA genes are very rare in the general population, with *BRCA1* mutations estimated to be observed in 1 of every 800 individuals and *BRCA2* mutations in 1 out of 500 individuals, corresponding to a frequency of 0.125% and 0.2% respectively (13). However, in certain populations, particular mutations are more common due to a founder effect, which occurs when small groups of people have remained isolated over a long time and the consequent interbreeding results in a normally rare mutation becoming more frequent within that population (31). The best-known examples of founder mutations are those in the Ashkenazi Jewish population, where three BRCA mutations (185delAG and 5382insC in *BRCA1* and 6174delT in *BRCA2)*, have an overall rate of 2.6% and account for approximately 10% of familial cases (11, 19, 23). Founder mutations are also prevalent in the Nordic countries. In Norway, four *BRCA1* mutations (1675delA, 816delGT, 3347delAG and 1135insA) account for the majority of hereditary breast- and ovarian cancer (HBOC) cases, and only one of these (1135insA) has been found in other ethnic groups. In Sweden, the most common BRCA mutation is 3171ins5 in *BRCA1*. This mutation, located on a conserved 3.7 cM haplotype thought to have originated 50 generations ago, accounts for 70% of all BRCA mutations in western Sweden. In Finland, 11 mutations account for 84% of all BRCA mutations. Some of these mutations, for example IVS11+3A>G in *BRCA1* and 9345+1G>A in *BRCA2*, can only be found in Finland (31). In Iceland, one variant has been reported in each BRCA gene. The *BRCA2* mutation, 999del5, is about ten times more frequent than the *BRCA1* mutation, G5193A, and is found in approximately one fourth of HBC families in Iceland. The frequency of the *BRCA2*-999del5 mutation in Iceland is around 0.5% and it is found in 7-8% of those diagnosed with BC in Iceland (32-34).

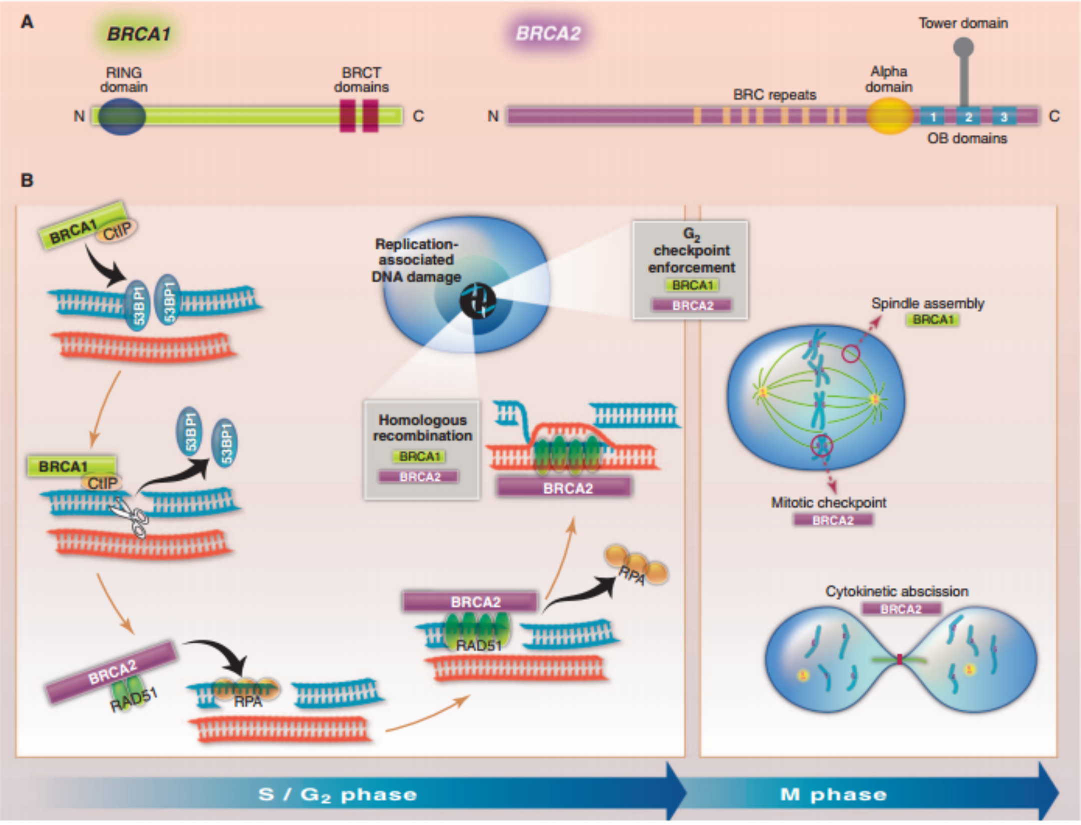


Figure 4: Known roles of BRCA1 and BRCA2. Figure 4a shows the main structural domains of the human BRCA1 and BRCA2 proteins and figure 4b shows how BRCA1 and BRCA2 contribute to homologous recombination (HR) and other cellular processes in different stages of the cell-cycle. BRCT domains at the C-terminal of BRCA1 recruit it to damaged sites where BRCA1 helps to initiate HR by displacing 53BP1 and thus triggering end resection (scissors). The single-stranded (ss)DNA is then coated with replication protein A (RPA). The oligonucleotide-binding (OB) domains of BRCA2 are capable of binding ssDNA and displacing RPA and the BRC repeats bind RAD51 recombinase and recruit it to the lesion. BRCA1 and BRCA2 also have roles in G2 checkpoint enforcement, along with roles in mitotic spindle assembly and cytokinetic abscission (adapted from Venkitaraman, 2014 (24)).

1.3.2 Other high-penetrance genes

Since the discovery of *BRCA1* and *BRCA2*, four other high-penetrance BC susceptibility genes have been identified. These genes are tumor protein P53 (*TP53*), phosphatase and tensin homolog (*PTEN*), cadherin-1 (*CDH1*) and serine/threonine kinase 11 (*STK11*). Like *BRCA1* and *BRCA2,* these genes are tumor-suppressors in which pathogenic mutations cause cancer syndromes (10-12, 18, 19). Germline mutations in *TP53* cause Li-Fraumeni syndrome (LFS). Individuals with LFS are at high risk of getting BC as well as other cancers (12). Approximately 30% of female *TP53* mutation carriers will develop BC before the age of 30 (11, 35), and they have a greater than 90% risk of getting BC by the age of 60 (18). TP53 has an essential role in cell-cycle control, which has earned it a status as the “guardian of the genome”. Impeding its function is valuable to a growing tumor and therefore it is not surprising that *TP53* is the most commonly mutated gene in human cancers (11, 35). Germline mutations in *PTEN* are the cause of Cowden-syndrome (CS), which is characterized by the formation of benign hamartoma tumors throughout the body. CS also increases risk of various cancers, with BC being the most common. Women with CS have a lifetime risk of BC of up to 50%. In 1998, germline mutations in *STK11* were identified as the cause of Peutz-Jeghers syndrome (PJS). PJS is a rare autosomal dominant disorder which causes growth of multiple benign polyps in the gastrointestinal tract and mucocutaneous pigmentations of the lips, buccal mucosa and digits. Individuals with PJS are at an increased risk for certain cancers, and women with PJS have a 32% risk of getting BC before the age of 30 (18, 19). The latest gene to be identified as a high-risk BC susceptibility gene is *CDH1*, in which germline mutations have been associated with hereditary diffuse gastric cancer (HDGC). According to estimations, women from HDGC families that carry a mutation in *CDH1* have a 39% risk of developing BC before the age of 80 (18). With the exception of *TP53*, the contribution of the loss of these genes to cancer pathogenesis is not all that well understood. The tumor-suppressor activity of *PTEN* is thought to be related to its function as a lipid phosphatase, which regulates the mTOR pathway, and to its function as a protein phosphatase, which has a role in cell-cycle arrest and inhibition of invasion. PTEN localization to the nucleus also seems to be necessary for DSB repair and it appears to regulate CDH1 tumor-suppression in the nucleus as well (36, 37). Loss of STK11 may possibly cause the mTOR pathway to become hyperactive in HER2 amplified BC (38).

1.4 Rare variants with moderate penetrance

The discovery of *BRCA1* and *BRCA2* and the increasing knowledge of their roles in DNA repair guided researchers towards searching for predisposing variants in other genes known to interact with *BRCA1*/*2* or act in the same pathways. Such studies have identified variants in several genes that are likely to contribute moderately to predisposition for BC. Some of these genes are now well established, most notably ataxia telangiectasia mutated (*ATM*), checkpoint-kinase 2 (*CHEK2*), BRCA1 interacting helicase 1 (*BRIP1*) and partner and localizer of *BRCA2* (*PALB2*). All of these genes are involved in DNA repair mechanisms and have been shown to confer a two- to fourfold risk of BC (10, 13, 19, 39). *ATM* encodes a checkpoint kinase that plays a vital role in both HR and in cell-cycle progression. In HR, damage recognition by ATM is required and in cell-cycle progression, phosphorylation of BRCA1 by ATM at the G1/S checkpoint is essential (28). Biallelic mutations in *ATM* cause ataxia telangiectasia (AT), a disorder characterized by progressive cerebellar ataxia, oculomotor apraxia, conjunctival telangiectasia, immunodeficiency and increased risk of cancer (11). Heterozygous mutations in *ATM* confer a twofold increase in BC risk (18, 19). CHEK2 also has roles in checkpoint control and HR, for example by phosphorylating BRCA1, facilitating its role in DSB repair. The most common mutation in *CHEK2* is 1100delC, which is seen in up to 1%-2% of some populations and increases BC risk by two- to threefold (18, 19, 28). The BRIP1 protein is a binding partner of BRCA1, interacting with its C-terminus BRCT domain, and was therefore investigated as a BC susceptibility gene. In heterozygous carriers with a strong family history of BC, pathogenic *BRIP1* mutations are associated with a RR of 2.0. Biallelic mutations in *BRIP1* result in Fanconi anemia type J (FANCJ) (11, 19, 24). PALB2 is a partner protein of BRCA1/2 that bridges formation of a BRCA1-BRCA2 complex that assists in their localization at the site of DNA-damage (24). Pathogenic mutations in *PALB2* were identified by studying BC families without mutations in *BRCA1/2*, and have been associated with a RR of 2.3 (11). Variants in other DNA repair genes have also been shown to moderately increase the risk of BC in certain populations, for example in *MRE11*, *RAD50* and *NBS1* which together form the MRN complex, a vital sensor of DNA damage (18, 23). In 2004, a Finnish study associated a *BARD1* mutation (Cys557Ser) with a moderate increase in BC risk (40). Additional association studies have been performed, and although some support the initial results (41-43), others have failed to confirm these findings (44-46). Recently, two nonsense mutations in *FANCM* (c.5101C>T and c.5791C>T) have been shown to moderately increase BC risk (47, 48) but these have yet to be confirmed in additional studies.

1.5 Common low-penetrence variants

Of the three groups of BC susceptibility variants, the group containing common low-penetrance variants is the most recent to emerge. In 2013, 72 loci harboring such variants had been identified, and in 2015 this number has increased to approximately 100 (13, 49). Some of these loci lie in regions with no known protein-coding genes, but others are located within or near genes. These nearby genes are involved in various biological processes, such as DNA repair, mammary gland development and the ER-pathway and can effect tumor growth and aggressiveness (figure 5) which indicates that these variants might contribute to BC pathogenesis in a more complex manner than high- and moderate-penetrance variants (12, 13). All of the common variants associated with BC risk confer risks of less than 1.5 times higher than the general population, and several studies have suggested that these risks combine with genetic or non-genetic risk factors in a multiplicative rather than an additive way. In addition, many of these variants are differentially associated with BC by ER-status (17, 49).

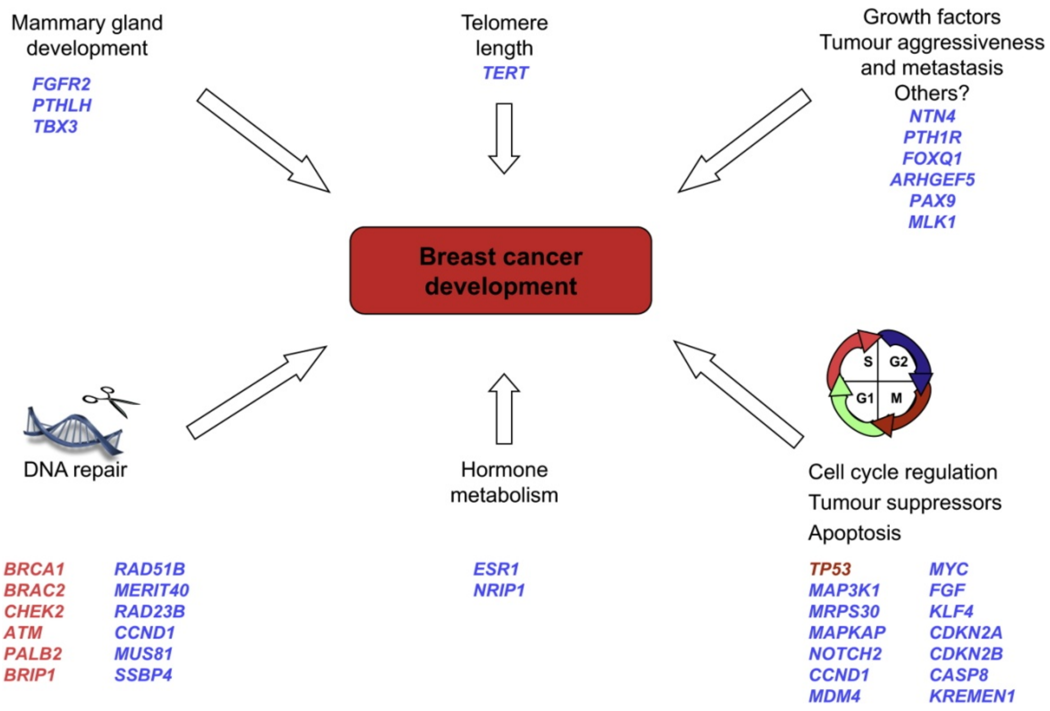


Figure 5: Mechanisms involved in breast cancer (BC) susceptibility. High- and moderate-penetrance genes are listed in red and BC susceptibility candidate genes found in or near common BC susceptibility loci are listed in blue (adapted from Ghoussaini et.al., 2013 (13)).

As the search for more common variants associated with BC risk continues, researchers are also looking into how this information could be of value in a clinical setting. This is a problematic task since these variants are common and the risks conferred by each variant are not high enough to individually be useful for risk prediction. In a recent study conducted by The Breast Cancer Association Consortium (BCAC: http://apps.ccge.medschl.cam.ac.uk/consortia/bcac//about/about.html), the value of using 77 BC susceptibility loci for risk stratification was investigated. Polygenic risk scores (PRS) were constructed to assess the combined effects of these variants on both overall BC risk and on the risk of ER-positive and ER-negative BC separately. The main findings were that for women in the lowest and highest quintiles of the PRS distribution, estimated lifetime risk of BC is 5.3% and 17.2%, respectively. The corresponding risks of ER-positive BC are 4.1% and 15.7%, while the highest lifetime-risk of ER-negative disease is only 2.4%. The PRS affected BC regardless of family history, indicating that polygenic inheritance is a factor in sporadic cases as well as familial. All in all, this study indicates that although each common variant individually has a low impact on risk, screening for these variants could be informative for BC prevention (17).

1.6 Identifying new predisposing variants

Since the 1990s, researchers have made great progress in explaining the genetic elements causing familial cancer by identifying several genes whose loss-of-function increases risk of BC. Even so, over half of the familial risk of the disease is still unexplained. Although high-risk mutations have such a big effect on the risk of breast cancer, they are very rare in the general population and therefore do not explain a large proportion of familial BC risk. In fact, the high-risk genes identified to date are only estimated to account for 20-25% of the familial risk of BC, with *BRCA1* and *BRCA2* accounting for 16-20% of the familial risk (10-12, 18). Genes with moderate penetrance add little to this proportion since they are also very rare and have a smaller effect. It is estimated that moderate -penetrance mutations in *CHEK2*, *ATM*, *BRIP1* and *PALB2* account for less than 3% of the familial risk of BC (10, 12). Finally, although variants with low penetrance only minimally increase risk of BC, due to their high prevalence in the general population and because of the high number of variants identified they explain a substantial portion of familial cases, or an estimated 14% (13). Figure 6 shows the contribution of known BC susceptibility genes and loci to the familial aggregation of BC.

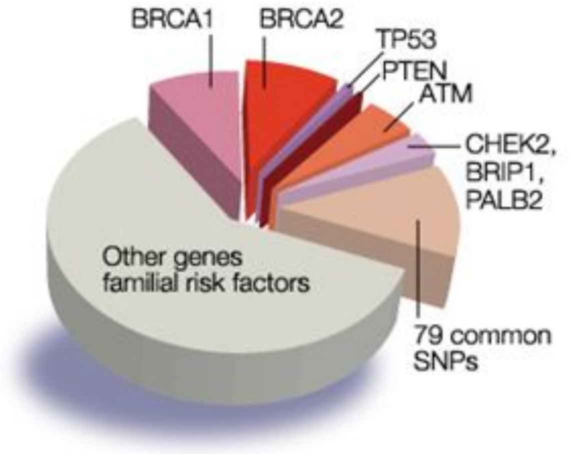


Figure 6: Contribution of known genes to familial aggregation of breast cancer (BC). High-risk genes identified so far only account for 20-25% of the familial risk of BC, and moderate- and low-risk genes together explain less than 20%. Approximately 55% of the familial risk of BC is therefore still unexplained (adapted from Discovery's Edge: Mayo clinic's online research magazine; http://www.mayo.edu/research/discoverys-edge/breast-cancer-predicting-individual-risk) (10-13)

When looking to identify the genetic factors that underlie the remaining familial risk of breast cancer, it is important to first gain understanding of how we have come to know what we do. The high-penetrance genes *BRCA1* and *BRCA2* were discovered through linkage analysis in high-risk BC families and subsequent positional cloning (21, 26, 27). Following the increased knowledge of the function of these genes in DNA repair, researchers successfully identified moderate-penetrance genes using the candidate-gene approach. This approach involves the direct interrogation of coding variants in candidate genes (in this case genes involved in the same pathways as *BRCA1/2*) in large series of genetically enriched cases and controls (10, 12, 13, 19, 23). Following this, researchers tried to associate BC to common single nucleotide polymorphism (SNPs) in candidate genes using a small number of cases and controls. This method initially reported many positive associations but few were convincingly replicated in subsequent studies (50). Improvements in genotyping and sequencing technologies of the human genome have lead to the more agnostic genome-wide association studies (GWAS) (10). Since these studies require thousands of cases and controls to have substantial statistical power to detect common variants that have small effects, this has lead to the formation of international multi-group collaborations such as the previously mentioned BCAC, which is a forum of investigators interested in the inherited risk of BC, and the CIMBA consortium (Consortium of Investigators of Modifiers of *BRCA1/2*), which focuses on identifying genetic modifiers of *BRCA* (13). In figure 3, the main methods used to identify susceptibility genes in each risk-group are written in bold text.

Genome-wide linkage (GWL) studies in families without BRCA mutations have not yet been successful in finding other high-penetrance genes, although they have in some cases succeeded in linking the cause of BC to certain chromosomal locations (12, 51). Since linkage studies may not be powered to detect very rare high-risk mutations this does not exclude the presence of additional high-penetrance genes. This does however suggest that they are likely to be very rare (12, 13, 19, 52). However, most of the missing heritability is most likely explained by a polygenic model, where BC develops because of the cumulative effects of multiple variants with moderate and/or low penetrance (23, 53). The remaining low-penetrance variants will most likely be found through large international collaborations and GWAS, since these have increasingly large numbers of samples from diverse locations and ethnic origins (10). Further moderately penetrant (or even highly penetrant) genes could be identified in population isolates with founder mutations of higher prevalence than the general population. This enrichment of certain mutations provides an advantage in gene identification studies, as was seen recently with the identification of the c.5101C>T mutation in *FANCM* in the Finnish population (47).

1.6.1 DNA repair genes and breast cancer susceptibility

In “The Hallmarks of Cancer”, a seminal paper by Douglas Hanahan and Robert A. Weinberg, six important biological capabilities acquired during the development of human tumors are detailed. In the same paper, genomic instability is named as the single characteristic that can enable the acquisition of these six capabilities. Genomic integrity is maintained by DNA monitoring and DNA repair mechanisms, so it is not surprising that mutations in genes that encode proteins that have vital roles in these mechanisms can greatly increase the risk of cancer (54). Defects in DNA repair mechanisms have been shown to substantially increase risk of several cancers. An example of this is Lynch syndrome, which is caused by germline mutations in DNA mismatch repair (MMR) genes. Lynch syndrome is associated with an increased risk of many cancers, with the greatest effect on the risk of colorectal cancer (up to 80% lifetime risk) and endometrial cancer (up to 60% lifetime risk) (55).

In the case of BC, the majority of the high- or moderate-risk mutations identified so far are protein-truncating mutations (also called loss-of-function (LOF) mutations) in genes directly involved in DNA repair (figure 6). As mentioned previously, moderate-risk mutations were identified by the systemic interrogation of genes known to play roles in DSB repair, following the increased knowledge of the functions of BRCA1 and BRCA2 (13). Additional DNA repair genes have been investigated without many more mutations being revealed, but it is still possible that they exist (12, 19). A recent success story is the discovery of the previously mentioned *FANCM* mutations. *FANCM* is a DNA repair gene, whose protein product activates a DNA damage response when encountering stalled replication forks. The technology used to identify these mutations is called whole exome sequencing (WES) and is a version of next generation sequencing (NGS) (47, 48).

1.6.2 Next Generation Sequencing (NGS)

NGS is a collective term for a set of massively parallel sequencing (MPS) techniques that make it possible to perform large numbers of sequencing reactions simultaneously, and thus make the sequencing process more time- and cost-effective than before. NGS techniques can be used to detect the full spectrum of DNA mutations, including SNPs, indels, copy number variations (CNVs) and chromosomal structural rearrangements. In addition, NGS can be used to study transcriptomes through RNA-sequencing. NGS can be designed to target parts of the genome, such as exomes (WES: Whole exome sequencing) or a subset of selected genes of interest (gene panels) or it can be used to sequence entire genomes (WGS: Whole genome sequencing). There are many different NGS platforms available that each has their own sequencing chemistry. Most platforms share the basic steps of the sequencing process though, which are DNA fragmentation, amplification of the fragments and their subsequent sequencing. Following the sequencing, the resulting sequence tags are computationally aligned to a reference genome and variants identified (56, 57). Although NGS technologies are relatively new, they are already being used in clinical settings as well as in basic research. In the clinic, parallel sequencing of specific genes is used to identify families at risk of BC. Multigene panels for known B loci are available but to this day a consensus among geneticists as to when these should be applied is lacking and their clinical validity has not been fully established (18, 49). In basic research, NGS technologies have been proven to be very useful when studying the genetics of human diseases, *e.g.* in the search for variants that predispose for BC and various other diseases (47, 58-60).

The analysis of NGS data can be a strenuous task, mainly because reads from NGS platforms tend to have more sequencing errors than older methods and the volume of data generated is quite extensive (61). In a typical NGS project, tens of thousands to millions of variants are identified depending on the extent to which the genome is sequenced. Identifying the variants most likely to contribute to a disease and reducing their number to one managable for validation is a challenging task. Common strategies include focusing on LOF variants and pathogenic missense variants. Determining which missense variants are likely to be pathogenic can be problematic, and therefore several computational methods that predict the function of variants, such as SIFT (http://sift.jcvi.org/) and PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/), have been developed. These tools predict the effect of missense mutations on the function of the protein, using various methods such as evolutionary sequence comparison, structural constraint and physiochemical features of amino acids (62, 63).

2 Aims

The general aim of this project was to identify novel BC susceptibility mutations in Icelandic HBC families not explained by mutations in *BRCA1* or *BRCA2*. Two different approaches were used:

1. In the first part of the project, WGS data from 4 BC cases in one HBC family were analyzed. The BC pattern in this family indicates an autosomal dominant inheritance pattern, which implies that a high-risk mutation is segregating within the family. Therefore, the aim of this part of the project was to identify such a mutation. The specific tasks of this part of the project were the following:
    a. Analyze WGS data and select candidate variants with possible large effects on gene function or expression; in this case LOF variants and/or variants predicted to disrupt the binding of a miRNA.
    b. Screen for candidate variants in a set of unselected Icelandic BC cases and control subjects to determine if they may increase risk of BC.
2. In the second part of the project, WGS data from 9 BC cases from three additional HBC families were added to the analysis. The aim of this part of the project was to identify moderate-risk variants that might contribute to the increased risk of BC in these families in a polygenic fashion. The specific tasks of this part of the project were the following:
    a. Analyze WGS data and select candidate variants that are predicted to be deleterious and are located in the coding region of genes that are known to be important for DNA repair.
    b. Determine the frequency of the candidate variants in a set of unselected Icelandic BC cases and controls to determine if they may increase risk of BC.

3 Materials and Methods

3.1 Sample selection

This project is part of a larger ongoing study, “A search for additional breast cancer genes”, at the Laboratory of Cell Biology, Department of Pathology at Landspitali University Hospital (LSH). This study has been approved by the National Bioethics Committee (reference number 11-105-V5 and 11-105-V5-S1) and the National Data Protection Authority (reference number 2001/523 and 2014/679). The individuals responsible for this study are Rósa Björk Barkardóttir, Aðalgeir Arason, Bjarni A. Agnarsson and Óskar Þ. Jóhannsson.

3.1.1 Samples selected for WGS

The WGS data that were analyzed in this project resulted from blood samples from 13 BC cases in 4 Icelandic HBC families. Tumor samples were also sequenced from 4 of these cases. The pedigrees of the families can be seen in figure 7, with arrows marking the cases from which DNA samples were sequenced. The selection of these families and individuals for sequencing was based on 4 criteria: 1) the BC cases within the family show a pattern of dominant inheritance, 2) the family is negative for founder mutations in *BRCA1* and *BRCA2*, 3) the individuals were relatively young when they were diagnosed with BC and 4) high quality DNA samples were available from 3 or more BC cases within the family.

3.1.2 Samples used for the screening of candidate variants

The allele frequency of candidate variants was estimated in DNA samples from a group of unselected BC patients and in controls. The unselected BC samples were collected from Icelandic BC patients diagnosed in the period 1987-2009 and agreeing to participate in the study “A search for additional breast cancer genes”. The BC diagnosis was verified by the ICR or in the registry of the Department of Pathology at LSH. In 1987-1999, blood samples from these BC patients were collected at LSH and in 2002-2009 they were collected at the Research service center located in Nóatún. DNA was isolated from the blood samples by staff members at the Laboratory of Cell Biology, Department of Pathology at LSH.

The control group consists of four different blood sample collections. Two of these collections were performed at the Laboratory of Cell Biology, Department of Pathology at LSH. The former consists of samples from healthy female and male donors with no family history of cancer and the latter consists of samples from an unselected group of individuals. The third collection was performed at the Icelandic Blood Bank and at the Icelandic Heart Association (IHA) and also consists of samples from unselected individuals. The fourth collection was performed at the IHA as a part of their REFINE-Reykjavik study (National Bioethics Committee reference number: 05-112) (64).

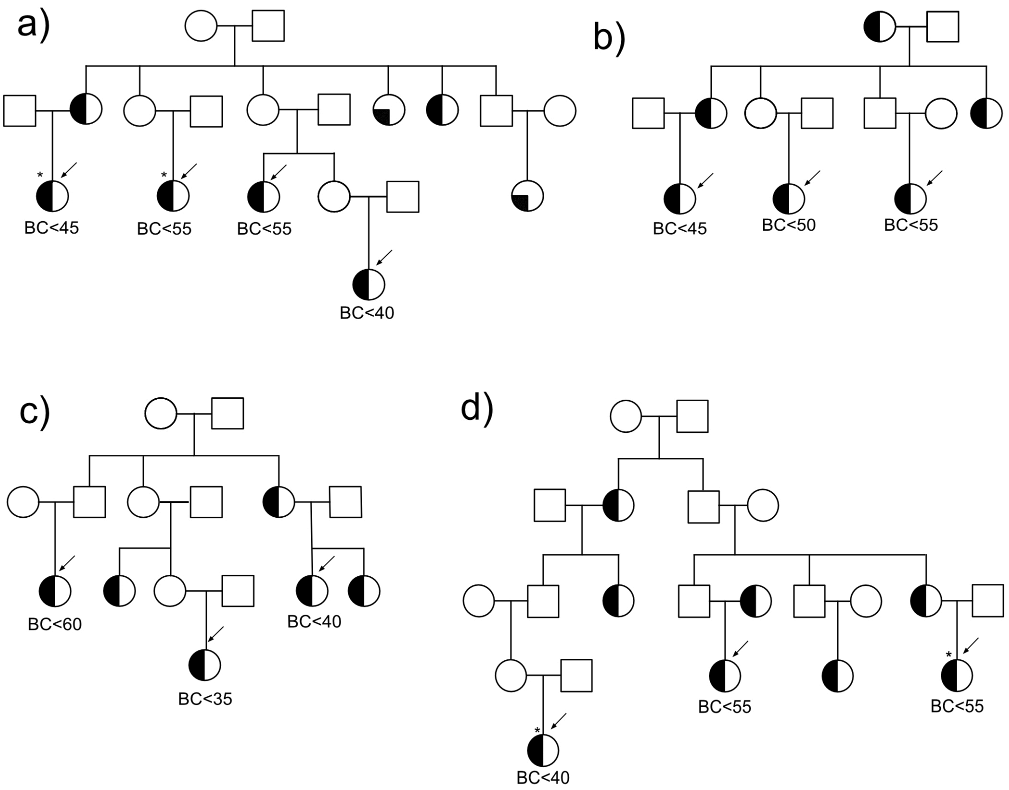


Figure 7: Pedigrees of the families from which the WGS data originated. Circles denote women and squares represent men. Circles that are half-filled with black represent BC diagnosis and circles quarter-filled with black represent OC diagnosis. The arrows point to women from whom blood samples were sequenced. Those from whom tumor samples have also been sequenced are marked with an asterisk. The age of first BC diagnosis is written below each of the labels for women whose samples were sequenced. In the first part of the project family a) was focused on with the aim of identifying a high-risk mutation. In the second part of the project, families b) through d) were added to the analysis with the aim of identifying moderate-risk variants.

3.2 Initial WGS analysis

The WGS of the DNA samples from the study participants, along with alignment and variant calling, was performed by Complete Genomics (http://www.completegenomics.com/). The resulting data were uploaded to Ingenuity® Variant Analysis™ (IVA) (http://www.ingenuity.com/products/variant-analysis) (65), which is a web-based, commercially available application that identifies and prioritizes candidate causal variants. IVA combines analytical tools and integrated content based on published biological evidence and knowledge of disease biology (66). In IVA, it is possible to apply various filters to help identify the variants that are most likely to contribute to a disease of interest. The filters used in this project were the following: 1) Confidence filter, 2) Common variants filter, 3) Predicted deleterious filter, 4) Genetic analysis filter and 5) Biological context filter. The confidence filter allows for filtering out variants that are potentially of low quality by adjusting the minimum call quality (CQ) and/or minimum read depth (RD). The CQ scores are based on phred (Phil's Read Editor) quality scores which measure the quality of the identification of the bases generated by DNA sequencing (67). The common variants filter allows you to "keep only" or "exclude" variants based on their frequency in the following databases; The 1000 Genomes Project – Global dataset (http://www.1000genomes.org/), Complete Genomics (CG) Public Genomes (http://www.completegenomics.com/public-data/) and the NIH Heart, Lung and Blood Institute (NHLBI) Exome Sequencing project (ESP) (http://evs.gs.washington.edu/EVS/). The predicted deleterious filter allows for identification of variants that are either predicted or observed to disrupt gene function or expression. In this filter you can either "keep only" or "exclude" variants based on their pathogenicity and their net gain-of-function or loss-of-function effects. The genetic analysis filter enables filtering based on inheritance models and genotypes, as well as allowing for filtering of the variants based on their frequency in the samples that are being analyzed. Finally, the biological context filter enables identification of variants that, based on current knowledge, are likely to be involved in a biological process of interest (68).

As described in the aims, two different approaches were applied in this project. The filtering process in IVA and the following work was adjusted to each part of the project and thus will be detailed separately.

3.2.1 WGS analysis of family a)

When searching for a high-risk mutation in family a), all of the five IVA filters described above were used with the following settings:

1) The confidence filter was set to keep only variants with call quality (CQ) of at least 20.
2) The common variants filter was set to exclude variants that had a frequency higher than 3% in any of the three databases mentioned above. Although mutations of high-risk are expected to have a frequency of < 1%, this filter was set to 3% to keep the possibility of identifying a more prevalent variant of lesser risk.
3) The predicted deleterious filter was set to keep only variants that were either predicted or observed to be *pathogenic*, *likely pathogenic* or of *uncertain significance* based on literature evidence linking the variant to a phenotype. This assessment is based on the American College of Medical Genetics and Genomics (ACMG) guidelines for the interpretation of sequence variants. In these guidelines, certain criteria for classifying pathogenicity are categorized as *very strong*, *strong*, *moderate* or *supporting*. For example, if a loss-of-function (LOF) variant is in a gene where LOF is a known mechanism of disease, it is considered very strong evidence for pathogenicity. For missense variants, if the same amino acid change has previously been established pathogenic, it is considered strong evidence for pathogenicity. There are also criteria for classifying benign variants, where allele frequency > 5% in any of the incorporated databases is considered

as stand-alone evidence for a variant being benign. The combination of criteria met by variants determine their classification as *pathogenic*, *likely pathogenic*, of *uncertain significance, benign* or *likely benign*. These combinations are detailed in the ACMG guidelines (69). This filter was also set to keep variants that were associated with a net gain-of-function (established in the literature or predicted by TargetScan (http://www.targetscan.org/) to disrupt a miRNA binding site) or associated with a net loss-of-function (frameshift variants, in-frame insertion or deletion (indel) variants, variants that change a stop-codon, missense variants that are not predicted tolerated by SIFT (http://sift.jcvi.org/) or PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/), variants likely to cause splice-site loss, variants that are deleterious to a miRNA, structural variants, variants that disrupt predicted or known promoters or transcription factor binding sites and variants that are within a region known or predicted by the VISTA database (http://enhancer.lbl.gov/) to be an enhancer binding site).

4) The genetic analysis filter was set to keep variants with a dominant inheritance pattern that were present in at least 2 out of the 6 samples that were sequenced in this family. These variants could be heterozygous or het-ambiguous (meaning that the variant is present on at least one chromosome but could possibly be homozygous though).
5) The biological context filter was set to keep only variants in genes that were known or predicted to affect BC and tumors and variants in genes that are within one hop downstream, in biological interaction, of such genes.

The variants that remained after the filtering process in IVA were then exported to Microsoft Office Excel 2007 (Excel) and further filtered. First of all, only variants that were present in at least two blood samples (two individuals) were kept. The CQ requirements were also made more stringent, and variants that didn’t have CQ ≥ 60 in at least one sample were removed. In this part of the project, it was decided to focus on LOF variants, predicted to cause truncation of a protein or altered splicing, as well as variants predicted to disrupt the binding of a miRNA to mRNA.

3.2.2 Combined analysis of all four families

In this part of the project, the focus was on variants located in the coding regions of genes that were directly annotated as double-strand break repair genes or DNA repair genes in the Gene Ontology AmiGO database (http://amigo.geneontology.org/amigo/). When searching for moderate-risk variants in all four families, only the first three IVA filters described above were used. The settings in these filters were the same as described above with the exception of the common variants filter where variants that had a frequency higher than 5% in any of the incorporated databases were now excluded. The variants were then exported from the predicted deleterious filter to Excel and only variants that were present in at least one blood sample (not only in a tumor sample) and had CQ ≥ 60 in at least one sample were kept.

3.3 Validation of candidate variants

Following the variant filtering in family a), the remaining variants were validated in their original sample(s). First, primers were designed, then polymerase chain reaction (PCR) was performed and the amplified region sequenced using Sanger-sequencing.

3.3.1 Primer design

Primers were designed by the following procedure. First, the genomic sequence flanking the variant location (500 bases upstream and downstream) was exported from the UCSC Genome Browser (GRCh37/hg19 assembly) (https://genome-euro.ucsc.edu/cgi-bin/hgGateway). The sequence was uploaded to Sequencher 5.0 (a commercial DNA sequence analysis software produced by the Gene Codes Corporation), which was used to visualize the sequence. Candidates for forward and reverse PCR primers were then selected and subsequently tested for possible 3´-complementarity, hairpin formation and self-annealing sites using the oligonucleotide properties calculator OligoCalc (http://www.basic.northwestern.edu/biotools/oligocalc.html) and their specificity was checked by using the NCBI Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). Sequences of primers used for Sanger-sequencing and their annealing temperatures are listed in table S1.

3.3.2 Polymerase Chain Reaction (PCR)

The polymerase chain reaction (PCR) was used to amplify targeted sequences of the genome for subsequent analysis. For each PCR reaction, the following recipe was followed:

PCR amplification

Reactant (concentration)	Amount
10x PCR Buffer	1.00 µL
dNTP mix (2.5 mM)	0.64 µL
Betaine (5 M)*	2.00 µL
Taq DNA Polymerase (5U/µL)	0.06 µL
Forward primer (20 µM)	0.20 µL
Reverse primer (20 µM)	0.20 µL
DNA	10 ng

*Used to increase specificity of primers when amplifying difficult targets

The total reaction volume was then brought up to 10 µL with ddH_2O.

The reactants used for the PCR were purchased from the following resources: MCLAB (http://www.mclab.com/) for 10x PCR buffer, dNTP mix and Taq DNA Polymerase, Eurofins Genomics (http://www.eurofinsgenomics.eu/) or TAG Copenhagen (http://tagc.dk/) for primers and Sigma-Aldrich (https://www.sigmaaldrich.com/) for Betaine EC No. 2034906. The PCR was performed in a 2720 Thermal Cycler from Applied Biosystems ® using the following program:

PCR program

Step	Temperature (°C)	Time (minutes)
1. Denaturing	94	3:00
2. Denaturing	94	0:30
3. Annealing	52-64	0:45
4. Elongation	72	0:45
Step 2-4 repeated 25-35 times		
5. Elongation	72	10
6. Hold	4	∞

3.3.3 Electrophoresis on agarose gels

Prior to sequencing, to check if the amplification was successful, PCR amplified DNA products were electrophorized on 1-2% agarose gels. The gels were made by the following procedure:

1-2% Agarose gel

Reactant (concentration)	Amount
1x TBE buffer	60 mL
Agarose low EEO	0.6 -1.2 g
Ethidium Bromide (EtBr; 10 mg/mL)	1.8 µL

5x TBE buffer was prepared by the following recipe:

5x TBE Buffer

Reactant (concentration)	Amount
Trizma® base	54 g
Boric acid	27.5 g
EDTA (0.5 M)	20 ml

Total volume brought up to 1 L with dH_2O.

For use: 200 mL 5x TBE buffer mixed with 800 mL dH_2O to make 1x TBE buffer.

First, 1xTBE buffer (Trizma® base from Sigma-Aldrich and boric acid from Merck: http://www.merck.com/index.html) and agarose (from Applichem; https://www.applichem.com/start.html) were mixed together and heated in a microwave oven for 70 seconds. EtBr was added to the melted agar and the mix was then cooled under cold water for a few seconds before it was poured into a cast. Then, 5.0 µL of each PCR product that was to be electrophorized were mixed with 1.0 µL of Blue/Orange 6x loading dye from Promega. A 1 kb DNA ladder (from Thermo Scientific; corporate.thermofisher.com) was run on each gel and 1xTBE was used as a running buffer. Each electrophoresis was carried out for approximately 30 minutes with an electric potential of 90-100V and the gel was subsequently visualized under UV light.

3.3.4 Sanger-sequencing

PCR amplified DNA products were sequenced using Sanger-sequencing to validate candidate variants. First, the DNA products were purified using the following procedure:

Purification reaction

Reactant (concentration)	Amount (µL)
ddH_2O	4.25
Fast AP (1U/µL)	0.5
Exonuclease I (2U/µL)	0.25
Amplified DNA product	2.0

Fast AP (Thermosentitive Alkaline Phosphatase) was purchased from Thermo Scientific (corporate.thermofisher.com) and Exonuclease I (E.Coli) from MCLAB. The purification reaction was performed in a 2720 Thermal Cycler (Applied Biosystems®) using the following program:

Purification program

Step	Temperature (°C)	Time (minutes)
1	37	15
2	85	15
3	4	∞

Next, a sequencing reaction was performed:

Sequencing reaction

Reactant (concentration)	Amount (µL)
ddH_2O	2.5
BigDye® Terminator v1.1, v3.1 5x Sequencing Buffer	1.0
BigDye® Terminator v1.1 Cycle Sequencing RR-100	0.5
Forward or reverse primer (20 µM)	0.1
Purified DNA product	1.0

The BigDye® Terminator v1.1 Cycle Sequencing Kit was purchased from Applied Biosystems® (http://www.appliedbiosystems.com/absite/us/en/home.html) and the sequencing reaction was performed in a 2720 Thermal Cycler (Applied Biosystems®) using the following program:

Sequencing program

Step	Temperature (°C)	Time (minutes)
1	96	0:10
2	50	0:05
3	60	4:00
Steps 1-3 repeated 35 times		
4	4	∞

Finally, 3.5 µL of BigDye® Cleaning Beads (MCLAB) and 20 µL of 85% ethanol (EtOH) were added to the entire sequencing reaction product. All was mixed together and then put on a magnetic plate for 3-5 minutes. With the samples still on the magnetic plate, the ethanol was removed and thrown away and the samples washed with an additional 50 µL of EtOH. The samples were then removed from the magnetic plate and 70 µL of 1x Elution buffer (MCLAB) added to each sample and let sit for 3-5 minutes. The samples were then put on the magnetic plate again for another 3-5 minutes and subsequently moved to a sequencing plate. The samples were sequenced in a 3130xl Genetic Analyzer (Applied Biosystems®) and the resulting sequences analyzed in Sequencher 5.0 (Figure 8).

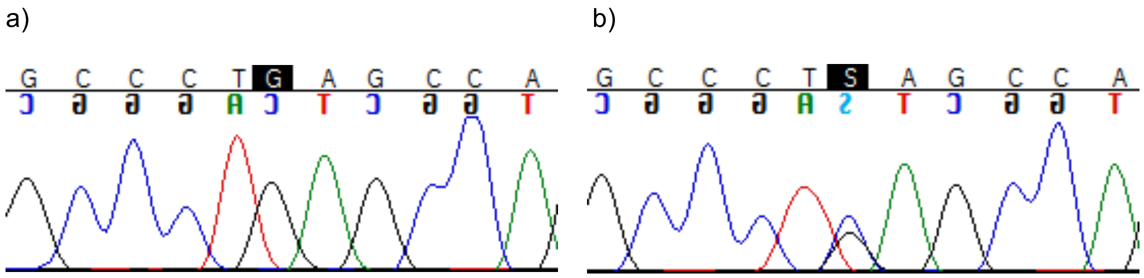


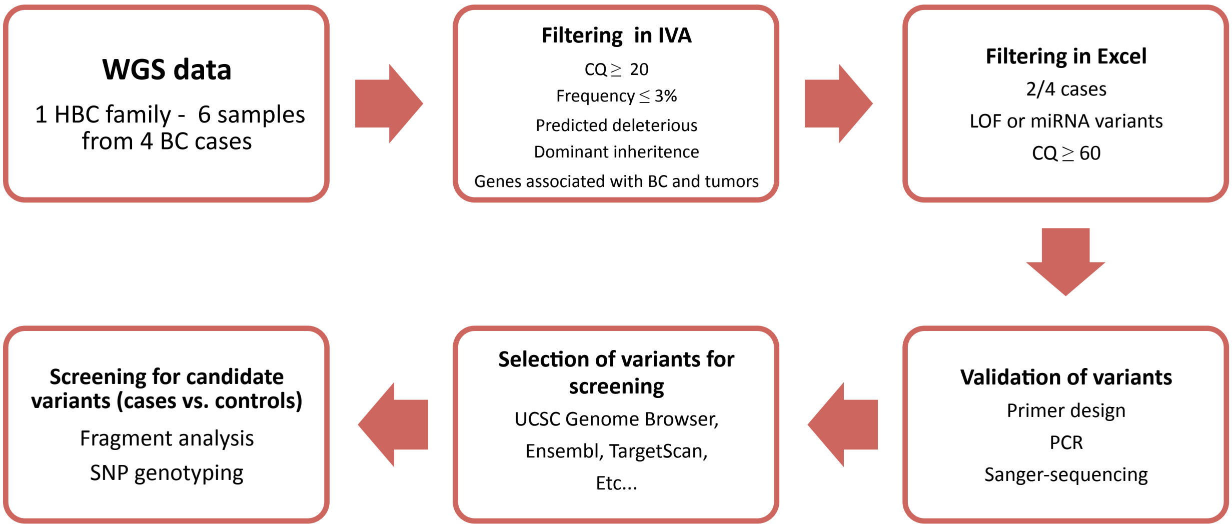
Figure 8: Sanger-sequencing analysis results from Sequencher 5.0. Figure a) shows a sequence from a reference sample where no variant is present. Figure b) shows a sequence where a heterozygous SNP is present; a cytosine (C) replaces a guanine (G) on one of the chromosomes. Cytosine bases are represented by blue peaks, guanine bases by black peaks, thymine (T) bases by red peaks and adenine (A) bases by green peaks.

3.4 Databases used for further assessment of candidate variants

Candidate variants were assessed using various databases. In the first part of the project these were LOF and miRNA variants and in the second part of the project these were variants located in double-strand break repair genes and general DNA repair genes.

3.4.1 Assessing LOF and miRNA variants

For general information on genes harboring candidate variants and for prediction of their possible effect on the protein product the UCSC Genome Browser (70), UniProt (http://www.uniprot.org/), Ensembl (http://www.ensembl.org/index.html), PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) and the Human Splicing Fincer v3.0 (http://www.umd.be/HSF3/index.html) were used. When assessing variants predicted to disrupt a miRNA binding site, the databases used were TargetScan (http://www.targetscan.org/), miRBase (http://www.mirbase.org/), miRanda – mirSVR (http://www.microrna.org/), DIANA microT v5.0 (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=microT_CDS/index) and DIANA TarBase v7.0 (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index). When comparing the expression of genes harboring candidate variants and their protein products, The Expression Atlas (https://www.ebi.ac.uk/gxa/home) and The Human Protein Atlas (http://www.proteinatlas.org/) were used. A schematic representation of the filtering process applied when searching for a high-risk variant in the first part of the project can be seen in figure 9.


WGS data

1 HBC family - 6 samples from 4 BC cases

Filtering in IVA

CQ ≥ 20

Frequency ≤ 3%

Predicted deleterious

Dominant inheritence

Genes associated with BC and tumors

Filtering in Excel

2/4 cases

LOF or miRNA variants

CQ ≥ 60

Validation of variants

Primer design

PCR

Sanger-sequencing

Selection of variants for screening

UCSC Genome Browser, Ensembl, TargetScan, Etc...

Screening for candidate variants (cases vs. controls)

Fragment analysis

SNP genotyping

Figure 9: The filtering process applied when searching for a high-risk mutation in the first part of the project. Whole Genome Sequencing (WGS) data from four BC cases in one Icelandic hereditary breast cancer (HBC) family were analyzed. The variants were filtered in Ingenuity® Variant Analysis™ (IVA) based on their call quality (CQ), frequency, predicted deleteriousness, inheritance pattern and association to breast cancer (BC). Loss-of-function (LOF) variants and variants predicted to disrupt miRNA binding present in at least 2 out of 4 BC cases were focused on. These variants were validated by Sanger-sequencing and then further assessed using various tools and databases. Based on this assessment, candidate variants were chosen for screening in a set of unselected BC cases and controls.

3.4.2 Assessing the deleteriousness of coding variants in DNA repair genes

IVA incorporates three different algorithms which assess the pathogenicity and deleteriousness of variants: IVA assessment, PolyPhen-2 and SIFT. To further assess the deleteriousness of variants located in DNA repair genes, three additional online software tools were used: PROVEAN (Protein Variation Effect Analyzer: http://provean.jcvi.org/index.php), Condel (Consensus Deleteriousness score: http://bg.upf.edu/fannsdb/) and MutationAssessor (http://mutationassessor.org/). If a variant was considered to be pathogenic or deleterious by at least 2 of these tools, it was kept in the analysis. Frequencies from an exome-chip performed on 2983 samples from the AGES-Reykjavik cohort (71) were provided by the IHA and variants found on the list with a frequency between 1-5% were kept for further studies. A schematic representation of the filtering process applied when searching for moderate-risk variants in the second part of the project can be seen in figure 10.

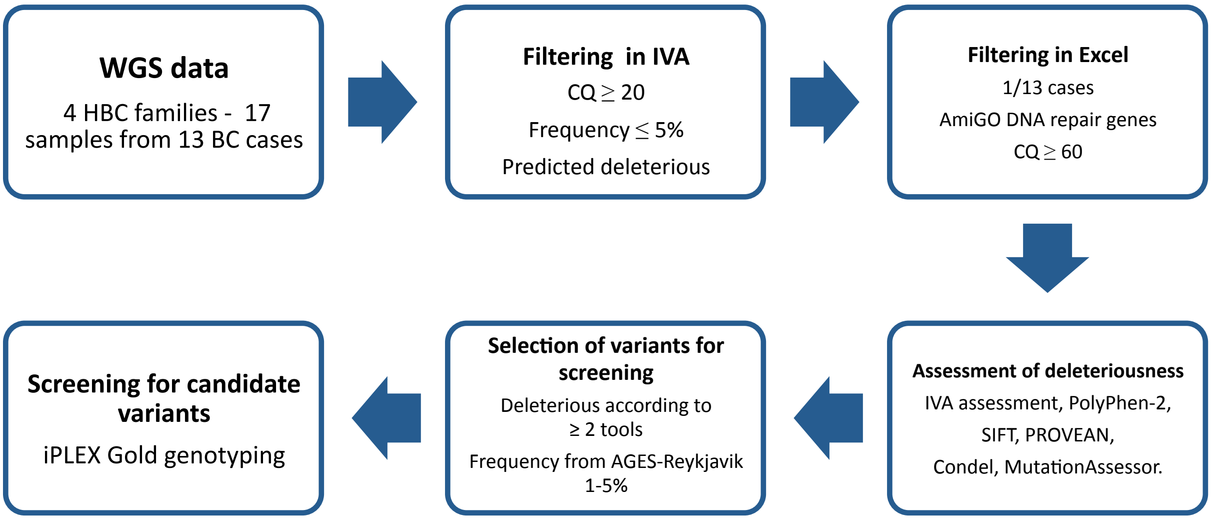

WGS data
4 HBC families - 17 samples from 13 BC cases
Filtering in IVA
CQ ≥ 20
Frequency ≤ 5%
Predicted deleterious
Filtering in Excel
1/13 cases
AmiGO DNA repair genes
CQ ≥ 60
Assessment of deleteriousness
IVA assessment, PolyPhen-2, SIFT, PROVEAN, Condel, MutationAssessor.
Selection of variants for screening
Deleterious according to ≥ 2 tools
Frequency from AGES-Reykjavik 1-5%
Screening for candidate variants
iPLEX Gold genotyping

Figure 10: The filtering process applied when searching for moderate-risk variants in the second part of the project. Whole Genome Sequencing (WGS) data from 13 BC cases in 4 Icelandic hereditary breast cancer (HBC) families were analyzed. The variants were filtered in Ingenuity® Variant Analysis™ (IVA) based on their call quality (CQ), frequency and predicted deleteriousness. Variants within the coding regions of genes directly annotated as double-strand break repair genes or general DNA repair genes were focused on. The deleteriousness of these variants was further assessed using five prediction tools. Variants that were deleterious according to at least two of these prediction tools and had a frequency between 1-5% in the Icelandic Heart Association's (IHA) AGES-Reykjavik cohort were screened for in a set of unselected BC cases.

3.5 Screening for variants in cases and controls

Three different methods were used for variant screening. In the first part of the project, indels were screened for by fragment analysis and SNPs were screened for using TaqMan® SNP genotyping assays. In the second part of the project, screening for multiple variants simultaneously using a genotyping technique called iPLEX was bought from the Mutation Analysis Facility at Karolinska

University Hospital in Stockholm, Sweden (http://www.maf.ki.se/). They used iPLEX® Gold, performed on a MassARRAY® system from Agena Bioscience (http://agenabio.com/), to perform the screening (72).

3.5.1 Fragment analysis

Primers were designed in the same way as described in section 3.3.1, except for one of the primers being marked with a FAM fluorescent dye at the 5' end. Table S2 lists the sequences of the primers used for fragment analysis and the length of the fragments generated. DNA fragments were amplified by PCR as described in section 3.3.2. Following the amplification, 0.3 µL of the PCR product were mixed with 9.6 µL of Super-DI™ formamide (MCLAB) and 0.1 µL of orange DNA Liz500 size standard (MCLAB). The samples were subsequently denatured at 94°C for 3 minutes (in a 2720 Thermal Cycler from Applied Biosystems™) and run in a 3130xl Genetic Analyzer (Applied Biosystems®). The resulting fragments were analyzed in the GeneMapper® software (Applied Biosystems®) (see figure 11).

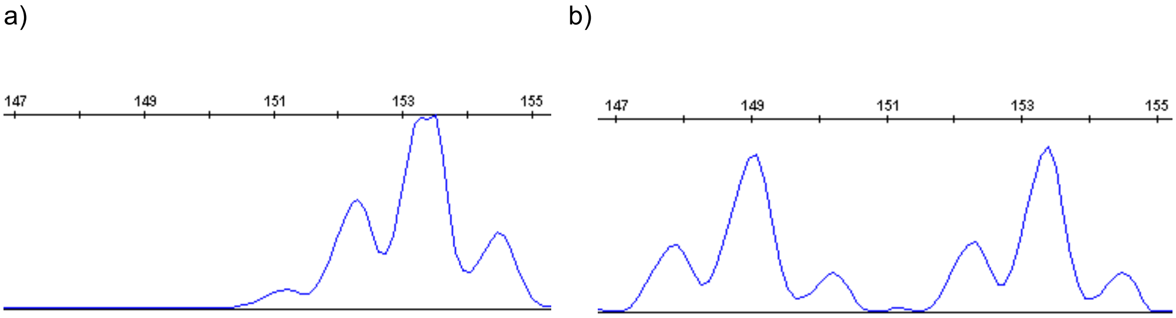


Figure 11: Fragment analysis results from GeneMapper®. Figure a) shows a sequence from a reference sample where no variant is present and all fragments that are amplified are 153 bases long. Figure b) shows a sequence where a heterozygous 4-base deletion is present, which results in a mutated 149 base fragment being amplified from one chromosome and the wild-type 153 base fragment from the other.

3.5.2 SNP genotyping with TaqMan® Assays

For genotyping of individual SNPs, single-tube TaqMan® genotyping assays and TaqMan® genotyping master mix were purchased from Applied Biosystems®. The genotyping reaction was prepared using the following recipe:

SNP genotyping reaction

Reactant (concentration)	Amount
TaqMan® genotyping master mix (2x)	5.0 µL
TaqMan® SNP genotyping assay mix (40x)	0.19 µL
ddH2O	4.81 µL
DNA (dried)	20 ng

The samples were sealed using Optical Adhesive Covers (Applied Biosystems®) and centrifuged for 5 minutes at 1200 rpm in an Eppendorf 5810 R centrifuge. The genotyping was performed in a 48-well StepOne Real-Time™ PCR system (Applied Biosystems®) using the following program:

StepOne Real-Time™ PCR program

Step	Temperature (°C)	Time (minutes)
1. AmpliTaq Gold enzyme activation	95	10:00
2. Denature	95	0:15
3. Anneal/extend	60	1:00

Steps 2 and 3 repeated 45 times

The results of the genotyping were then analyzed in the StepOne™ software (Applied Biosystems®) (see figure 12).

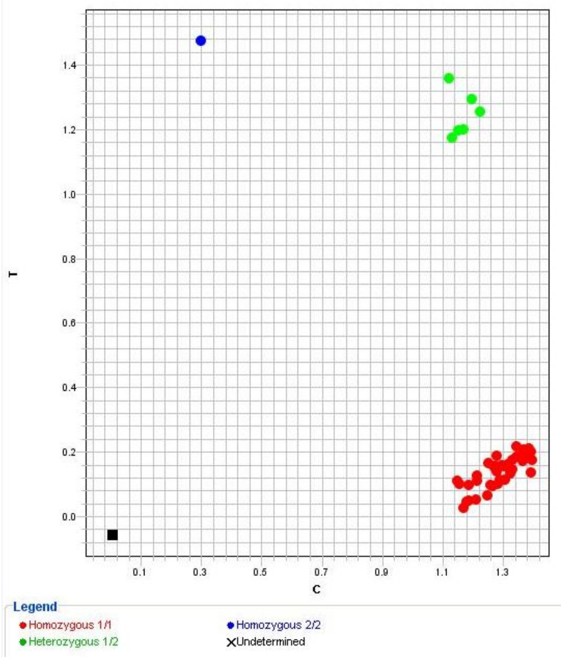


Figure 12: SNP genotyping analysis from the StepOne™ software. The results from the genotyping of 47 samples for a C>T variant. The strength of the C allele amplification is shown on the x-axis and the T allele on the y-axis. Each red dot represents a sample where only the C allele is amplified and therefore these samples have a homozygous wild-type genotype. The green dots represent samples that have equal amplification of each of the alleles and are therefore heterozygous for the mutation. The blue dot only has amplification of the T allele and therefore is a homozygous carrier of the mutation. The black dot represents a water-sample, used as a negative control.

3.6 Statistical analysis

Following variant screening, allele frequencies were calculated in Microsoft Excel using the following equation, where possible genotypes are AA, Aa and aa and N is the total number of alleles screened:

$$p = \frac{AA * 0.5Aa}{N}$$

Fisher's exact significance tests and Hardy-Weinberg equilibrium tests were performed using the R-Project for Statistical Computing. An example of Fisher's exact test performed in R-project using the EpiR package (https://cran.r-project.org/web/packages/epiR/epiR.pdf):

```
> GID8table <- matrix(c(12,11,538,643),ncol=2,byrow=TRUE)

> colnames(GID8table) <- c("BC-cases","Control")

> rownames(GID8table) <- c("delG","nomutation")

> print(GID8table)

         BC-cases Control

delG             12      11

nomutation      538     643

> fisher.test(GID8table)


        Fisher's Exact Test for Count Data


data:  GID8table

p-value = 0.5349

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

 0.5218977 3.2890776

sample estimates:

odds ratio

  1.303522
```

Example of Hardy-Weinberg equilibrium test performed in R-project using the HardyWeinberg package (https://cran.r-project.org/web/packages/HardyWeinberg/HardyWeinberg.pdf):

```
   > GID8_controls <- c(AA=316,Aa=11,aa=0)

   > HW.test <- HWChisq(GID8_controls,cc=0,verbose=TRUE)

   Chi square test for Hardy Weinberg equilibrium

   Chi2  =   0.09569983  p  value  =   0.7570523  D  =   0.09250765  f  =
0.01710731


   > GID8_cases <- c(AA=263,Aa=12,aa=0)

   > HW.test <- HWChisq(GID8_cases,cc=0,verbose=TRUE)


   Chi square test for Hardy Weinberg equilibrium

   Chi2 =  0.136814 p value =  0.7114691 D =  0.1309091 f =  0.02230483
```

4 Results

4.1 WGS analysis - Searching for a high-risk mutation in a breast cancer family with strong resemblance of dominant inheritance

The aim of the first part of the project was to identify a high-risk mutation in an Icelandic HBC family, the pedigree of which can be seen in figure 1a. The BC pattern in the family indicates autosomal dominant inheritance of a high-risk mutation. To identify a mutation with high risk in the family, WGS data from 6 samples from 4 BC cases was analyzed (both blood and tumor samples were sequenced from two of the cases).

The WGS data was initially analyzed using Ingenuity® Variant Analysis™ (IVA) (http://www.ingenuity.com/products/variant-analysis), a web-based application. In IVA, variants were filtered based on their call quality, frequency, predicted deleteriousness, inheritance model and biological context. Only variants with CQ ≥ 20 were kept, which is the default setting and usually balances sensitivity and specificity well enough to efficiently identify causal variants (68). Variants with a known frequency of ≥ 3% in the public Complete Genomics, 1000 Genomes (global) or the NHLBI ESP exome datasets were excluded from the analysis. Variants observed or predicted to be *pathogenic*, *likely pathogenic* or of *uncertain significance* were kept. This classification is based on the American College of Medical Genetics and Genomics (ACMG) guidelines for the interpretation of sequence variants, where certain criteria are categorized as *very strong*, *strong*, *moderate* or *supporting* evidence for pathogenicity. An example of a variant with very strong evidence for pathogenicity is if a loss-of-function (LOF) variant lies in a gene where LOF is a known mechanism of disease and an example of strong evidence for pathogenicity (of missense variants) is if the same amino acid change has previously been established as being pathogenic. The combination of criteria met by variants ultimately determines their classification. Variants predicted to cause a net loss- or gain-of-function and present in at least 2 out of the 6 sequenced samples, were kept. Finally, variants in genes known or predicted to affect BC and tumors were kept. The filtering process in IVA left 771 variants that were exported to Microsoft Office Excel for further analysis.

In Excel, variants that were either only present in tumor samples or 1 blood sample and one tumor sample were removed, leaving only variants that were present in blood samples from at least 2 BC cases. For the purpose of identifying high-risk variants, it was decided to focus on the variants likely to have the greatest effect on protein function or gene expression. Therefore, only variants predicted to result in protein truncation or to disrupt the binding of a miRNA to mRNA were kept, leaving 33 variants. Based on a growing experience of analyzing WGS data in our lab, the stringency of the analysis was increased and variants that did not have CQ ≥ 60 in at least 1 sample were removed. This left 24 variants, 15 indels (insertion/deletion polymorphisms) and 9 SNPs (single nucleotide polymorphisms), to be validated by Sanger-sequencing in their original samples (table 1).

Table 1: LOF and miRNA variants detected in WGS data from the family.

Type of variant	Predicted effect of variant	Gene	Gene region	Location	Transcript variant	Protein variant
Indel	LOF	*TRMT44*	Splice site	Chr4:8472809	c.1928-2_1929delAGAG	
		ZNF534	Exon	Chr19:52942693	c.2021dupC	p.*675fs
		ZNF488	Exon	Chr10:48370726	c.194delC	p.A65fs*14
		PRIM2	Exon	Chr6:57398186	c.889_890insA	p.N298fs
		CHST15	Exon	Chr10:125780752	c.1366_1367insCC	p.R456fs
		GPR27	Exon	Chr3:71804240	c.1040_1041insG	p.D347fs
		CCDC48	Exon	Chr3:128758687	c.1793delG	p.C598fs
		MEX3B	Exon	Chr15:82336814	c.396_397insA	p.H133fs
		KRTAP10-12	Exon	Chr21:46117660	c.544_545insT	p.A182fs
		COPZ2	Exon	Chr17:46115085	c.57_58dupCC	p.Q20fs*34
	miRNA binding disrupted	*MTMR3*	3'UTR	Chr22:30422037	c.*247_*248insT	
		GID8	3'UTR	Chr20:61576850	c.*586delG	
		DICER1	3'UTR	Chr14:95555386	c.*1564_*1565insT	
		STAC2	3'UTR	Chr17:37366899	c.*1645_*1646insC	
		SLC9A3R2	3'UTR	Chr16:2088379	c.*394_*395delAT	
SNP	LOF	*INMT*	Exon	Chr7:30795466	c.791G>C	p.*264S
		TROAP	Exon	Chr12:49724121	c.1493C>G	p.S498*
		ASPSCR1	Splice site	Chr17:79972949	c.1354-2A>T	
	miRNA binding disrupted	*HOXA5*	3'UTR	Chr7:27181092	c.*362T>G	
		DLG2	3'UTR	Chr11:83170723	c.*138A>G	
		MAPK1IP1L	3'UTR	Chr14:55531526	c.*181A>C	
		DCAF7	3'UTR	Chr17:61666687	c.*153C>T	
		MYT1	3'UTR	Chr20:62872106	c.*305G>T	
		PPP6R1	3'UTR	Chr19:55741371	c.*525G>A	

Green: validated variants. Blue: Validated variants, also present in controls. Red: Not validated. Black: Amplification or Sanger sequencing unsuccessful. SNP: Single nucleotide polymorphism. Indel: Insertion or deletion. LOF: Loss-of-function. 3'UTR: 3' untranslated region. Chr: Chromosome.

Of the 24 variants, 13 were successfully validated in the discovery samples and 5 were not present in those samples. For the remaining 6 variants, the PCR amplification or the Sanger-sequencing was unsuccessful. Out of the 13 validated variants, 5 were also found in ≥ 2 out of six control samples that were included in the validation process. Since this indicates that these variants have a high frequency in the Icelandic population, they were excluded from further analysis. At this point, 8 validated variants remained; 4 indels and 4 SNPs. The 4 validated indels (colored green in table 1) were all selected for screening in a separate and larger set of BC cases and controls by fragment analysis. Since SNP genotyping is much more expensive than fragment analysis, further information was gathered on the SNPs from various databases before making a decision of whether or not to include them in a case-control study. Further information was also gathered on the 6 variants for which the amplification or sequencing were unsuccessful before more resources were spent trying to verify them.

The databases used to gather information on these 10 variants (3 indels and 7 SNPs) were the following: The UCSC Genome Browser (https://genome-euro.ucsc.edu/cgi-bin/hgGateway), UniProt (http://www.uniprot.org/), Ensembl (http://www.ensembl.org/index.html), PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) and the Human Splicing Fincer v3.0 (http://www.umd.be/HSF3/index.html) were used for gathering general information on the genes harboring the variants and for prediction of their possible effect on the protein product. TargetScan (http://www.targetscan.org/), DIANA microT v5.0 (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=microT_CDS/index) and DIANA TarBase v7.0 (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index) were used to further assess variants predicted to disrupt a miRNA binding site. The Expression Atlas (https://www.ebi.ac.uk/gxa/home) and The Human Protein Atlas (HPA: http://www.proteinatlas.org/) were used to compare the expression of the genes harboring the candidate variants and their protein products in BC and in normal breast tissue.

4.1.1 Insertion or deletion variants (indels) predicted to lead to protein truncation or disrupted miRNA binding

4.1.1.1 COPZ2

COPZ2 (coatomer protein complex, subunit zeta 2) is located on chromosome 17q21.32 (70). This gene contains 11 exons and encodes for a protein that is 210 amino acids (aa) long (73, 74). The COPZ2 protein is a subunit of the coatomer complex which is associated with Golgi non-clathrin-coated vesicles. These vesicles mediate protein transport from the endoplastic reticulum (ER) to the trans-Golgi network. The coatomer complex is essential for transport of dilysine-tagged proteins. By sequence similarity, it is likely that the zeta subunit is involved in the assembly of the coat (73). Although it has been shown that *COPZ2* is down-regulated in various cancer cell-lines, it does not display tumor-suppressive activities. It does however harbor miRNA-152 in one of its introns, which is a tumor-suppressive miRNA that is silenced in tumor cell lines along with *COPZ2* and has been shown to be silenced by hypermethylation in endometrial cancer. *COPZ2* down-regulation in cancer cell-lines is therefore more likely to be a by-product of the oncogenic silencing of miRNA-152 (75, 76).

The variant, c.57_58dupCC, was present in 2 out of 4 sequenced BC cases in the family according to the WGS data but due to background in the Sanger-sequencing, it could not be confirmed. The variant is an insertion of two cytosine (C) bases in exon 3, causing the introduction of a stop codon 54 aa into the protein product. The variant has not been previously reported and therefore lacks frequency information. Since COPZ2 is not likely a tumor-suppressor protein and the variant is not expected to lead to a loss of miRNA-152 expression, no further efforts were made trying to verify its existence.

4.1.1.2 MTMR3

MTMR3 (myotubularin related protein 3) is located on chromosome 22q12.2 (70). This gene contains 20 exons and encodes for a 1198 aa long protein. The MTMR3 protein is a phosphatase that acts in the metabolism of lipids with a phosphoinositol headgroup (73, 74). MTMR3 has been shown to promote growth of colorectal cancer (CRC) cells while a recent study indicates that increased expression of *MTMR3* through the down-regulation of its regulatory miRNA hsa-miR-100 has anti-cancer effects in BC cell lines (77, 78).

The variant, c.*247_*248insT, is present in 2 of the 4 sequenced BC cases in the family according to the WGS data but due to the nature of the sequence surrounding the variant (poly-T site) it could not be confidently confirmed. It is located at the end of a predicted binding site of hsa-miR-208a and hsa-miR-208b, spanning bases 241-247 of the 3'UTR region of *MTMR3* (79). According to DIANA-TarBase, this has not been confirmed to be a real miRNA target site (80). The variant is an insertion of a tyrosine (T) nucleotide to a sequence of 13 T nucleotides, starting at the end of the predicted miRNA binding site. There are two reports in dbSNP of variants adding a T to this poly-T sequence (rs139430397 and rs200777264), but neither has a known frequency (81). Since the variant does not change the sequence of the predicted miRNA binding site, further efforts were not made to confirm its existence.

4.1.1.3 SLC9A3R2

SLC9A3R2 (solute carrier family 9, subfamily A (NHE3, cation proton antiporter 3), member 3 regulator 2) is located on chromosome 16p13.3. This gene contains 6 exons and encodes for a protein called Na(+)/H(+) exchange regulatory cofactor NHE-RF2 (NHERF-2) (73, 74). This is a scaffolding protein that connects to plasma membrane proteins and helps link them to the cytoskeleton and regulate their surface localization (73). NHERF-2 is a negative regulator of endothelial proliferation and recruits the tumor-suppressor PTEN to PDGFR (platelet derived growth factor) to restrict activation of PI3 kinase. This suggests that NHERF-2 has a helper role in PTEN tumor suppression (82, 83). In a more recent study, NHERF-2 was shown to interact with estrogen receptor alpha (ERα). Over-expression of NHERF-2 was shown to increase ERα activation in MCF7 BC cells and increase cell proliferation and tumor formation in mice. The same study found NHERF-2 to be up-regulated in 50% of BC tumor samples. The authors concluded that NHERF-2 is a co-activator of ERα and that it possibly participates in the development of estrogen-dependent BC (84).

The variant, c.*394_*395delAT, is present in 2 of the 4 sequenced BC cases in the family according to the WGS data but due to background in the Sanger-sequencing, it could not be confirmed to be present. The location of the variant in the 3' UTR of the gene was predicted by TargetScan 5.2 (the version incorporated by IVA) to be a binding site for hsa-miR-494. This binding site is not predicted by the most recent version of TargetScan (version 6.2) or by DIANA-MicroT, and *SLC9A3R2* is not a validated target of hsa-miR-494 according to DIANA-TarBase v7.0 (79, 80, 85). The variant is recorded in dbSNP (rs200451810) and has a frequency of 0.7% in the 1000 Genomes EUR dataset (European populations) (81, 86, 87). Based mainly on the lack of evidence behind the binding of hsa-miR-494 to the 3'UTR region of *SLC9A3R2,* work with this variant was not continued.

4.1.2 Single nucleotide polymorphisms (SNPs) leading to protein-truncation or disrupted miRNA binding

4.1.2.1 INMT

INMT (indolethylamine N-methyltransferase) is located on chromosome 7p14.3 (70). This gene contains 3 exons and encodes for a 263 aa protein. The INMT protein functions as a thioether S-methyltransferase and is predicted (by sequence similarity) to play a role in the detoxification of selenium compounds (73, 74). According to the EMBL-EBI Expression Atlas, *INMT* expression is low in normal breast tissue and even lower in breast cancer cell lines (88). Deregulated expression of *INMT* associated with *PTEN* loss has been reported in endometrial cancer and down-regulation of *INMT* has been shown to be correlated with more aggressive disease in pancreatic cancer (89).

The variant, c.791G>C, is a stop-loss variant that was validated to be present in 2 out of the 4 sequenced BC cases in the family. It is listed in dbSNP (rs61741736) and has a frequency of 0.93% in CG Public Genomes, 3.4% in the 1000 Genomes EUR (European) dataset and 3.60% in the ESP (population of European ancestry). The variant was not excluded by IVA because the 1000 Genomes global frequency (incorporated by IVA) is 1.2% and when the analysis was performed, the ESP European frequency was not available (81, 86, 87, 90). It is easy to assume that such a variant would result in an extension of the reading frame until the next in-frame stop-codon is encountered. If this was the case, then this variant would result in the addition of 20 aa to the protein product. However, too few human stop-loss mutations have been studied too allow for such a conclusion. In yeast, nonstop mRNAs (lacking stop codons) are removed by "nonstop mRNA decay" or their protein products are degraded by the proteosome. Studies indicate that in humans, the expression of nonstop mRNAs is generally unaltered while translation is blocked before the synthesis of full-length polypeptides is completed. A meta-analysis on 119 disease causing stop-loss mutations from the Human Gene Mutation Database (HGMD) indicates that the distance from the mutated stop codon to the next stop codon is a "key determinant of whether a given nonstop mutation will come to clinical attention". The analysis showed that there is a significant excess of disease causing stop-loss mutations that have in-frame stop codons 150-199 nucleotides downstream of the mutated stop codon (compared to control DNA sequences). This indicates that if the next in-frame stop codon is more than 150 bases downstream, there is an increased chance that the mutation causes a clinically significant phenotype. This study also found that there is a scarcity of disease-causing mutations with alternative stop codons 0-49 nucleotides downstream from the mutation, meaning that such mutations are less likely to have a clinical effect (91). These results make it less likely that the *INMT* c.791G>C variant, which has an in-frame stop codon 60 nucleotides downstream, causes decreased levels of INMT and a clinically significant phenotype. Although studies indicate that INMT is possibly down-regulated in BC and other cancers, due to lack of knowledge regarding how the loss of *INMT* might contribute to disease predisposition and uncertainty that the variant causes loss of INMT function, it was decided not to include this variant in a case-control study.

4.1.2.2 TROAP

TROAP (Trophinin associated protein) is located on chromosome 12q13.12 (70). It encodes a 778 aa protein called tastin. Tastin is thought to be involved, along with bystin and trophinin, in a cell adhesion complex likely to be involved in the initial attachment of a blastocyst to the uterus (73). Tastin may also be required for spindle assembly and centrosome integrity during mitosis, and is widely expressed in rapidly proliferating tumor cell lines. Loss of tastin expression causes mitotic block, inhibiting cell proliferation (92). According to the Human Protein Atlas (HPA), expression of *TROAP* is seen in low to medium levels in normal breast tissue. In BC tissue, low staining is seen in 2 out of 11 samples while the rest show no staining (93).

The variant, c.1493C>G, was confirmed to be present in 2 out of 4 sequenced BC cases. The variant, which has not been reported before, is a nonsense variant that turns a serine residue at position 498 into a stop-codon resulting in a loss of approximately one-third of the protein product. The wild-type tastin protein contains 4x33 aa approximate tandem repeats (nearly identical sequences of 33 aa are repeated four times) between aa 516-647, although alternatively spliced isoforms that lack this part of the protein do exist (73, 74). Since the literature suggests that tastin is essential for cell-cycle progression (making it unlikely to be a tumor suppressor), it was decided not to include the variant in a case-control study.

4.1.2.3 MAPK1IP1L

MAPK1IP1L (MAPK-interacting and spindle-stabilizing protein-like or mitogen-activated protein kinase 1 interacting protein 1-like) is located on chromosome 14q22.3 and encodes for a 245 aa protein (70, 73, 74). Not much is known about the function of the protein, except that it is involved in maintaining spindle integrity during cytostatic factor (CSF) arrest in the second metaphase of meiosis in mouse oocytes (94). No information is available in HPA on the expression of the protein in normal breast tissue, but in BC expression is high in 1/9 samples, medium in 6/9 samples and low in 2/9 samples (93). In the EMBL-EBI Expression Atlas, the expression of *MAPK1IP1L* in normal breast tissue from two studies is recorded. One of the studies shows low expression of the gene and the other study shows medium expression (88).

The variant, c.*181A>C, was validated to be present in 3 out of the 4 sequenced BC cases. The variant is listed in dbSNP (rs45513892) and has a frequency of 2.5% in the 1000 Genomes global dataset (incorporated by IVA) and 3.7% in the 1000 Genomes EUR dataset (81, 86). It is located in the 3'UTR of *MAPK1IP1L*, site predicted by TargetScan 5.2 to be a binding site for hsa-miR-219-5p. This binding site is however not predicted by the most recent version of TargetScan (version 6.2) or by DIANA-MicroT. According to DIANA-TarBase v7.0, *MAPK1IP1L* is not a validated target of hsa-miR-219-5p (79, 80, 85). Based on lack of information regarding the function of the protein in humans and scarce evidence behind the binding of hsa-miR-219-5p to the 3'UTR of *MAPK1IP1L*, this variant was not included in a case-control study.

4.1.2.4 DCAF7

DCAF7 (DDB1 and CUL4 associated factor 7) is located on 17q23.3 (70), in a region frequently amplified in BC (95). This highly conserved gene encodes for a 342 aa protein involved in craniofacial development and is possibly involved in skin development (73). DCAF7 has multiple WD40 repeats which enable the assembly of multiprotein complexes, and DCAF7 has been shown to be part of a nuclear complex along with ZNF703, PHB2 and NCOR2. ZNF703 amplification has been shown to have a role in the oncogenesis of luminal B breast tumors (96). According to the EMBL-EBI Expression Atlas, *DCAF7* has low expression in normal BC tissue. It is, however, more highly expressed in various BC cell-lines (88).

The variant, c.*153C>T, was validated to be present in 3 of the 4 sequenced BC cases. It is listed in dbSNP (rs72845886) and has a reported frequency of 5.8% in the 1000 Genomes EUR dataset (81, 86). When the analysis was performed, the only known frequency for the variant was 2.57% from the 1000 Genomes global dataset. Therefore, the variant was not excluded in the IVA filtering process.

The variant is located in the 3'UTR of *DCAF7*, in a site predicted by TargetScan and DIANA-MicroT to be a conserved binding site of hsa-miR-193b-3p (79, 85). The binding of hsa-miR-193b-3p to *DCAF7* mRNA has been indirectly validated according to DIANA-TarBase 7.0 (80). In the publication cited by TarBase, over-expression of mir193b in a melanoma cell-line was shown to result in down-regulation of DCAF7 and repress cell proliferation (97). In addition to this, the mutated base seems to be the most conserved base in the predicted binding site; receiving a GERP score of 5.73 (GERP scores range from -12.36 and up to 6.18 for the most conserved bases). Based on the evidence listed here, it was decided to include this variant in a case-control analysis.

4.1.2.5 *ASPSCR1*

ASPSCR1 (alveolar soft part sarcoma chromosome region, candidate 1) is located on chromosome 17q25.3 (70). This gene contains 17 exons and encodes for a 647 aa protein called Tether containing UBX domain for GLUT4. It is a tethering protein that controls the amount of glucose transporter type 4 (GLUT4) available at the cell surface in response to insulin stimulation. *ASPSCR1* has been identified as a proto-oncogene, based on a translocation which forms an ASPSCR1-TFE3 fusion protein that has been found in patients with alveolar soft part sarcoma and renal cell carcinoma. This translocation has not been reported in BC (73). According to HPA, the ASPSCR1 protein is expressed in low to medium levels in both normal and cancerous breast tissue (93). The EMBL-EBI Expression Atlas reports relatively low expression levels of *ASPSCR1* in normal breast tissue and the vast majority of BC cell lines are reported to have similar expression of *ASPSCR1* as normal breast tissue (88).

The variant, c.1354-2A>T, is present in 3 of the 4 sequenced BC cases in the family, according to the results of the WGS but due to an impure PCR product, indicating non-specific binding of the primers, the variant could not be confirmed. The variant, which is listed in dbSNP (rs199665633), has a frequency of 0.58% in the ESP (population of European ancestry) and 0.4% in the 1000 Genomes EUR dataset (81, 86, 90) and is located in an acceptor splice site in front of exon 13. According to the Human Splicing Finder the variant most probably affects splicing, although there is an alternative in-frame splice site located 6 bases downstream (98). If this alternative splice site would be used, it would result in the loss of two aa (proline and glutamine) from the protein product. The ASPSCR1 protein has 9 alternative protein-coding transcripts, only one of which contains exon 13. No information on the expression of different transcripts could be found, but the canonical isoform does not include exon 13. The main functional domain of the protein (UBX domain) is encoded by exons 9-12 and a portion of exon 14 (which is exon 13 in the canonical isoform) (73, 74). Based on the literature, ASPSCR1 is an unlikely tumor suppressor and the variant is unlikely to affect the function of the protein in a substantial manner. Therefore, it was decided not to continue further with this variant.

4.1.2.6 *MYT1*

MYT1 (myelin transcription factor 1) is located on chromosome 20q13.33 (70). This gene encodes for a 1121 aa zinc-finger DNA-binding protein which may have a role in the development of the nervous system (73). This protein has not been studied in the context of cancer, except that increased expression has been reported in high-grade human brain tumors (99). According to HPA, the MYT1 protein is expressed in low levels in both normal and cancerous breast tissue (93). The EMBL-EBI Expression Atlas reports no expression of *MYT1* in normal breast tissue and the vast majority of BC cell lines. In the few BC cell lines that do have expression of *MYT1*, the expression of the gene is low (88).

The variant, c.*305G>T, is present in 3 out of the 4 sequenced BC cases in the family, according to the results of the WGS but since no PCR amplification was achieved by the designed primer pair, it could not be confirmed. The variant is recorded in dbSNP (rs34316071) and has a reported frequency of 2.1% in the 1000 Genomes global dataset (incorporated by IVA) and 4.5% in the 1000 Genomes EUR dataset (81, 86). It is located in a binding site of hsa-miR-146a and hsa-miR-146b-5p, as predicted by TargetScan (79). This binding site is also predicted by DIANA-MicroT but according to TarBase, *MYT1* is not a validated target of either of these miRNA molecules (80, 85). Based on MYT1 mainly having a role in the nervous system, and that the variant is not in a validated miRNA binding site, it was decided not to continue further with this variant.

4.1.2.7 PPP6R1

PPP6R1 (protein phosphatase 6, regulatory subunit 1) is located on chromosome 19q13.42 (70). It encodes for an 881 aa protein which is a regulatory subunit of protein phosphatase 6 (PP6) and may function as a scaffolding PP6 subunit (73). No studies were found that link *PPP6R1* to cancer. According to HPA, the protein is expressed at low levels in normal breast tissue while expression in BC ranges from being not detected to medium expression (93). In the EMBL-EBI Expression Atlas, one study reports no expression of *PPP6R1* while another reports low expression of the gene. Eight studies have looked at the expression of the gene in BC cell lines and they show that although *PPP6R1* is generally up-regulated, it is not highly expressed (88).

The variant, c.*525G>A, is present in 2 out of the 4 sequenced BC cases in the family, according to the results of the WGS but due to background in the Sanger-sequencing, it could not be confirmed to be present. It is recorded in dbSNP (rs77942969) and has a recorded frequency of 0.5% in the 1000 Genomes global dataset and 1.0% in the 1000 Genomes EUR dataset (81, 86). It is located in the 3'UTR of *PPP6R1*, in a site predicted by TargetScan to be a binding site of hsa-miR-506 and hsa-miR-124 (79). This binding is, however, not predicted by DIANA-MicroT and according to TarBase, *PPP6R1* is not a validated target of either of these miRNA molecules (80, 85). Based on the limited information available on the function of PPP6R1 and lack of evidence behind the binding of the miRNA molecules to the region where the variant is located, it was decided not to continue further with this variant.

4.2 Screening for potential high-risk mutations

Based on the information gathered on the 10 variants listed above, 1 SNP (c.*153C>T in *DCAF7*) was considered of sufficient interest to be included in a case-control study, in addition to the 4 validated indel variants previously selected. Each of the indels was screened for by fragment analysis in ~300 cases and ~300 controls to get a rough estimate of their frequencies in these groups. The SNP was screened by using a TaqMan$^{®}$ genotyping assay in ~1000 total cases and controls, or as the quantity of the assay mix allowed. The results of the screening for the 5 variants can be seen in table 2.

Table 2: Results of the initial screening for selected variants in BC cases and controls. Indels were screened by fragment-analysis and the SNP was screened by a TaqMan$^{®}$ genotyping assay. The primers that were used can be seen in supplementary table 1.

Gene	Transcript Variant	MAF in controls (n*)	MAF in BC cases (n*)	p-value**
ZNF488	c.194delC	0.008 (314)	0.015 (339)	0.304
ZNF534	c.2021dupC	0.010 (300)	0.006 (351)	0.527
GID8	c.*586delG	0.017 (327)	0.022 (275)	0.535
TRMT44	c.1928-2_1929delAGAG	0.000 (292)	0.0014 (363)	1.000
DCAF7	c.*153C>T	0.061 (270)	0.058 (693)	0.83

* Number of genotyped samples.
**p-value from Fisher's exact test, performed in the R-Project (version 2.15.1) using the epiR package.

Chi-square tests for Hardy-Weinberg Equilibrium (HWE) were performed using the R-project to see if the results of the genotyping of the variants were as expected. Deviation from HWE was not significant for any of the variants. The results of the screening for the 5 candidate variants didn't show a significant difference in the frequency of any of the variants in BC cases versus controls. For three variants, *ZNF534* c.2021dupC, *GID8* c.*586delG and *DCAF7* c.*153C>T, the difference in frequencies between the BC cases and the controls would not reach significance even if all available cases and controls were to be screened (around 1500 BC cases and 6000 controls) and the frequencies of the variants remained the same. These variants are therefore unlikely to be high-risk BC mutations and were not studied further. For the other two variants, *ZNF488* c.194delC and *TRMT44* c.1928-2_1929delAGAG, the frequency of the variants was higher in BC cases than controls and the difference would reach significance if all available samples were to be screened and the frequencies of the variants remained the same. For *ZNF488* c.194delC, this would result in a p-value of 0.0009281 and for *TRMT44* c.1928-2_1929delAGAG the p-value would be 0.001597. These variants were therefore included in further studies.

4.2.1 The *ZNF488* c.194delC variant

The results of the initial screening indicated that if *ZNF488* c.194delC was a predisposing mutation, it would be one associated with a moderate increase in risk (OR based on initial screening = 1.88). It was decided to screen more samples for the *ZNF488* variant and see if the frequencies remained the same. When the number of screened BC cases and controls had been doubled, the frequencies in the groups had evened out (see table 3). Chi-square tests for Hardy-Weinberg Equilibrium (HWE) revealed that the results of the screening did not deviate significantly from HWE. It was decided not to continue working with this variant.

Table 3: Screening for *ZNF488* c.194delC in additional BC cases and controls. The screening was performed using fragment analysis and the sequences of the primers that were used can be seen in supplementary table 1.

Gene	Trancript Variant	MAF in controls (n*)	MAF in BC cases (n*)	p-value**
ZNF488	c.194delC	0.011 (660)	0.011 (688)	1.000

*Number of genotyped samples.
** p-value from Fisher's exact test, performed in the R-Project (2.15.1) using the epiR package.

4.2.2 The *TRMT44* c.1928-2_1929delAGAG variant

Since no carriers of the *TRMT44* c.1928-2_1929delAGAG variant were found in the controls, it's not possible to calculate its OR. However, based on its very low frequency and that it was not identified in any controls, it could be a very rare high-risk mutation. The variant was only detected in one BC case during the screening process. When looking at the family history of this positive BC case, it proved to be strong (four 1st and 2nd-degree relatives with BC), providing further evidence of possible high risk. Based on this, it was decided to perform a segregation analysis of *TRMT44* c.1928-2_1929delAGAG in the sequenced family to see if it segregated with the BC cases as well as screen for the variant in more unselected BC cases to see if more positive BC cases were found, and then see if those cases also belonged to families with history of BC. The segregation analysis revealed that three out of six diagnosed BC cases in the sequenced family and one out of two OC cases were positive carriers of the variant (see figure 13).

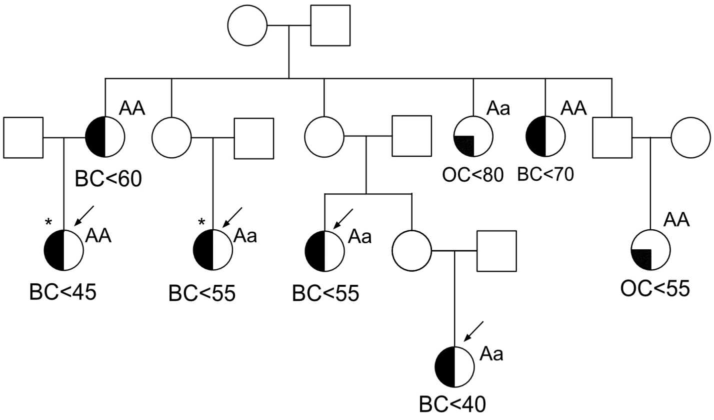


Figure 13: Segregation analysis of *TRMT44* c.1928-2_1929delAGAG in the sequenced family. Circles denote women and squares represent men. Circles that are half-filled with black represent BC diagnosis and circles quarter-filled with black represent OC diagnosis. The arrows point to women whose blood samples were whole-genome sequenced. Individuals with sequenced tumor samples are marked with an asterisk. The age of first cancer diagnosis is written below each of the women from whom samples were analyzed. Women that are heterozygous carriers of the *TRMT44* c.1928-2_1929delAGAG variant are marked "Aa" and women that are not carriers are marked "AA". Three out of six BC cases in the family and one out of two OC cases are carriers.

The results of the screening for the *TRMT44* c.1928-2_1929delAGAG variant in 434 additional unselected BC cases didn't reveal additional carriers (see table 4). Chi-square tests for HWE revealed that the results of the screening in BC-cases did not deviate significantly from HWE. No positive control samples were identified, so a HWE test was not applicable in the control group. Since the results of the segregation analysis and screening in BC cases and controls do not add strength to the original findings, it is not very likely *TRMT44* c.1928-2_1929delAGAG is a high-risk mutation.

Table 4: Screening for *TRMT44* c.1928-2_1929delAGAG in additional BC cases. The screening was performed using fragment analysis and the sequences of the primers that were used can be seen in supplementary table 1.

Gene	Transcript Variant	MAF in controls (n*)	MAF in BC cases (n*)	p-value**
TRMT44	c.1928-2_1929delAGAG	0.000 (292)	0.00063 (797)	1.000

*Number of genotyped samples.
** p-value from Fisher's exact test, performed in the R-Project (version 2.15.1) using the epiR package.

4.3 WGS analysis - Searching for moderate-risk variants in four families with multiple breast cancer cases suggestive of segregating a high-risk genetic factor

Based on the results of the search for a high-risk BC mutations in the family analyzed in the first part of this MSc project (figure 1a), along with the results of a recent M.Sc. project by Edda Sigríður Freysteinsdóttir, "*A search for a cancer susceptibility gene in a high risk breast cancer family without a mutation in BRCA1 and BRCA2*", where the family in figure 1b was the focus of the study, and unpublished analysis of families in figure 1c-d by the supervisors of this MSc study, it is not very likely that these BC families can be explained by high-risk mutations in protein coding genes. Since several studies have suggested that the majority of the missing BC heritability is due to polygenic inheritance (53), it was decided to continue by focusing the search in the families towards finding moderate-risk mutations whose combined effects might explain the increased risk of BC. It was also decided that in this part of the project, variants in DNA repair genes would be put in priority. This decision was made based on numerous previous findings linking mutations in known DNA repair genes to BC susceptibility (13).

WGS data from 17 samples from 13 BC cases was analyzed in IVA (both blood and tumor samples were sequenced from four of the cases). In IVA, variants were filtered based on their call quality, frequency and predicted deleteriousness. Only variants with CQ ≥ 20 were kept. In this part of the project, we wanted to be able to identify variants with a moderate effect on BC predisposition. Since such variants are generally not as rare as high-risk mutations, variants with a known frequency of up to 5% (instead of ≤ 3%) in the incorporated databases were kept. Variants observed or predicted to be *pathogenic*, *likely pathogenic* or of *uncertain significance* were kept, along with variants predicted to cause a net loss- or gain-of-function. Since the search was not focused on a specific inheritance model in this part of the project and only variants in selected genes were considered, the genetic analysis filter and the biological context filter were not applied. The filtering process in IVA left 19.003 variants that were exported to Microsoft Office Excel to be further analyzed.

In Excel, variants that were only present in tumors were removed. The CQ conditions were made more stringent, and only variants with CQ ≥ 60 in at least 1 sample were kept. Variants within any of 312 genes directly annotated as double-strand break repair genes or DNA repair genes in the Gene Ontology AmiGO database were kept. In addition to its own assessment of deleteriousness, IVA incorporates two tools that predict the impact of missense variants on protein function; SIFT (100) and PolyPhen-2 (101). To increase the chances of selecting variants that were truly deleterious, it was decided to assess the remaining variants by three additional web-based function prediction tools; MutationAssessor (102), PROVEAN (Protein Variation Effect Analyzer) (103) and Condel (Consensus Deleteriousness score) (104). These tools were selected based on their different approaches for assessing the predicted impact of variants. The algorithm applied by MutationAssessor mainly takes into account the evolutionary conservation of specific residues within protein families and their subfamilies (102). PROVEAN is also based on conservation, but the scoring system is not only determined by the position of the amino acid where the variation is observed but also by the surrounding sequence (103). The Condel score consists of a weighted average of the scores of MutationAssessor and FATHMM (Functional Analysis Through Hidden Markov Models). FATHMM incorporates Hidden Markov Models, which are powerful probabilistic models capable of capturing position-specific information within a multiple sequence alignment of homologous sequences (35). A variant had to be deleterious or damaging according to at least two out of these six prediction tools to be further considered in this analysis. A total of 40 variants met all these criteria.

For economic reasons, it was decided that only variants with available frequencies from an exome-chip performed on 2983 samples from the AGES-Reykjavik (provided by the Icelandic Heart Association) would be considered. By making use of this information, it was financially possible to screen for more variants than otherwise would have been possible. Calculations were performed to evaluate the frequency of the variants that we had statistical power to identify as predisposing. The results were that with the control frequencies from the AGES-Reykjavik cohort, the ~1500 total BC samples we have available to screen and with the aim of identifying variants with an RR as low as 1.5, the frequency of the variants could be as low as 1% in the general population for the association to be significant (theoretical p-value = 0.0476). Therefore, variants with frequencies between 1-5% in the Icelandic population were kept. Out of the 16 variants that now remained, one variant (p.G998E in the gene *PALB2*) had been screened for in large cohorts of BC cases and controls (n=1846 and 2168, respectively) and revealed to be a benign variant in a study published by Rahman *et.al.* in 2007 (105). This left 15 candidate variants to be screened for in BC cases. These variants are detailed in table 5.

Table 5: Variants in DNA repair genes selected for case-control study and their predicted functional impact.

Chromosome	Gene	Transcript Variant	Protein Variant	dbSNP ID	Cases	Mutation Assessor	PROVEAN	Condel	SIFT	PolyPhen-2	IVA Assessment
15q26.1	*TICRR*	c.2240T>C	p.V747A	12905387	1	medium	Deleterious	N	Damaging	Possibly Damaging	Uncertain Significance
11q22.3	*ATM*	c.2572T>C	p.F858L	1800056	1	medium	Deleterious	N	Tolerated	Benign	Likely Benign
10q11.23	*ERCC6*	c.2741C>T	p.T914M	142580756	1	high	Deleterious	D	Damaging	Probably Damaging	Uncertain Significance
11q22.3	*ATM*	c.3161C>G	p.P1054R	1800057	1	medium	Deleterious	D	Damaging	Probably Damaging	Uncertain Significance
17q23.3	*POLG2*	c.1247G>C	p.G416A	17850455	1	medium	Neutral	D	Tolerated	Probably Damaging	Uncertain Significance
15q22.31	*PIF1*	c.850T>C	p.C284R	118062397	1	medium	Deleterious	D	Damaging	Probably Damaging	Uncertain Significance
19q13.33	*LIG1*	c.1226G>A	p.R409H	4987068	1	low	Deleterious	N	Damaging	Benign	Uncertain Significance
6p21.33	*MDC1*	c.5648G>A	p.R1883Q	28994875	1	medium	Neutral	N	Damaging	Probably Damaging	Uncertain Significance
14q11.2	*PARP2*	c.704A>G	p.D235G	3093921	3	medium	Deleterious	N	Tolerated	Benign	Uncertain Significance
14q21.2	*FANCM*	c.4799C>T	p.T1600I	61746943	1	low	Deleterious	D	Damaging	Benign	Uncertain Significance
14q11.2	*SUPT16H*	c.1244C>G	p.A415G	61739513	2	medium	Deleterious	D	Tolerated	Benign	Likely Benign
5q13.2	*CDK7*	c.854C>T	p.T285M	34584424	1	high	Deleterious	D	Damaging	Probably Damaging	Uncertain Significance
8q24.3	*RECQL4*	c.2395G>A	p.V799M	34293591	1	medium	Neutral	N	Damaging	Probably Damaging	Likely Benign
1q43	*EXO1*	c.836A>G	p.N279S	4149909	1	medium	Deleterious	N	Damaging	Probably Damaging	Uncertain Significance
13q33.1	*ERCC5*	c.760A>G	p.M254V	1047769	2	medium	Deleterious	N	Damaging	Benign	Likely Benign

MutationAssessor: Medium or high functional impact was considered as deleterious/damaging. PolyPhen-2: Probably damaging variants were considered as deleterious/damaging. Condel: D = Deleterious, N = Neutral.

4.4 Screening for potential moderate-risk variants

The 15 candidate variants were screened for simultaneously using a SNP genotyping technique called iPLEX® Gold, performed on a MassARRAY® system from Agena Bioscience (72). This method of screening was chosen because it allows for the genotyping of up to 40 variants in one reaction in a cost-effective manner (http://agenabio.com/products/applications/genotyping-and-mutation-detection/). This system is not available in Iceland, and based on communication with several facilities in the USA and Sweden which offer iPLEX® Gold genotyping, the Mutation Analysis Facility (MAF) at Karolinska University Hospital in Stockholm, Sweden (http://www.maf.ki.se/) was the most experienced and cost-effective. It was therefore decided to send BC samples there for screening. Due to budgetary reasons, not all available BC cases were screened during this round of screening. It was decided to screen 540 BC cases for the 15 candidate variants since the results of the screening of this many samples could serve as a good indicator to see if any of the variants were interesting candidates for screening in the rest of the available BC cases (p-value < 0.01 for for the most rare variant that was included given an expected RR of ~2.0). Unfortunately, one variant (*ERCC5* c.760A>G) didn't fit in the design of the iPLEX® Gold reaction, and for another variant, (*RECQL4* c.2395G>A), the assay failed. These variants have since been included in another iPLEX® Gold project, for which results are expected later this year. The screening was successful for the other 13 variants, the results of which can be seen in table 6. The results indicate that none of the candidate variants predispose to BC.

Table 6: Screening for selected variants in DNA repair genes in BC cases and controls. Control frequencies were determined by exome-chip on samples from the AGES-Reykjavik cohort and were provided by the Icelandic Heart Association (IHA). Frequencies in BC cases were determined by iPLEX® Gold genotyping, performed at the Mutation Analysis Facility at Karolinska University Hospital in Stockholm, Sweden.

Gene	**Trancript variant**	**MAF in controls (n=2983)**	**MAF in BC Cases (n=540)**	**OR****	**p-value****
PIF1	c.850T>C	0.025	0.023	0.91	0.751
TICRR	c.2240T>C	0.013	0.015	1.14	0.665
ERCC6	c.2741C>T	0.017	0.015	0.88	0.795
POLG2	c.1247G>C	0.022	0.023	1.04	0.824
ATM	c.2572T>C	0.015	0.013	0.86	0.682
ATM	c.3161C>G	0.021	0.023	1.10	0.648
MDC1	c.5648G>A	0.028	0.018	0.62	0.050
PARP2	c.704A>G	0.030	0.030	0.98	1.000
CDK7	c.854C>T	0.035	0.043	1.23	0.216
EXO1	c.836A>G	0.044	0.044	0.99	1.000
LIG1	c.1226G>A	0.025	0.016	0.61	0.065
SUPT16H	c.1244C>G	0.034	0.038	1.12	0.526
FANCM	c.4799C>T	0.031	0.031	1.03	0.848

*Number of genotyped samples.

**OR and p-value from Fisher's exact test, performed in the R-Project (version 2.15.1) using the epiR package.

5 Discussion

In the first part of this project, WGS data from four BC cases belonging to one Icelandic HBC family were analyzed with the aim of identifying a high-risk mutation that could explain the increased risk of BC within the family. The main focus of the search was on variants predicted to alter protein function or gene expression through protein-truncation or disrupted binding of miRNA, resulting in five candidate variants taken forward to a screen in BC cases and controls. None of the variants showed a significant difference in frequency between the groups. In the second part of the project, WGS data from nine BC cases belonging to three additional HBC families were added to the analysis with the aim of identifying moderate-risk variants that contribute to the increased risk of BC in a polygenic fashion. Variants in genes that have a role in DNA repair were focused on, resulting in 13 candidate variants being screened for. None of the variants showed a significant difference in frequency between BC cases and controls.

5.1 Project background

In the four HBC families that were analyzed in this project, the number of BC cases and the average age of diagnosis indicate that hereditary factors play an important role. The families and the sequenced individuals were selected for WGS based on the following criteria: 1) the BC history within the family shows a pattern of dominant inheritance, 2) the family is negative for Icelandic founder mutations in *BRCA1* and *BRCA2*, 3) the individuals were relatively young when they were diagnosed with BC and 4) high quality DNA samples were available from three or more BC cases within the family. Three out of the four families (figure 1b-d) had previously been analyzed in a GWL analysis by Arason *et.al*. (51). In the study, the three families were analyzed alongside six other non-*BRCA1/2* high-risk BC families, with the aim of finding whether new BRCA-like genes existed in Icelandic families with an autosomal dominant inheritance pattern. One of families (not included in this project) exhibited suggestive linkage signals at three chromosomes (2p, 6q and 14q), but the study did not reveal evidence of one or two high-risk loci associated with BC. There is a possibility that the study missed a high-risk locus due to lack of statistical power (the mutation being confined to one family and the family being too small for significant linkage) or due to the presence of phenocopies (affected cases not carrying the high-risk mutation segregating in the other cases in the family). Such a mutation could possibly be reminiscent of the Icelandic *BRCA1* founder mutation c.5074G>A, which is estimated to be carried by <100 Icelanders and has been shown to segregate in two HBC families that previously had shown linkage to the 17q locus that harbors *BRCA1* (32). However, the study by Arason *et.al.* indicated that the strong family history of the nine families included in the study is not likely to be explained by one or a few high-risk mutations, but rather by two or more low- or moderate-risk variants that contribute to BC in a polygenic fashion (51). One of the three families included in the GWL study (figure 1b) had also been the focus of another M.Sc. project, which aimed at identifying a high-risk mutation. That project resulted in the identification of a very rare moderate-risk variant that partly explains the BC clustering in the family. The supervisors of this M.Sc. project have also analyzed the other two families (figure 1c-d) and for both of them, preliminary data do not indicate that the BC risk within these families is explained by segregation of a high-risk mutation.

The family focused on in the first part of the project (figure 1a) had not been analyzed before. Therefore, the aim was to identify a high-risk mutation that could explain the BC clustering within that family. In the second part of the project, the other three families were added to the analysis with the aim of identifying mutations with moderate effects on risk, which together could cause increased risk of BC within the families.

5.2 Filtering of variants detected by WGS

The WGS data were analyzed in IVA and in each part of the project the filtering pipeline was adjusted according to the aim. In the first part of the project, the aim was to identify a high-risk mutation. We therefore searched for predicted deleterious variants that were present in at least 2 of the 4 sequenced BC cases and were inherited in a dominant fashion. Although a high-risk mutation would most likely be very rare in the general population, the filters were set to keep variants with a known frequency of up to 3% in the first part of the project. The reason for this was that we also wanted to keep the possibility of identifying moderate-risk variants with higher frequency that might segregate within the family. IVA allows filtering of variants according to their phred-scaled quality scores (67). The default setting of the call quality (CQ) filter is to keep variants with a phred-score of ≥ 20 and according to IVA, this setting generally balances sensitivity and specificity well enough to identify causal variants efficiently (68). A CQ score of 20 corresponds to a 99% chance of the base being called correctly or alternatively, a 1 in a 100 chance of an incorrect call. However, through previous work with WGS data from CG at our laboratory we had only been successful in verifying the results of the WGS for variants with a CQ of at least 60. Therefore, it was decided to keep only variants with CQ ≥ 60, in at least one sample, corresponding to a 99.9999% chance of a correct base-call. It could be argued that these call quality conditions were too strict and might lead to the exclusion of real (and possibly predisposing) variants from the analysis. Indeed, when looking at published studies where IVA has been used to analyze NGS data, most have kept the default setting of the CQ filter (106-108). However, given that 5 out of the 18 variants that were successfully sequenced in the validation process turned out to be errors in the WGS, these conditions are still not so strict that the need to validate the results of the WGS is eradicated. Interestingly, all five variants that could not be validated were indels while all of the six SNPs for which the sequencing was successful were validated to be present. This suggests that stricter CQ conditions might be appropriate for indels than for SNPs. IVA was also used to identify variants in genes that were known or predicted to affect BC and variants known or predicted to affect genes that play a role in the same biological pathways as known BC genes. Therefore, although only variants in genes that act in known or predicted BC-pathways were kept, the possibility of identifying a high-risk variant residing within a gene that hasn't been implicated in BC pathogenesis before was kept open. Finally, only variants predicted to cause loss-of-function (LOF) or to disrupt miRNA binding (which generally has a gain-of-function effect) were chosen. LOF variants, which most often lead to the truncation of a protein, include insertions and deletions that cause a shift in the reading frame of the mRNA (frameshift variants), nonsense variants that result in a premature stop-codon or the loss of a stop-codon (also called stop-loss or readthrough variants) and variants that disrupt splice-sites which can lead to intron-retention or missing exons (109). The reason for focusing on LOF variants is that most high-risk BC susceptibility mutations identified to date cause LOF (49). MiRNAs are approximately 22 nucleotide RNAs which bind to complementary sites on mRNAs, generally causing mRNA cleavage or translational repression (110) although recent studies show that they are able to stimulate gene expression as well (111). Over the last few years, increased interest in the regulatory effects of miRNAs has lead to the discovery of variants in miRNA binding sites that predispose to various cancers (112), such as a SNP in a binding site for Let-7 in the 3'UTR of the *KRAS* oncogene which confers an increased risk of non small cell lung cancer (NSCLC) (113). Although variants in miRNA binding sites have not been shown to have more than a moderate effect on cancer risk (112), we didn't want to leave out the possibility of such a variant having an important effect on the risk of BC in the family. This step left 24 variants on the candidate list, of which 13 were successfully validated, 6 could not be validated because the PCR amplification or Sanger-sequencing was unsuccessful and, as previously mentioned, 5 variants were not present in the original samples.

As the aim of the second part of the project was to identify variants with a moderate effect on BC risk, some adjustments were made to the filtering process. Because such variants are likely to be more frequent in the general population than high-risk mutations, variants with a known frequency of up to 5% were kept. It was decided that in this part of the project, only variants in genes involved in DNA repair would be further considered. This decision was based on the past success of the candidate gene approach, focusing on genes acting in the same pathways as the *BRCA* genes, in identifying variants with moderate effects on BC risk (12). Indeed, all moderate-risk BC genes identified so far are directly involved in DNA repair (13). To increase the chances of selecting truly deleterious variants, in addition to IVA's assessment of deleteriousness and the incorporated SIFT and PolyPhen-2 predictions, three additional online function-prediction tools (PROVEAN, Condel and MutationAssessor) were used to assess each variant and only variants that were predicted deleterious by at least two separate tools were kept. By making use of control frequencies from IHA's AGES-Reykjavik cohort, we could afford selecting a higher number of variants for screening than we otherwise could have since selected variants only had to be screened for in BC cases. However, this meant postponing the screening of candidate variants not included in the IHA's genotyping project. Based on the selection criteria listed above, 15 variants with an estimated frequency of 1-5% in the Icelandic population were selected for screening in a group of BC cases. The reason for confining the frequency to 1-5% was that we expected this range to be within our power to detect statistical significance (at level 0.05), given the number of samples available for testing and given the aim to identify variants of moderate risk. A lower frequency would need a larger number of samples or a higher risk effect of the variant, and a higher frequency would not be considered meaningful in the search of moderate-risk variants. Ideally, when testing the frequencies of the selected variants, a Bonferroni correction should be applied (49) but given the size of this study, the aim was rather to get an idea of the effects of the variants on BC risk. If the results indicated that any of the variants were causative then the specific effects of those variants would be studied further.

5.3 Selection and screening of candidate variants

In the first part of the project, validated LOF and miRNA binding site indels were screened for in BC cases and controls. These variants were *TRMT44* c.1928-2_1929delAGAG, *ZNF534* c.2021dupC, *ZNF488* c.194delC and *GID8* c.*586delG. The results of the screening indicated that none of these are high-risk mutations, although it cannot be ruled out that the *TRMT44* variant predisposes to BC. Apart from the sequenced family, the *TRMT44* variant was only identified in one BC case and no controls. Furthermore, segregation analysis in the family revealed that three of six BC cases and one of two OC cases were carriers of the variant. If the *TRMT44* variant confers high BC-risk, this family would need other unrelated factors to explain the BC in the non-carriers. However, when the BC risks associated with a variant are only two- to threefold, it would not necessarily segregate with the disease (12) and therefore *TRMT44* c.1928-2_1929delAGAG might have a moderate effect on BC risk. This could be demonstrated by screening for the variant in all available unselected BC cases (~1500 cases) and controls (~6000), which would result in a p-value of 0.001597 given that the frequency of the variant remained that same. The effect of the variant on BC susceptibility could also be studied by identifying more carriers, looking at whether these carriers have a family history of BC and if so, performing a segregation analysis to see if the variant segregates with BC more often than would be expected by chance. In fact, the single carrier that was identified in the group of unselected BC cases does belong to a family with a strong history of BC. However, since *TRMT44* c.1928-2_1929delAGAG was unlikely to be a high-risk mutation, and also because additional samples would need to be collected and consents provided, segregation analysis was not performed in the family of the carrier.

In hindsight, the indel variants probably should have been assessed further before they were screened for. Such an assessment likely would have lead to only the *TRMT44* variant being screened for. *TRMT44* (tRNA methyltransferase 44 homolog) encodes for a 757 amino acid (aa) protein likely to be a tRNA methyltransferase that functions in the cytoplasm (73). The gene has not been associated with BC, but a missense variant within this gene is likely to cause Partial Epilepsy with Pericentral Spikes (PEPS), which is a Mendelian form of idiopathic epilepsy (114). *TRMT44* c.1928-2_1929delAGAG is predicted to result in an alteration of the length of an exon (98), shifting the reading frame and introducing a premature stop-codon, thereby removing the predicted functional domain of the protein (73). Although the variant causes LOF, it is important to keep in mind that not all LOF variants are disease-causing. In fact, 2636 individuals sequenced by deCODE Genetics each carried, on average, 149 LOF variants. Of those variants, 1.4 was only seen in 1 or 2 of the sequenced individuals and thus is a very rare variant (115). This study indicated that even though a variant is very rare and causes LOF, it is not necessarily associated with a disease. The other three variants (*ZNF534* c.2021dupC, *ZNF488* c.194delC and *GID8* c.*586delG) might have been considered of less importance for a case-control study. The *ZNF534* (zinc finger protein 534) c.2021dupC variant lengthens the 674 aa protein by only 3 aa (73). A study mentioned in chapter 4.1.2.1 found that if the distance from a mutated stop-codon to the next in-frame stop codon is less than 49 nucleotides the mutation is less likely to have a clinical effect. Given that this also holds true for frameshift variants that cause the loss of a stop-codon, the *ZNF534* variant (which introduces a stop-codon only 5 nucleotides downstream from the wild-type stop-codon) is unlikely to cause a clinically significant phenotype. Although the *ZNF488* (zinc finger protein 488) c.194delC variant shortens a 340 aa protein down to 79 aa and therefore is likely to cause LOF, studies show that the expression of the gene and its protein product is either low or not detected in normal breast tissue. Studies also indicate slight up-regulation of the gene and protein in BC tissue compared to normal breast tissue (88, 93), indicating that the loss of *ZNF488* would not be beneficial to the development of BC. Finally, the *GID8* (Glucose-induced degradation protein 8 homolog) c.*586delG variant is the deletion of a guanine (G) at the end of a predicted miRNA binding site for miR-342. According to DIANA-TarBase, binding of this miRNA to the 3'UTR of *GID8* has not been experimentally validated (80). Furthermore, the variant doesn't change the sequence of the predicted binding site since the next downstream base is also a G (70), making it unlikely to affect miRNA binding.

Validated SNPs and variants that could not be validated due to unsuccessful amplification or Sanger-sequencing were assessed using various databases. Following this, one additional variant, *DCAF7* c.*153C>T, was selected for screening in cases and controls. This decision was based on the following evidence: 1) DCAF7 is part of a nuclear complex with ZNF703 which has a role in the oncogenesis of luminal B breast tumors which indicates that DCAF7 might have a oncogenic role in BC as well, 2) the variant is situated in the most conserved base of a predicted binding site for hsa-miR-193b-3p, a miRNA which has been shown to down-regulate *DCAF7*, suggesting that this predicted miRNA binding site is likely to be real and that the variant could disrupt binding of hsa-miR-193b-3p to *DCAF7* mRNA and 3) *DCAF7* is more highly expressed in BC cell lines than normal BC tissue suggesting that disruption of *DCAF7* downregulation might have an oncogenic role. However, the results of the screening did not indicate that the variant increases risk of BC. Since then, we have come across evidence for why this variant is less likely to have an effect on the expression of *DCAF7* than was originally thought. There are two predicted binding sites of hsa-miR-193b-3p in the 3'UTR of *DCAF7*: the one where the variant is present and another one ~50 nucleotides upstream. By making use of mirSVR scores (available on www.microrna.org) it is possible to rank miRNA target sites by predicted down-regulation, and the upstream binding site receives a considerably lower mirSVR score than the binding site in which the variant is present (-0.4517 vs. -0.0159, respectively). According to the microrna.org site, the mirSVR score for the upstream binding site is considered a good score while the other one is not (116). In the study that was cited by DIANA-TarBase as indirect validation of the binding of hsa-miR-193b-3p to *DCAF7*, over-expression of the miRNA was shown to result in down-regulation of *DCAF7.* No efforts were made to identify the specific mechanism of the regulation and therefore it is possible that binding of the miRNA to the upstream binding site, rather than the binding site in which the variant is located, is responsible for the down-regulation (97). In addition, when the decision to screen for the *DCAF7* variant was made, we were not aware of its high frequency in European populations (5.8% in the 1000 Genomes EUR dataset). Since it is highly unlikely that such a prevalent variant has more than a low effect on BC risk, this variant probably would not have been chosen for a case-control study.

In the second part of the project, thirteen of the fifteen candidate variants were screened for simultaneously in a set of BC cases and their frequencies compared to those from the AGES-Reykjavík cohort. The other two candidate variants could not be included due to technical difficulties and will be screened for as part of another project. The results of the screening indicated that none of the variants is likely to have a moderate or high effect on BC risk. The variants were screened for in 540 unselected BC cases and the AGES-Reykjavik frequencies are based on the screening of 2983 individuals. To significantly detect a moderate-risk variant (OR ≥ 1.5) in this round, using the frequencies from AGES-Reykjavik and given a significance cut-off of p = 0.0038 (corrected for 13 tests using a Bonferroni correction) we would need to screen 10.000 unselected BC cases for the variants (theoretical p-value from Fisher's exact test for a variant with a frequency of 1% and an OR of 1.5, based on 10,000 BC samples and 2,983 control samples = 0.003647, calculated in R-project). If all available control samples (~6,000 samples) were screened and added to the AGES-Reykjavik frequency, the number of available BC samples for screening (~1,500) would still need to be doubled to reach a significant score (theoretical p-value from Fisher's exact test for a variant with a frequency of 1% and an OR of 1.5, based on 3,000 BC samples and 8,983 control samples = 0.002304, calculated in R-project). Therefore, if the screening had indicated that a variant was likely to predispose to BC, we would not have been able to significantly demonstrate its effect on BC risk without expanding our cohort considerably and/or establishing a collaboration with colleagues abroad with large sets of BC cases and controls and/or with deCODE Genetics that have developed a method to impute genotypes of detected variants into > 100,000 Icelanders. Alternatively, a variants effect could be demonstrated by using other tactics, *e.g.* showing that the variant segregates with BC more often than would be expected by chance and performing functional studies in cell-lines or animal models.

It is important to keep in mind that although the decision to make use of frequency data from the AGES-Reykjavik cohort provided us with the opportunity to screen for more variants than we otherwise could have due to financial reasons, it also put limits on the study. The samples from the AGES-Reykjavik cohort were genotyped by an exome-chip, which means that frequencies are only available for recorded SNPs located in (or close to) exons. Therefore, previously unknown coding SNPs, all indel variants and variants in non-coding regions were excluded from the analysis. The decision to rely on function-prediction tools to assess the deleteriousness of variants is also a caveat, since most such tools only predict the deleteriousness of missense variants, but place no assessment on SNPs that introduce stop-codons (nonsense variants) or disrupt splice-sites. The variants that were excluded because of these limits will be individually assessed as part of another ongoing project at the Laboratory of Cell Biology.

5.4 Where are the variants causing the high risk of BC in the families included in this study?

The GWL study by Arason *et. al.* indicated that there is no novel BRCA-like gene to be found in the nine families incorporated in that study (51). The study did not eliminate the possibility of high-risk mutations segregating within some or all of the families, but suggested that they would likely each be confined to a single family and therefore missed by the GWL study due to lack of statistical power. Another possibility is that the increased BC risk in some or all of the families incorporated in that study is caused by polygenic inheritance. The results of the first part of this M.Sc. project do not exclude that a high-risk mutation is segregating within family a). Such a mutation might be identified *e.g.* by lowering the CQ conditions or including variants that are present in both of the tumor samples that were sequenced, since such a variant is unlikely to be a somatic event and was likely missed by the WGS of the blood samples from the same individuals. However, based on the results of the first part of the project, as well as on preliminary data from analyses of families b) through d) and studies that have found that the majority of the missing heritability of BC is likely due to multiple variants with low to moderate effects (53, 117), the focus in the second part of the project was set on identifying variants with a moderate effect on risk that together could explain the increased risk of BC. Since no such variants were identified in this project, it is natural to wonder how they might be identified. As mentioned previously, all the variants that were screened for in the second part of the project were missense variants in genes that have a role in DNA repair. The fact that all known moderate-risk BC genes have a role in DNA repair could be due to ascertainment bias and does not exclude the possibility of moderate-risk variants being identified in genes that don‘t have a role in DNA repair (12). Therefore, although searching for moderate-risk variants in DNA repair genes is a logical first step, variants in other genes should be interrogated in future analyses of the WGS data. It is plausible that rare non-truncating variants, such as missense variants, play a part in BC susceptibility (12). High-risk missense variants have been identified in *BRCA1* and *BRCA2*, and there is good evidence of missense variants in several other genes, such as ATM and CHEK2, that confer an increased risk of BC (49, 118). Even so, LOF variants are more likely to have a large effect on protein function than missense variants (49) and therefore such variants should be included in future analyses of these families. The strict CQ conditions are another factor that could have caused us to miss causal variants. Since control frequency from the AGES-Reykjavik cohort was only available for reported variants, the CQ conditions could have been left at the default setting of 20, without the risk of screening for variants that were errors in the WGS.

Taking the above factors into account, work has started on a new analysis of the WGS data. The focus is still on variants in DNA repair genes and LOF variants are given priority. Predicted deleterious missense variants for which frequency from AGES-Reykjavik was not available, variants in promoter regions (as predicted by RegulomeDB: http://regulomedb.org/) and variants in binding sites of known BC tumor-suppressing miRNAs (119) are also given priority. The study is ongoing and the hope to find additional genetic factors that explain high BC risk in Icelandic non-*BRCA1* and *BRCA2* families is still present.

6 Conclusions

In the first part of this project, analysis of WGS data from an Icelandic HBC family (figure 1a) identified twenty-four loss-of-function (LOF) and miRNA binding site variants. Following validation and further assessment of the variants, five were considered candidate high-risk mutations and were screened for in a series of unselected BC cases and controls. The results of the screening revealed that none of the variants confer a significantly increased risk of BC, although it cannot be ruled out that one of the candidate variants, *TRMT44* c.1928-2_1929delAGAG, is a very rare predisposing variant. In the second part of the project, analysis of WGS data from four Icelandic HBC families (figure 1a-d) with a focus on exonic variants in DNA repair genes revealed 15 candidate moderate-risk variants. Thirteen of the candidate variants were successfully screened for in a series of unselected BC samples. The frequencies of the variants were compared to their control frequencies from AGES-Reykjavik, which revealed that none of them are likely to increase the risk of BC.

The pattern of BC cases in the family analyzed in the first part of the project suggests segregation of a high-risk mutation, and therefore the aim was to identify such a mutation. However, given that no high-risk mutation was missed in the first part of the project, the results indicate that the BC clustering in this family is more likely to be caused by the conjoined effects of two or more low- to moderate-risk variants. Taking the results of the first part of the project into account, WGS data from three additional families (which are also suggestive of segregating a high-risk genetic factor) were added to the analysis in the second part of the project. These three families had been previously studied in a genome-wide linkage (GWL) study, in which no evidence was revealed of a mutual high-risk mutation segregating within them. There is a possibility that the GWL study missed a high-risk locus due to lack of statistical power, but subsequent analyses have not been successful in identifying high-risk mutations within the families. Although the possibility of individual high-risk mutations segregating within the families has not been ruled out, all four families were analyzed together in the second part of the project with the aim of identifying variants of moderate-risk that contribute in a polygenic fashion to the increase in BC risk seen in the families. The analysis was limited to exonic variants in DNA repair genes and did not reveal any predisposing variants.

The use of control frequencies from the IHA and online function-prediction tools placed several limits on the study, *e.g.* because frequencies were only available for recorded exonic SNPs and the tools only assess missense variants. Further analyses are needed, and the next immediate step is to study variants in DNA repair genes for which frequency from AGES-Reykjavik was not available or were not assessed by function-prediction tools. These include LOF variants, predicted deleterious missense variants, variants located in promoter regions and variants located in binding sites of known BC tumor-suppressing miRNAs.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136(5):E359-86.
2. Tryggvadóttir L, Ólafsdóttir EJ, Jónasson JG. Krabbameinsskrá Íslands hjá Krabbameinsfélagi Íslands; 2015 [cited 2015 05/07]; Available from: www.krabbameinsskra.is.
3. Sinn HP, Kreipe H. A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition. Breast Care. 2013;8(2):149-54.
4. McCart Reed AE, Kutasovic JR, Lakhani SR, Simpson PT. Invasive lobular carcinoma of the breast: morphology, biomarkers and 'omics. Breast Cancer Res. 2015;17(1):12.
5. The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.
6. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.
7. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. P Nat Acad Sci USA. 2001;98(19):10869-74.
8. Weigelt B, Baehner FL, Reis-Filho JS. The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. J Pathol. 2010;220(2):263-80.
9. The American Cancer Society. What are the risk factors for breast cancer? 2015 [cited 2015 10/07]; Available from: http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-risk-factors.
10. Mavaddat N, Antoniou AC, Easton DF, Garcia-Closas M. Genetic susceptibility to breast cancer. Mol Oncol. 2010;4(3):174-91.
11. Lalloo F, Evans DG. Familial breast cancer. Clin Genet. 2012;82(2):105-14.
12. Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. Nat Genet. 2008;40(1):17-22.
13. Ghoussaini M, Pharoah PD, Easton DF. Inherited genetic susceptibility to breast cancer: the beginning of the end or the end of the beginning? Am J Pathol. 2013;183(4):1038-51.
14. Peto J, Mack TM. High constant incidence in twins and other relatives of women with breast cancer. Nat Genet. 2000;26(4):411-4.
15. Barh D, Gunduz M. Noninvasive Molecular Markers in Gynecologic Cancers. Boca Raton: CRC Press, Taylor & Francis Group; 2015.

16. Lu Y, Ek WE, Whiteman D, Vaughan TL, Spurdle AB, Easton DF, Pharoah PD, Thompson DJ, Dunning AM, Hayward NK, et al. Most common 'sporadic' cancers have a significant germline genetic component. Hum Mol Genet. 2014;23(22):6112-8.
17. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, Wang Q, Dennis J, Dunning AM, Shah M, et al. Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer I. 2015;107(5).
18. Economopoulou P, Dimitriadis G, Psyrri A. Beyond BRCA: new hereditary breast cancer susceptibility genes. Cancer Treat Rev. 2015;41(1):1-8.
19. Shiovitz S, Korde LA. Genetics of breast cancer: a topic in evolution. Ann Oncol. 2015;26(7):1291-9.
20. Harris TJ, McCormick F. The molecular pathology of cancer. Nat Rev Clin Oncol. 2010;7(5):251-65.
21. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. 1994;266(5182):66-71.
22. Couch FJ, Wang X, McGuffog L, Lee A, Olswold C, Kuchenbaecker KB, Soucy P, Fredericksen Z, Barrowdale D, Dennis J, et al. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS Genet. 2013;9(3):e1003212.
23. Shuen AY, Foulkes WD. Inherited mutations in breast cancer genes--risk and response. J Mammary Gland Biol. 2011;16(1):3-15.
24. Venkitaraman AR. Cancer suppression by the chromosome custodians, BRCA1 and BRCA2. Science. 2014;343(6178):1470-5.
25. van den Broek AJ, Schmidt MK, van 't Veer LJ, Tollenaar RA, van Leeuwen FE. Worse breast cancer prognosis of BRCA1/BRCA2 mutation carriers: what's the evidence? A systematic review with meta-analysis. PLoS One. 2015;10(3):e0120189.
26. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D, et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science. 1994;265(5181):2088-90.
27. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G. Identification of the breast cancer susceptibility gene BRCA2. Nature. 1995;378(6559):789-92.
28. Roy R, Chun J, Powell SN. BRCA1 and BRCA2: different roles in a common pathway of genome protection. Nat Rev Cancer. 2012;12(1):68-78.
29. Petrucelli N, Daly MB, Feldman GL. Hereditary breast and ovarian cancer due to mutations in BRCA1 and BRCA2. Genet Med. 2010;12(5):245-59.

30. Sawyer SL, Tian L, Kahkonen M, Schwartzentruber J, Kircher M, Majewski J, Dyment DA, Innes AM, Boycott KM, Moreau LA, et al. Biallelic mutations in BRCA1 cause a new Fanconi anemia subtype. Cancer Discov. 2015;5(2):135-42.

31. Ferla R, Calo V, Cascio S, Rinaldi G, Badalamenti G, Carreca I, Surmacz E, Colucci G, Bazan V, Russo A. Founder mutations in BRCA1 and BRCA2 genes. Ann Oncol. 2007;18 Suppl 6:vi93-8.

32. Bergthorsson JT, Jonasdottir A, Johannesdottir G, Arason A, Egilsson V, Gayther S, Borg A, Hakanson S, Ingvarsson S, Barkardottir RB. Identification of a novel splice-site mutation of the BRCA1 gene in two breast cancer families: screening reveals low frequency in Icelandic breast cancer patients. Hum Mutat. 1998;Suppl 1:S195-7.

33. Johannesdottir G, Gudmundsson J, Bergthorsson JT, Arason A, Agnarsson BA, Eiriksdottir G, Johannsson OT, Borg A, Ingvarsson S, Easton DF, et al. High prevalence of the 999del5 mutation in icelandic breast and ovarian cancer patients. Cancer Res. 1996;56(16):3663-5.

34. Thorlacius S, Sigurdsson S, Bjarnadottir H, Olafsdottir G, Jonasson JG, Tryggvadottir L, Tulinius H, Eyfjord JE. Study of a single BRCA2 mutation with high carrier frequency in a small population. Am J Hum Genet. 1997;60(5):1079-84.

35. Sorrell AD, Espenschied CR, Culver JO, Weitzel JN. Tumor protein p53 (TP53) testing and Li-Fraumeni syndrome : current status of clinical applications and future directions. Mol Diagn Ther. 2013;17(1):31-47.

36. Song MS, Carracedo A, Salmena L, Song SJ, Egia A, Malumbres M, Pandolfi PP. Nuclear PTEN regulates the APC-CDH1 tumor-suppressive complex in a phosphatase-independent manner. Cell. 2011;144(2):187-99.

37. Hollander MC, Blumenthal GM, Dennis PA. PTEN loss in the continuum of common cancers, rare syndromes and mouse models. Nat Rev Cancer. 2011;11(4):289-301.

38. Andrade-Vieira R, Xu Z, Colp P, Marignani PA. Loss of LKB1 expression reduces the latency of ErbB2-mediated mammary gland tumorigenesis, promoting changes in metabolic pathways. PLoS One. 2013;8(2):e56567.

39. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K, et al. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat Genet. 2006;38(11):1239-41.

40. Karppinen SM, Heikkinen K, Rapakko K, Winqvist R. Mutation screening of the BARD1 gene: evidence for involvement of the Cys557Ser allele in hereditary susceptibility to breast cancer. J Med Genet. 2004;41(9):e114.

41. Karppinen SM, Barkardottir RB, Backenhorn K, Sydenham T, Syrjakoski K, Schleutker J, Ikonen T, Pylkas K, Rapakko K, Erkko H, et al. Nordic collaborative study of the BARD1 Cys557Ser allele in 3956 patients with cancer: enrichment in familial BRCA1/BRCA2 mutation-negative breast cancer but not in other malignancies. J Med Genet. 2006;43(11):856-62.

42. Stacey SN, Sulem P, Johannsson OT, Helgason A, Gudmundsson J, Kostic JP, Kristjansson K, Jonsdottir T, Sigurdsson H, Hrafnkelsson J, et al. The BARD1 Cys557Ser variant and breast cancer risk in Iceland. PLoS Med. 2006;3(7):e217.

43. Gonzalez-Hormazabal P, Reyes JM, Blanco R, Bravo T, Carrera I, Peralta O, Gomez F, Waugh E, Margarit S, Ibanez G, et al. The BARD1 Cys557Ser variant and risk of familial breast cancer in a South-American population. Mol Biol Rep. 2012;39(8):8091-8.

44. Ding DP, Zhang Y, Ma WL, He XF, Wang W, Yu HL, Guo YB, Zheng WL. Lack of association between BARD1 Cys557Ser variant and breast cancer risk: a meta-analysis of 11,870 cases and 7,687 controls. J Cancer Res Clin. 2011;137(10):1463-8.

45. Jakubowska A, Cybulski C, Szymanska A, Huzarski T, Byrski T, Gronwald J, Debniak T, Gorski B, Kowalska E, Narod SA, et al. BARD1 and breast cancer in Poland. Breast Cancer Res Tr. 2008;107(1):119-22.

46. Johnatty SE, Beesley J, Chen X, Hopper JL, Southey MC, Giles GG, Goldgar DE, Chenevix-Trench G, Spurdle AB. The BARD1 Cys557Ser polymorphism and breast cancer risk: an Australian case-control and family analysis. Breast Cancer Res Tr. 2009;115(1):145-50.

47. Kiiski JI, Pelttari LM, Khan S, Freysteinsdottir ES, Reynisdottir I, Hart SN, Shimelis H, Vilske S, Kallioniemi A, Schleutker J, et al. Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. P Nat Acad Sci USA. 2014;111(42):15172-7.

48. Peterlongo P, Catucci I, Colombo M, Caleca L, Mucaki E, Bogliolo M, Marin M, Damiola F, Bernard L, Pensotti V, et al. FANCM c.5791C>T nonsense mutation (rs144567652) induces exon skipping, affects DNA repair activity and is a familial breast cancer risk factor. Hum Mol Genet. 2015.

49. Easton DF, Pharoah PD, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, Devilee P, Meindl A, Couch FJ, Southey M, et al. Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med. 2015;372(23):2243-57.

50. Pharoah PD, Tyrer J, Dunning AM, Easton DF, Ponder BA. Association between common variation in 120 candidate genes and breast cancer risk. PLoS Genet. 2007;3(3):e42.

51. Arason A, Gunnarsson H, Johannesdottir G, Jonasson K, Bendahl PO, Gillanders EM, Agnarsson BA, Jonsson G, Pylkas K, Mustonen A, et al. Genome-wide search for breast cancer linkage in large Icelandic non-BRCA1/2 families. Breast Cancer Res. 2010;12(4):R50.

52. Easton DF. How many more breast cancer predisposition genes are there? Breast Cancer Res. 1999;1(1):14-7.

53. Gracia-Aznarez FJ, Fernandez V, Pita G, Peterlongo P, Dominguez O, de la Hoya M, Duran M, Osorio A, Moreno L, Gonzalez-Neira A, et al. Whole exome sequencing suggests much of non-BRCA1/BRCA2 familial breast cancer is due to moderate and low penetrance susceptibility alleles. PLoS One. 2013;8(2):e55681.

54. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70.

55. Zhang T, Boswell EL, McCall SJ, Hsu DS. Mismatch repair gone awry: Management of Lynch syndrome. CRC Cr Rev Oncol-Hem. 2015;93(3):170-9.

56. Desmedt C, Voet T, Sotiriou C, Campbell PJ. Next-generation sequencing in breast cancer: first take home messages. Curr Opin Oncol. 2012;24(6):597-604.

57. Hilbers FS, Vreeswijk MP, van Asperen CJ, Devilee P. The impact of next generation sequencing on the analysis of breast cancer susceptibility: a role for extremely rare genetic variation? Clin Genet. 2013;84(5):407-14.

58. Connor AA, Katzov-Eckert H, Whelan T, Aronson M, Lau L, Marshall C, Charames GS, Pollett A, Gallinger S, Lerner-Ellis J. Identification of a novel MSH6 germline variant in a family with multiple gastro-intestinal malignancies by next generation sequencing. Fam Cancer. 2015;14(1):69-75.

59. Nikopoulos K, Gilissen C, Hoischen A, van Nouhuys CE, Boonstra FN, Blokland EA, Arts P, Wieskamp N, Strom TM, Ayuso C, et al. Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy. Am J Hum Genet. 2010;86(2):240-7.

60. Rosa-Rosa JM, Gracia-Aznarez FJ, Hodges E, Pita G, Rooks M, Xuan Z, Bhattacharjee A, Brizuela L, Silva JM, Hannon GJ, et al. Deep sequencing of target linkage assay-identified regions in familial breast cancer: methods, analysis pipeline and troubleshooting. PLoS One. 2010;5(4):e9976.

61. Koboldt DC, Ding L, Mardis ER, Wilson RK. Challenges of sequencing human genomes. Brief Bioinform. 2010;11(5):484-98.

62. Flanagan SE, Patch AM, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Bioma. 2010;14(4):533-7.

63. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics. 2013;14 Suppl 3:S7.

64. Svansdottir E, Denollet J, Thorsson B, Gudnason T, Halldorsdottir S, Gudnason V, van den Broek KC, Karlsson HD. Association of type D personality with unhealthy lifestyle, and estimated risk of coronary events in the general Icelandic population. Eur J Prev Cardiol. 2013;20(2):322-30.

65. Ingenuity® Variant AnalysisTM. "Data were analyzed through the use of QIAGEN's Ingenuity® Variant Analysis™ software from QIAGEN Redwood City.". Ingenuity® Variant AnalysisTM; 2015 [cited 2015 22.07]; Available from: http://www.ingenuity.com/products/variant-analysis.

66. Ingenuity® Variant AnalysisTM. Ingenuity® Variant AnalysisTM Quick-Start Guide. Ingenuity® Variant AnalysisTM; 2014 [22.07]; Available from: http://www.ingenuity.com/wp-content/uploads/2014/04/1080849_QSGuide_Ingenuity_VariantAnalysis_0314_lr.pdf.

67. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998;8(3):175-85.

68. Ingenuity® Variant AnalysisTM. Filters and analytics overview. 2015 [cited 2015 23.07]; Available from: http://ingenuity.force.com/variants/VariantTutorials#.

69. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.

70. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006.

71. Harris TB, Launer LJ, Eiriksdottir G, Kjartansson O, Jonsson PV, Sigurdsson G, Thorgeirsson G, Aspelund T, Garcia ME, Cotch MF, et al. Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. Am J Epidemiol. 2007;165(9):1076-87.

72. Gabriel S, Ziaugra L, Tabbaa D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr Protoc Hum Genet. 2009;Chapter 2:Unit 2.12.

73. Uniprot Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):D204-12.

74. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):D662-9.

75. Shtutman M, Baig M, Levina E, Hurteau G, Lim CU, Broude E, Nikiforov M, Harkins TT, Carmack CS, Ding Y, et al. Tumor-specific silencing of COPZ2 gene encoding coatomer protein complex subunit zeta 2 renders tumor cells dependent on its paralogous gene COPZ1. P Nat Acad Sci USA. 2011;108(30):12449-54.

76. Tsuruta T, Kozaki K, Uesugi A, Furuta M, Hirasawa A, Imoto I, Susumu N, Aoki D, Inazawa J. miR-152 is a tumor suppressor microRNA that is silenced by DNA hypermethylation in endometrial cancer. Cancer Res. 2011;71(20):6450-62.

77. Zheng B, Yu X, Chai R. Myotubularin-related phosphatase 3 promotes growth of colorectal cancer cells. ScientificWorldJournal. 2014;2014:703804.

78. Gong Y, He T, Yang L, Yang G, Chen Y, Zhang X. The role of miR-100 in regulating apoptosis of breast cancer cells. Sci Rep. 2015;5:11650.

79. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120(1):15-20.

80. Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, Anastasopoulos IL, Maniou S, Karathanou K, Kalfakakou D, et al. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res. 2015;43(Database issue):D153-9.

81. Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. 2015 [cited 2015 5/8]; (dbSNP Build ID: {142}):[Available from: http://www.ncbi.nlm.nih.gov/SNP/.

82. Bhattacharya R, Wang E, Dutta SK, Vohra PK, E G, Prakash YS, Mukhopadhyay D. NHERF-2 maintains endothelial homeostasis. Blood. 2012;119(20):4798-806.

83. Takahashi Y, Morales FC, Kreimann EL, Georgescu MM. PTEN tumor suppressor associates with NHERF proteins to attenuate PDGF receptor signaling. EMBO J. 2006;25(4):910-20.

84. Meneses-Morales I, Tecalco-Cruz AC, Barrios-Garcia T, Gomez-Romero V, Trujillo-Gonzalez I, Reyes-Carmona S, Garcia-Zepeda E, Mendez-Enriquez E, Cervantes-Roldan R, Perez-Sanchez V, et al. SIP1/NHERF2 enhances estrogen receptor alpha transactivation in breast cancer cells. Nucleic Acids Res. 2014;42(11):6885-900.

85. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Res. 2013;41(Web Server issue):W169-73.

86. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.

87. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010;327(5961):78-81.

88. Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N, et al. Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Res. 2014;42(Database issue):D926-32.

89. Larkin SE, Holmes S, Cree IA, Walker T, Basketter V, Bickers B, Harris S, Garbis SD, Townsend PA, Aukim-Hastie C. Identification of markers of prostate cancer progression using candidate gene expression. Brit J Cancer. 2012;106(1):157-65.

90. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP). Seattle, WA [cited 2015 24.07]; Available from: http://evs.gs.washington.edu/EVS/.

91. Hamby SE, Thomas NS, Cooper DN, Chuzhanova N. A meta-analysis of single base-pair substitutions in translational termination codons ('nonstop' mutations) that cause human inherited disease. Hum Genomics. 2011;5(4):241-64.

92. Yang S, Liu X, Yin Y, Fukuda MN, Zhou J. Tastin is required for bipolar spindle assembly and centrosome integrity during mitosis. Faseb j. 2008;22(6):1960-72.

93. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.

94. Lefebvre C, Terret ME, Djiane A, Rassinier P, Maro B, Verlhac MH. Meiotic spindle stability depends on MAPK-interacting and spindle-stabilizing protein (MISS), a new MAPK substrate. J Cell Biol. 2002;157(4):603-13.

95. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet. 1999;23(1):41-6.

96. Sircoulomb F, Nicolas N, Ferrari A, Finetti P, Bekhouche I, Rousselet E, Lonigro A, Adelaide J, Baudelet E, Esteyries S, et al. ZNF703 gene amplification at 8p12 specifies luminal B breast cancer. EMBO Mol Med. 2011;3(3):153-66.

97. Chen J, Feilotter HE, Pare GC, Zhang X, Pemberton JG, Garady C, Lai D, Yang X, Tron VA. MicroRNA-193b represses cell proliferation and regulates cyclin D1 in melanoma. Am J Pathol. 2010;176(5):2520-9.

98. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009;37(9):e67.

99. Armstrong RC, Migneault A, Shegog ML, Kim JG, Hudson LD, Hessler RB. High-grade human brain tumors exhibit increased expression of myelin transcription factor 1 (MYT1), a zinc finger DNA-binding protein. J Neuropathol Exp Neurol. 1997;56(7):772-81.

100.Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81.

101.Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.

102.Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118.

103.Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7(10):e46688.

104.Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011;88(4):440-9.

105.Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet. 2007;39(2):165-7.

106.Kan Z, Zheng H, Liu X, Li S, Barber TD, Gong Z, Gao H, Hao K, Willard MD, Xu J, et al. Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome Res. 2013;23(9):1422-33.

107.Jones MA, Ng BG, Bhide S, Chin E, Rhodenizer D, He P, Losfeld ME, He M, Raymond K, Berry G, et al. DDOST mutations identified by whole-exome sequencing are implicated in congenital disorders of glycosylation. Am J Hum Genet. 2012;90(2):363-8.

108.Rienhoff HY, Jr., Yeo CY, Morissette R, Khrebtukova I, Melnick J, Luo S, Leng N, Kim YJ, Schroth G, Westwick J, et al. A mutation in TGFB3 associated with a syndrome of low muscle mass, growth retardation, distal arthrogryposis and clinical features overlapping with Marfan and Loeys-Dietz syndrome. Am J Med Genet A. 2013;161A(8):2040-6.

109.Berget SM. Exon recognition in vertebrate splicing. J Biol Chem. 1995;270(6):2411-4.

110.Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281-97.

111.Vasudevan S. Posttranscriptional upregulation by microRNAs. Wiley Interdiscip Rev RNA. 2012;3(3):311-30.

112.Pelletier C, Weidhaas JB. MicroRNA binding site polymorphisms as biomarkers of cancer risk. Expert Rev Mol Diagn. 2010;10(6):817-29.

113.Chin LJ, Ratner E, Leng S, Zhai R, Nallur S, Babar I, Muller RU, Straka E, Su L, Burki EA, et al. A SNP in a let-7 microRNA complementary site in the KRAS 3' untranslated region increases non-small cell lung cancer risk. Cancer Res. 2008;68(20):8535-40.

114.Leschziner GD, Coffey AJ, Andrew T, Gregorio SP, Dias-Neto E, Calafato M, Bentley DR, Kinton L, Sander JW, Johnson MR. Q8IYL2 is a candidate gene for the familial epilepsy syndrome of Partial Epilepsy with Pericentral Spikes (PEPS). Epilepsy Res. 2011;96(1-2):109-15.

115.Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Holm H, Saemundsdottir J, Helgadottir HT, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47(5):435-44.

116.Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol. 2010;11(8):R90.

117.Wen H, Kim YC, Snyder C, Xiao F, Fleissner EA, Becirovic D, Luo J, Downs B, Sherman S, Cowan KH, et al. Family-specific, novel, deleterious germline variants provide a rich resource to identify genetic predispositions for BRCAx familial breast cancer. BMC Cancer. 2014;14:470.

118.Tavtigian SV, Chenevix-Trench G. Growing recognition of the role for rare missense substitutions in breast cancer susceptibility. Biomark Med. 2014;8(4):589-603.

119.van Schooneveld E, Wildiers H, Vergote I, Vermeulen PB, Dirix LY, Van Laere SJ. Dysregulation of microRNAs in breast cancer and their potential role as prognostic and predictive biomarkers in patient management. Breast Cancer Res. 2015;17:21.

Supplementary data

Table S1: Sequences of primers used for Sanger-sequencing. Names of the genes harboring the variants and primer annealing temperatures are also listed.

Gene	Sequence 5’→ 3’	Annealing temperature (°C)
TRMT44	F: TCGTGCTGGAGTGCATAACG	55
	R: ACTCCGTACCTTGGAACACC	
ZNF534	F: GTCTTCAGTCGGAATTCACG	55
	R: TGACCTCATGATCCACCTGC	
ZNF488	F: ATCGGCTGAAAACAGATGGC	58*
	R: TGCTCTCCACGTGTCTTCG	
PRIM2	F: AGCAGCACTTTCTTATGGTG	55
	R: TTGTCTGGATCCATCTTTCC	
CHST15	F: TCTGTTCCTATGCTGAAACG	55*
	R: TCCATGAGAAAGTGACAGAAGC	
GPR27	F: GCTGTGCAAGATGTTCTACG	55*
	R: CCTCCCTCATAAACCAATGC	
CCDC48	F: CTCTGTGGGTGAAGGTGAGC	55*
	R: AGCTGTCAGTGGTTGGGCTG	
MEX3B	F: CACTTGGATGGTGGTCTGC	55
	R: GTAAAATCAAAGCGCTGCGG	
KRTAP10-12	F: CCGTGTCAACAGTCCTGCTG	58*
	R: GACTCATAGTGCCCAGTGG	
COPZ2	F: AGTTCTGAGTTGGCTGCT	x
	R: AGAACGAGCAAGAGGACG	
MTMR3	F: TGTACAGAGTGACAGATTTGG	x
	R: CGGAAGAAACAAGCCATCC	
GID8	F: GCTTTCTGTTAGCTTAGGCAG	55*
	R: ACGCCCCAAAGACAAAAGG	
DICER1	F: AATATGAGACACCTCTGCTC	55
	R: GACTTGTAGGCACTCTTCAC	
STAC2	F: AGAGACACAGAGCAGATGG	55*
	R: AAACACAGACCCTCGTACC	
SLC9A3R2	F: CCTGTGGCAGCAAGATAGG	x
	R: CGGGGAGGAAATGGTTTGC	
INMT	F: GAACAGCTCCTACACAGTCC	64*
	R: GTTAGACAGTATCCATTCCTCAC	
TROAP	F: CTCTAAATGGAGGCTCTTCTCTG	64*
	R: TACTTCAAGCTGTTCCTGGAGG	
ASPSCR1	F: AGACATCCTGGGACAGTGCT	x
	R: AAGCGTGTTCTCTGCTCTGG	
HOXA5	F: CAAGTCACCTCTACAACAGC	64*
	R: GATCTGCTTTCTGTTCATCTC	
DLG2	F: CTTCCTTCATACTGCAATGTC	58*
	R: CATCTGGATTCCCTCAAAGG	
MAPK1IP1L	F: TGCTGGTTTCACTATTAGAGG	58*
	R: ATTTGCCAATGAAGTTGCAG	
DCAF7	F: TCAGAGTGTAGTGTTGGTGG	64*
	R: CTCAACACAACGCCTGAC	
MYT1	F: GTGGTGGCCCTATCTGTGTG	x
	R: AAGATGCCACTCACACCACC	
PPP6R1	F: ACTTGGCAACAGGGCTGG	x
	R: AAGGTTGCCACCCACGTG	

F: Forward primer. R: Reverse primer. Asterisk (*): Betaine was used in the PCR reaction. X: Primers performed inadequately.

Table S2: Sequences of primers used for fragment analysis. Names of the genes harboring the variants, primer annealing temperatures and the length of the generated fragments are also listed.

Gene	Sequence 5’→ 3’	Annealing temperature (°C)	Length of fragment
TRMT44	F: FAM-TCGTGCTGGAGTGCATAACG	55	153
	R: AGCTCGTTGGCTACTTCTGC		
ZNF534	F: FAM-ACTGGAGTGAAGCCTTACAG	62	188
	R: TGACCTCATGATCCACCTGC		
ZNF488	F: FAM-GCGACTTAGCGAACCTGAGC	58*	189
	R: TGCTCTCCACGTGTCTTCG		
GID8	F: FAM-TCATGTGTGAGGGCATTGAG	55*	274
	R: ACGCCCCAAAGACAAAAGG		

F: Forward primer. R: Reverse primer. Asterisk (*): Betaine was used in the PCR reaction.