



**A search for novel genes on chromosomes 2p, 6q
and 14q in an Icelandic high-risk breast cancer
family**

Óskar Örn Hálfðánarson

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



HÁSKÓLI ÍSLANDS

**Leit að áhrifastökkbreytingum í genum á völdum svæðum á
litningum 2p, 6q og 14q í fjölskyldu með háa tíðni
brjóstakrabbameins**

Óskar Örn Hálfðánarson

Ritgerð til meistaragráðu í Líf- og læknavísindum

Umsjónarkennari: Rósa Björk Barkardóttir

Meistaránámsnefnd: Aðalgeir Arason og Inga Reynisdóttir

Læknadeild

Heilbrigðisvísindasvið Háskóla Íslands

Júní 2013

A search for novel genes on chromosomes 2p, 6q and 14q in an Icelandic high-risk breast cancer family

Óskar Örn Hálfðánarson

Thesis for the degree of Master of Science

Supervisor: Rósa Björk Barkardóttir

Masters committee: Aðalgeir Arason and Inga Reynisdóttir

Faculty of Medicine

School of Health Sciences

June 2013

Ritgerð þessi er til meistaragráðu í Líf- og læknávisindum og er óheimilt að afrita ritgerðina á nokkurn hátt nema með leyfi rétthafa.

© Óskar Örn Hálfðánarson 2013

Prentun: Háskólaprent

Reykjavík, Ísland 2013

Ágrip

Um það bil 5-10% greindra einstaklinga með brjóstakrabbamein tilheyra fjölskyldum með háa tíðni meinsins. Um helmingur fjölskyldnanna er án tengsla við stökkbreytingar í þekktum krabbameinsgenum á borð við *BRCA1* og *BRCA2* og kallast BRCAx-fjölskyldur. Í undanfara þessa verkefnis var sýnt fram á tengsl svæða á litningum 2p, 6q og 14q við brjóstakrabbamein í einni íslenskri BRCAx-fjölskyldu (70234). Í heildina eru 554 gen innan svæðanna þriggja en í þessu verkefni, skilgreint sem fyrsti hluti rannsóknarinnar, var ákveðið að raðgreina 274 gen. Markmið verkefnisins var að finna stökkbreytingar í einhverjum þessara gena sem teldust líklegar til þess að valda aukinni hættu á myndun brjóstakrabbameins.

Roche 454 raðgreiningarniðurstöður fjögurra sýna úr meðlimum 70234 mynduðu grunn verkefnisins. Kímlínubreytileikar sem voru sameiginlegir með sýnunum fjórum voru teknir fyrir og lagt var mat á hvaða breytileikar væru líklegastir til þess að hafa áhrif á virkni þeirra gena sem báru þá. Fyrst var horft til breytinga af þremur gerðum; breytinga sem valda hliðrun á lesamma, þeirra sem kalla fram ótímabæran stöðvunartákna og splæsibreytinga. Því næst var horft til próteinkóðandi basabreytinga sem leiða til amínósýruskipta. SIFT og polyphen2 voru notuð til þess að leggja mat á mögulega skaðsemi slíkra breytinga. Að lokum voru breytingar á öðrum svæðum skoðaðar. Kandídatbreytingar voru skimaðar í óvöldum sjúklingahópi og viðmiðunarhópi og einnig í völdum fjölskylduefnivið. Kíkvaðratpróf var notað til þess að leggja mat á hvort tölfræðilega marktækur munur væri á samsætutíðni milli hópa.

Heildarfjöldi sameiginlegra kímlínu breytileika var 1540. Þar af voru 146 breytileikar staðsettir á próteinkóðandi svæðum. Skimað var fyrir fjórum próteinkóðandi breytileikum og tveimur utan slíkra svæða, þar af einni splæsibreytingu. Ekki reyndist vera tölfræðilega marktækur munur á samsætutíðni þessara breytinga milli hópa.

Engar stökkbreytingar fundust sem líklegar eru til þess að skýra aukna tilhneigingu til myndunar brjóstakrabbameins í fjölskyldu 70234. Næsta skref er að raðgreina þau 280 gen innan svæðanna þriggja sem ekki voru raðgreind í þessum fyrsta hluta rannsóknarinnar.

Abstract

It has been estimated that approximately 5-10% of breast cancer cases arise within high-risk hereditary breast cancer families. A little less than half of these familial cases have not shown linkage to pathogenic mutations within known cancer genes such as *BRCA1* and *BRCA2*. Those families are generally referred to as BRCAx families. In a previous genome wide search for breast cancer linkage, performed at our laboratory, three highly suggestive signals were found at chromosomes 2p, 6q and 14q in one Icelandic BRCAx family (70234). There are a total of 554 genes to be found within the three regions combined. However, this project, defined as the first phase of a sequencing study on family 70234, revolved around the sequencing of 274 out of those 554 genes. The aim of the project was to identify pathogenic mutations in one or more of these genes in family 70234.

Resequencing data, obtained via the Roche 454 sequencing platform, from four samples from family 70234 was the base of this project. Germline variants that were shared across all four samples were identified and evaluated for their possible pathogenicity. Firstly, all frameshift, nonsense and splice-site mutations among the shared variants were identified. Secondly, non-synonymous variants were identified and evaluated. SIFT and polyphen2 were used for the prediction of which non-synonymous SNPs were most likely to have a detrimental effect on the protein products of the genes harbouring them. Thirdly, shared variants within other regions of the genome were considered. Candidate variants were screened for in groups of controls and unselected breast cancer cases, as well as in a group of selected samples from other Icelandic hereditary breast cancer families. A chi-square test was used to evaluate whether there was a statistically significant difference in allele frequency between groups.

The total number of shared germline variants identified was 1540. The number of variants within protein coding regions was 146. Four candidate protein coding variants were screened for and their allele frequency within the groups was estimated. Two non-coding variants were screened for as well, thereof one splice-site mutation. None of the candidate variants turned out to have a statistically significant difference in allele frequency when the groups were compared.

We were not successful in identifying mutations that are likely to explain the increased breast cancer risk for members of family 70234. The next steps involve the sequencing of the 280 genes within the regions on chromosomes 2p, 6q and 14q that were not included in this phase of the study.

“The more comfortable we become with being stupid, the deeper we will wade into the unknown and the more likely we are to make big discoveries.”

Martin A. Schwartz

“(...) and wherever the knowledge takes us, the knowledge will empower us to do more.”

Michael Stratton

Acknowledgements

Special thanks must go to **Rósa Björk Barkardóttir**, **Aðalgeir Arason** and **Inga Reynisdóttir** for giving me the opportunity to participate in this project and work in a very exciting, and expanding, field of genomic research. Their excellent guidance, fruitful discussions and endless support were second to none.

To my wonderful co-workers at the Department of Cell Biology, **Anna Marzellíusardóttir**, **Edda Olguðóttir**, **Eydís Þórunn Guðmundsdóttir**, **Guðrún Jóhannesdóttir** and **Þorbjörg Einarsdóttir**, who brighten up the darkest of days, I wish to express my sincere gratitude for everything. A cup of coffee never tastes as good as it does in your company. **Anna** deserves a big thank you, for all of her help in the final stages of this project. Also, a very special thank you must go to **Guðrún** for being my everlasting lighthouse in the lab.

I would like to thank **Elísabet Guðmundsdóttir**, formerly at Roche NimbleGen Iceland Llc but now at Matis, for her help in the sequence capture procedure. Furthermore, my gratitude goes to **Ólafur Friðjónsson** at Matis for all of his help and willingness to educate a novice like myself.

My thanks go out to **Vilmundur Guðnason** and **Guðný Eiríksdóttir** at the Icelandic Heart Association for granting us access to a portion of their samples. I would also like to thank **Bjarni A Agnarsson** for all of his help on this project.

Last but not least, I want to thank my family and friends for all of their endless support over the years and their patience while I was working on this thesis. Special thanks must also go to my **mom** and **dad**, whose support I can always count on. Without them I would not be where I am today or literally anywhere, for that matter.

So, from the bottom of my heart, thank you all!

This project was supported by the Research Fund of Landspítali University Hospital, the Nordic Cancer Union, the Icelandic Research Fund (Rannís) and the Icelandic association “Walking for Breast Cancer” (Göngum Saman).

Table of contents

Ágrip.....	3
Abstract	5
Acknowledgements.....	7
Table of contents	8
List of figures	10
List of tables	11
Abbreviations.....	12
1 Introduction	15
1.1 Breast cancer	15
1.2 Familial breast cancer.....	16
1.3 High penetrance genes.....	17
1.3.1 <i>BRCA1</i> and <i>BRCA2</i>	17
1.3.2 Other high penetrance genes	19
1.3.3 Moderate penetrance genes.....	19
1.3.4 Low penetrance alleles.....	20
1.4 Hereditary breast cancer in Iceland.....	21
1.5 Identifying new breast cancer susceptibility alleles	23
1.5.1 Genetic linkage studies	23
1.5.2 Genome Wide Association Studies (GWAS).....	23
1.5.3 The candidate gene approach	23
1.6 Next Generation Sequencing	24
1.6.1 Alignment and variant detection.....	25
1.6.2 Prioritizing identified variants	26
2 Aims.....	28
3 Material and Methods	29
3.1 Sample selection	29
3.1.1 Samples selected for targeted resequencing	29
3.1.2 Samples used for the screening of candidate variants	29
3.2 Data generation.....	29
3.2.1 Targeted sequence capture.....	29
3.2.2 454 resequencing of the targeted regions.....	30
3.3 Data analysis.....	30
3.3.1 Alignment.....	30
3.3.2 Variant calling	31

3.3.3	Variant annotation	31
3.3.4	Identification of candidate variants.....	32
3.4	Genotyping of candidate variants.....	32
3.4.1	Fragment analysis.....	32
3.4.2	SNP genotyping	34
3.5	Calculations and statistical analysis	36
4	Results.....	37
4.1	Data analysis.....	37
4.1.1	Alignment and variant calling.....	38
4.1.2	Variant annotation	38
4.1.3	Identification of candidate causal variants.....	40
4.2	Validation of candidate variants	42
4.3	Case-control analysis of candidate variants.....	42
4.3.1	<i>FANCM</i>	43
4.3.2	<i>TSPYL1</i>	43
4.3.3	<i>SRD5A2</i>	43
4.3.4	<i>CAPN14</i>	44
4.3.5	<i>SLC22A16</i>	44
4.3.6	<i>BAZ1A</i>	44
5	Discussion.....	46
5.1	Premise of the study	46
5.2	The sequencing procedure	47
5.3	Variant identification and annotation.....	48
5.4	Variant analysis pipeline	49
5.5	Genotyping did not confirm any candidate variants as likely to be causal.....	51
5.6	Where are the causative variants?	51
6	Conclusions	54
	References	55

List of figures

Figure 1. Worldwide incidence and mortality rates of breast cancer.	11
Figure 2. A schematic representation of <i>BRCA1</i> and <i>BRCA2</i>	14
Figure 3. Chromosomal positions of possible BC susceptibility alleles.....	17
Figure 4. Cosegregation of haplotypes at three chromosomal regions in family 70234.	18
Figure 5. An overview of genes in BRCA related functional pathways	20
Figure 6. The various types of functional variants within the human genome.....	22
Figure 7. Analysis of two samples from a fragment analysis reaction using GeneMapper	30
Figure 8. Analysis of SNP genotyping data.	32
Figure 9. An overview of the analysis pipeline used to identify and evaluate shared variants in family 70234	33
Figure 10. The allele frequency distribution of the 73 previously reported non-synonymous SNPs identified within all of the sequenced samples from family 70234.....	36
Figure 11. The variant prioritization strategy used in the search for novel BC genes in family 70234..	37

List of tables

Table 1. The targeted boundaries in bp based on the hg19 reference assembly.....	27
Table 2. A list of possible values for the annotation of all variants with ANNOVAR.	28
Table 3. A list of possible values for exonic variant annotations with ANNOVAR.	29
Table 4. A list of primer sequences used for fragment analysis of candidate variants.....	30
Table 5. A list of primer and probe sequences used for the genotyping of candidate SNPs.....	32
Table 6. The number of individual variant types among the 1540 shared variants within the three chromosomal regions sequenced.....	36
Table 7. The number of shared protein coding variants in family 70234 and their predicted effect on the coding sequence of the genes that harbour them.....	36
Table 8. The candidate variants selected for further analysis.....	39
Table 9. The allele frequency of candidate variants in BC cases versus controls	42

Abbreviations

ATM	Ataxia Telangiectasia Mutated
ACF	ATP-dependent Chromatin Assembly Factor
BARD1	BRCA1 Associated RING domain 1
BAZ1A	Bromodomain Adjacent to Zinc Finger Domain, 1A
bp	Basepairs
BC	Breast Cancer
BRCA1	Breast Cancer 1, Early Onset
BRCA2	Breast Cancer 2, Early Onset
BRCC45	BRCA1/BRCA2-containing Complex Subunit 45
BRIP1	BRCA1 Interacting Protein C-terminal Helicase 1
BWA	Burrows-Wheeler Aligner
BWA-SW	Burrows-Wheeler Aligner, Smith-Waterman Alignment
CAPN14	Calpain 14
CASP8AP2	Caspase 8 Associated Protein 2
CDH1	Cadherin 1, type 1, E-cadherin (epithelial)
CHEK2	Checkpoint Kinase 2
CIMBA	Consortium of Investigators of modifiers of BRCA1 and BRCA2
CSP1	Regulator of Calcineurin 1
CTIP	CTBP-interacting Protein
dbSNP	Database of Single Nucleotide Polymorphisms
DHT	Dihydrotestosterone
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
DSB	Double-Strand Breaks
dsDNA	Double Stranded DNA
ER	Estrogen Receptor
FANCM	Fanconi Anemia, Complementation Group M
FGFR2	Fibroblast Growth Factor Receptor 2
GWAS	Genome wide association study
GWS	Genome Wide Search
HBC	Hereditary Breast Cancer
HER2	Human Epidermal Growth Factor Receptor 2
HR	Homologous Recombination

IGV	Integrative Genomics Viewer
IHA	Icelandic Heart Association
LD	Linkage Disequilibrium
LOF	Loss of Function
LOH	Loss of Heterozygosity
LSP1	Lymphocyte-specific Protein 1
MAF	Minor Allele Frequency
MAP3K1	Mitogen-activated Protein Kinase Kinase Kinase 1, E3 Ubiquitin Protein Ligase
MAQ	Mapping and Assembly with Quality Aligner
Mb	Megabase
MERIT40	Mediator of RAP80 Interactions and Targeting Subunit of 40 kDa
miRNA	microRNA
NBS1	Nibrin
NCBI	National Center for Biotechnology Information
ncRNA	Non-coding RNA
NGS	Next Generation Sequencing
NHEJ	Non-Homologous End Joining
nsSNP	Non-synonymous SNP
OCT	Organic Cation Transporter
PALB2	Partner and Localizer of BRCA2
PC	Prostate Cancer
PCR	Polymerase Chain Reaction
PR	Progesterone Receptor
PTEN	Phosphatase and Tensin Homolog
RAD50	DNA Repair Protein RAD50 Homolog
RAD51	DNA Repair Protein RAD51 Homolog
RAD51L1	DNA Repair Protein RAD51 Homolog 2
RefSeq	NCBI Reference Sequence Collection Database
RNA	Ribonucleic Acid
RR	Relative Risk
SLC22A16	Solute Carrier Family 22, member 16
SNP	Single Nucleotide Polymorphism
SPAST	Spastin
SRD5A2	Steroid-5-Alpha-Reductase, Alpha Polypeptide 2

ssDNA	Single Stranded DNA
sSNP	Synonymous SNP
STK11	Serine/Threonine Kinase 11
TOPBP1	Topoisomerase (DNA) II Binding Protein 1
TOX3	Tox High Mobility Group Box Family Member 3
TP53	Tumour Protein p53
TSPYL1	Testis-specific Y-encoded-like Protein 1
UCSC	University of California, Santa Cruz
UTR	Untranslated Region
VCF	Variant Call Format
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

1 Introduction

1.1 Breast cancer

Breast cancer (BC) is the most commonly diagnosed cancer type among women worldwide, accounting for 23% of all female cancers. In 2008 the number of new breast cancer cases was estimated to be approximately 1.38 million. Incidence rates vary between global regions and span a range from 19.3 to 89.9 per 100.000 women (Figure 1). They are generally higher in developed regions, where the number of new cases each year is normally over 70-80 per 100.000 women (1). Although incidence rates in developing countries are lower than in developed countries they seem to be increasing (2). Along with being the most common cancer type, BC is also the leading cause of cancer related death among women, both in developing and developed countries. However, due to better survival in developed regions, global variation in BC mortality rates spans a smaller range, from 6 to 19 per 100.000 women (1). Annually, the number of new cases in Iceland between 2006 and 2010 was 89.8 per 100.000 and the mortality rate was 16.4 per 100.000. There has been a significant increase in incidence rates when the numbers of recent years are compared to e.g. the year 1959 when the number of annual incidence was 36.1 per 100.000. Despite the increasing incidence of BC the mortality rates have not fluctuated as drastically, with the annual mortality in 1959 being 14.8 per 100.000 (3).

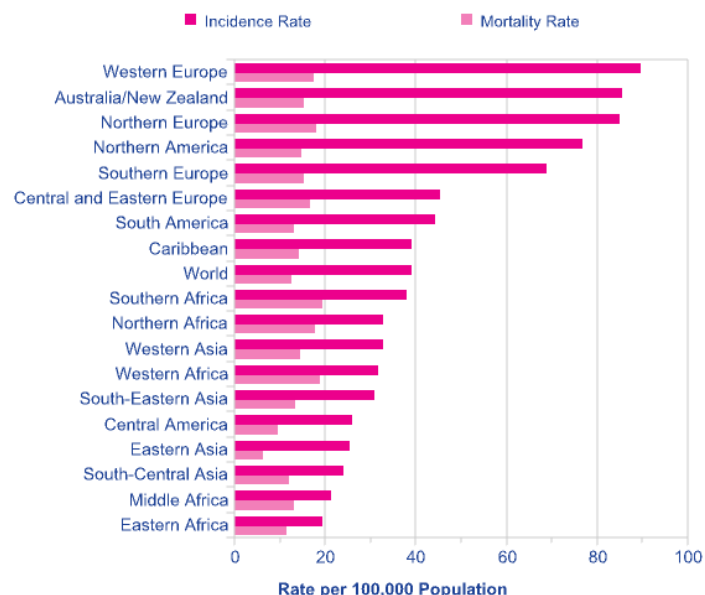


Figure 1. Worldwide incidence and mortality rates of breast cancer. There is a considerable variation in incidence rate between different regions of the globe. Generally they are higher in developed countries and lower in developing countries. Mortality rates also show a variation between regions but the range is not as drastic (Cancer Research UK, <http://www.cancerresearchuk.org/cancer-info/cancerstats/world/breast-cancer-world/>).

Rather than being a disease of a single specific origin, BC is a collection of tumours of different nature and with varying prognosis for patients. Therefore, not all BC tumours can be considered equal. The classification of breast tumours can be based on a number of different factors such as the origin of the tumour within the organ, i.e. whether it is lobular or ductal, the size of the tumour and histological grade. Tumours are also classified based on their expression of certain hormonal receptors; the estrogen receptor (ER), the progesterone receptor (PR) as well as the human epidermal growth factor receptor 2 (HER2) (4). The classification of BC tumours has an impact with regard to treatment options and overall survival of patients diagnosed with the disease.

Analyses of gene expression patterns in BC tumours have been used as a basis for yet another classification system that has been emerging in the last decade or so. This system divides tumours into five distinct molecular BC subtypes. These are the luminal A, luminal B, HER2, basal-like and normal-like subtypes. Each subtype displays a different expression pattern of a given group of genes (5, 6). There is an overlap, although incomplete, between tumours that are classified on the basis of HER2 and hormonal receptor expression and gene expression patterns. The HER2 subtype contains tumours that overexpress the HER2 gene, luminal A tumours are ER and PR positive and luminal B tumours generally express the ER and sometimes they are also PR and HER2 positive, although this varies. The basal subtype contains tumours that are also known as triple negative tumours, i.e. they do not express any of the three hormonal receptors. Finally, the normal subtype is comprised of tumours that bear a resemblance to normal tissue (5-8). Generally the best prognosis is for patients with tumours that belong to the luminal A subtype while the worst prognosis is for patients diagnosed with triple negative tumours of the basal subtype (6, 8).

The molecular pathology classification system of BC tumours is however a dynamic field that is constantly evolving. This is evident by the further classification and characterization of yet another subgroup of breast tumours; the claudin-low subtype of triple negative tumours, as well as the fact that the existence of the normal-like subtype of tumours is today being questioned by some researchers who claim that it is more likely to be an artefact than a real BC tumour subtype (4, 9, 10). Further refining and classifications of these distinct subgroups are likely to be made in future studies.

A number of factors have been identified as possible inducers of BC risk; both environmental and genetic factors. Obviously, gender matters since it has been estimated that out of every 150 cases only one BC case will occur in a male (11). Age is also an important factor as the risk of developing the disease increases with age. Other risk factors include ethnicity, alcohol consumption, low physical activity, obesity and exposure to sex hormones, both endogenous and exogenous (12). However, out of the recognized risk factors, the most potent one, besides age, is believed to be a family history of the disease (12, 13).

1.2 Familial breast cancer

Twin studies have shown that the BC risk for a monozygotic twin of a co-twin that has been diagnosed with breast cancer is high, which suggests that the contribution of genetic factors to the development of disease is important (14). First-degree relatives of BC patients are also under increased risk of developing the disease, with the elevation in risk being approximately two-fold. As the number of

affected relatives rises, the risk increases further. E.g. the overall lifetime risk for a western woman with no family history of BC is 7.8%, compared to a lifetime risk of 21.1% for those women who have two affected relatives. For women with affected relatives that are diagnosed at an early age, the risk is increased even further (12). However, most BC cases are sporadic and are estimated to account for approximately 75-80% of all cases (13, 15). A case is classified as being sporadic when the diagnosed individual has no family history of the disease. In 20-25% of cases individuals have a family history of the disease, with one or more family members affected. Approximately one fourth of those familial cases can be further defined as hereditary BC (HBC). Within HBC families a multiple number of cases cluster together, the disease has a dominant appearance and is mainly believed to be caused by germline mutations (12). Around 25-40% of these HBC families have been implicated with germline mutations in known cancer genes, with mutations within *BRCA1* and *BRCA2* being the most common. The remaining HBC families are generally referred to as BRCAx families (16, 17).

Three distinct classes of BC susceptibility alleles have been linked to the observed increase in risk for HBC. The first class includes high penetrance genes that harbour rare variants that confer a high increase in BC risk. The second class harbours moderate penetrance genes, also with rare variants that confer a moderate increase in risk. The third allele class is that of common but low penetrance variants that confer a small increase in BC risk (18).

1.3 High penetrance genes

The identification of two tumour suppressor genes in the 1990's, *BRCA1* and *BRCA2*, was a major breakthrough in the field of BC research. Of the genes that have been linked with BC predisposition, these are the ones that are most well known and studied. Both genes frequently display a loss of heterozygosity (LOH), characterised by the loss of the wild-type allele in tumours, and they have been shown to contribute to an increase in risk for not only breast cancer, but e.g. ovarian and prostate cancer as well. Other high-risk BC susceptibility genes that have been identified are e.g. *TP53*, *PTEN*, *STK11* and *CDH1*. The relative risk for carriers of pathogenic mutations within any of these genes, compared to non-carriers, ranges from 5 to over 20. But although these mutations confer high risk, they are quite rare in the general population and therefore each mutation only explains a small fraction of the increased BC risk (18).

1.3.1 *BRCA1* and *BRCA2*

The breast cancer 1, early onset gene, or *BRCA1*, is a large gene, containing 22 exons, and is located on chromosome 17q (19). It codes for a protein whose function has been implicated with the maintenance of genomic integrity through the regulation of cell-cycle progression and DNA repair via non-homologous end joining (NHEJ) and homologous recombination (HR). It's involvement in the HR pathway is believed to be as a signal mediator to elicit a response from effector molecules that initiate the mending of the damaged DNA (20).

The breast cancer 2, early onset gene, *BRCA2*, is located on chromosome 13q. This gene contains 27 exons, coding for a protein even larger than *BRCA1* (21). The encoded protein participates in DNA repair of double-strand breaks (DSBs) through the HR pathway, where it functions as an effector

molecule that initiates the repair process (20). A schematic representation of the BRCA genes can be seen in figure 2.

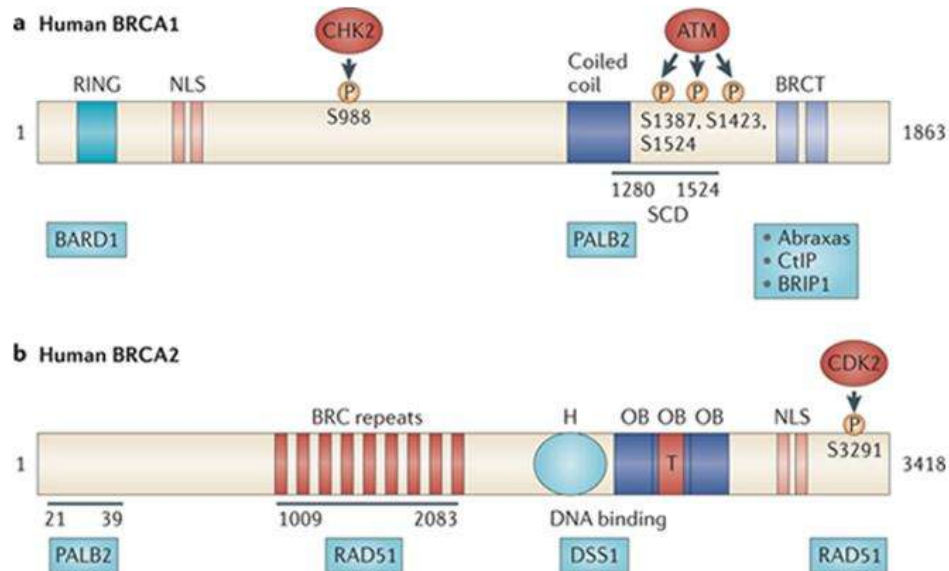


Figure 2. A schematic representation of *BRCA1* and *BRCA2*. The functional domains of the *BRCA1* (a) and *BRCA2* (b) proteins. (a) The N-terminus of *BRCA1* contains a RING domain that associates with BRCA1-associated RING domain protein 1 (BARD1) as well as a nuclear localization sequence (NLS). CHK2 phosphorylates a serine residue in position 988 in the protein. The C-terminus of *BRCA1* contains a coiled-coil domain that associates with the PALB2 protein, a SQ/TQ cluster domain (SCD) that harbours ATM phosphorylation sites and a BRCT domain associating with 3 proteins; Abraxas, CtIP and BRIP1. (b) The N-terminus of *BRCA2* has been shown to bind to PALB2. The central region of the protein harbours 8 BRC repeats that bind RAD51, a helical DNA binding domain and 3 oligonucleotide binding (OB) - folds surrounding a tower domain (T). The C-terminus of *BRCA2* contains a NLS and a CDK2 phosphorylation site that binds to RAD51 (adapted from Roy, Chun & Powell, 2011).

Pathogenic mutations within these two genes have been shown to incur an approximate 10- to 20-fold increase in relative risk (RR) BC risk for mutation carriers. These mutation carriers are diagnosed with the disease at a younger age compared to women with sporadic BC (18, 22). The disease causing mutations are primarily protein truncating variants that cause the encoded proteins to become inactive and thus disrupt their biological functions (18).

A considerable variation in the frequency of these high-risk mutations has been observed when individual populations are compared to each other. In some populations, such as in the Ashkenazi Jewish and the Icelandic populations, the prevalence of certain founder mutations is much higher than in the general Caucasian population. In most populations, mutations in *BRCA1* are generally more prevalent than mutations in *BRCA2*; with their frequency being 1.5- to 2-fold higher. This is not the case in the Icelandic population where *BRCA2* explains a higher proportion of BC cases than *BRCA1* (23).

Pathogenic germline variants within *BRCA1* and *BRCA2* generally result in a high overall lifetime risk of BC for carriers. It has however been observed that there is a variation in risk between mutation carriers (24). *BRCA1* and *BRCA2* mutation carriers have been studied with regard to possible factors that could explain this observed variation in risk between individuals. This is for example the collective aim of international research groups that together form the Consortium of Investigators of modifiers of *BRCA1* and *BRCA2* (CIMBA, <http://ccge.medschl.cam.ac.uk/consortia/cimba/index.html>). Genome wide association studies (GWAS) on unselected BC cases have identified alleles that have been associated with a low increase in BC risk. These alleles have been genotyped in *BRCA1* and *BRCA2* mutation carriers in studies that have for example revealed associations between an increase in risk for *BRCA2* mutation carriers and SNPs in *FGFR2*, *CSP1*, *RAD51*, *MAP3K1*, *TOX3* and 2q35. The alleles in *TOX3* and 2q35 have also been associated with risk in *BRCA1* carriers (25).

Studies on genes that have a role to play within BRCA-related pathways have also revealed possible associations between certain haplotypes and modification of risk in *BRCA1* and *BRCA2* mutation carriers. These genes were considered good candidates because they interact, directly and indirectly, with *BRCA1* and *BRCA2*. For *BRCA1*-associated BC there is an indication for the modifying effect of haplotypes at *ATM*, *BRCC45*, *BRIP1*, *CTIP*, *MERIT40*, *NBS1*, *RAD50* and *TOPBP1*. For *BRCA2*-associated BC, there is currently available evidence for modifying effects of haplotypes at *BARD1* and *RAD51* (26).

1.3.2 Other high penetrance genes

Other germline mutations that have been affiliated with a high increase in BC risk have been identified within genes that are connected to certain inherited syndromes that predispose their carriers to cancer. The *TP53* gene is associated with Li-Fraumeni syndrome and pathogenic mutations within this gene are linked with a high increase in both breast and ovarian cancer risk. The encoded protein has been shown to partake in cell-cycle control (27-29). Cowden syndrome is an example of a phenotype associated with germline mutations in the *PTEN* gene that has been shown to significantly increase the risk of BC (30). The gene *STK11* has similarly been associated with a cancer predisposing syndrome called Peutz-Jeghers. Patients suffering from Peutz-Jeghers are at an increased risk of developing BC (31).

A number of genetic linkage studies have not been successful in identifying new high penetrance genes. While this does not exclude the possibility that further genes of that class exist it strongly suggests that most of the remaining portion of unexplained hereditary BC risk is due to disease causing variants of a different nature, e.g. variants within genes that confer a moderate or low increase in risk. If several such variants of that nature would accumulate within an individual they could be viewed as powerful moderators of BC risk (32).

1.3.3 Moderate penetrance genes

A few genes have been implicated with BC susceptibility that confer a moderate increase in BC risk. Among these genes, that have been identified through mutation screening in candidate gene studies, are *ATM* (33), *BRIP1* (34), *PALB2* (35, 36) and *CHEK2* (37). The pathogenic mutations within the

genes that have been identified in these studies share certain characteristics with pathogenic variants within the high penetrance genes in that they have turned out to be quite rare and uncommon in the general population and most of them are loss of function (LOF) variants, i.e. they result in premature protein truncation. They do however differ with regard to the increased risk they confer; moderate-risk variants increase the RR 2- to 4-fold compared to the 5- to 20-fold observed for the high-risk variants (25). The encoded proteins of *ATM*, *BRIP1*, *PALB2* and *CHEK2* are all associated with DNA repair pathways. Variants within moderate penetrance genes have been estimated to account for approximately 5% of the hereditary BC risk (18, 25).

1.3.4 Low penetrance alleles

Since the susceptibility alleles that associate with a high or moderate increase in BC risk explain less than half of HBC cases it has been suggested that the majority of the remaining HBC and familial BC risk might be explained by a polygenic model. In recent years, GWAS have identified several genomic regions that seem to harbour yet another group of risk variants. The variants of this class confer only a small increase in risk, defined by an estimated RR below 2, and they are predominantly common SNPs that are carried by a high proportion of the general population (16). Some of these common low-risk SNPs have been shown to be located in regions either within or in close proximity to known genes such as *FGFR2*, *TOX3/TNRC9*, *MAP3K1*, *LSP51* and *RAD51L1*. Others, such as SNPs in regions at chromosomes 2q35 and 8q24, are positioned far away from the nearest known genes (25, 38-40).

Notably, some of these genes are involved in cellular pathways that involve the regulation of cell growth and signalling which differs from the pathways that previously reported BC susceptibility alleles participate in; which primarily involve the repair process of damaged DNA (38). This might suggest a different mechanism of action for the low risk variants, where their effect on BC risk could be mediated through the activation of oncogenes, e.g. genes that promote cell growth (18).

A number of chromosomal regions within the human genome have been proposed to harbour possible BC susceptibility alleles spread across all classes of penetrance type. Figure 3 gives an overview of these loci with regard to chromosomal positioning.

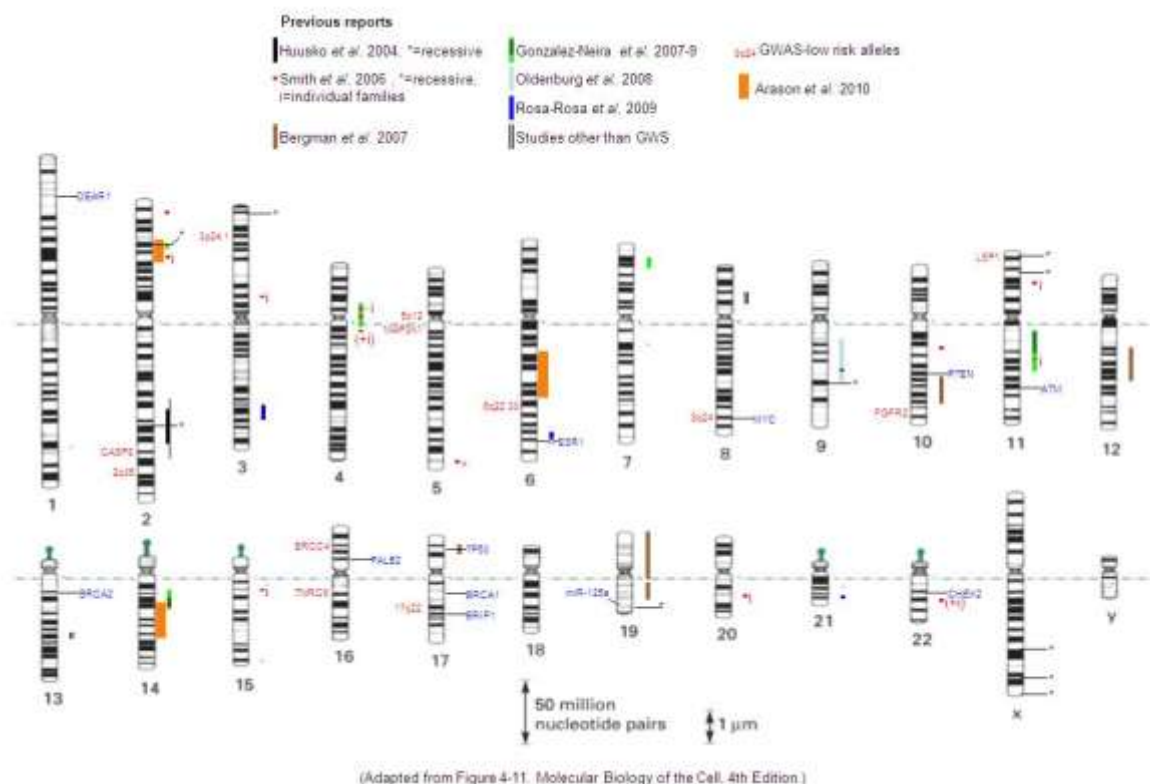


Figure 3. Chromosomal positions of possible BC susceptibility alleles. This figure gives an overview of the genomic locations of loci that have been identified as possible susceptibility loci with regard to BC. This figure was adapted in 2011 by Adalgeir Arason from Figure 4-11 of Molecular Biology of the Cell, 4th edition.

1.4 Hereditary breast cancer in Iceland

HBC in Icelandic families has been of great interest to researchers. Families that belong to founder populations, such as the Icelandic population, are valuable due to the fact that genetic heterogeneity of such populations might be reduced (41, 42). An example of this is the involvement of researchers studying the Icelandic population in the discovery of the *BRCA2* gene (21, 43) and more recently the identification of a link between the *BRIP1* gene and increased susceptibility to ovarian cancer (44). Studies on Icelandic high-risk BC families can therefore have an impact and be helpful in the search and identification of new BC susceptibility alleles.

In the early 1990's it was reported that markers on chromosome 17q showed linkage to breast-ovarian cancer in Icelandic families (45). This report was in agreement with previously published reports of the BC disease linkage to a gene located in this region (46, 47), that had been assigned the name *BRCA1* and was later identified by positional cloning (19, 48). The search for pathogenic mutations within the *BRCA1* gene in Icelandic high-risk BC families led to the identification of a recurrent splice-site mutation in exon 17 (49).

The *BRCA2* gene was localised to chromosome 13q in 1994 (43) and subsequently identified in 1995 (21). Mutational analysis of the *BRCA2* gene revealed a recurrent Icelandic founder mutation in Icelandic families; a deletion of 5 nucleotides, starting at nucleotide 999, in exon 9 (999del5) and the

mutation carrying families all segregate a common *BRCA2* haplotype (50, 51). These two mutations are the only *BRCA1/2* disease causing mutations published thus far in the Icelandic population. The 999del5 mutation in *BRCA2* is much more prevalent compared to the *BRCA1* mutation and is believed to explain increased risk in half of the Icelandic HBC families (50, 52, 53). With regard to other Icelandic HBC families, it seems unlikely that other mutations will be found in these two genes (54). The focus of researchers has therefore shifted towards other possible genes that might be linked to an increase in HBC risk.

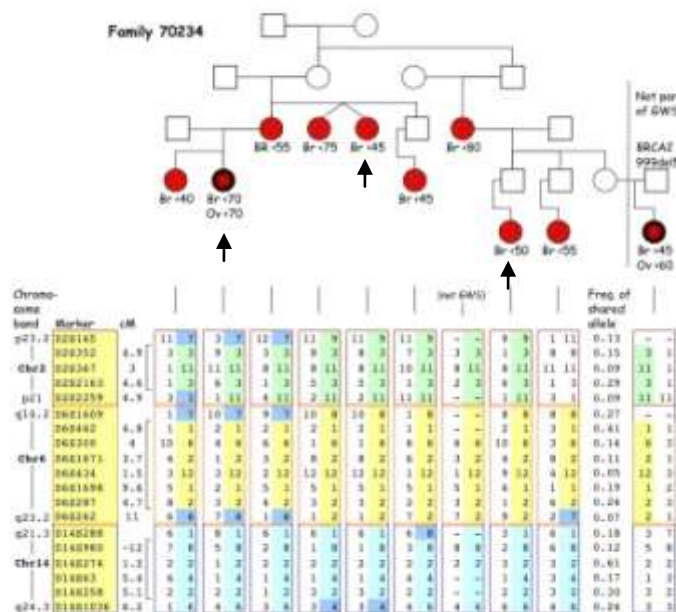


Figure 4. Cosegregation of haplotypes at three chromosomal regions in family 70234. Information about approximate age at diagnosis of cancer is shown below pedigree symbols (Br for breast and Ov for ovarian). One woman inherited a 999del5 BRCA2 mutation from her father, not otherwise blood-related to this family. She was not included in the LOD score calculations (Arason et al., 2010). The arrows indicate the cases selected for the targeted resequencing.

1.5 Identifying new breast cancer susceptibility alleles

Since known mutations in already identified BC genes are thought to account for less than half of the HBC families, the search continues for new BC susceptibility genes.

1.5.1 Genetic linkage studies

Employing the genetic linkage method in HBC families where BC cases cluster together can be useful. The most obvious example of the success is the identification of *BRCA1* and *BRCA2* (21, 46). It has been proposed that other BC susceptibility genes, that have yet to be identified, are not likely to confer a high increase in risk, but rather that the emerging landscape is likely to involve a large number of susceptibility alleles with each allele only conferring a small or moderate increase in the risk of developing BC (57). If this hypothesis turns out to be true, it is unlikely that linkage studies will be a successful method in identifying those genes. It can however not be declared that all high penetrant genes have been identified. The loci harbouring such genes could be identified by linkage analysis in large HBC families (42).

1.5.2 Genome Wide Association Studies (GWAS)

When it became increasingly obvious that genetic linkage studies would not be likely to yield further successful results with regard to the identification of new BC genes other measures had to be made. The strategy behind GWAS revolves around the identification of genetic markers that associate with risk by typing SNPs that are relatively common. The correlation among SNPs that are in linkage disequilibrium (LD) with the disease in question is used to identify the risk factor. Large international consortiums have in part driven this area of BC research forward in recent years. By utilising combined datasets of a large number of cases and controls researchers have been able to identify genomic associations of loci to BC where each confers a small increase in risk, as has been discussed (18, 25).

1.5.3 The candidate gene approach

Candidate gene approaches have been useful in identifying variants that confer a moderate or low increase in BC risk (25). This methodology involves the investigation of handpicked genes that can be of interest for a number of reasons. For a researcher trying to identify new BC susceptibility alleles, genes that function within certain pathways may be of a special interest. E.g. the genes that code for the proteins that participate in the DNA damage and repair pathways alongside *BRCA1* and *BRCA2*. Figure 5 gives an overview of the genes that participate in functional networks connected to *BRCA1* and *BRCA2* and can be viewed as candidate genes with regard to BC susceptibility. This method has been shown to be successful on a number of occasions, such as in identifying a number BC susceptibility genes that confer a moderate increase in risk; *CHEK2*, *ATM*, *PALB2* and *BRIP1* (58).

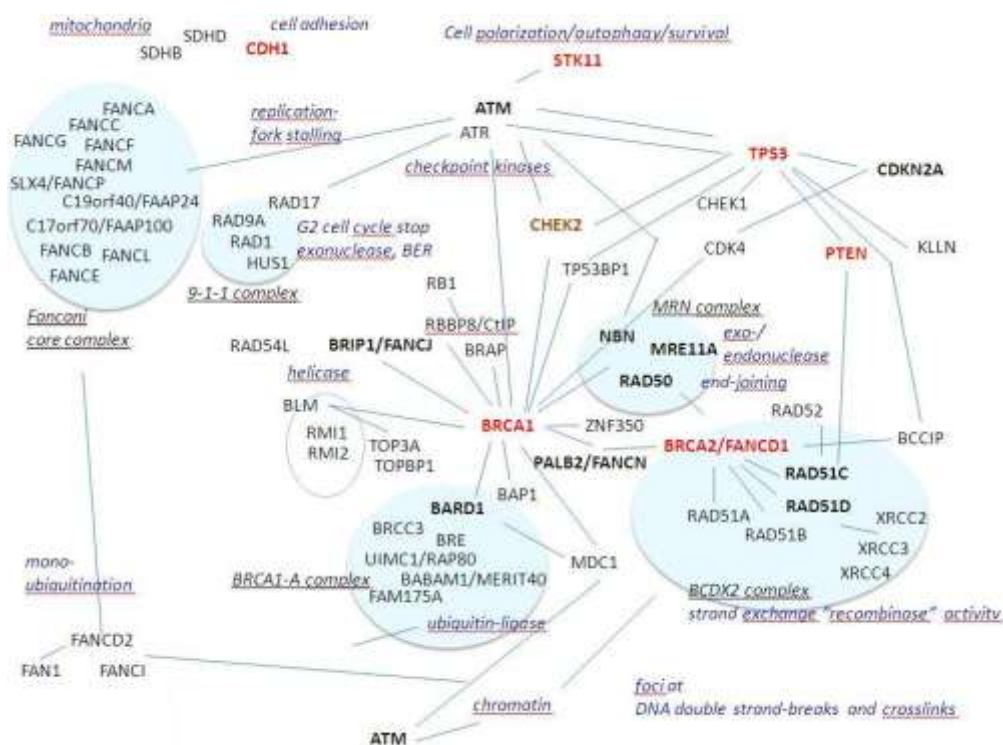


Figure 5. An overview of genes in BRCA related functional pathways. The genes within these pathways encode proteins that may be of interest to scientists studying BC candidate genes. This figure was adapted in 2013 from a slide from Ake Borg.

1.6 Next Generation Sequencing

In the past decade, new sequencing technologies have emerged that enable researchers to dive into individual genomes at great depth at sequence-level resolution in the search for disease causing variants. These new technologies have been coined as “Next Generation Sequencing” (NGS) and include the sequencing of whole genomes, entire exomes or more targeted applications such as the sequencing of specific chromosomal regions and genes (58-60).

Most costly is the sequencing of whole genomes. However, whole genome sequencing (WGS) provides the most extensive information of the genome since it includes the sequencing of all protein coding regions as well as non-coding regions. Covering the entire genome allows the identification of not only the variants that alter the protein coding sequences, and the regions near the exons of genes, but also the ones that affect other genomic regions, e.g. enhancers, intronic and intergenic regions. But the increased sequencing capacity not only covers the fact that the entire DNA sequence itself is brought into light. It also opens up the potential of discovering chromosomal rearrangements, which is beneficial when studying the landscape of a cancer genome (59).

Targeted sequence capture of specific genomic regions followed by NGS is becoming more readily available and as such has been shown to be an attractive option for studies revolving around specific pre-defined regions of the human genome. By targeting sub regions of the genome, the data generation is less costly compared to WGS (61). Utilizing this approach, any section of the genome can be targeted by oligonucleotides that are designed to retrieve the desired sequences from a pond

of genomic DNA. The DNA is then sequenced via any of the NGS platforms. This is an attractive option for researchers who have a reason to believe that risk variants lie within a specific region of the genome or within specific genes (62, 63).

There are a variety of NGS platforms available today. Increasing competition between the companies on the market today is resulting in improvements with regard to sequence output, read length and cost reduction as mentioned above. Each manufacturer uses its own specialized chemistry to perform the sequencing reaction. But although the details for each platform may differ from the other, the different approaches share certain traits in the sequencing preparations. These initial steps involve turning double stranded DNA (dsDNA) to single stranded DNA (ssDNA) followed by fragmentation of the ssDNA. Synthetic adapters are then attached to the ssDNA fragments to create a DNA library to be sequenced. Each DNA fragment is then amplified to generate multiple copies of each fragment. The amplification step is supposed to ensure that there are enough signals for the sequencing reaction to determine the DNA sequences via the optical system of the sequencing instruments. The sequencing of the amplified DNA library then allows for millions of sequencing reaction to happen in parallel (64).

1.6.1 Alignment and variant detection

Due to the digital nature of NGS and the amount of data generated, bioinformatics knowledge has become increasingly more important in genomic studies. Computational tools for the analysis of data from the NGS systems have been developing rapidly in recent years. These computational tools include many different software programs that serve a specific purpose in the data analysis workflow, spanning different areas such as the alignment of the sequencing reads to a reference genome, visualization of the aligned reads, variant calling, annotation of called variants and predictions of their functional effect with regard to their genomic positioning (59). Among the programs that have been developed are the Burrows-Wheeler Aligner (BWA) (65), Bowtie (66) and the Mapping and Assembly with Quality aligner (MAQ) (67) for the alignment procedure. Tablet (68), CIRCOS (69) and the Integrative Genomics Viewer (IGV) (70) can be used for visualization. Variant calling can be performed using SamTools (71) and VarScan (72) as an example. Finally, ANNOVAR (73) and SnpEff (74) can be used to annotate the identified variants. The programs mentioned above are all open source programs and most of them require computational knowledge to a certain degree, e.g. knowledge of the Linux/Unix operating systems. A host of commercial programs are also becoming increasingly more available to those who prefer such programs for their data analysis. These programs have the benefit of being considerably more user friendly but with the drawback that their use is dependent upon the user purchasing access. It therefore may increase the cost of the entire process. Examples of companies that offer such commercial tools are DNAnexus (<https://dnanexus.com/>), Geospiza, Inc (<http://www.geospiza.com/>) and Ingenuity Systems (<http://www.ingenuity.com/>). Some companies that provide the sequencing service, i.e. Roche/454, also provide their customers with a workflow containing alignment and variant calling using their in-house programs, e.g. the GS reference mapper software (<http://454.com/products/analysis-software/index.asp>).

1.6.2 Prioritizing identified variants

As new sequencing technologies alter the stage of genomic research with regard to data acquisition then a whole set of new challenges await those that take on the task of analysing the wealth of data being generated. Analysing these datasets can be daunting due to the overwhelming amount of identified variants within individual genomes. Therefore, the pursuit of the specific causal variants and the genes that harbour them has to be well structured. Deploying a sophisticated variant prioritization strategy can thus be important for the process of identifying a new BC susceptibility gene (75). The various types of functional classes of genetic variants can be seen in figure 6.

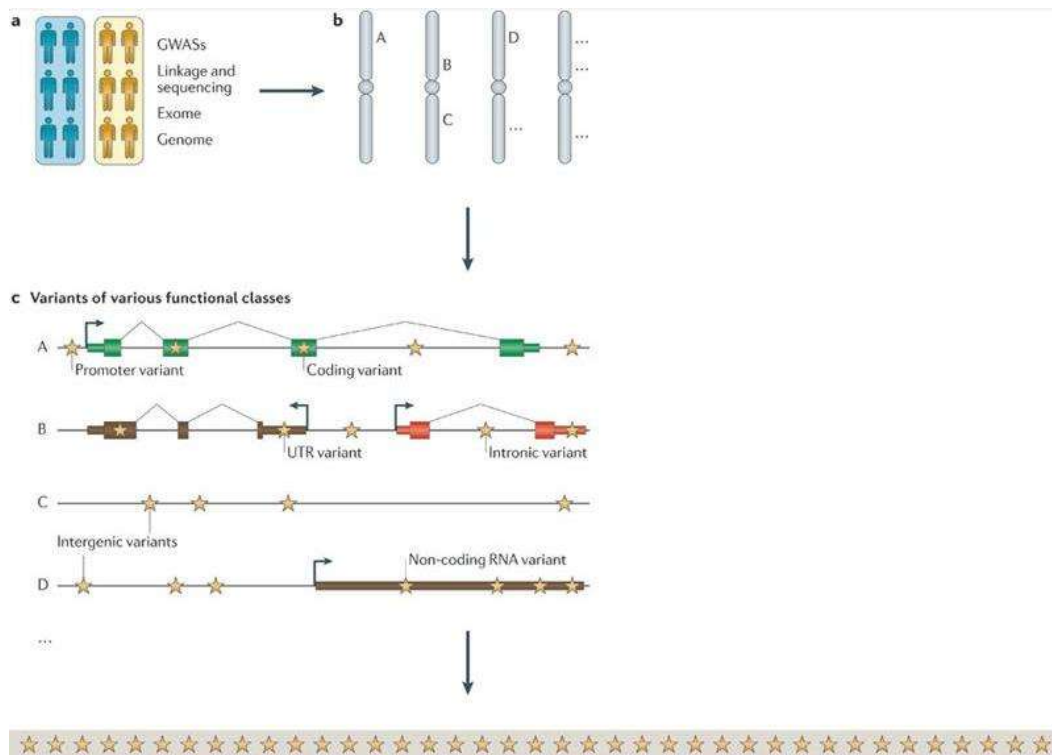


Figure 6. The various types of functional variants within the human genome. Genetic approaches (a) e.g. linkage analysis followed by re-sequencing can be used to identify regions of interest (b) and the different types of genetic variants within these regions. (c) It depends on the sequencing approach used which variant types can be identified, e.g. whole genome sequencing allows for the identification of all genetic variants while exome sequencing excludes most intronic and intergenic regions (adapted from Cooper & Shendure, 2011).

Sequencing experiments that target the protein coding regions of genes are performed under the assumption that the causative variants are embedded within the targeted regions. Usually, in those kinds of experiments, the exon-intron boundaries are captured as well to include possible splice-site variants that might disrupt splicing. Given the aforementioned assumption, and the fact that most of the previously identified BC susceptibility genes that confer a high or moderate increase in risk harbour protein truncating variants, a reasonable strategy for variant prioritization would be to focus on LOF variants as a first step in variant analysis (76, 77). The variant types that are classified as LOF variants are nonsense mutations, coding insertion or deletions that cause a frameshift in protein

translation and splice-site variants of any kind. The next steps would involve an analysis of non-synonymous SNPs (nsSNPs). The nsSNPs are coding variants that cause the substitution of one specific amino acid residue to another, in contrast to synonymous SNPs (sSNPs) that do not result in a change in the protein sequence. Predicting the effect of nsSNPs on protein function and structure is not easy. Therefore, distinguishing between a set of rare variants of this type to identify a causal one is very challenging. It has however been suggested that this variant type could be an integral part in polygenic mechanisms that increase an individuals susceptibility to BC, highlighting the importance of this area of the analysis of NGS data (16, 18).

Current methods that have been developed to evaluate the possible deleteriousness of nsSNPs share certain characteristics in their predictions. For example, variation within evolutionary conserved regions is predicted to be more likely to have a deleterious effect compared to changes in less conserved regions. Some methods also take into account the different biochemical properties of different amino acids where radical changes are thought to be more likely to cause a harmful effect on protein function (75). Two of the most popular functional prediction tools are SIFT, which bases its predictions on the assumption that essential amino acid residues are likely to be evolutionary conserved (78) and Polyphen/Polyphen2, which bases its predictions on a similar assumption but also takes into account the structure of the amino acid residues (79). These tools are open-source and can be acquired online, without charge.

2 Aims

This study is based on a previous study performed in our laboratory that identified linkage of BC to three chromosomal regions in an Icelandic high-risk BRCAx family. The candidate regions are located at chromosomes 2p, 6q and 14q. The original positions on chromosomes 2p and 14q were modified due to observed overlap with previously published candidate positions. The original regions combined contain a total number of 554 protein coding genes. The modified regions contain 274 protein coding genes. This project involved the analysis of raw NGS data of the 274 genes within the modified regions. The aim was to identify one or more new mutations that increase BC risk.

The specific tasks were as follows:

1. Since this was the first time next generation sequencing was used to generate data for a project performed at our lab, one of the main tasks of this project was to design and set up a variant analysis pipeline to be used for future projects of similar or greater magnitude.
2. Analyse data from the resequencing of 274 genes, in DNA samples from selected BC cases from one HBC family, and look for possible causal mutations within that might explain the observed increase in BC risk in this family.
3. Candidate variants that are identified as possible causal mutations will be confirmed in family 70234 to be genuine germline variants. They will then be screened for in other high-risk BRCAx families as well as in a set of controls and unselected BC cases.
4. If a candidate variant is confirmed as a causal mutation, the gene that harbours it will be sequenced, using the Sanger sequencing method, in a set of selected high-risk BC families from Iceland, Sweden and Finland.

3 Material and methods

3.1 Sample selection

This project is a part of a larger ongoing study approved by the Icelandic Data Protection Authority (2001/523 and 2002/463) as well as the National Bioethics Committee of Iceland (99/051, 99/051_FS1 and 11-105-S1). All samples used in this study had been previously collected and the DNA isolated by the staff at the Laboratory of Cell biology, Department of Pathology at Landspítali University Hospital except for the samples that originate from the Icelandic Heart Association (IHA).

3.1.1 Samples selected for targeted resequencing

The project was based on NGS data from four DNA samples. All four samples, were from three breast cancer patients (two isolated from blood samples and two from tumour tissue), of a family which showed strong linkage between breast cancer and delimited regions at chromosomes 2p, 6q and 14q (55). A pedigree of family 70234 is shown in figure 4 and the BC patients, from whom the DNA samples originate, are marked in the pedigree. All three cases shared the segregating disease-associated haplotypes at chromosome 2p, 6q and 14q.

3.1.2 Samples used for the screening of candidate variants

The allele frequency of candidate variants was estimated in DNA samples of three different groups; controls, unselected BC patients and familial BC cases. The control samples originated from four different samplings. The first sampling, performed at the Laboratory of Cell biology, Department of Pathology at Landspítali University Hospital, was from a group consisting of healthy female and male blood donors with no family history of cancer. The second sampling was performed at the Icelandic Blood Bank and the IHA from an unselected group of controls. The third control sampling was performed at the Laboratory of Cell biology, Department of Pathology at Landspítali University Hospital, also from an unselected group of individuals. The fourth control group originates from the AGES-Reykjavik study at the IHA. The DNA samples from BC patients were collected from all Icelandic BC patients who were diagnosed in the period of 1987 to 2009 and agreed to participate in the BC research study, “A search for additional breast cancer genes”, conducted by the BC study group at Landspítali (presently responsible researches are: Aðalgeir Arason, Bjarni A Agnarsson, Óskar Þ. Jóhannsson and Rosa B Barkardóttir). The familial BC samples were DNA samples selected from affected members of 37 HBC families, not accounted for by recurrent Icelandic *BRCA1* or *BRCA2* mutations. The youngest case with available DNA samples was selected from each branch of the family if DNA samples were available.

3.2 Data generation

3.2.1 Targeted sequence capture

The sequence capture was done in collaboration with Elisabet Guðmundsdóttir at NimbleGen, the Icelandic daughter company of Roche. The 385K Roche NimbleGen Inc arrays were used for the sequence capture. The design of the microarray capture probes was performed by Haukur

Gunnarsson MSc. and Rosa B Barkardottir, with the assistance of the bioinformatics team at NimbleGen. The microarray capture probes used were intended to capture the exons and exon/intron boundaries of the genes within the regions of interest, as well as 400 bp upstream and downstream of the genes in question. The estimated total length of the sequences captured was one megabase (Mb). The positions of the sequences targeted on each chromosome can be seen in table 1.

Table 1. The targeted boundaries in bp based on the hg19 reference assembly.

Chromosome	Positions
2p	28.603.574
	36.395.113
6q	90.595.826
	125.080.605
14q	34.459.447
	47.247.975

3.2.2 454 resequencing of the targeted regions

The NimbleGen sequence capture array used in this study is optimized for the 454 sequencing platform. The sequencing service was bought from Matís and performed by the staff of Matís under the supervision of Ólafur Friðjónsson. Matís uses a 454 sequencing instrument and the Titanium GS FLX chemistry from Roche. In brief, purified DNA samples of interest are hybridised onto DNA capture beads. Each capture bead contains a unique single-stranded DNA library fragment. The DNA libraries are then amplified in an emulsion PCR within a microreactor. The beads are then loaded onto a 454 PicoTiterPlate™ containing 1.6 million wells. The wells are loaded with the library beads, one bead within each well, and a layer of enzyme beads (containing sulfurylase and luciferase). A loaded PicoTiterPlate is placed into the sequencing instrument and the sequencing reaction is initiated (80).

3.3 Data analysis

3.3.1 Alignment

The sequencing reads were aligned to the human genome reference sequence, hg18, using the GS reference mapper software, that accompanies the 454 instrument at Matís. The Tablet software (68) was used for visualisation and manual inspection of the aligned reads. For comparison, alignment of the sequencing reads was also performed in-house, at our lab at the Department of Pathology, using two open-source software programs; BWA-SW (65) and Bowtie (66). The hg19 version of the human genome reference was used for these alignments. The IGV software (70) was used for visualisation of the aligned reads from BWA-SW and Bowtie.

3.3.2 Variant calling

Variant calling on the GS reference mapper alignment was done using the same software. The coordinates of the variants identified were then converted from hg18 to hg19, using the LiftOver tool from the UCSC Genome Browser (81), to facilitate further downstream analysis of the data. SamTools (71) was used for variant calling on the alignments from BWA-SW and Bowtie.

3.3.3 Variant annotation

Annotation of identified variants within the sequenced regions was performed using the open-source ANNOVAR software (73), providing them with genomic context. Using ANNOVAR we were able to generate gene-based, filter-based and region-based annotation of the variants. First we performed gene-based annotation to identify the location of the variants within the genome. This was done by annotating against the NCBI reference sequence collection database (RefSeq) (82), which results in two output files being generated. The first file contains annotations for all variants (table 2), e.g. whether a variant is an exonic or intronic variant, what genes they reside within or if they are intergenic etc.

Table 2. A list of possible values for the annotation of all variants with ANNOVAR.

Annotation	Explanation
Exonic	variant overlaps a coding exon
Splicing	variant is within 2-bp of a splicing junction
ncRNA	variant overlaps a transcript without a coding annotation in the gene definition
UTR5	variant overlaps a 5' untranslated region
UTR3	variant overlaps a 3' untranslated region
Intronic	variant overlaps an intron
Upstream	variant overlaps 1-kb region upstream of transcription start site
Downstream	variant overlaps 1-kb region downstream of transcription end site
intergenic	variant is in intergenic region

The second output file contains the predicted effect of exonic variants, e.g. whether they result in frameshift or amino acid changes etc (table 3). ANNOVAR was then used for filter-based annotation to identify which variants had been previously reported to dbSNP build 135, and which of these reported variants had a frequency in the 1000 genomes dataset (83).

Table 3. A list of possible values for exonic variant annotations with ANNOVAR.

Annotation	Explanation
Frameshift insertion	a nucleotide insertion that causes frameshift changes in the protein coding sequence
Frameshift deletion	a nucleotide deletion that causes frameshift changes in the protein coding sequence
Stop-gain	a variant that leads the creation of a stop codon at the variant site
Stop-loss	a variant that leads to the elimination of a stop codon at the variant site
Nonframeshift insertion	an insertion of 3 or multiples of 3 nucleotides that do not cause frameshift changes
Nonframeshift deletion	a deletion of 3 or multiples of 3 nucleotides that do not cause frameshift changes
Nonsynonymous SNP	a single nucleotide change that causes an amino acid change
Synonymous SNP	a single nucleotide change that does not cause an amino acid change

Region-based annotation was performed for the purposes of identifying variants within specific genomic regions. In our case we wanted to know which of our variants were within predicted miRNA target sites. ANNOVAR was used to run this query against miRNA target site predictions from the TargetScan database (84).

Microsoft Office Excel 2007 and VarSifter (85) were used for the viewing and further filtering of the annotated variants.

3.3.4 Identification of candidate variants

Variants were grouped together and evaluated on the basis of their functional effect (see table 2 and 3). SIFT (78) and Polyphen2 (79) were used for the prediction of the possible deleteriousness of nsSNPs.

3.4 Genotyping of candidate variants

Genotyping was performed for variants deemed to be possible candidate causal variants. Estimates of their allele frequency within the sample groups were generated by performing fragment analysis for indels (insertions and deletions) and SNP genotyping for SNPs. Validation of the candidate mutations within family 70234 was performed by screening for them in members of the family.

3.4.1 Fragment analysis

For candidate indels, fragment analysis was performed to screen for the variants within the sample material. Forward and reverse primers were designed and then purchased from Eurofins through their website (<http://www.eurofins.com/en.aspx>). For the design, the DNA sequence of the gene in question was retrieved from the UCSC Genome Browser and imported into the Sequencher 5.0 sequence analysis software from Gene Codes. Possible primer sequences were located manually and then blasted, using the NCBI Primer-Blast webpage to evaluate their specificity. We also checked for possible self-complementarity of candidate primers by utilising the Oligo Calc: Oligonucleotide Properties Calculator (<http://www.basic.northwestern.edu/biotools/oligocalc.html>). In each primer set, one primer was tagged with a FAM fluorescent dye on the 5' end. Table 4 lists all primers used in fragment analyses in this project.

Table 4. A list of primer sequences used for fragment analysis of candidate variants. Also listed are the genes that harbour the variants, the orientation of the primers (forward (F) and reverse (R)) and the product length of the amplified fragments.

Target		Sequences	Product length (bp)
<i>TSPYL1</i>	F	5'-FAM-GAG GTG AAG ACA GGA AAG TG-3'	175
	R	5'-CCT CCA ATC TAT CCT CCT CC-3'	
<i>BAZ1A</i>	F	5'-FAM-TCA CGA ATC TGA CTT TGT CC-3'	173
	R	5'-TCC TAC ATT CTC CTG AGT GC-3'	
<i>FANCM</i>	F	5'-GCA TTG ATA AGA AAT CAG TTT TCC AG-3'	151
	R	5'-FAM-GCA TCT TCT TCA GAA AGT TCT-3'	

Genomic DNA was used to amplify each fragment, which length was defined by the number of base pairs (bp) between the forward and reverse primers. 10 ng of DNA was used per reaction. Amplification was performed in 10 µl total reaction volume containing 7.1 µl H₂O, 1 µl 10X Taq buffer with (NH₄)₂SO₄, 0.8 µl 25mM MgCl₂, 0.64 µl 10 mM dNTP, 0.2 µl of each primer from a 20 pmol solution and 0.06 µl of 5 U/µl Taq DNA polymerase. The PCR programme reaction was as follows: 1 cycle of initial denaturation at 94°C for 3 minutes followed by 35 cycles of denaturation at 94°C for 45 seconds, annealing at 55°C for 45 seconds and elongation at 72°C for 45 seconds. Finally there was a final elongation step at 72°C for 10 minutes. The PCR apparatus used for these reactions was the Applied Biosystems® 2720 thermal cycler.

After amplification the PCR products were mixed with Super-DI™ formamide and orange DNA size standard (Liz500). The optimized ratio of components used is listed below:

Components	Volume per reaction
Sample	0.5 µl
Size Standard	0.1 µl
Formamide	9.4 µl
Total:	10 µl

The samples were then denatured at 95°C for 3 minutes and then immediately put on ice for another 3 minutes. Finally, the samples were run on an ABI 3130xl genetic analyser (Applied Biosystems). The GeneMapper® software was used to analyse and evaluate the data that was generated (Figure 7).

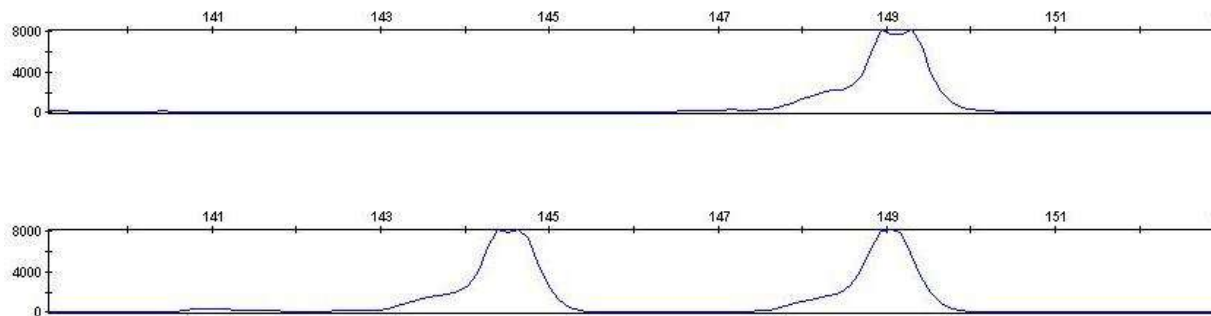


Figure 7. Analysis of two samples from a fragment analysis reaction using GeneMapper®. The upper portion of the figure shows how a wild type allele is represented by a single curve. The lower portion gives an example of a heterozygous carrier of a 4 bp deletion, where one curve represents the wild type allele and the other curve the mutated allele.

Commercial sources of reagents used in all PCR reactions and fragment analyses were as follows: Taq buffer (Fermentas), MgCl₂ (Fermentas), dNTP (Fermentas), primers (Eurofins), Taq DNA polymerase (MCLAB), Super-DI™ formamide (MCLAB) and Liz500 orange DNA size standard (MCLAB).

3.4.2 SNP genotyping

Custom, single-tube TaqMan® reagent-based assays, purchased from Applied Biosystems, were used to perform SNP genotyping to screen for candidate SNPs in our samples. The DNA sequences covering the variants were submitted to the custom TaqMan® assay design tool, on the Applied Biosystems webpage, for primer and probe design. The sequences were retrieved from the UCSC Genome Browser. Primers and probes used in this part of the study can be viewed in table 5.

Table 5. A list of primer and probe sequences used for the genotyping of candidate SNPs.

Target		Sequences	Product length (bp)
<i>SRD5A2</i>	F	5'-GCA CAC GGA GAG CCT GAA G-3'	70
	R	5'-CAG CTC CTG CAG GAA CCA-3'	
	Probe 1	5-VIC-CTG CCA GCC CGC G-3'	
	Probe 2	5-FAM-CTG CCA ACC CGC G-3'	
<i>CAPN14</i>	F	5'-GAG CCC CAA GGA GAA GAT TCT G-3'	76
	R	5'-TCC AAG TCC CTA AGT CCA ATA CCA-3'	
	Probe 1	5-VIC-TTC TGA GGA AAG ACA ATG A-3'	
	Probe 2	5-FAM-CTT CTG AGG AAA GAC AGT GA-3'	
<i>SLC22A16</i>	F	5'-GGC GTC TGT CCA TTT GCA TTC-3'	138
	R	5'-ACC CAA CAG CAC AGG ATA AAG G-3'	
	Probe 1	5-VIC-CTT TTT TGC AGT TGG AAC C-3'	
	Probe 2	5-FAM-TTT TGC AGC TGG AAC C-3'	

The setup per reaction was: 10 ng of genomic DNA, 4.75 µl H₂O, 5 µl TaqMan® genotyping master mix (2x) and 0.25 µl TaqMan® SNP genotyping assay mix (40x). The TaqMan® genotyping master mix contains AmpliTaq Gold® DNA polymerase, dNTPs, ROX™ passive reference and buffer components. The samples were run on a 48-well StepOne™ Real-Time PCR system (Applied Biosystems) and the program can be seen below:

Real-time PCR program		
AmpliTaq Gold Enzyme Activation	PCR (40 cycles)	
HOLD	Denature	Anneal/Extend
10 min at 95°C	15 sec at 95°C	1 min at 60°C

Data was collected and analysed on the StepOne™ software v.2.0 (Figure 8). Reagents used for the genotyping of SNP candidates were purchased from Applied Biosystems.

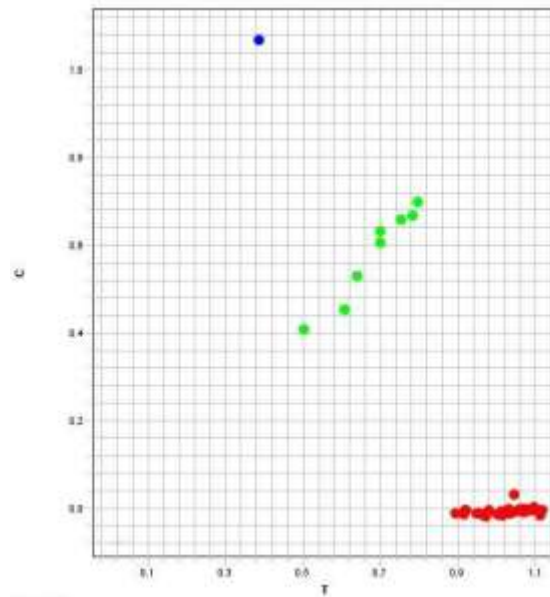


Figure 8. Analysis of SNP genotyping data. An example of SNP genotyping data being analysed in the StepOne™ software. The red dots represent carriers of a wild-type allele, the green dots represent heterozygous carriers of a mutated allele and the blue dot represents a homozygous carrier of a mutated allele.

3.5 Calculations and statistical analysis

Allele frequency and Hardy-Weinberg calculations were done using Microsoft Office Excel 2007. Genotyping data was collected into Excel. The allele frequency for each variant was calculated using the equation below:

Calculating the allele frequency (p) of allele A, where possible genotypes are AA, Aa and aa:

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

In order to check if calculated allele frequencies deviated from the Hardy-Weinberg equilibrium, the following equation was used to calculate the expected genotype frequencies:

$$p^2 + 2pq + q^2 = 1$$

The Chi-Square test was used to compare the observed and expected allele frequencies and determine if the difference between groups was statistically significant. Statistical analysis was performed using R (86).

4 Results

In this study NGS of three regions on chromosomes 2p, 6q and 14q was performed (table 1). They had all been associated with BC in a previous GWS of BC linkage in family 70234, an Icelandic high-risk BRCAx family (55). Within the original regions identified in the GWS there were 554 protein coding genes to be found. Due to overlapping of the position coordinates on chromosomes 2p and 14q with positions from published studies they were modified accordingly (42, 56). The modified regions contained 274 genes in total. Following are the results from the analysis of the targeted resequencing data from these genes.

4.1 Data analysis

The analysis pipeline designed and used in the search for candidate causal mutations in family 70234 is outlined in figure 9 below.

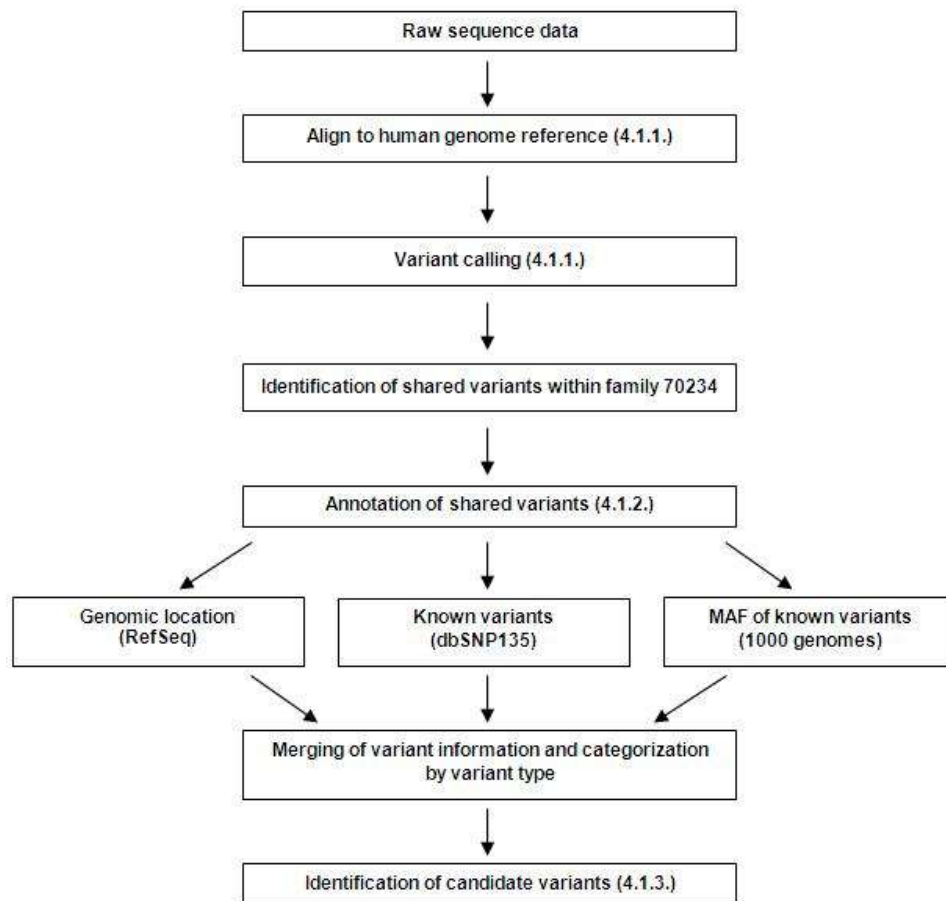


Figure 9. An overview of the analysis pipeline used to identify and evaluate shared variants in family 70234. First the raw sequence reads generated by the 454 instrument were aligned to the human genome reference sequence (see chapter 4.1.1.). Variant calling was then performed (see chapter 4.1.2.) and the variants shared among all 4 samples identified. Those variants were then annotated (see chapter 4.1.2.) and variant information from public databases retrieved. The shared variants were then merged into one file containing information about each variant. Finally candidate variants were identified (see chapter 4.1.3.).

4.1.1 Alignment and variant calling

On average 458.759 reads were generated across all four samples that were sequenced. The average read length was 336 bp. The reads were aligned to the hg18 version of the human genomes reference using the GS Reference Mapper. The percentage of sequence generated reads that aligned to the reference sequence on average was 98.6% across the four samples. On average 95.86% aligned uniquely to the reference sequence. Variant calling on the aligned reads was also performed using the GS Reference Mapper software. The genomic coordinates of the variant calls were then converted from hg18 to hg19 coordinates using the LiftOver tool on the UCSC Genome Browser webpage (<http://genome.ucsc.edu/cgi-bin/hgLiftOver>).

For comparison, and in an effort to generate a variant file in the Variant Calling Format (VCF) developed for the 1000 genomes project (87), the sequencing reads were also aligned using two other aligners; BWA-SW and Bowtie 2. Those alignments were made using the hg19 version of the human genome reference sequence. SamTools was used for variant calling on the BWA-SW and Bowtie 2 generated alignments. Variant calling with SamTools revealed a considerably greater number of variants, which could clearly be classified as homopolymer errors, compared to the number of variant calls made by the GS Reference Mapper. Homopolymer errors were therefore less of an issue when dealing with the GS Reference Mapper generated variants since that software is optimized for 454 generated data and is a part of the standard workflow at Matís. Variants of low quality, e.g. homopolymer errors were filtered out so the variant list generated by the GS reference mapper variant call should only have included high quality variant calls but calls of low quality should have been excluded. However, each insertion and deletion of interest had to be evaluated independently to exclude the possibility that they were in fact sequencing artefacts. This was done by looking at their genomic positions i.e. whether they were positioned within homopolymer regions and by looking at the sequencing reads covering the variant position.

Variant calling with the GS reference mapper identified an average number of 2698 variants per sample. Since the family members that were sequenced share the haplotypes on chromosomes 2p, 6q and 14q it was expected that causative variants would be shared among the samples that were sequenced. A total of 1540 variants were identified that were shared across all four samples. On average there were 31 reads covering the variant sites of the shared variants.

4.1.2 Variant annotation

Annotation of the variants with ANNOVAR provided them with genomic context. Gene-based annotation was done using data from the RefSeq database. The results from the annotation can be viewed in table 6, which shows the number of each variant type, i.e. whether a variant is within an exon, intron or in an intronic region etc.

Table 6. The number of individual variant types among the 1540 shared variants within the three chromosomal regions sequenced.

Variant type	Number of variants
Exonic	146
Splicing	1
ncRNA	95
5'UTR	37
3'UTR	222
Intronic	864
Upstream	76
Downstream	27
Intergenic	72
Total number	1540

Comparison with dbSNP135 resulted in the identification of all previously published variants. Of the total of 1540 shared variants, 1496 had been reported to the database by previous studies. An analysis of the 1000 genomes dataset (an update from April 2012) revealed that 1419 of the 1496 known variants had a reported allele frequency in the dataset, so the allele frequency of the 1419 variants was retrieved.

The majority of the shared variants were of the intronic variant type, or 864 out of the total of 1540 shared variants. The proportion of intronic variants thus was 56.1%. The variant type class with the second highest number were the 3'untranslated regions (3'UTR) variants, accounting for 14.4%. Protein coding variants within targeted exons were 146. Table 7 gives an overview of the number of each variant type, classified on the basis of their functional effect on the protein coding sequence. Briefly, out of the total number of 146 protein coding variants 74 turned out to be nsSNPs and 71 were sSNPs. Only one protein coding indel was identified; a non-frameshift insertion of three nucleotides.

Table 7. The number of shared protein coding variants in family 70234 and their predicted effect on the coding sequence of the genes that harbour them.

Protein coding variants	
Predicted effect	Number of variants
Frameshift insertion	0
Frameshift deletion	0
Stop-gain	0
Stop-loss	0
Nonframeshift insertion	1
Nonframeshift deletion	0
Nonsynonymous SNP	74
Synonymous SNP	71
Total number	146

An analysis of the 74 identified nsSNPs revealed that 73 had been previously reported to dbSNP and had a frequency within the 1000 genomes dataset. Therefore, only one SNP of this variant type could be considered as novel; a SNP within the *CASP8AP2* gene on chromosome 6q that results in a change from methionine to valine. When analysed with respect to their reported allele frequency, five out of these 73 SNPs had a frequency below 5%. A total of 15 variants had an allele frequency below 10%. The allele frequency distribution of the 73 known nsSNPs can be viewed in figure 10 below.

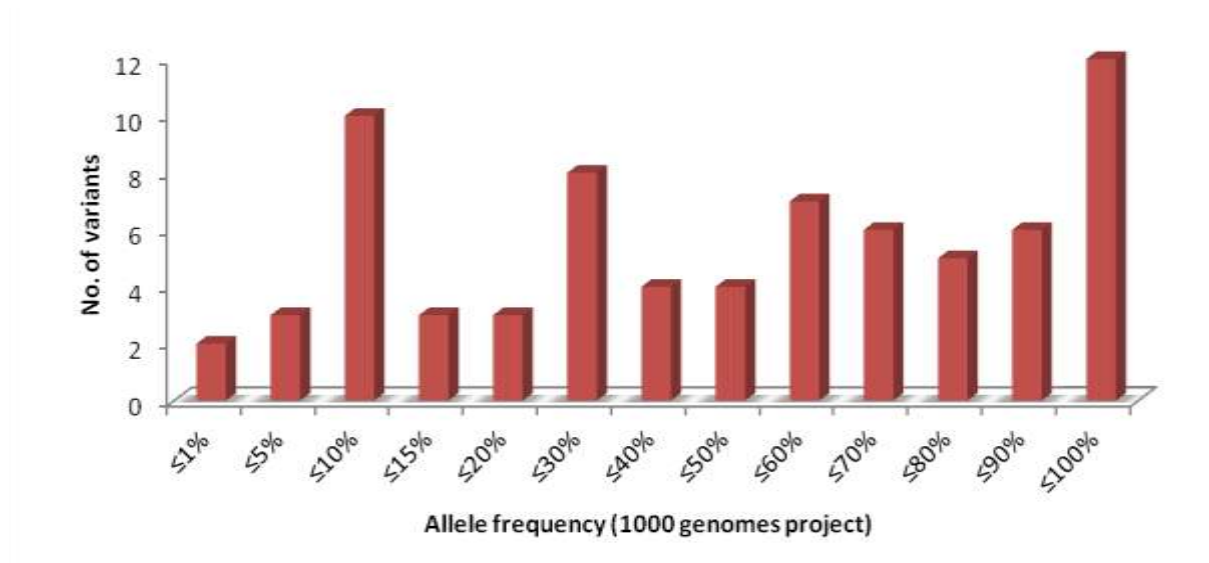


Figure 10. The allele frequency distribution of the 73 previously reported non-synonymous SNPs identified within all of the sequenced samples from family 70234.

4.1.3 Identification of candidate causal variants

At the start of this project, our main working hypothesis was twofold. Firstly, that it might be expected that a causal variant identified within the targeted regions would be a variant of the high-risk allele class. Should we be successful in identifying e.g. a high-risk variant within any of the regions then the two remaining regions might be expected to harbour two modifying low-risk variants that would, when in collaboration with the high-risk variant, increase the risk of getting the disease. Secondly, that it might be expected that three moderate-risk variants would be identified, one within each targeted region. On the basis of this hypothesis an analysis pipeline was designed with the primary emphasis of identifying all shared LOF variants. LOF variants are those that have the most obvious effect on the protein products of the genes that harbour them; frameshift indels, stop-gain or stop-loss mutations and splicing mutations.

Figure 11 depicts the prioritization strategy used in the search for causal variants within the targeted regions in family 70234.

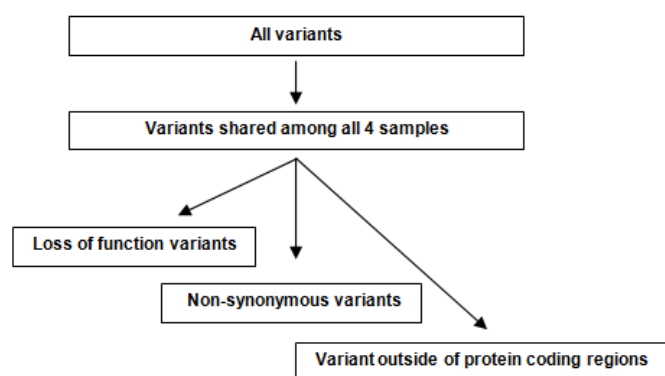


Figure 11. The variant prioritization strategy used in the search for novel BC genes in family 70234. First, the main emphasis was placed on looking for LOF variants that have the most obvious effect on the protein. Other protein coding variants were then examined, e.g. non-synonymous SNPs. Finally, an effort was made to identify variants outside of the protein coding regions of the targeted genes and evaluate their possible effect.

Inspection of the annotated variants within the sequencing data revealed only one variant fulfilling the LOF criteria; a deletion of four nucleotides at the 3' splice site of exon 18 within the *FANCM* gene on chromosome 14q. It therefore quickly became clear that other measures would have to be taken due to the fact that LOF variants seemed to be almost completely absent in the analysed sequences. Due to the nature and location of the *FANCM* variant i.e. because it is a splice site deletion in a Fanconi anemia gene, it was considered to be an attractive candidate for further inspection.

The next step was to take a closer look at other variant types identified within the protein coding regions of the targeted genes. All of the shared protein coding variants were SNPs, except for one shared indel; an insertion of three nucleotides in the *TSPYL1* gene on chromosome 6q. The insertion does not alter the reading frame and is therefore classified as a non-frameshift insertion. The mutation leads to an insertion of a valine amino acid residue into the protein sequence.

SIFT and Polyphen2 were used to estimate which of the 74 nsSNPs were most likely to have a detrimental effect on the encoded proteins of the genes harbouring the variants. A query was run on each program and the results were compared and analysed with respect to each other. SIFT predicted six variants to be deleterious while Polyphen2 predicted eight variants to be deleterious. Only those variants predicted by both programs to be deleterious were considered for further downstream analysis and they turned out to be three: a nucleotide substitution of G to A in the *SRD5A2* gene on chromosome 2p resulting in the amino acid change of alanine to threonine, a substitution of T to C in the *CAPN14* gene on chromosome 2p resulting in a change from arginine to serine and a substitution of A to G in the *SLC22A16* gene on chromosome 6q that leads to an amino acid change of valine to alanine. The novel variant in the *CASP8AP2* gene was not predicted to be deleterious by either program.

The region-based annotation of predicted miRNA target sites revealed two variants; a SNP in the gene *SPAST* on chromosome 2p and a 4 bp insertion in the 3'UTR of the *BAZ1A* gene on chromosome 14q, located in the target site for miR-137. It was decided that out of these two variants,

the insertion in *BAZ1A* would be analysed further. The decision to leave the SNP in *SPAST* out of further inspection was taken because it was easier to make an argument for the inclusion of the *BAZ1A* indel in further analysis i.e. an indel is perhaps more likely to alter a binding site than a SNP. Furthermore, SNP genotyping is more expensive than fragment analysis, which can be used to screen for an indel.

The analysis of the data therefore resulted in the selection of six variants to be validated and checked further (table 8).

Table 8. The candidate variants selected for further analysis. The table lists the genes that harbour the variants, the variant type and its predicted effect (if protein coding), the reference number in dbSNP and its reported MAF. Also listed are the chromosomal positions of the candidate variants, as well as the average read depth and the percentage of sequence reads that the variants were present in.

Gene	Type	Effect	dbSNP ID	MAF	Mutant sites identified (hg19)			Read depth	
					Chromosome	Start	End	Average depth	Average % Variant
<i>SRD5A2</i>	Non-synonymous	p.A49T	rs9282858	0.018	2	31805826	-	23	61.8
<i>CAPN14</i>	Non-synonymous	p.N312S	rs76523220	0.039	2	31417290	-	43	55.1
<i>SLC22A16</i>	Non-synonymous	p.V252A	rs723685	0.082	6	110763875	-	48	49.3
<i>TSPYL1</i>	Nonframeshift insertion	p.E174del	rs56100880	n/a	6	116600472	116600474	46	81.3
<i>BAZ1A</i>	miRNA target site	-	rs57991301	n/a	14	35222631	35222634	55	54.3
<i>FANCM</i>	Splice site variant	-	rs34252356	n/a	14	45654415	45654418	29	44.9

4.2 Validation of candidate variants

Before primers were ordered, the variants that had been selected as candidates had to be evaluated in the sequence alignments with regard to the sequence read depth covering the variant site and the proportion of reads containing each variant. The requirement for a variant call to be deemed genuine was that the read depth covering the variant site had to be equal or above 10 reads and the candidate variant had to be present in at least 15% of the reads covering that specific site. In an effort to avoid being too stringent, and possibly excluding genuine candidates using this analysis filter, we evaluated each variant on the basis of whether it fulfilled these requirements after it had been identified as a possible candidate causal variant. All six variants met the aforementioned requirements and are listed in table 8 with relevant information. Lastly, after seeing that the aligned data suggested that the candidates were genuine variants, they were screened for in selected members of family 70234 to confirm their segregation within the family.

4.3 Case-control analysis of candidate variants

The allele frequency of the selected candidate variants was estimated in a group of unselected BC cases and controls to estimate the likelihood of any of them being a BC risk variant. They were also screened for in selected familial samples from other Icelandic high-risk BRCAx families.

4.3.1 *FANCM*

The Fanconi anemia, complementation group M (*FANCM*) gene is located on chromosome 14q and, as the name suggests, is part of the Fanconi anemia complementation group of genes. Other members of this group are for example *FANCD1* (*BRCA2*), *FANCI* (*BRIP1*) and *FANCF* (*PALB2*). The *FANCM* encoded protein, along with the other known Fanconi anemia proteins, is a participant in DNA damage response (88).

The *FANCM* variant has a reference number in dbSNP (rs34252356) but as of yet its allele frequency has not been reported in the 1000 genomes dataset. The allele frequency of the variant in *FANCM* was 12.7% in the control group of 635 samples. The allele frequency was 13.1% in a group of 1521 unselected BC samples ($p = 0.739$).

The screening in the BC familial material revealed that the allele frequency within the families was 14.5% in 83 samples from 37 HBC families.

4.3.2 *TSPYL1*

The *TSPYL1* gene, or Testis-specific protein Y-like-1, is positioned on chromosome 6q. *TSPYL1* does not contain an intron but is composed of a single, large exon (5259 bp). The entire composition of the gene is however not completely known, e.g. it contains an undefined promoter region (89). The gene's function is not well characterized although the encoded protein has been implicated with chromatin remodelling activity (90) through its relations with the TSPY-SET-NAP1L1 family of chromatin modifiers (91). Mutations within the *TSPYL1* gene have been associated with Sudden Infant Death with Dysgenesis of the Testes, or SIDDIT, that results in the death of infants before 12 months of age (92).

The variant identified in family 70234 has been reported to dbSNP (rs56100880) but does not have a frequency in the 1000 genomes dataset. The allele frequency of the variant within our controls turned out to be 68.9% (240 samples). Within the unselected BC case group the frequency was 70.9% based on screening of 380 samples ($p = 0.603$).

Familial screening revealed that the allele frequency of the *TSPYL1* variant was 68.1% in the HBC material.

4.3.3 *SRD5A2*

SRD5A2 is a gene located on chromosome 2p and encodes for the steroid-5-alpha-reductase type 2 protein that is involved in steroid metabolism (93). To date, two steroid-5-alpha-reductases have been identified. Their reported role is to convert one potent androgen, testosterone, to another; dihydrotestosterone (DHT) (94).

The A49T variant identified in family 70234 has been previously reported to dbSNP (rs9282858) with a MAF of 1.8% in the 1000 genomes dataset, but the frequency of the variant in the European portion of the dataset is a little bit higher, or 4%. Variation within the *SRD5A2* gene has been associated with an increased risk of developing prostate cancer (PC). Some reports have concluded that there is little evidence for the involvement of rs9282858 in increased PC risk (95). However, a recent report has implicated this variant with a small increase in prostate cancer risk (96). It has been

suggested that the variant affects the encoded protein of *SRD5A2* by increasing its enzymatic activity (97). In the 230 control samples that were screened for this variant the allele frequency was 2.8%. The frequency in 184 unselected BC samples was 3.5% ($p = 0.570$).

The allele frequency within the 37 high-risk BC families was 1.8%.

4.3.4 CAPN14

The *CAPN14* gene is located on chromosome 2p. The little that is known about this gene is that it belongs to a family of calpains, which are calcium-activated cysteine proteases involved in a number of cellular processes such as cell division and apoptosis (98). The function of this particular calpain protein is however unknown (99).

The N312S variant in *CAPN14* has been previously reported to dbSNP (rs76523220) and has an allele frequency of 3.9% both in the entire 1000 genomes dataset as well as the European data. The allele frequency within the control group was 3.1% in 229 samples, compared to 3.9% in a total of 181 unselected BC samples genotyped ($p = 0.526$).

The variant had an allele frequency of 3.6% in the 83 samples genotyped belonging to the HBC material.

4.3.5 SLC22A16

The *SLC22A16* gene is located on chromosome 6q and belongs to a family of organic cation transporters (OCTs). Other OCT family members are e.g. *SLC22A1*, *SLC22A2*, *SLC22A3*, *SLC22A4* and *SLC22A5* (100). There have been published data suggesting that the encoded protein of *SLC22A16* is involved in mediating the cellular influx of some anticancer agents such as bleomycin (101) and doxorubicin (102).

The variant identified in family 70234 has been previously reported (rs723685) and has an allele frequency of 8.2% in the whole 1000 genomes dataset. Looking only at the European portion of the 1000 genomes data the reported allele frequency is 9.0%. The IHA had previously genotyped this variant in 3219 samples from individuals of their AGES-Reykjavik study, the allele frequency being 8.2%. The samples from the AGES-Reykjavik study originate from the Reykjavik study, an established population-based cohort (103). The variant was also genotyped in 405 control samples at our lab and had an allele frequency of 7.7%. When these two sets of data were combined the allele frequency of rs723685 was estimated to be 8.4% in a total of 3624 control samples. In the unselected BC group the variant was genotyped in 844 samples, the allele frequency being 8.6% ($p = 0.757$).

The allele frequency in the 37 HBC families was 9.0% in the 83 samples genotyped.

4.3.6 BAZ1A

The bromodomain adjacent to zinc finger domain, 1A gene, or *BAZ1A*, is a gene located on chromosome 14q. This gene is also known as *hACF1*, among other synonyms, and encodes for ACF1/BAZ1A, a subunit of the human ATP-dependent chromatin assembly factor (ACF) that belongs to the ISWI family of chromatin remodelling complexes (104). The protein product of *BAZ1A* has been

implicated with the maintenance of genome integrity through its participation in the cells response to DNA damage (105, 106).

The BAZ1A insertion has been reported to dbSNP (rs57991301) but lacks a reported MAF in the 1000 genomes dataset. The allele frequency of the variant in a control group of 312 samples was 32.5%. The frequency in the unselected BC group, consisting of 537 samples, was 32.7% ($p = 0.949$).

The allele frequency of the variant was 34.3% in the HBC material.

The allele frequency of the candidate variants within controls and unselected BC cases is summarized in table 9, below.

Table 9. The allele frequency of candidate variants in BC cases versus controls. The difference between groups was not statistically significant for any of the variants.

Gene	Control (n)	Unselected BC (n)	<i>p</i>
<i>FANCM</i>	12.7% (635)	13.1% (1521)	NS
<i>TSPYL1</i>	68.9% (240)	70.9% (380)	NS
<i>SRD5A2</i>	3.0% (317)	2.8% (356)	NS
<i>CAPN14</i>	3.2% (316)	3.3% (500)	NS
<i>SLC22A16</i>	8.4% (3624)	8.6% (844)	NS
<i>BAZ1A</i>	32.5% (312)	32.7% (537)	NS

5 Discussion

In this project, targeted sequence capture followed by 454 NGS was performed on four samples from three members of family 70234, an Icelandic high-risk BC family. The family has not been affiliated with mutations in known BC susceptibility genes and has therefore been defined as a BRCAx family. The targeted sequences were within three chromosomal regions, that have shown linkage to BC in a previous study (55). Their genomic locations are at chromosomes 2p, 6q and 14q. Together, the three regions harbour 554 protein coding genes. However, due to overlapping of the positions at chromosomes 2p and 14q with previously published positions (42, 56) the original positions were modified accordingly and the exploration of the genes within the linked regions divided into two separate phases of study. The first phase, presented in this thesis, involved the sequencing of the 274 genes located within the modified positions.

The specific aims of this part of the study were mainly twofold. Firstly, to implement a variant analysis pipeline that would also be of use in future projects of this nature. Secondly, to analyse the data from the sequenced samples from family 70234 and identify causal mutations that would be likely to explain the increased BC risk seen in members of the family. Out of a total of 1540 shared variants among the samples, six were selected as candidates for further analysis. The allele frequency of the variants was compared between controls and unselected cases and none of the six candidates had a statistically significant difference in allele frequency between groups.

5.1 Premise of the study

The pedigree of family 70234 (Figure 5) is reminiscent of pedigrees of families that harbour a known pathogenic mutation in a high penetrance BC susceptibility gene such as *BRCA2* (50). Our estimates indicate that approximately 50% of the women within this family are diagnosed with BC and that the average age at diagnosis is 53 years. The kind of familial clustering of BC as is evident in this family seems highly unlikely to occur by chance. It therefore suggests that the increased susceptibility to BC is due to a high penetrance mechanism of some sort. The fact that regions on three chromosomes provided highly suggestive signals of BC linkage indicates that BC susceptibility in family 70234 does not solely depend on one high penetrance gene. A more plausible mechanism would involve variants at each of these regions; together contributing to the increased risk. This kind of polygenic mechanism has been suggested to be rather common in BRCAx families (55).

It is presumably hard to imagine upfront how a polygenic model of this sort would present itself. However, one could hypothesize how the inner workings of such a model might be. By ruling out that there is one gene conferring a high increase in risk leaves us with two other options with regard to the regions at 2p, 6q and 14q. Previously reported high penetrance genes have a relative risk of 5-20 (18) and the pedigree of family 70234 suggests a high penetrance mechanism. One plausible mechanism would be that within each region lies a gene, harbouring a variant that confers a moderate increase in BC risk. Together, the moderate risk variants would confer a high increase in risk with a combined relative risk above 5. Each variant of this kind would presumably be quite rare in the general population (55).

Another possible model would also involve the participation of three variants; one at each region. In this model, one region might harbour a variant that confers a high increase in risk. The other two regions would then harbour variants that would serve as modifiers for the high penetrance gene. Each of these two would confer a low increase in risk on its own. This model involves the participation of low penetrance alleles but previously reported variants of this allele class, that have been identified through GWAS, have been shown to be quite common in the general population (16). Houlston and Peto have estimated the number of cases required to be able to identify a low risk susceptibility allele through a GWAS. Their estimations are dependent on the frequency and the conferred relative risk of the allele. For detection of a low penetrance allele with a frequency of 5% and conferring a relative risk of 2 in an unselected population would require 800 cases. To be able to detect a similar susceptibility allele with a frequency of 1% in the same unselected population, approximately 3700 cases would be required (107). A GWAS on 1600 Icelandic BC cases was performed by Stacey et al. in an effort to identify new BC susceptibility alleles (40). Based on the calculations made by Houlston and Peto (107), the study by Stacey et al. should have identified low penetrance BC risk alleles with a frequency above 5%. The study did not associate SNPs on chromosomes 2p, 6q and 14q with increased BC susceptibility. Therefore, it might have been expected that the causative variants at chromosomes 2p, 6q and 14q in family 70234 would turn out to be variants with an allele frequency below 5%.

5.2 The sequencing procedure

The costs of NGS application have been continually dropping in recent years (60). NGS is therefore steadily becoming a realistic tool to be used by the common researcher. It is an effective way to interrogate many genes within genomes of individuals that are at high risk of developing a particular disease, allowing for the identification of disease causal variants.

In the present study, the protein coding genes within three chromosomal regions were sequenced using the 454 sequencing platform. As has been mentioned, the positions on chromosomes 2p and 14q overlapped with previously published candidate positions from other studies. It was therefore decided to modify the original chromosomal positions in family 70234 accordingly. This decreased the number of protein coding genes to be sequenced from 554 to 274. This was mainly done to limit the cost of the sequencing procedure, because although the cost of using the NGS technologies have been decreasing, it is still a relatively expensive procedure and was even more expensive at the beginning of this study, in late 2010, compared to what it is today. The protein coding regions of the genes were targeted since most pathogenic mutations in BC susceptibility genes, that have been identified to date and confer a moderate or high increase in risk, are protein truncating mutations (25).

The sequence reads generated by the 454 instrument were aligned on three different software programs. However, we opted for the alignment generated at Matís by the GS reference mapper software. The decision to generate three alignments was in part taken so the alignments could be compared to one another. However, the main reason behind the decision was that it could provide more possibilities with regard to the file formats of the variant files created downstream, i.e. we wanted to be able to generate a variant file in the Variant Call Format (VCF). At the time, the pipeline at Matís

was not equipped to be able to generate a variant file in this format. After variant calling had been performed on the BWA-SW and Bowtie2 alignments, it quickly became obvious that these alignments were fraught with homopolymer errors. It would have required a certain amount of time and tedious work to be able to filter these files to get the same degree of quality as the GS reference mapper generated variant files offered. Therefore, we ignored these other two alignments and concentrated solely on the GS Mapper alignment from Matís.

The decision to use the 454 sequencing platform was in part influenced by the fact that both Nimblegen and Matís are located in Iceland. The entire process, from the targeted sequence capture to the sequencing of the captured regions, could be carried out in close proximity to our lab. The 454 sequencing system has certain advantages as well. It is based on the pyrosequencing technology that relies on the detection of light emission representing the nucleotides that are incorporated and reveal the DNA sequence. The read length of the 454 system separates it from other sequencing systems such as the Illumina, which is probably the most used commercial platform today (108). The average read length of the sequence reads generated in this project was 336 bp. These long reads make the alignment process more reliable compared to systems that generate shorter reads, as e.g. the Illumina (108).

5.3 Variant identification and annotation

From a pool of variants within each sequenced sample, 1540 germline variants, shared across all samples, were identified. Their genomic context can be viewed in table 6. The majority of them are located within intronic regions while only 146 were protein coding. The intronic regions were not directly targeted per se but that does not rule out the possibility that some intronic regions might have crept in and got captured on the side. Also, the exon/intron boundaries were targeted, which might explain a small proportion of the observed intronic variation. Furthermore, to improve the overall coverage of the targeted regions, a small proportion of the probes designed to capture the sequences were non-unique, i.e. they allowed for non-specific capture of sequences outside of the target regions. Therefore, although the vast number of intronic variants might be somewhat surprising, there are a number of possible explanations for the capture of intronic sequences. Also, the level of genetic variation within introns has been shown to be higher compared to the protein coding exons of genes. This observed difference in genetic diversity has been attributed to evolutionary forces, e.g. natural selection that contributes to the eradication of strongly deleterious mutations within populations (83). This is also reflected in whole genome and whole exome sequencing studies. A rough estimate of the distribution rate of DNA variants spread across the entire human genome reveals that if we assume that within an individual genome there are ≈ 4 million SNPs to be found (109) then the rate of SNP distribution would be approximately one SNP in every 770 bp. Some reports have published an estimate of the total number of coding SNPs (cSNPs) that can be found, on average, within an individual human exome. These estimates range from ≈ 17.000 cSNPs to ≈ 21.000 cSNPs (62, 110). If we estimate the size of an individual human exome as being ≈ 30 Mb then the rate of variant distribution within the human exome should range from approximately one variant per 1400-1700 bp.

This is of course a rough estimate, but reflects nicely on the different nature of introns and exons with regard to DNA variant distribution.

A query against the dbSNP database (version 135) revealed that the majority of shared germline variants identified in the study were known and previously reported to the database, or 97.1%. This was anticipated since the majority of variants within individual genomes are present in dbSNP (83). Estimates further indicate that the pilot phase of the 1000 genomes project has identified over 95% of common polymorphisms, defined as variants with an allele frequency above 5% (111). Out of the previously reported variants in this study 89.0% had an allele frequency within the 1000 genomes dataset above 5% and could thus be classified as common variants. A total of 43 variants had an allele frequency below 5%, or 2.9% of the total number of known variants. The number of known variants without allele frequency information was 77, or 5.1% of the total number of known variants. A total of 45 variants were identified that had not been previously reported to the dbSNP database. Interestingly, when the proportion of insertion and deletions are compared between the two groups of variants, known and unknown, we see that the majority of unknown variants are classified as being either of these two mutation types, or 73.3%, compared to only approximately 5% of the known variants. This vast difference might reflect on the drawbacks of the 454 sequencing platform, namely its difficulties in sequencing homopolymer regions. Since all indels were evaluated individually as they came up as possible candidates, but were not looked at in detail beforehand, we did not estimate the number of homopolymer artefacts among these unknown insertions and deletions. However, the fact that homopolymer errors did not trouble our analysis of the variants at all shows that the filters applied in the GS reference mapper software for the variant calling served their purpose well to filter out variant calls of low quality. It must however also be acknowledged that by identifying shared variants across all 4 samples the number of potential artefacts was likely reduced even further. However, this high number of indels within the unknown variant group might also reflect on the fact that public variant databases are not as complete for short indels compared to SNPs as was revealed by the pilot phase of the 1000 genomes project where 50% of common short indels identified were novel compared to 9% of common coding SNPs (83). As the 1000 genomes project progresses, more and more of the common variation will be identified and available in public repositories.

5.4 Variant analysis pipeline

When trying to identify the cause of a disease through NGS, assumption will have to be made about the nature of the variant or variants that the researcher expects to identify. This has to be done to provide a filtering strategy that prioritizes the search for the disease causing variants (112). In this project it was expected that the causative variants would have a low population frequency and they should be shared by all samples, since it had been established that these individuals share haplotypes at the targeted regions. Because it was expected that the variants would be highly or moderately penetrant then a case could be made that they most likely reside within protein coding regions. Furthermore, reported BC susceptibility alleles conferring a high or moderate increase in risk are predominantly protein truncating variants (25). Our priority was thus first to identify all protein truncating LOF mutations by employing a variant prioritization strategy similar to the one described in

two studies performed by Snape et al. (76, 77). Therefore, the first analysis step involved applying a filter to identify all nonsense mutations, frameshift insertions and deletions and splice site mutations. Only one variant that could be classified as a LOF was identified, which was a *FANCM* splicing deletion. Out of the total number of protein coding variants, 74 turned out to be non-synonymous and 71 synonymous. Only one insertion was detected but it did not alter the reading frame since it was an insertion of three bp. The ratio between nsSNPs and sSNPs was approximately 1:1 in our study, which is in agreement with other reports (110, 111). Due to the fact that the non-frameshift insertion in *TSPYL1* was the only protein coding variant identified that was not a SNP and because the variant in question was an indel and could be evaluated in the sample material via fragment analysis, which is not as costly as e.g. SNP genotyping, it was decided that this three bp insertion would be analysed further.

After the first round of evaluation it became obvious that the prioritization strategy would have to be altered since the primary filters used did not result in the identification of clear candidate LOF variants. It was decided that the next step would involve further analysis of the nsSNPs. These variants do not have an obvious affect on the protein products of the genes that harbour them to the same degree as LOF variants do. For this reason we had to decide on what grounds this list of 74 nsSNPs would be filtered, to decide which one among them would be most likely to have a detrimental effect. Genetic variants can be ranked based on a number of different factors that include their predicted effect on the function and structure of proteins and their positioning with regard to conservation. A number of tools are available that can be utilized for the determination of these factors (78, 79, 113-115). Evaluating nsSNPs to determine their effect on the function of proteins is not an easy thing to undertake. When considering this for the 74 nsSNPs in this project the first choice was to estimate the allele frequency of each variant directly. This would however have required a massive financial outlay and was therefore not a realistic option. By utilizing publicly available information in biological databases and some of the tools mentioned above, the list of nsSNPs could be filtered based on a couple of assumptions about the nature of these SNPs. The nsSNPs that have been reported to affect protein function are usually rare or novel (110). Collecting and merging information from the dbSNP database and the 1000 genomes project allowed us to rank the nsSNPs in order, from high to low allele frequency. A total of five variants had an allele frequency below 5% (Figure 10) and one was novel, i.e. it had neither been submitted to dbSNP nor did it have a frequency in the 1000 genomes project data. SIFT and Polyphen2 provided prediction on which SNPs would be most likely to be deleterious. SIFT predictions are based on the assumption that essential amino acid residues, that are important for the function of the encoded proteins they are part of, should be evolutionary conserved. SNPs that alter conserved amino acid residues in proteins are therefore predicted to be deleterious (78). Polyphen2 predictions are in part based on a similar assumption with regard to conserved amino acid residues through evolutionary processes but also take into account the structure and function of the protein itself. It does that by taking into consideration evolutionary conservation, as well as evaluating the sequence of the protein and structural information with regard to where the substituted residue is within the protein and the nature of the amino acids (79). All 74 nsSNPs were included in the functional prediction analysis and only those predicted by both programs to be damaging were

considered as possible candidates. This filter was used to narrow down the results and identify the ones most likely to be real causative variants. Three SNPs were deemed by both programs to be deleterious. Of these three, two had a MAF below 5% and one had a 9% allele frequency and could therefore be classified as a common variant. All of these predicted deleterious SNPs were included in the screening in the sample material. The decision to include the common variant in *SLC22A16* in further analysis was taken because we wanted to cover some kind of middle ground in our filtering steps and did not want to apply filters that would turn out to have been too stringent. Therefore, this SNP was included as a candidate variant based on the results from the functional prediction programs but we turned a blind eye towards its frequency.

At this stage all protein coding variants had been evaluated and prioritized as well as splice site variants. Although we did not cover other regions in the sequence capture to the extent that we could actively target e.g. intronic regions in the variant analysis pipeline, we nevertheless wanted to make an effort to analyse other regions than the protein coding exons and the exon/intron boundaries. With this in mind an annotation was performed using ANNOVAR in an effort to identify those variants that were within predicted miRNA target sites. Two variants were identified, one SNP and one insertion of four bp in the 3'UTR of a gene on chromosome 14q, *BAZ1A*. As has been mentioned, screening for indels is financially more attractive than screening for SNPs and therefore it was decided that this insertion would be analysed further with the rationale being that an insertion of four bp in a miRNA target site could potentially disrupt the binding of the miRNA to its target and thus affect the expression of the gene normally targeted by the miRNA.

5.5 Genotyping did not confirm any candidate variants as likely to be causal

Six variants from members of family 70234 were genotyped in the sample material of unselected BC cases and controls. The only LOF variant identified in the sequencing data was the splice site variant in *FANCM*. This gene belongs to the Fanconi anemia complementation group and encodes a protein that participates in DNA damage repair (88). What made the variant in this gene especially interesting, along with the nature of the variant itself, is the fact that four Fanconi anemia genes have been implicated with increased susceptibility to BC (116). However, neither the *FANCM* variant nor any of the other candidate variants that were genotyped turned out to be likely causal variants with respect to the observed increase in BC risk in family 70234.

5.6 Where are the causative variants?

There are a number of possible reasons that could explain why we were not successful in identifying the variants responsible for the increased BC susceptibility in family 70234 in this first phase of the study.

The original chromosomal regions identified in the genome wide linkage study contained 554 genes in total. This phase of the study however only targeted 274 protein coding genes. The modification of the regions on chromosomes 2p and 14q were based on previously published reports of BC linkage of positions that intersected with the positions identified in family 70234. It is a possibility

that some of the genes that were not included in the NGS process carry mutations that would be of interest. However, the region linked to BC susceptibility on chromosome 6q was not modified. So the question lingers; why were we unable to identify a causal variant within that region?

The target capture of the 6q region was performed based on the original positions identified in the GWS that identified the linkage in family 70234 (55) but not on modified positions as was the case with 2p and 14q. Therefore, in theory at least, we should have identified the variant causing the increased BC susceptibility in this phase of the study. Possible explanations for the lack of identification of causative variants might include that the reads representing regions with large insertions or deletions in the genomic sequence might not align to the reference sequence due to lack of homology between the two. However, this scenario seems unlikely. The average proportion of the generated sequence reads that mapped to the reference genome sequence (98.6%) was considered to be acceptable and implies that it is unlikely that the causative variant remains unknown due to poor alignment of the reads covering the region. In our downstream analysis we did not perform a local re-alignment around indels which can improve the variant calling by eliminating false positive calls and identify insertions or deletions that remain hidden due to misalignment of the sequencing reads (117). Although, by applying a filter that requires variants to be shared by all samples reduces the number of false positives but then there is still the question of revealing possible hidden insertions and deletions in the data. This might also seem an unlikely explanation but should not be ruled out entirely. At the time of the writing of this thesis we have implemented a local realignment step in our processing and analysis of sequencing data that is worked on at the department. This procedure is however optimized for sequencing instruments that generate shorter sequence reads than the 454 instrument, e.g. data from the Illumina platform. Another possibility that should be kept in mind is that the filter, requiring variants to be shared among all four samples, might have been too stringent. A causative variant could reside in a genomic area that was perhaps not adequately enriched in one of the four samples. Such a scenario would result in the variant only being called in three of the samples and therefore not passing the analysis filter. Future re-analysis of the data should take this possibility into account.

As was mentioned above, protein coding regions were primarily targeted for sequencing. Therefore, there are two possible explanations for why no causative variants were identified and confirmed. One possibility is that the region containing the causative variant on chromosome 6q is not covered in the sequence capture, e.g. intronic regions. Evaluating non-coding variants is a challenge as their characterization remains troublesome (110). Whole genome sequencing is the only approach that allows for extensive analysis of all of the non-coding regions of the genome, although some whole-genome sequencing studies have nonetheless focused primarily on coding variants in their variant analysis pipeline (118, 119). As time passes, however, it is likely that our understanding of the nature of disease causing non-coding variants will grow and their analysis and identification become easier (120). The other possibility is that the real causal variant on chromosome 6q is in fact protein coding and is among the 74 identified nsSNPs in the 454 sequencing data but that our analysis pipeline was not sophisticated or robust enough to identify it as a candidate variant for screening and further analysis. One novel nsSNP was identified, located at the 6q region, but was not considered further after the functional prediction programs both predicted it to be neutral. In hindsight, the decision

to leave the novel variant out of the screening in the sample material might seem quite naive, although it is supported by the prioritization strategy we were following at the time. It is therefore not a foregone conclusion that the causal variant at the 6q region is a coding nsSNP that has been identified in the samples that were sequenced. With this in mind we are continually re-evaluating our analysis pipeline and making an effort to improve and polish it, for the benefit of the next phase of the study on family 70234 and other sequencing projects that are currently under way.

6 Conclusions

The work presented in this thesis represents the first steps in identifying the causal variants underlying the increased BC susceptibility of members of family 70234, an Icelandic high-risk BRCA1 family. In this phase of the study we were not successful in identifying a causal variant. The next steps will therefore involve the sequencing of the rest of the protein coding genes that lie within the original positions. This is a total of 280 genes.

Since sequencing costs have been dropping considerably, whole exome sequencing (WES) is becoming more of a feasible option today compared to 2-3 years ago. Therefore it was decided that the second phase of the study would involve WES rather than targeting just the 280 genes left out in the first phase. At the time of writing, the sequencing data has arrived at the department and analysis of the data awaits.

The variant analysis pipeline that will be pursued in the second phase will be constructed and based on the experience we gained when analysing the 454 data in the first phase of the study. Briefly, after sequence reads have been aligned to the hg19 version of the human genome the data will be processed through a workflow based on the framework presented by Depristo et al. (117) to maximize the quality of the variant and genotype calls. All identified variants will be annotated and the shared variants in regions 2p, 6q and 14q will be sought out. Known variants will be identified and their allele frequency retrieved from the 1000 genomes project data. All shared variants within the regions will then be ranked according to their allele frequency and variants with a MAF below 5% will be identified for further analysis. We will retrieve conservation scores for all of the shared variants from e.g. GERP (113, 114) and rank them according to the score they are given. We will look for possible LOF variants and functional predictions for all nsSNPs will also be retrieved. The decision of which variants will be screened for will be based on the aforementioned information as well as e.g. pathway analysis and further curation of the literature. If one or more of the susceptibility alleles we identify eventually turn out to be low penetrance alleles then a large sample size will be required in each sample group, so that we will be able to confirm the contribution to the development of the disease. With this in mind we are continually expanding our sample collection, although it will be difficult to collect enough samples to be able to provide our statistical analysis with enough power to determine whether the difference in allele frequency of low risk variants, which tend to be common in the population, between groups is statistically significant.

Finally, we will also carry on structuring and re-evaluating our analysis pipeline continually to make it as robust and precise as it can possibly be. This includes improving the analysis of variants that are not located within the protein coding regions of the human genome.

References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*. 2010;127(12):2893-917. doi: 10.1002/ijc.25516.
2. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin*. 2011;61(2):69-90. doi: 10.3322/caac.20107. Epub 2011 Feb 4.
3. Tryggvadottir L, Olafsdottir EJ, Jonasson JG. Icelandic Cancer Registry at Icelandic Cancer Society editors: Retrieved 11.03.2013 <http://www.cancerregistry.is>.
4. Colombo PE, Milanezi F, Weigelt B, Reis-Filho JS. Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction. *Breast Cancer Res*. 2011;13(3):212. doi: 10.1186/bcr2890.
5. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001;98(19):10869-74.
6. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*. 2003;100(14):8418-23. Epub 2003 Jun 26.
7. Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, et al. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*. 2006;7:96.
8. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. *N Engl J Med*. 2009;360(8):790-800. doi: 10.1056/NEJMra0801289.
9. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res*. 2010;12(5):R68. doi: 10.1186/bcr2635. Epub 010 Sep 2.
10. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. *Mol Oncol*. 2011;5(1):5-23. doi: 10.1016/j.molonc.2010.11.003. Epub Nov 24.
11. Hill TD, Khamis HJ, Tyczynski JE, Berkel HJ. Comparison of male and female breast cancer incidence trends, tumor characteristics, and survival. *Ann Epidemiol*. 2005;15(10):773-80.
12. Oldenburg RA, Meijers-Heijboer H, Cornelisse CJ, Devilee P. Genetic susceptibility for breast cancer: how many more genes to be found? *Crit Rev Oncol Hematol*. 2007;63(2):125-49. Epub 2007 May 10.
13. Lalloo F, Evans DG. Familial breast cancer. *Clin Genet*. 2012;82(2):105-14. doi: 10.1111/j.399-0004.2012.01859.x. Epub 2012 Apr 13.
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. Mitrunen K, Hirvonen A. Molecular epidemiology of sporadic breast cancer. The role of polymorphic genes involved in oestrogen biosynthesis and metabolism. *Mutat Res*. 2003;544(1):9-41.
16. Fanale D, Amodeo V, Corsini LR, Rizzo S, Bazan V, Russo A. Breast cancer genome-wide association studies: there is strength in numbers. *Oncogene*. 2012;31(17):2121-8. doi: 10.1038/onc.2011.408. Epub Sep 26.
17. Lux MP, Fasching PA, Beckmann MW. Hereditary breast and ovarian cancer: review and future perspectives. *J Mol Med (Berl)*. 2006;84(1):16-28. Epub 2005 Nov 11.

18. Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. *Nat Genet.* 2008;40(1):17-22.
19. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science.* 1994;266(5182):66-71.
20. Roy R, Chun J, Powell SN. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat Rev Cancer.* 2011;12(1):68-78. doi: 10.1038/nrc3181.
21. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, et al. Identification of the breast cancer susceptibility gene BRCA2. *Nature.* 1995;378(6559):789-92.
22. Meindl A, Ditsch N, Kast K, Rhiem K, Schmutzler RK. Hereditary breast and ovarian cancer: new genes, new treatments, new concepts. *Dtsch Arztebl Int.* 2011;108(19):323-30. doi: 10.3238/arztebl.2011.0323. Epub 2011 May 13.
23. Szabo CI, King MC. Population genetics of BRCA1 and BRCA2. *Am J Hum Genet.* 1997;60(5):1013-20.
24. Rebbeck TR, Domchek SM. Variation in breast cancer risk in BRCA1 and BRCA2 mutation carriers. *Breast Cancer Res.* 2008;10(4):108. doi: 10.1186/bcr2115. Epub 008 Jul 25.
25. Mavaddat N, Antoniou AC, Easton DF, Garcia-Closas M. Genetic susceptibility to breast cancer. *Mol Oncol.* 2010;4(3):174-91. doi: 10.1016/j.molonc.2010.04.011. Epub May 21.
26. Rebbeck TR, Mitra N, Domchek SM, Wan F, Friebel TM, Tran TV, et al. Modification of BRCA1-Associated Breast and Ovarian Cancer Risk by BRCA1-Interacting Genes. *Cancer Res.* 2011;71(17):5792-805. doi: 10.1158/0008-5472.CAN-11-0773. Epub 2011 Jul 28.
27. Garber JE, Goldstein AM, Kantor AF, Dreyfus MG, Fraumeni JF, Jr., Li FP. Follow-up study of twenty-four families with Li-Fraumeni syndrome. *Cancer Res.* 1991;51(22):6094-7.
28. Malkin D. Germline p53 mutations and heritable cancer. *Annu Rev Genet.* 1994;28:443-65.
29. Wong SS, Lozano G, Gaff CL, Gardner RJ, Strong LC, Aittomaki K, et al. Novel p53 germline mutation in a patient with Li-Fraumeni syndrome. *Intern Med J.* 2003;33(12):621.
30. Nelen MR, Padberg GW, Peeters EA, Lin AY, van den Helm B, Frants RR, et al. Localization of the gene for Cowden disease to chromosome 10q22-23. *Nat Genet.* 1996;13(1):114-6.
31. Hemminki A, Markie D, Tomlinson I, Avizienyte E, Roth S, Loukola A, et al. A serine/threonine kinase gene defective in Peutz-Jeghers syndrome. *Nature.* 1998;391(6663):184-7.
32. Antoniou AC, Easton DF. Models of genetic susceptibility to breast cancer. *Oncogene.* 2006;25(43):5898-905.
33. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet.* 2006;38(8):873-5. Epub 2006 Jul 9.
34. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, 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. Epub 2006 Oct 8.
35. Erkkö H, Xia B, Nikkila J, Schleutker J, Syrjäkoski K, Mannermaa A, et al. A recurrent mutation in PALB2 in Finnish cancer families. *Nature.* 2007;446(7133):316-9. Epub 2007 Feb 7.

36. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet.* 2007;39(2):165-7. Epub 2006 Dec 31.
37. Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, et al. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet.* 2002;31(1):55-9. Epub 2002 Apr 22.
38. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature.* 2007;447(7148):1087-93.
39. Milne RL, Benitez J, Nevanlinna H, Heikkinen T, Aittomaki K, Blomqvist C, et al. Risk of estrogen receptor-positive and -negative breast cancer and single-nucleotide polymorphism 2q35-rs13387042. *J Natl Cancer Inst.* 2009;101(14):1012-8. doi: 10.93/jnci/djp167. Epub 2009 Jun 30.
40. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet.* 2007;39(7):865-9. Epub 2007 May 27.
41. Huusko P, Juo SH, Gillanders E, Sarantaus L, Kainu T, Vahteristo P, et al. Genome-wide scanning for linkage in Finnish breast cancer families. *Eur J Hum Genet.* 2004;12(2):98-104.
42. Smith P, McGuffog L, Easton DF, Mann GJ, Pupo GM, Newman B, et al. A genome wide linkage search for breast cancer susceptibility genes. *Genes Chromosomes Cancer.* 2006;45(7):646-55.
43. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science.* 1994;265(5181):2088-90.
44. Rafnar T, Gudbjartsson DF, Sulem P, Jonasdottir A, Sigurdsson A, Jonasdottir A, et al. Mutations in BRIP1 confer high risk of ovarian cancer. *Nat Genet.* 2011;43(11):1104-7. doi: 10.038/ng.955.
45. Arason A, Barkardottir RB, Egilsson V. Linkage analysis of chromosome 17q markers and breast-ovarian cancer in Icelandic families, and possible relationship to prostatic cancer. *Am J Hum Genet.* 1993;52(4):711-7.
46. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science.* 1990;250(4988):1684-9.
47. Narod SA, Feunteun J, Lynch HT, Watson P, Conway T, Lynch J, et al. Familial breast-ovarian cancer locus on chromosome 17q12-q23. *Lancet.* 1991;338(8759):82-3.
48. Solomon E, Ledbetter DH. Report of the committee on the genetic constitution of chromosome 17. *Cytogenet Cell Genet.* 1990;55(1-4):198-215.
49. Bergthorsson JT, Jonasdottir A, Johannesdottir G, Arason A, Egilsson V, Gayther S, et al. 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.
50. Gudmundsson J, Johannesdottir G, Arason A, Bergthorsson JT, Ingvarsson S, Egilsson V, et al. Frequent occurrence of BRCA2 linkage in Icelandic breast cancer families and segregation of a common BRCA2 haplotype. *Am J Hum Genet.* 1996;58(4):749-56.
51. Thorlacius S, Olafsdottir G, Tryggvadottir L, Neuhausen S, Jonasson JG, Tavtigian SV, et al. A single BRCA2 mutation in male and female breast cancer families from Iceland with varied cancer phenotypes. *Nat Genet.* 1996;13(1):117-9.

52. Johannesdottir G, Gudmundsson J, Bergthorsson JT, Arason A, Agnarsson BA, Eiriksdottir G, et al. High prevalence of the 999del5 mutation in Icelandic breast and ovarian cancer patients. *Cancer Res.* 1996;56(16):3663-5.
53. Thorlacius S, Sigurdsson S, Bjarnadottir H, Olafsdottir G, Jonasson JG, Tryggvadottir L, et al. Study of a single BRCA2 mutation with high carrier frequency in a small population. *Am J Hum Genet.* 1997;60(5):1079-84.
54. Arason A, Jonasdottir A, Barkardottir RB, Bergthorsson JT, Teare MD, Easton DF, et al. A population study of mutations and LOH at breast cancer gene loci in tumours from sister pairs: two recurrent mutations seem to account for all BRCA1/BRCA2 linked breast cancer in Iceland. *J Med Genet.* 1998;35(6):446-9.
55. Arason A, Gunnarsson H, Johannesdottir G, Jonasson K, Bendahl PO, Gillanders EM, et al. Genome-wide search for breast cancer linkage in large Icelandic non-BRCA1/2 families. *Breast Cancer Res.* 2010;12(4):R50. doi: 10.1186/bcr2608. Epub 010 Jul 16.
56. Gonzalez-Neira A, Rosa-Rosa JM, Osorio A, Gonzalez E, Southey M, Sinilnikova O, et al. Genomewide high-density SNP linkage analysis of non-BRCA1/2 breast cancer families identifies various candidate regions and has greater power than microsatellite studies. *BMC Genomics.* 2007;8:299.
57. Antoniou AC, Pharoah PP, Smith P, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *Br J Cancer.* 2004;91(8):1580-90.
58. Stratton MR. Journeys into the genome of cancer cells. *EMBO Mol Med.* 2013;22(10):201202388.
59. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet.* 2010;11(10):685-96. doi: 10.1038/nrg2841.
60. Ku CS, Naidoo N, Pawitan Y. Revisiting Mendelian disorders through exome sequencing. *Hum Genet.* 2011;129(4):351-70. doi: 10.1007/s00439-011-0964-2. Epub 2011 Feb 18.
61. Lin X, Tang W, Ahmad S, Lu J, Colby CC, Zhu J, et al. Applications of targeted gene capture and next-generation sequencing technologies in studies of human deafness and other genetic disabilities. *Hear Res.* 2012;288(1-2):67-76. doi: 10.1016/j.heares.2012.01.004. Epub Jan 14.
62. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature.* 2009;461(7261):272-6. doi: 10.1038/nature08250. Epub 2009 Aug 16.
63. Walsh T, Lee MK, Casadei S, Thornton AM, Stray SM, Pennil C, et al. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proc Natl Acad Sci U S A.* 2010;107(28):12629-33. doi: 10.1073/pnas.1007983107. Epub 2010 Jun 28.
64. Mardis ER. A decade's perspective on DNA sequencing technology. *Nature.* 2011;470(7333):198-203. doi: 10.1038/nature09796.
65. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010;26(5):589-95. doi: 10.1093/bioinformatics/btp698. Epub 2010 Jan 15.
66. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357-9. doi: 10.1038/nmeth.923.
67. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 2008;18(11):1851-8. doi: 10.101/gr.078212.108. Epub 2008 Aug 19.

68. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, et al. Tablet--next generation sequence assembly visualization. *Bioinformatics*. 2010;26(3):401-2. doi: 10.1093/bioinformatics/btp666. Epub 2009 Dec 4.
69. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res*. 2009;19(9):1639-45. doi: 10.101/gr.092759.109. Epub 2009 Jun 18.
70. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2012;19:19.
71. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8.
72. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*. 2009;25(17):2283-5. doi: 10.1093/bioinformatics/btp373. Epub 2009 Jun 19.
73. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164. doi: 10.1093/nar/gkq603. Epub 2010 Jul 3.
74. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6(2):80-92. doi: 10.4161/fly.19695.
75. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet*. 2011;12(9):628-40. doi: 10.1038/nrg3046.
76. Snape K, Hanks S, Ruark E, Barros-Nunez P, Elliott A, Murray A, et al. Mutations in CEP57 cause mosaic variegated aneuploidy syndrome. *Nat Genet*. 2011;43(6):527-9. doi: 10.1038/ng.822. Epub 2011 May 8.
77. Snape K, Ruark E, Tarpey P, Renwick A, Turnbull C, Seal S, et al. Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. *Breast Cancer Res Treat*. 2012;134(1):429-33. doi: 10.1007/s10549-012-2057-x. Epub 2012 Apr 18.
78. 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. doi: 10.38/nprot.2009.86. Epub Jun 25.
79. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248-9. doi: 10.1038/nmeth0410-248.
80. Droege M, Hill B. The Genome Sequencer FLX System--longer reads, more applications, straight forward bioinformatics and more complete data sets. *J Biotechnol*. 2008;136(1-2):3-10. doi: .1016/j.jbiotec.2008.03.021. Epub Jun 21.
81. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*. 2006;34(Database issue):D590-8.
82. Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res*. 2009;37(Database issue):D32-6. doi: 10.1093/nar/gkn721. Epub 2008 Oct 16.

83. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-73. doi: 10.1038/nature09534.
84. 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.
85. Teer JK, Green ED, Mullikin JC, Biesecker LG. VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer. *Bioinformatics*. 2012;28(4):599-600. doi: 10.1093/bioinformatics/btr711. Epub 2011 Dec 30.
86. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2013.
87. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-8. doi: 10.1093/bioinformatics/btr330. Epub 2011 Jun 7.
88. Huang M, Kim JM, Shiotani B, Yang K, Zou L, D'Andrea AD. The FANCM/FAAP24 complex is required for the DNA interstrand crosslink-induced checkpoint response. *Mol Cell*. 2010;39(2):259-68. doi: 10.1016/j.molcel.2010.07.005.
89. Javaher P, Stuhmann M, Wilke C, Frenzel E, Manukjan G, Grosshenig A, et al. Should TSPYL1 mutation screening be included in routine diagnostics of male idiopathic infertility? *Fertil Steril*. 2012;97(2):402-6. doi: 10.1016/j.fertnstert.2011.11.002. Epub Dec 2.
90. de Andrade TG, Peterson KR, Cunha AF, Moreira LS, Fattori A, Saad ST, et al. Identification of novel candidate genes for globin regulation in erythroid cells containing large deletions of the human beta-globin gene cluster. *Blood Cells Mol Dis*. 2006;37(2):82-90. Epub 2006 Sep 6.
91. Biason-Lauber A. Control of sex development. *Best Pract Res Clin Endocrinol Metab*. 2010;24(2):163-86. doi: 10.1016/j.beem.2009.12.002.
92. Puffenberger EG, Hu-Lince D, Parod JM, Craig DW, Dobrin SE, Conway AR, et al. Mapping of sudden infant death with dysgenesis of the testes syndrome (SIDDT) by a SNP genome scan and identification of TSPYL loss of function. *Proc Natl Acad Sci U S A*. 2004;101(32):11689-94. Epub 2004 Jul 23.
93. Redler S, Tazi-Ahnini R, Drichel D, Birch MP, Brockschmidt FF, Dobson K, et al. Selected variants of the steroid-5-alpha-reductase isoforms SRD5A1 and SRD5A2 and the sex steroid hormone receptors ESR1, ESR2 and PGR: no association with female pattern hair loss identified. *Exp Dermatol*. 2012;21(5):390-3. doi: 10.1111/j.1600-0625.2012.01469.x.
94. Li X, Huang Y, Fu X, Chen C, Zhang D, Yan L, et al. Meta-analysis of three polymorphisms in the steroid-5-alpha-reductase, alpha polypeptide 2 gene (SRD5A2) and risk of prostate cancer. *Mutagenesis*. 2011;26(3):371-83. doi: 10.1093/mutage/geq103. Epub 2010 Dec 21.
95. Pearce CL, Van Den Berg DJ, Makridakis N, Reichardt JK, Ross RK, Pike MC, et al. No association between the SRD5A2 gene A49T missense variant and prostate cancer risk: lessons learned. *Hum Mol Genet*. 2008;17(16):2456-61. doi: 10.1093/hmg/ddn145. Epub 2008 May 10.
96. Li Q, Zhu Y, He J, Wang M, Zhu M, Shi T, et al. Steroid 5-alpha-reductase type 2 (SRD5A2) V89L and A49T polymorphisms and sporadic prostate cancer risk: a meta-analysis. *Mol Biol Rep*. 2013;1:1.
97. Zhao D, Wu W, Xu B, Niu X, Cui H, Zhang Y, et al. Variants in the SRD5A2 gene are associated with quality of semen. *Mol Med Report*. 2012;6(3):639-44. doi: 10.3892/mmr.2012.965. Epub Jun 25.

98. Dear TN, Boehm T. Identification and characterization of two novel calpain large subunit genes. *Gene*. 2001;274(1-2):245-52.
99. National center for biotechnology information. Retrieved 11.03.2013 <http://www.ncbi.nlm.nih.gov/gene/440854>.
100. Koepsell H, Lips K, Volk C. Polyspecific organic cation transporters: structure, function, physiological roles, and biopharmaceutical implications. *Pharm Res*. 2007;24(7):1227-51. Epub 2007 May 1.
101. Aouida M, Poulin R, Ramotar D. The human carnitine transporter SLC22A16 mediates high affinity uptake of the anticancer polyamine analogue bleomycin-A5. *J Biol Chem*. 2010;285(9):6275-84. doi: 10.1074/jbc.M109.046151. Epub 2009 Dec 25.
102. Okabe M, Unno M, Harigae H, Kaku M, Okitsu Y, Sasaki T, et al. Characterization of the organic cation transporter SLC22A16: a doxorubicin importer. *Biochem Biophys Res Commun*. 2005;333(3):754-62.
103. Harris TB, Launer LJ, Eiriksdottir G, Kjartansson O, Jonsson PV, Sigurdsson G, et al. Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. *Am J Epidemiol*. 2007;165(9):1076-87. Epub 2007 Mar 10.
104. Racki LR, Yang JG, Naber N, Partensky PD, Acevedo A, Purcell TJ, et al. The chromatin remodeller ACF acts as a dimeric motor to space nucleosomes. *Nature*. 2009;462(7276):1016-21. doi: 10.1038/nature08621.
105. Lan L, Ui A, Nakajima S, Hatakeyama K, Hoshi M, Watanabe R, et al. The ACF1 complex is required for DNA double-strand break repair in human cells. *Mol Cell*. 2010;40(6):976-87. doi: 10.1016/j.molcel.2010.12.003.
106. Sanchez-Molina S, Mortusewicz O, Bieber B, Auer S, Eckey M, Leonhardt H, et al. Role for hACF1 in the G2/M damage checkpoint. *Nucleic Acids Res*. 2011;39(19):8445-56. doi: 10.1093/nar/gkr435. Epub 2011 Jul 11.
107. Houlston RS, Peto J. The search for low-penetrance cancer susceptibility alleles. *Oncogene*. 2004;23(38):6471-6.
108. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol*. 2012;2012:251364. (doi):10.1155/2012/251364. Epub 2012 Jul 5.
109. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53-9. doi: 10.1038/nature07517.
110. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, et al. Genetic variation in an individual human exome. *PLoS Genet*. 2008;4(8):e1000160. doi: 10.1371/journal.pgen..
111. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65. doi: 10.1038/nature11632.
112. Ng SB, Nickerson DA, Bamshad MJ, Shendure J. Massively parallel sequencing and rare disease. *Hum Mol Genet*. 2010;19(R2):R119-24. doi: 10.1093/hmg/ddq390. Epub 2010 Sep 15.
113. Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods*. 2010;7(4):250-1. doi: 10.1038/nmeth0410-250.

114. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 2005;15(7):901-13. Epub 2005 Jun 17.
115. Johnston JJ, Teer JK, Cherukuri PF, Hansen NF, Loftus SK, Chong K, et al. Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am J Hum Genet.* 2010;86(5):743-8. doi: 10.1016/j.ajhg.2010.04.007. Epub May 6.
116. Litim N, Labrie Y, Desjardins S, Ouellette G, Plourde K, Belleau P, et al. Polymorphic variations in the FANCA gene in high-risk non-BRCA1/2 breast cancer individuals from the French Canadian population. *Mol Oncol.* 2013;7(1):85-100. doi: 10.1016/j.molonc.2012.08.002. Epub Sep 11.
117. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43(5):491-8. doi: 10.1038/ng.806. Epub 2011 Apr 10.
118. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med.* 2010;362(13):1181-91. doi: 10.056/NEJMoa0908094. Epub 2010 Mar 10.
119. Rios J, Stein E, Shendure J, Hobbs HH, Cohen JC. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Hum Mol Genet.* 2010;19(22):4313-8. doi: 10.1093/hmg/ddq352. Epub 2010 Aug 18.
120. Dolled-Filhart MP, Lee M, Jr., Ou-Yang CW, Haraksingh RR, Lin JC. Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. *ScientificWorldJournal.* 2013;2013:730210.(doi):10.1155/2013/730210. Epub 2013 Jan 13.