HELSINGIN YLIOPISTO
BIO- JA YMPÄRISTÖTIETEELLINEN TIEDEKUNTA

Master’s thesis
Genetics

A Case-Control Study of SERPINA3 c.918-1G>C Variant on Breast Cancer Risk in Southern Finland Population

Niina Aho

10.03.2022

Supervisor: Maija Suvanto
Examiner: Heli Nevanlinna

UNIVERSITY OF HELSINKI
FACULTY OF BIOLOGICAL AND ENVIRONMENTAL SCIENCE

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty	Koulutusohjelma – Utbildingsprogram – Degree Programme
Bio- ja Ympäristötieteellinen tiedekunta	Genetiikan ja Molekulaaristen Biotieteiden Maisteriohjelma

Tekijä – Författare – Author

Niina Aho

Työn nimi – Arbetets titel – Title

Tapaus-verrokkitutkimus Serpina3 c.918-1G>C muunnoksen rintasyöpärikin arvioimiseksi Etelä-Suomen väestössä

Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track

Genetiikka ja Genomiikka

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Pro Gradu tutkielma	03/22	35 + 12

Tiivistelmä – Referat – Abstract

Rintasyöpä on yleisin naisilla esiintyvä syöpä ja vuonna 2020 se oli viidenneksi yleisin kuolemaan johtava syöpä. Suomessa diagnosoitiin vuonna 2019 yli 5000 uutta rintasyöpä tapausta, joista 94 % esiintyi naisilla ja 6 % miehillä. Tähän mennessä korkean riskin rintasyöpägeenit kuten *BRCA1*, *BRCA2* ja *TP53* on tunnistettu sekä lisäksi monia keskikorkean riskin geenejä. Silti yhdessä kaikki löydetyt geenit selittävät vain noin puolet familiaarisista rintasyöpätapauksista. Lisäksi kaikki löydetyt rintasyövän alttiusgeenit ovat yhteydessä DNA:n korjausmekanismiin.

Serpina3 erottuu joukosta ei-DNA-korjausmekanismigeeninä vain geeninä, joka koodaa Serpin-superperheeseen kuuluvaa proteaasi-inhibiittoria. Serpina3 on yhdistetty moniin sairauksiin ja erityisesti muutokset sen ilmentymistasossa on liitetty kasvaimen etenemiseen monissa syövissä, myös rintasyövässä. Aikaisempi tutkimus kuitenkin esittää, että *Serpina3* c.918-1G>C variantti on rintasyövälle altistava Pohjois-Suomen väestössä.

Tässä pro gradu- työssä tutkitaan tapaus-verrokkitutkimuksena *Serpina3* c.918-1G>C variantin liittymistä rintasyöpään Etelä-Suomen väestössä. Lisäksi *Serpina3* c.918-1G>C muutoksen kantajien kasvaimen histologia ja solumarkkerit analysoitiin. Tutkimuksessa käytettiin rintasyöpäpotilailta kerättyä DNA:ta sekä DNA:ta veren luovuttajilta ja biopankista. Rintasyöpäpotilaat kattoivat sekä familiaariset että valikoimattomat tapaukset. *Serpina3* c.918-1G>C variantin esiintyvyyttä kartoitettiin genotyypittämällä potilaat ja kontrollit. Genotyypitys tehtiin käyttämällä TaqMan reaaliaikaista PCR:ää ja kantajat varmistettiin Sanger sekvensoinnilla. Tilastollisia testejä käytettiin aineiston analysoinnissa.

Tutkittua *Serpina3* c.918-1G>C varianttia ei löydetty tilastollisesti merkittävästi enemmän ($p>0.05$) rintasyöpätapauksista kuin kontrolleista. Varianttia löydettiin 0.23 % familiaarista tapauksista ja 0.36 % valikoimattomista tapauksista, yhteensä 0.28 % kaikista rintasyöpätapauksista. Yleisyys väestökontrolleissa oli 0.27 %. Kasvaimen histologia oli duktaalinen 73 % kantajilla ja 9 % kasvain oli lobulaarinen. Toisin sanoen kantajien kasvainten histologia mukaili tavallista jakaumaa. Kaikki kantajat olivat HER2 negatiivisia ja kaikki paitsi yksi olivat sekä ER- että PR-positiivisia. Solun jakautumisen markkeriainetta, Ki-67, löydettiin noin puolelta kantajista.

Yhteenvetona tämän tutkimuksen perusteella voidaan esittää, että *Serpina3* c.918-1G>C ei ole rintasyövälle altistava geenimuutos ainakaan Etelä-Suomen väestössä.

Avainsanat – Nyckelord – Keywords

Rintasyöpä, Serpina3, tapaus-verrokkitutkimus

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors

Maija Suvanto ja Heli Nevanlinna

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Faculty of Biological and Environmental Sciences	Koulutusohjelma – Utbildingsprogram – Degree Programme Master's Programme in Genetics and Molecular Bioscience	
Tekijä – Författare – Author Niina Aho		
Työn nimi – Arbetets titel – Title A Case Control Study of Serpina3 c.918-1G>C Variant on the Breast Cancer Risk in Southern Finland Population		
Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track Genetics and Genomics		
Työn laji – Arbetets art – Level Master's Thesis	Aika – Datum – Month and year 03/2022	Sivumäärä – Sidoantal – Number of pages 35 + 12

Tiivistelmä – Referat – Abstract

Breast cancer is the most prevalent cancer in women worldwide and in 2020 it was the fifth deadliest. In Finland 2019 more than 5000 breast cancer cases were diagnosed, 94% in women and 6% in men. Until now, the high-risk breast cancer susceptibility genes have been identified including *BRCA1*, *BRCA2* and *TP53* as well as many of the moderate risk genes. Still, together all the identified genes explain only approximately half of the familial breast cancer cases. Furthermore, all the known breast cancer susceptibility genes are linked to the DNA repair mechanism.

Serpina3 stands out as a non-DNA repair gene but as a gene that encodes a protease inhibitor which belongs to the serpin superfamily. *Serpina3* has been associated with various diseases before and especially changes in its expression levels are linked to the tumor prognosis in many cancers including breast cancer. However, a previous study proposed that Serpina3 c.918-1G>C is a susceptibility variant for breast cancer in the Northern Finland population.

This thesis a case-control study to investigate whether Serpina3 c.918-1G>C variant is associated with breast cancer in the Southern Finland population. In addition, the tumor histology and cellular markers of Serpina3 c.918-1G>C carriers were examined. This study utilized DNA collected from breast cancer patients as well as DNA from blood donors and healthy biobank controls. Breast cancer patients included both familial and unselected cases. The prevalence of Serpina3 c.918-1G<C variant was studied by genotyping the cases and controls. Genotyping was done by TaqMan real-time PCR and carriers were further confirmed by Sanger sequencing. Moreover, statistical tests were used in the data analyses.

The studied Serpina3 c.918-1G>C variant was not found to be significantly ($p>0.05$) enriched in the breast cancer cases. The variant was found in 0.23 % of familial and 0.36 % of unselected cases, altogether in 0.28 % of all studied breast cancer cases, the frequency in controls was 0.27 %. The tumor histology was found to be ductal in 73 % of the Serpina3 c.918-1G>C variant carriers and only 9 % had lobular tumor. In other words, the tumor histology followed the usual distribution. All the carriers had a HER2 negative tumor and all except one case were both ER and PR positive. About half of the carriers expressed the cellular proliferation marker Ki67.

As a conclusion, the results from this study do not suggest Serpina3 c.918-1G>C as a breast cancer risk variant at least in the Southern Finland population.

Avainsanat – Nyckelord – Keywords
Serpina3, Breast cancer, case-control study

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Maija Suvanto and Heli Nevanlinna

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

Contents

1 Introduction 1

2 Cancer 1

2.1 Cancer Epidemiology. 1
2.2 Cancer Biology and Progression. 1
2.3 Cancer Genes. 3
2.4 Hereditary Cancer Susceptibility. 5

3 Breast cancer 7

3.1 Breast Cancer Epidemiology. 7
3.2 Biology of Breast Cancer. 7
3.3 Subtypes and Cellular Markers. 8
3.5 Breast Cancer Susceptibility. 10
3.4 Breast Cancer Susceptibility Genes. 10
3.6 Finnish Founder Mutations. 13

4 Serpina3 15

4.1 Serpina3 Structure. 15
4.2 Cellular Function. 16
4.3 Link to Disease. 17
4.4 Serpina3 918-1G>C. 18

5 Aims of the Study 19

6 Materials and Methods 20

6.1 Patient Series. 20
6.2 TaqMan Real-Time PCR. 22
6.3 Sequencing. 24
6.4 Statistical Analysis. 26

7 Results 27

7.1 Power Analysis. 27
7.2 Fisher Test. 28
7.3 Histology and Cellular Markers. 29

8 Discussion 30

9 Conclusion 33

10 Acknowledgement 34

References 35

Abbreviations

ACT	Alpha 1-antichymotrypsin
ATM	Serine/threonine kinase
BC	Breast cancer
BRCA1	Breast cancer gene 1
BRCA2	Breast cancer gene 2
BSCS	Breast cancer stem cell
c-Abl	Tyrosine protein kinase
CHEK2	Checkpoint kinase 2
CI	Confident interval
ECM	Extracellular matrix
EGFR	Epidermal growth factor receptor
ER	Estrogen receptor
HER2	Human epidermal growth factor receptor 2
IDC	Invasive ductal carcinoma
ILC	Invasive lobular carcinoma
Ki67	Marker of cell proliferation
MAF	Minor allele frequency
MLH1	MutL homolog 1
MMP-9	Matrix metallopeptidase 9
MQ	Milli-Q purified water
MSC	Mammary stem cell
MSH1	DNA mismatch repair protein
OR	Odds ratio
P53	Tumor protein 53
PALB2	Partner and localizer of BRCA2
PAR4	Protease activated receptor 4
PR	Progesterone receptor
PTEN	Phosphatase and tensin homolog
qPCR	quantitative PCR
Ras	Rat sarcoma virus
RB	Retinoblastoma
Rb1	Retinoblastoma gene

RCL	Reactive center loop
SNP	Single nucleotide polymorphism
TGF-B	Transforming growth factor beta
TN	Triple negative
TP53	Gene that codes tumor protein 53

Cancer

Cancer epidemiology

Cancer is the number one worldwide cause of death in 2020 (WHO, 2020). It includes a large group of diseases that can affect almost any part of the human body. Cancer is a result of a group of abnormally proliferating cells which no longer follow the rules of normal tissue maintenance. Instead, they learn to thrive with one focus: to develop more copies of themselves. Further, they have a tendency to evolve to invade other body parts and to metastasise distant organs (Weinberg, 2014, p.32). In 2020 cancer caused over 10 million deaths worldwide and most new cancer cases were diagnosed as breast, lung and colon cancers. In general, one third of all deaths is caused by tobacco use, obesity, low physical activity or alcohol use (WHO, 2020). In Finland, approximately 35000 people are diagnosed with cancer each year, the highest diagnostic rate being in lung and prostate cancer in men and in breast cancer in women. However, over 90% of diagnosed cases are alive after five years from primary diagnosis (Finnish Cancer registry, 2021).

Cancer biology and progression

Cancer is a multi-step disease or a family of diseases that affect higher multicellular organisms. In general, cancer is defined as abnormal cell growth caused by a variety of changes in gene regulation resulting in unbalanced cell proliferation and death, ultimately leading to a tissue invasion and metastasis of distant organs (Weinberg, 2014).

Over one billion years of evolution multicellular organisms have developed an effective and sophisticated control machinery, that tightly regulates cell proliferation inducing apoptosis to cells that escape it. All cells in human body follow the highly regulated cell cycle which contains a growth phase (G1), DNA synthesis (S), second growth phase (G2) and mitotic phase where cells divide (Weinberg, 2014). For cells to escape the machinery spesific skills are required. For normal cells to turn into a malignant cancer cell requires six hallmark capabilities, each one of them helping to shut down or bypass the control mechanism. These hallmarks (Figure 1.) include self-sufficiency of growth signals, insensitivity to anti-growth signals, evading cell death, limitless replication, enhanced angiogenesis and ability to tissue invasion and metastasis (Weinberg et al. 2000).

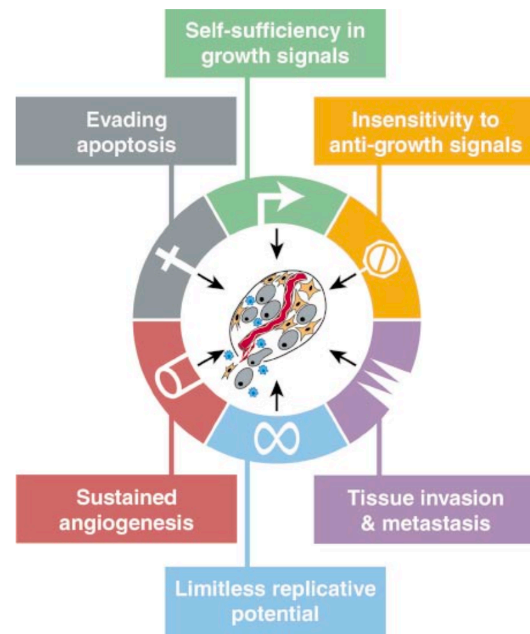

Figure 1. Here is presented the known six hallmarks of cancer. Image: Weinberg et al. 2000

Acquired growth signals was the first discovered capability in cancer cells. When normal cells are unable to proliferate without these signals cancer cells instead produce their own growth signals and are not dependent on the stimulation coming from normal tissue. This way they liberate themselves from exogenous signals and create a positive feedback loop called autocrine signaling. Moreover, cancer cells have a tendency to overexpress the cell surface receptors that transduce the signals into the cell leading to hyperresponsiveness of growth signals. Anti-growth signals are equally important as growth signals to maintain tissue homeostasis. Anti-growth signals are mostly soluble growth inhibitors and immobilized inhibitors on the cell surface. These signals operate with the cell cycle clock and can stop cells from proliferating in two ways: to push them from G1 phase to G0 state, which is quiescent, or in the case of highly specific cells e.g., neuronal cells, they are pushed to post mitotic state. In both scenarios the cells no longer proliferate. For cancer cells it is crucial to evade these signals. Most cancers have mutations in retinoblastoma (RB) protein and p53 protein that give the cancer cells the advantage to avoid the G1/S checkpoint and hence keep proliferating. In the cell cycle RB transduces extracellular growth inhibition signals into the cell and is responsible for the G1/S checkpoint. P53 instead receives input from inside the cell, determining if the genome of the cell is intact enough for further steps; if not, p53 can induce apoptosis (Weinberg et al. 2000 and Weinberg et al. 2011).

The third major hallmark of cancer cells is their ability to evade apoptosis. The apoptosis machinery is roughly divided into two elements: sensors and effectors. Sensors monitor both intra- and extracellular environment and effectors execute the apoptosis. The main strategy of cancer cells to evade apoptosis is by mutating the *TP53* tumor suppressor gene. As mentioned in the previous paragraph p53 can induce apoptosis in case of DNA damage or hyper proliferation. By eliminating p53, cancer cells eliminate the crucial damage sensor (Weinberg et al. 2000 and Weinberg et al. 2011).

For cancer cells to be able to form a tumor they need to pass the replication barrier. When culturing cells it can be noticed that in normal cells repeated replication cycles lead to senescence which is a viable non-proliferating state. Tumor cells instead have the ability to replicate infinitively, and they are not restricted by cell-cell contact inhibition like non-cancerous cells (Hartwell, 2014, p.632). What gives cancer cells this advantage can be found in their telomeres. Telomeres are tandem hexanucleotide repeats found in the end of chromosomes; in normal cells they shorten after every replication until they fail to protect the chromosome's end leading to apoptosis. Cancer cells express high amount of the special DNA polymerase which adds repeats to the end of telomeres. By extending the telomeres the cancer cell has the advantage to continue to replicate infinite number of times. Despite the ability of a cancer cell to proliferate in high rate or replicate indefinitely to keep going, it needs oxygen and nutrients just like normal cells (Weinberg et al. 2000 and 2011). The growing need of oxygen and nutrition are fulfilled by tumor associated angiogenesis meaning the formation of new blood vessels. In normal human adult the formation of new blood vessels is quiescent and only activated in case of physical changes such as pregnancy. Cancer cells on the other hand have the ability to keep the angiogenesis going in order to get enough nutrients (Nishida et al. 2006). When cancer progresses a part of the primary tumor will invade tissues nearby as well as metastasize to distant organs by circulation. Escaping the primary tumor site enables those cancer cells to find a new location where there is plenty of nutrients and initially very little competition (Weinberg et al. 2010).

Cancer genes

All cancer cells contain several abnormal sets of chromosomes with translocations, deletions and insertions causing genome instability. Genome instability is in most cases due to the deficiency of DNA repair mechanism. In addition to genetic changes cancer cells go through epigenetic changes like chromatin remodeling, DNA methylation and histone modifications which affect the gene expression by either activating or silencing it without altering the DNA sequence (Strachan et al. 2015). Each individual cancer patient has a different set of mutations and epigenetic changes. However, several cancer-critical genes have been identified. The cancer-critical genes are further divided into two groups: oncogenes and tumor suppressor genes depending on whether they drive cancer progression by gain- or loss of function mutation (Alberts et al. 2008).

Oncogenes are mutated genes that act in a dominant way in cancer, since the effect is dominant only one mutated allele is sufficient to cause cancer phenotype or progression. Before mutation oncogenes

are called proto-oncogenes which typically encode proteins for cell cycle. In normal cell proto-oncogenes keep the cell cycle going further and help for its progression. When a proto-oncogene is mutated and turns into an oncogene the cell cycle progression is accelerated (Strachan et al. 2015). This is due to the mutation causing an increase in the gene expression causing the cell to complete more mitotic cycles than normally. This is called a dominant gain-of-function mutation. The underlying mechanism in conversion from proto-oncogene to oncogene can include the following: point mutation, insertion or deletion and translocation as well as gene amplification that results in extra chromosomal copy of the proto-oncogene. The most common proto-oncogenes are *Ras* which encodes an intracellular protein, *c-Abl* fused with *bcr* gene which together encode a hybrid protein and *EGFR* which encodes a transmembrane receptor (Hartwell et al. 2015).

Tumor suppressor genes instead act in recessive way in cancer progression i.e., both copies in a cell need to be mutated to cause an abnormal cancerous phenotype. Typically, these mutations are either point mutations or small deletions (Alberts et al. 2008, p.1109). Two normal copies of a tumor suppressor gene encode a protein which serve as protector against genome instability or slows down the cell cycle progression. Mutations in tumor suppressor genes are loss of function mutation resulting in faster proliferating cell and higher mutation accumulation rate. The two most known tumor suppressor genes are *RB1* and *TP53* which both act on cell division and cell death (Hartwell et al. 2015, p. 648). Further, tumor suppressor genes can be divided into two subgroups: gatekeepers and caretakers (Levitt et al. 2002). Gatekeepers are genes that act directly as rate limiters for tumor growth by regulating cell proliferation or death. Caretakers on the other hand act on tumor initiation indirectly by maintaining genome stability. A mutated caretaker gene promotes genetic instabilities and accelerate the normal cell to neoplastic cell conversion. The previously mentioned *RB1* is an example of gatekeeper gene and the breast cancer genes *BRCA1* and *BRCA2* that function in DNA repair are examples of caretaker genes (Kinzler et al. 1997 and Levitt et al. 2002).

Cancer starts by a single somatic or inherited mutation that gives the cell for example a growth advantage or disturbs its DNA repair system. A second mutation in another key gene followed up by more mutations (preferably 6-12) creates a malignant cancer cell. This one malignant cell serves as progenitor for the cancer cell lineage (Hartwell et al. 2015). The mutations which give cells a growth advantage are called driver mutations and result from positive selection in the cancer evolution. These mutations are usually behind the proto-oncogene to oncogene switch. The rest of the mutations which do not contribute to the development of the cancer but are harbored by the progenitor of cancer cell clones are called passenger mutations (Stratton et al. 2009).

Hereditary cancer susceptibility

In most cases cancer results from several somatic mutations acquired by time, however a small proportion (5-10%) is due to inherited germline mutation. Inherited germline mutation does not directly mean that the person will have cancer, but it increases the risk of a tumor formation. These inherited mutations are mostly in tumor suppressor genes and can be found either in gatekeepers or caretakers (Negrini. 2010). Gatekeeper genes with one inherited mutation need only one additional somatic mutation to result in the loss of function of the gene and to initiate tumor formation. This is Knudson's two hit theory and based on that individuals who inherit one mutated gatekeeper gene have higher risk (>10^3) to form cancer than individuals who need two somatic mutations (Knudson. 2011). On the other hand, inherited mutated caretaker gene which have indirect effect on tumor progression requires tree additional mutations: mutated second caretaker allele as well as both mutated gatekeeper alleles. This leads to 5-50-fold increase compared to non-inherited cases (Kinzler et al. 1997).

It has been proposed that most of the inherited germline mutations are on the caretaker genes and that hereditary cancer evolves from genomic instability (Kinzler et al. 1997 and Negrini et al. 2010). Moreover, genomic instability in hereditary cancers further drives the accumulation of key mutations in oncogenes and tumor suppressor genes enhancing the acquisition of other hallmarks of cancer. As a comparison the general theory presents that in sporadic cancers the first mutation would rather happen in oncogenes favoring the activation of growth signals which would lead to accelerated proliferation and further to DNA damage and genomic instability (Negrini et al. 2010). In general, sporadic cancer rarely results from mutated caretaker gene since it would require altogether four mutation: both caretaker alleles and both gatekeeper alleles (Kinzler et al. 1997).

In sporadic cancer and hereditary cancer several differences can be spotted including patient's young age, several sick relatives (often expressing more than one type of cancer) and finding two or more primary tumors in the same person. However, in some cases the inherited susceptibility can be difficult to determinate for example when a mutation *BRCA1* or *BRCA2* breast cancer gene is inherited through the males in the family. All of the known germline mutations have dominant inheritance pattern and show partial penetrance. Partial penetrance means that not all carriers get cancer, but that it requires harboring other additional key somatic mutations as well (Lääketieteellinen genetiikka, 2016).

Several genes have been linked to hereditary cancers. The most well-known ones are *BRCA1* and *BRCA2* which are breast cancer susceptibility genes. Another well-known hereditary cancer linked genes are *MLH1* and *MSH2* which are both found in Lynch syndrome and which participate in mismatch repair (Negrini et al. 2010 and Strachan et al. 2015, p.397).

Breast cancer

Breast cancer epidemiology

In 2020 over two million new breast cancer cases were diagnosed making breast cancer the most common cancer (11,7%) worldwide and fifth most deadly. Breast cancer is mostly found in women, but it can affect men as well although it is very rare (0,5-1% of all the cases) (Global cancer observatory, 2020). In Finland in 2019 more than 5000 new cases were diagnosed. The overall prognosis of breast cancer cases between 2017-2019 was high with 91 % survival rate. The survival rate was measured five years after initial diagnosis (Finnish cancer registry, 2019).

Breast cancer biology

Mammary gland is built of two components: parenchyma and stroma (figure 2). Parenchyma is a glandular tissue that consists of lobules and ducts, lobules function as milk storage and ducts as tubes when milk is transported to the nipple. Stroma is the supportive tissue around parenchyma and in mammary gland it is mostly composed of adipose tissue but also of extracellular matrix (ECM), blood vessels and fibroblasts (Wiseman et al., 2002).

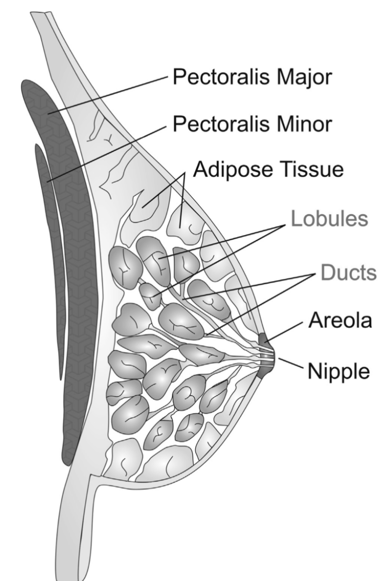

Figure 2. Representative picture of human mammary gland and its anatomy. All the main structures of the breast are shown. Picture: Feng et al. 2018)

The human mammary gland originates during embryonic development and is built of epithelial tubes with a branched network type of structure embedded within a stroma (Huebner et al., 2014). During development, the mammary gland experiences a series of changes in its shape, size, and function which are strongly associated with the following key developmental time points: puberty, pregnancy and lactation. In order to undergo the needed morphological changes, the mammary glands are highly responsive to hormone signaling (Feng et al., 2018).

Two different theories of breast cancer formation have been proposed: sporadic clonal theory and breast cancer stem cell (BCSC) theory. The sporadic clonal theory suggests that any epithelial breast cell might be the target of epigenetic changes or random mutations and that the cell with an advantageous mutation will thrive due to natural selection (Bombonati et al., 2010). The breast cancer

stem cell theory on the other hand suggests that cancer initiates from a small population of undifferentiated cells, for instance mammary stem cells (MSC) or progenitor cells (Sin et al., 2017). This theory is supported by the similar phenotypic features of MSC and BCSC, similar cellular markers and a long-life span and thus enough time to accumulate multiple genetic alterations. MSC and BSCS also share the important ability to self-renewal and differentiation (Reyay et al., 2001 and Smalley et al., 2003). Furthermore, a study where human breast cancer cells were grown in immunocompromised mice revealed that only cancer cell population expressing spesific cellular markers ($CD44^{+}/CD24^{-/low}/lineage^{-}$) were able to form tumors. Tumors initiated from this cell population were heterogenic and contained a large diversity of different cell types. Interestingly, the key features of this cell population, for instance self-renewal and differentiation, are the same as those in MSC which further supports the BSCS theory (Al-Hajj et al., 2003).

Breast cancer subtypes

Breast cancer is a heterogeneous disease as it can present in any part of the mammary gland. The broadest classification of breast cancer is based on the affected cell type being either carcinoma or sarcoma. Carcinoma originates from the epithelial cells (e.g., ducts and lobules) and sarcoma from stromal components which makes it a much rarer breast cancer type (Schatten, 2013).

Carcinomas are divided into non-invasive and invasive breast cancer. The non-invasive ductal carcinoma in situ (DCIS) is one of the most common types of breast cancer. DCIS originates inside the ducts and it has a high potential to develop into an invasive breast cancer (Breastcancer, 2020). The other non-invasive breast cancer is lobular carcinoma in situ (LCIS). In LCIS, abnormal cells are growing on the lining of the lobules without spreading further to the lobules. LCIS is not classified as a breast cancer but it increases the risk of invasive breast cancer development later on. Invasive breast cancer has the characteristic ability to spread outside of lobules and ducts and evade the outer tissue. The group is divided into two: invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC). IDC is the most common breast cancer type covering up to 80% of all diagnosed cases. IDC originates from the ducts and it consists of a few distinguishable subtypes such as for instance medullary carcinoma, mucinous carcinoma, tubular carcinoma and papillary carcinoma (Feng et al., 2018 and Schatten, 2013). ILC instead develops from the lobules in the mammary gland and is the second most common breast cancer type (15%) and is usually found in women over 60 years (Orvieto et al., 2008).

Lobular and ductal carcinomas are separated by the location of the tumor and further divided into more spesific molecular subtypes. The main molecular subtypes are Luminal A breast cancer, Luminal B breast cancer, HER2 enriched breast cancer and triple negative breast cancer (Prat et al. 2015). The difference is based on the distinct biomarker expression of estrogen receptor (ER), progesterone receptor (PR), growth hormone marker (HER2) as well as proliferation marker (Ki-67) (Table 1). These biomarkers are relevant because their existence/absence guide the chemotherapy response as well as prognosis in patient with hormone receptor positive tumor (Colomer et al., 2017). The luminal A breast cancer is hormone receptor positive (ER and PR) and has the highest prevalence as well as the best prognosis. The 5-year survival rate is approximately 85 %. Luminal B instead has lower prevalence and an intermediate prognosis, the 5-year survival rate is roughly 80 %. The main difference between Luminal A and B is that type B expresses high levels of proliferation marker Ki-67. The Her2 subtype has a poor prognosis due to its negative characteristic to the hormone receptors however, it is positive for the growth hormone marker HER2. Triple negative breast cancer is more prevalent in younger women and slightly more common than the HER2 subtype. TN breast cancer also has the poorest prognosis out of all the subtypes. Poor prognosis results from the complete lack of the biomarkers which makes the tumor unsensitive to the endocrine therapy as well as HER2 treatment. (Eroles et al., 2011 and Feng et al., 2018).

Table 1. Here, the different subtypes, molecular profiles, prevalence and prognosis of breast cancer are categorizer

Subtypes	Molecular Profile	Prevalence	Prognosis
Luminal A	ER+, PR+, HER-, Low Ki-67	60–70 %	Good
Luminal B	ER+, PR+, HER+/-, High Ki-67	10-20%	Intermediate
HER2	ER-, PR-, HER+	5-15%	Poor
Triple Negative	ER-, PR-, HER-	15-20%	Poor

Clinically, breast cancer is also evaluated based on TNM system (**t**umor size, spreading to the lymph **n**odes and **m**etastasis). Based on this system breast cancer is divided into different stages (-0, I, II, III, IV) the stage 0 represents a small tumor which has not spread to the lymph nodes or metastasized. The stage IV however means large, already metastasized tumor. This way the TNM system also correlates with the survival of the patients (Kalli et al., 2018 and Cancer Society of Finland).

Inheritance of breast cancer

Most of the breast cancer cases (80%) are sporadic which means that they arise spontaneously due to somatic mutations. The rest (15-20%) of all breast cancer cases are familial. In general, familial cases have at least one first- or second degree relative with breast cancer, the more relatives with the disease the higher the individual's risk becomes (Wendt et al., 2019). The most important high risk breast cancer susceptibility genes are *BRCA1*, *BRCA2, TP53* and *PTEN.* The heritance of high-risk genes e.g., *BRCA1* and *BRCA2* increases the risk to develop other cancers for example ovarian and prostate cancer (Jonsson et al., 2019). In familial cases, the individual is usually diagnosed at a younger age and the chance for bilateral breast cancer is greater than in sporadic cases (Wendt et al., 2019).

Breast cancer susceptibility genes

3.1 High risk breast cancer susceptibility genes

BRCA1 and *BRCA2* are the most prevalent familial breast cancer genes. *BRCA1* and *BRCA2* are tumor suppressor genes with wide involvement in DNA repair system, genome instability and checkpoint control. Both breast cancer susceptibility genes were discovered almost at the same time, *BRCA1* in 1994 and *BRCA2* in 1995. *BRCA1* is located in chromosome 17q21 and *BRCA2* in chromosome 13q12.3. Up until now, over 1500 dominantly inherited mutations in BRCA1/BRCA2 have been found and associated with cancer (Rebbeck et al., 2018 and Abu-Helalah et al., 2020). High number of the *BRCA1* and *BRCA2* alleles have loss of function mutations of which the majority leads to a truncated protein. (Yoshida et al., 2004).

Together BRCA1 and BRCA2 are responsible for approximately 20-40% of all the familial breast cancer cases and the estimated cumulative risk to develop breast cancer by the age of 80 is 72% for *BRCA1* and 69% for *BRCA2* (Coignard et al., 2021). In addition, the *BRCA1* or *BRCA2* mutation carriers have roughly 11-fold increased relative risk to develop breast cancer when compared to general population (Easton et al., 2015). The prevalence of *BRCA1* and *BRCA2* mutations varies depending on the population. In Finland the BRCA1/BRCA2 mutations are found in 21% of familial breast cancer cases. (Vehmanen et al., 1997 and Eerola et al., 2000). In males, the mutations in *BRCA1* gene are rarely responsible of breast cancer, however, mutations in *BRCA2* gene are found in 8% of all male breast cancer cases and in 44% of the familial cases in Finland (Syrjäkoski et al., 2014).

The histopathology of *BRCA1* and *BRCA2* associated breast cancer show different characteristics. Individuals with *BRCA1* tumors are usually diagnosed with grade tree ductal carcinomas (74%) or with medullary subtype features such as high mitotic count. In addition, *BRCA1* tumors are associated with early onset triple negative breast cancer. *BRCA2* tumors show less distinguishable characteristics, often resembling sporadic tumors. Nevertheless, *BRCA2* tumors are more likely to have a high grade and reduced tubular formation (Spurdle et al., 2014 and Honrado et al., 2006).

Germline mutation in *TP53* has been found to be strongly associated with a rare hereditary autosomal dominant disorder called Li-Fraumeni syndrome (LFS) (Wendt et al., 2019). Individuals with LFS have a very high risk for breast cancer, leukemia, osteosarcoma and tissue sarcoma. The risk for women is nearly 100% and for males approximately 75%. Breast cancer is the most common tumor among LFS patients carrying *TP53* mutations and the risk by the age of 60 in women is as high as 85%. The median age for breast cancer onset is 34 years (Schon et al., 2018). In breast cancer patients the mutations in *TP53* are associated with poor prognosis and most of the tumors are either TN or HER2 positive (Huszno et al., 2018).

PTEN gene encodes a phosphatase involved in P13K/AKT-mTOR intracellular signalling pathway that regulates the cell cycle, *PTEN* is also classified as a tumor suppressor gene (Wendt et al., 2019). Mutations in *PTEN* are linked to PTEN Hamartoma Tumor syndrome that includes a spectrum of heritable disorders: Cowden syndrome (CS), Proteus and Proteus-like syndrome, and Bannayan-Riley-Ruvalcaba syndrome. CS is caused by a germline mutation in *PTEN* inherited in an autosomal dominant manner (Ngeow et al., 2017). Patients with CS have 67 to 77% increased cumulative risk to develop breast cancer by the age of 60. In contrast, in the general population, the risk to develop breast cancer is 12% (Bubien et al., 2013, Niuwenhuis et al., 2014). A meta-analysis study (Li et al., 2017) discovered that breast cancer patients with the loss of *PTEN* in the tumor had also often lymph node metastasis, larger tumor size and were diagnosed with triple negative breast cancer.

3.2 Moderate risk breast cancer genes

The moderate-risk variants possess a relative risk of 2-4 to develop breast cancer, with minor allele frequency 0.005-0.01 (Figure 3). The first identified moderate risk breast cancer gene was *CHEK2*. *CHEK2* encodes checkpoint kinase 2 which is involved in cell cycle arrest, DNA repair and apoptosis, in addition it phosphorylates other tumor suppressor genes for instance *TP53* and *BRCA1*. *CHEK2* is

a tumor suppressor gene, and it was first discovered in LFS patients, its protein-truncating mutation c.1100delC has been later associated with increased breast cancer risk in familial patients. The breast cancer risk is approximately three-fold higher among patients with family history, and the cumulative risk to develop a breast cancer by the age of 70 is 37% (Weischer., 2007). In Finland, 12.1 % of the CHEK2 mutation carriers were also associated with bilateral breast cancer (Vahteristo et al., 2002). Similar results were found later in the Italian population where 41.2% of the *CHEK2* mutations carriers were associated with a high rate of bilateral breast cancer. Moreover, majority of the tumors were found to be luminal A (Toss et al., 2021).

PALB2 gene encodes a protein known as **p**artner **a**nd **l**ocalizer of **B**RCA**2** that functions in genome maintenance by interacting with BRC2. *PALB2* is also associated with breast cancer. In Finland, the variant c.1592delT is found in 1% of the breast cancer cases causing a four-fold increased risk to develop breast cancer (Erkko et al., 2007). In a meta-analysis the relative risk for *PALB2* mutation carriers was revealed to be 5.3. Furthermore, the risk to develop breast cancer by the age of 70 is 33% for female *PALB2* mutation carriers with no family history and 58% with family history (Antoniou et al., 2014 and Wendt et al., 2019).

The *ATM* gene encodes a protein serine/threonine kinase that functions in DNA repair, cell growth control and phosphorylation of other proteins such as *BRCA1*, *TP53* and *CHEK2*. Several biallelic mutations found in Finland in *ATM* gene are associated with autosomal recessive disease ataxia-telangiectasia (AT) and sensitivity to ionizing radiation (Allinen et al., 2002). Heterozygote mutations in *ATM* gene are also associated with breast cancer (Renwick et al., 2006). Familial breast cancer patients heterozygous for the *ATM* mutations were found to develop breast cancer with threefold increased risk compared to the general population. In addition, *ATM* mutation carriers were also found to be associated with bilateral breast cancer (26.3 %) and the most common tumor type was luminal B/HER2 negative (Toss et al., 2021).

Other moderate breast cancer susceptibility genes are *BARD1*, *RAD51C*, *RAD51D* and *FANCM*. According to recent studies the tree former genes are especially highly associated with triple negative breast cancer. The increased risk to TN breast cancer for *BARD1* is 5.92-8.15, for *RAD51D*: 6.97-7.35 and for *RAD51C*: 2.64-5.25 (Shimelis et al., 2018 and Breast Cancer Association Consortium, 2021). FANCM has also been suggested to associate with triple negative breast cancer. Studies made in Finland have shown an increased association of two FANCM variants (c.5101C>T and c.5791C>T) with TN breast cancer. The combined analysis of c.5101C>T and c.5791C>T variants

suggested a tree fold increased breast cancer risk in the TN subgroup (Kiiski et al., 2014 and Kiiski et al., 2017). However, the precise, mutation specific risks remain to be established. In the recent breast cancer association consortium study FANCM mutations showed also some association with ER negative breast cancer (Breast Cancer Association Consortium, 2021).

3.3 Low penetrance breast cancer variants

The low penetrance variants associated with breast cancer are usually SNPs with minor allele frequency >0.05 and relative risk below 1.5 (Figure 3) (Wendt et al., 2019). The low penetrance variants are identified in genome-wide association studies (GWAS) where DNA from large set of breast cancer patients is compared to healthy individuals. Many of the already found SNPs function in breast cancer cell growth, mammary gland development and DNA repair (Ghoussaini et al., 2012). In total, it has been calculated that all low penetrance variants account for 18 % of familial relative breast cancer risk (Michailidou et al., 2017).

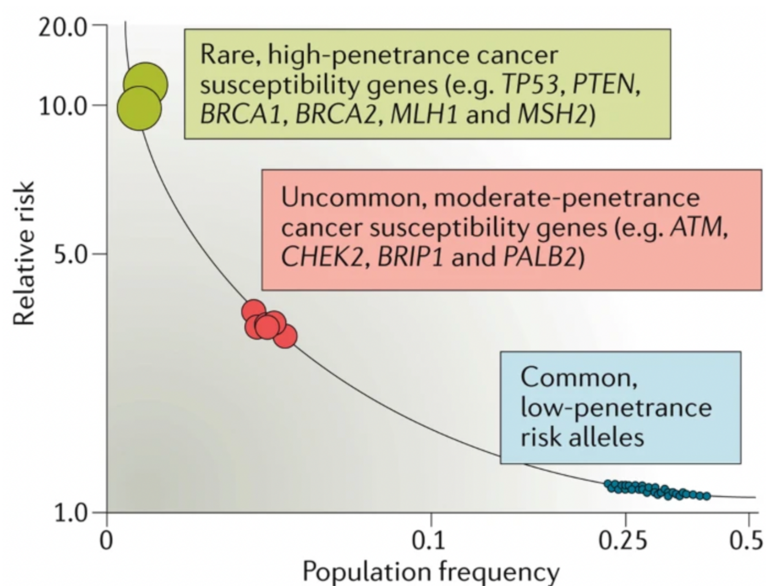

Figure 3. The picture shows the relative risk and population frequency for the tree breast cancer gene-categories. The high-risk genes are illustrated as green and have high relative risk. In red is marked the moderate genes with relative risk between 2-4 and on blue the low penetrance variants. Picture: Sud et al., 2017

Finnish founder mutations

Finland is an isolated population with a unique genetic background shaped by small population size and bottlenecks. The population is more homogeneous than many other populations and is hence enriched for some disease-causing variants while others have disappeared (Kääriäinen et al. 2017). The separately expanded Western and Eastern populations in Finland harbour rare deleterious variants alleles which show 20-times increased frequency when compared to other countries (Locke et al. 2019). Moreover, recent fine-scale mapping of population of Finland revealed genetic subgrouping throughout the whole country proposing strong genetic differences defined by region. The regional differences showed concordance with Finland's population history (Kerminen et al. 2017).

Only a handful of Finnish founder mutations in breast cancer susceptibility genes (*BRCA1/2*, *CHEK2* and *PALB2)* have been found. Several mutations in *BRCA1* and *BRCA2* genes have been identified but only few are exclusive to Finland. Moreover, in Finland the *BRCA1* and *BRCA2* mutations show variation in allele frequencies between different subpopulations due to the geographical clustering of the founder mutations. One of the unique Finnish founder mutations is the c.4097-2A>G variant in *BRCA1* gene which originates from the Ostrobothnia/Central Finland region and is the most common *BRCA1* variant in the country. In *BRCA2* gene the variant c.7480C>T is unique for Finnish population. However, the most prevalent *BRCA2* variant c.771_775delTCAAA is also found in high frequencies in Iceland (Sarantaus et al. 2000).

In *PALB2* gene the c.1592delT variant has been identified as Finnish founder mutation. The one nucleotide deletion forms a premature stop codon which leads to a greatly increased breast cancer risk, more specifically a tenfold increased risk in familial cases and fourfold in unselected cases. The *PALB2* c.1592delT variant explains roughly 1% of unselected breast cancer cases in Finland when in comparison *BRCA1/2* together explain around 1,8% (Erkko et al. 2007). The *CHEK2* variant c.1100delC is a frameshift mutation which is most frequently found in Finland and Netherlands. In Finland a heterozygous mutation of c.1100delC is associated with a roughly four-fold increased breast cancer risk in both females and males (Hallamies et al. 2017 and Vahteristo et al. 2002).

Serpina3

Serpina3 (Serpin peptidase inhibitor A member 3) protein is also known as a1-antichemotrypsin (ACT) and belongs to the serpin superfamily and functions as serpin protease inhibitor (Kelly-Robinson et al. 2021). *Serpina3* is located on a q-arm of chromosome 14 more specifically at 14q32.1 and it is part of the serpin gene cluster. Serpina3 acts as acute-phase protein during inflammation and its expression has been linked to several human diseases including Alzheimer's disease and cystic fibrosis as well as liver, prostate and colon cancers (Baker et al. 2007). Moreover, several studies suggest an association between single nucleotide polymorphisms (SNPs) in *Serpina3* gene and altered gene expression, possibly contributing to a disease. (Chelbi et al. 2011). SERPINA c.918-1G>C germ line variant has been suggested as a breast cancer susceptibility allele in a study from Northern Finland (Koivuluoma et al., 2020). In this thesis a link between *Serpina3* c.918-1G>C and breast cancer risk is studied in a larger case-control study in Southern Finland population to screen the prevalence of the variant among breast cancer patients and controls.

Structure

Serpina3 has a typical serpin structure: three beta sheets, eight alpha helixes and active site with reactive center loop (RCL) located on the top of the protein structure (Sanrattana et al. 2019). The reactive loop is critical for Serpina3 activity and even one nucleotide changes are known to affect its function (Sánchez-Navarro et al., 2021). Serpina3 keeps the native conformation until a serine protease binds to the RCL triggering dramatic changes in the serpin structure. The serine protease binding can induce two scenarios: Serpina3 to rapidly insert its RCL into the enzyme's body as antiparallel beta sheet while covalently attaching protease to itself or, if RCL insertion process is slow, the protease detaches resulting in an inactive Serpina3 protein and still active protease. The covalently bound protease-Serpin complex is further removed from the plasma by the liver. Both of these cases cause irreversible inhibition, meaning that Serpina3 is no longer able to inhibit other serine proteases due to the large conformational change which cause disruption of the protease active site.

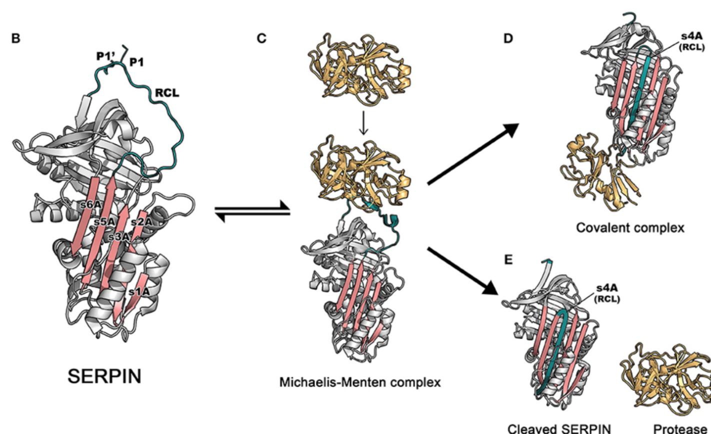

Figure 4. Serpina3 inhibition function illustrated. In B, Serpina3 is in the native conformation. In C, the substrate is binding to the RCL. In D, the protease is covalently linked (inhibited) into Serpina3 protein after a rapid structural change. In E, the process of inserting RCL into the enzyme to inhibit protease has been slow resulting detachment of the protease and an inactive Serpina3 protein.

In addition, Serpina3 has DNA binding activity which is known to be separate from its protease inhibition activity, this character is unique and not found from other serpins. However, so far it is unknown how Serpina3 utilizes the DNA binding ability (Kelly-Robinson et al. 2021 and Baker et al. 2007).

Cellular Function

Serpina3 is part of the serpin superfamily containing roughly 37 different serine protease inhibitors which all have a nucleophilic serine in the active site. The encoded Serpina3 protein is secreted into the circulation by liver where it functions as acute-phase protein inhibiting serine proteases during both chronic and acute inflammation. As acute-phase protein Serpina3 has several roles in maintaining body homeostasis; it regulates a variety of biological processes including wound healing, coagulation and inflammation (Fig. 5). Protease targets for Serpina3 are mast cell chymase, pancreatic chymotrypsin and human glandular kallikrein 2 and 3 which are prostate spesific antigens. However, the main inhibitory target is Cathepsin G which is a pro-inflammatory enzyme produced by neutrophils (Baker et al. 2007).

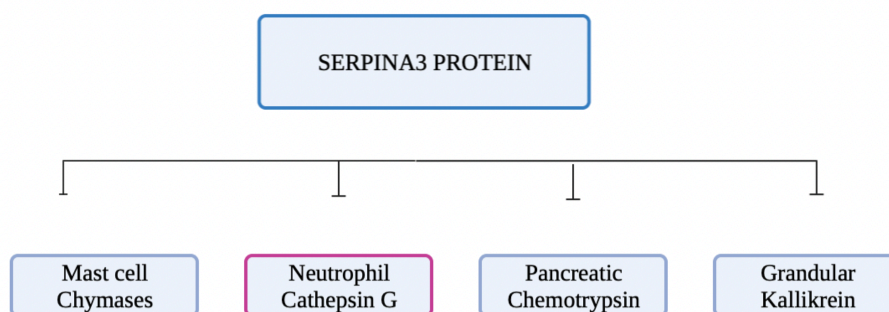

Figure 5. Here is a summary of Serpina3 main cellular function. Serpina3 regulates mast cell chymase, Cathepsin G, pancreatic chymotrypsin and grandular kallikrein. From these the red circled Cathepsin G is the main cellular target.

Cathepsin G is released at the inflammation site and it participates for instance in wound repair by activating matrix metallopeptidase (MMP-9) which degrades ECM as well as release growth factors (Wilson et al. 2009). Additionally, MMP-9 is capable of activating latent TGF-B (transforming

growth factor beta) which is a key player in breast cancer metastasis (Moore-Smith et al. 2017). Cathepsin G directly interacts with protease-activated receptor (PAR4) on the surface of platelets inducing aggregation and by recruiting more neutrophils at the site. Moreover, Cathepsin G possesses a role in pathogen degradation, and it has a pro-apoptotic activity (Siming et al. 2018 and Meyer-Hoffert et al. 2005). Based on these by regulating Cathepsin G Serpina3 plays a role in apoptosis, downregulation of inflammation, coagulation and ECM remodeling.

Link to human diseases

Serpina3 has a wide range of functions in the cell and its normal activity is important. Imbalance in equilibrium between Serpina3 and target proteases has the potential to lead to a disease. Mutations causing the deficiency of Serpina3 result in excess of target proteases effecting downstream pathways. On the other hand, Serpina3 overexpression leads to over-inhibition of target proteases which disturbs the cellular downstream pathways. So far, Serpina3 has been linked to many cancers and its role has been studied from many different aspects. The most studied features are *Serpina3* expression levels in different tumors as well as its SNPs. (Baker et al. 2007)

Several papers indicate that *Serpina3* expression levels correlate with tumor prognosis and metastasis. Overexpression of Serpina3 has been detected for example in endometrial, colon, breast and placental tumor cells. However, the literature shows inconsistent and contradictory results about the effect of *Serpina3* expression levels in tumor prognosis (Zhou et al. 2019). Recently, a study of colon cancer reported that high level of Serpina3 is linked to clinicopathological features and correlates to migration and invasiveness of cancer cells promoting metastasis (Cao et al. 2018). Similar evidence was shown in a study about triple negative breast cancer where highly expressed *Serpina3* was associated with earlier relapse and poorer prognosis. The study was based on TGF-B derived signature (Katayama et al. 2018). On the other hand, a study about invasive breast cancer proposed that the mRNA levels of Serpina3 could be used as a marker of good prognosis. The study provided evidence that the higher expression levels are linked to better survival in patients with HR+ or PR+ breast cancer (Yamamura et al. 2004).

Effects of SNPs on *Serpina3* function has been studied as well, but not as much as its expression levels. SNPs on *Serpina3* gene have been linked to a placental disease, skin pustules and Alzheimer. It has been shown that a SNP in the *Serpina3* promoter region (rs1884082) corresponds clearly to the basal expression level of the gene which is two times higher with TT genotype compared to GT.

Further, the genotype TT was linked to an incidence of vascular intrauterine growth restriction (IUGR) in pregnant women (Chelbi et al. 2011). Previously, two SNPs on *Serpina3* (rs373526796 and rs771543687) have been identified which both leads to a premature stop codon formation and lower levels of Serpina3. A clinical case study by Kantaputra et al., linked the lower expression levels of *Serpina3* to skin pustule formation and study by Frey et al., connected the lowered expression to general pustular psoriasis (Frey et al. 2020 and Kantaputra et al., 2021). Furthermore, extensive studies in Alzheimer disease have suggested that expression pattern of several SNPs should be studied to provide a clear proof of Serpina3 role in disease formation (Baker et al. 2007).

Serpina c.918-1G>C variant

The studied Serpina3 c.918-1G>C SNP variant shuts down the valid splice acceptor of exon 4 and is predicted to form a new one right next to the original one (Fig 6). The one nucleotide change from G to C results in total of two nucleotide deletion and a frameshift mutation leading to a premature stop codon (p.Arg306ArgfsTer3). Premature stop codon appears before the RCL sequence which lead to the conclusion that the *Serpina3* with c.918-1G>C variant does not have the RCL and thus cannot participate in serpin protease inhibition (Koivuluoma et al. 2020).

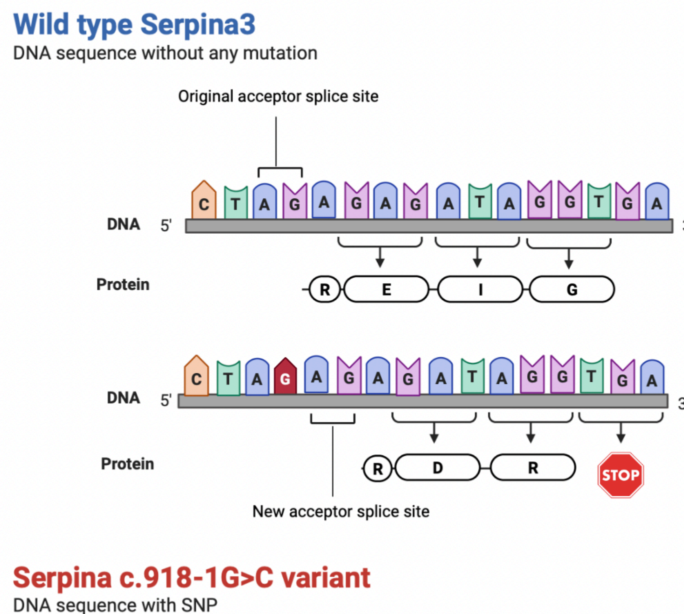

Figure 6. A comparison of wild type Serpina3 DNA sequence (upper) and a Serpina3 c.918-1G>C SNP variant with new predicted splice site(lower).

Aims of the study

This study has two main aims.

1. To investigate whether Serpina3 c.918-1G>C variant is associated with breast cancer in patient series collected from Helsinki region. The analysis is done by genotyping the DNA from patients as well as from population controls for the variant. Results are further examined in contrast to previous studies concerning Serpina3 c.918-1G>C (Koivuluoma et al. 2020)

2. To examine tumor histology and cellular markers of Serpina3 c.918-1G>C variant carriers.

Materials and methods

Patient series

The putative new breast cancer variant *Serpina3* c.918-1G>C was studied by genotyping germline DNA samples from Finnish familial and unselected breast cancer patients as well as from population controls. In total in this study 3399 breast cancer patients were screened for the variant and 3168 population controls (Table 2). In the unselected series only patients with invasive breast cancer were included into the analysis and in familial series patients with invasive breast cancer and patients with in situ breast cancer was included. Information of histological analysis was collected from pathology reports.

Table 2. Table shows the overall number of breast cancer patients and healthy population controls genotyped in this study.

Study cohort	N (breast cancer)	N (controls)
Helsinki	3399	3168

Unselected series

In the unselected cohort the patients are newly diagnosed and not selected based on the diagnosis age or family history. The unselected patients belong to tree separate cohorts. These cohorts are collected in Helsinki at the Department of Oncology (1997-1998 and 2000), the Department of Surgery (2001-2004) and the Breast Surgery Unit (2006-2010) (Syrjäkoski et al. 2000, Fagerholm et al., 2008 and Kaunisto et al., 2013). In total 2513 unselected patients were included to this study. Part of the patients (N=380) included to the unselected cohort have family history with breast or ovarian cancer and are also included into the familial cohort.

Familial Series

Patients included to the familial series were collected in Helsinki University Central Hospital at the Departments of Oncology and Clinical genetics until 2015 (Eerola et al., 2000 and Vahteristo et al., 2001). Additionally, 380 familial cases from the unselected series were included. In total 1259 familial patients were included in this study. The familial patient series included families with at least three breast or ovarian cancer cases in the first- or second-degree relatives or patients with one first

degree relative diagnosed with breast or ovarian cancer. In this study all patients with familial background are pooled together.

Control series

The control samples have been collected as two separate series. First collected samples are healthy blood donors from Helsinki region collected by SPR (N=1271) and later collected controls are cancer free population controls from Helsinki Biobank (N=1897). All genotyped controls are healthy females.

Ethical statement

The study was approved by the Ethics committee of the Helsinki University Hospital and informed consent was obtained from all patients. In this study all patient samples and data have been handled anonymously.

TaqMan genotyping

TaqMan genotyping detects spesific variants in the genome, in this case single nucleotide polymorphism (SNP) of Serpina3 c.918-1G>C variant. The method was used to screen all breast cancer patients as well as population controls for the c.918-1G<C variation.

The TaqMan assay includes an unlabeled PCR primer pair, custom made allele specific TaqMan probes with either VIC or FAM dye in the 5’ end (Table 5) and nonfluorescence quencher with minor groove binder on the 3’ end.

In the real time PCR, the double stranded template DNA is first denaturized allowing the correct custom-made probe with complementary nucleotide to the SNP to hybridize with the template. At this state, the quencher molecule keeps the fluorescent dye inactive. During PCR the unlabeled PCR primer pairs anneal to the spesific locus of the genomic template DNA and are expanded by Taq polymerase. The assay is based on the 5’ end nuclease activity of the Taq polymerase. When polymerase reaches the probe hybridized with spesific target allele, the polymerase cleaves the dye molecule. Dye molecule released from the quencher generates a fluorescence signal which is detected by the qPCR machine. Each cycle of qPCR enhances exponentially the fluorescence signal. The fluorescence signal is determinated for each sample resulting in robust allelic discrimination and allowing the detection SNP for each sample (Woodward, 2014).

In this genotyping study the following reagents were used: 80x Custom TaqMan SNP genotyping Assay (Thermo Fisher, Walters, MA, USA), TaqMan Genotyping Master Mix (Thermo Fisher), MQ water and patient genomic DNA (10 ng/µl) extracted from blood (Table 3). The genotyping was conducted on the 96 well fast amplicon plates (Thermo Fisher). For each sample 9 µl of TaqMan reaction mixture was used with 1 µl of DNA with total reaction volume of 10 µl per well. For genotyping the 7500 Fast Real Time PCR System machine was utilized using 40 amplification rounds.

Table 3. TaqMan reaction mixture scaled for one 96 well plate.

TaqMan reaction mixture (96x)	
Master Mix	550 µl
MQ water	426 µl
80x assay	14 µl

Table 4. qPCR program for genotyping

qPCR program	
95 ℃	10 min
92 ℃	15 s
60 ℃	1 min

Table 5. Table shows the taqman reporter sequences used in the genotyping assay. In both sequences the Serpina3 c.918-1G>C SNP is marked in bolded font. Also, the dye (VIC, FAM) reporter sequences carry is shown.

TaqMan reporter sequences

Sequence	Dye
CTCACCTATCTCT**C**TAGAAAA	VIC
CACCTATCTCT**G**TAGAAAA	FAM

The real-time PCR presents genotyping results as allelic discrimination plot (Figure. 7). In the plot on the x-axis is allele1 and on the y-axis allele 2, the genotypes of the samples are shown as circles and their color and position represent the alleles sample carries. In this plot homozygous for wildtype (GG genotype) are shown as red circles and heterozygous carriers (GC genotype) as green circles, homozygous CC genotype was not observed in this study. The undetermined samples, in other words samples which genotype could not be determined due to for instance too little DNA or evaporation would be shown as Xs in the plot. The patient samples were run against positive and negative control.

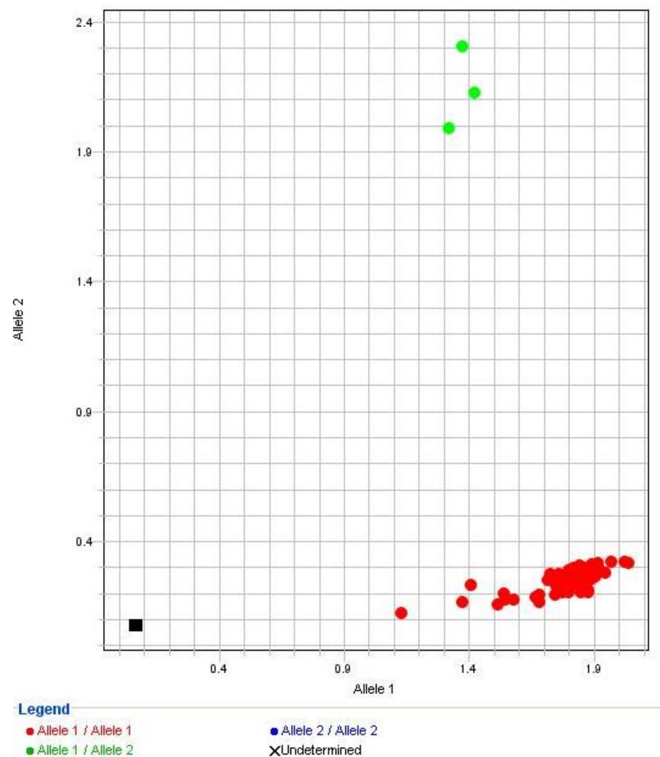

Figure 7. Allelic discrimination plot presents the genotypes of the patient samples. Allele 1 is on the x-axis and allele 2 on the y-axis. The genotypes are shown as circles. Red circles have GG genotype and green circles GC genotype.

Serpina3 c.910-1G>C Carrier's sequencing

Serpina3 c.918 -1G>C carriers identified by real-time PCR were Sanger-sequenced to confirm the carrier status. Primers were designed by Primer3 and produced by Metabion International (Planegg, Bayer, Germany).

Table 6. In this table the primers are shown for both forward (F) and reverse (R) sequences as well as their melting temperature. In addition, the expected product length is shown in base pairs (bp).

Primers

Sequence	Tm	Product
F: 5' GACAGGGGTTAAGAAATTGAGGA 3'	58.1	396 bp
R: 5' ACCCAAGTTCAAGAGTACCCA 3'	58.6	

Gradient PCR was run to optimize the best reaction condition to two different volumes of DNA (3 µl and 6 µl) as well as for eleven temperatures from scale of 47 °C to 66°C. The PCR was done as 25 µl reaction using 10 ng/µl of template with forward and reverse primers at the concentration of 20 µM of each and 2,5 µl 10x standard reaction buffer (Biotools, Jupiter, FL, USA), 5U/µl DNA-polymerase (Biotools) and 25 µM of each dNTP. The gradient PCR program was ran as shown in Table 8.

PCR products were run on 2% agarose gel for 35 minutes at 200V, 2 µl of 100bp DNA ladder (Biotools) was used as a marker. For the gel run 5 µl of product was mixed with 2 µl of 6x DNA loading dye (New England Biolabs, Ipswich, MA, USA). Ready gel was visualized under UV light (Biorad). The most suitable annealing temperature for primers is 58 °C coupled with 6 µl of DNA.

Table 7. A mixture for 1x gradient PCR sample

Gradient PCR mix (1x)	
MQ water	21,25 µl
Buffer	2,5µl
dNTP	0,15 µl
Forward primer	0,25 µl
Reverse primer	0,25 µl
DNA Polymerase	0,1 µl

Table 8. Gradient PCR program

Gradient PCR program	
95 °C	10 min
95 °C	1 min
47 °C – 65 °C	1 min
72 °C	1 min
72 °C	10 min
8 °C	forever

The Serpina 918-1G>C variant carriers PCR was performed with the primers tested above. The same reaction mixture was used as in gradient PCR (Table 8) with 5 ul of 10ng/µl carrier genomic DNA. Moreover, the same PCR program was followed with 58 ℃ as annealing temperature and products confirmed with gel ran as described earlier. The PCR products were cleaned with A'SAP treatment. In the reaction mixture the exonuclease I degrades unused primers and alkaline phosphatase excess dNTP. This step is necessary for further Sanger-sequencing procedure. Exonuclease I and alkaline phosphatase (ArcticZymes, Tromso, Norway) was mixed in 1:1 ratio and 2µl of mix was added into 5µl of PCR product. The samples were heated up to 37 ℃ for 15 minutes followed up by heat inactivation at 80 ℃ for 5 min.

Cleanup procedure was followed up by sequencing. The sequencing was done as 10 µl reaction using 1 µl of cleaned up PCR product, 1 µl 3,2µM forward or reverse primer, 2 µl 5x sequencing buffer (Big Dye Terminator, Thermo Fisher) and 0,5 µl Ready Reaction mixture (RR-100, Thermo Fisher). Sequencing was performed by using both forward and reverse primers to confirm carriers. The sequencing program was followed as described in table x. To determine the sequence, the samples were analyzed in Finland Institute of Molecular Medicine (FIMM) by capillary electrophoresis. Further, the sequences from FIMM were interpreted by ApE – A plasmid editor v.3.0.8 (Wayne Davis).

Table 9. Sequencing mixture calculated for one reaction.

Sequencing mixture	
MQ water	5,5 µl
Sequencing buffer	2 µl
Primer (F/R)	1µl
Ready Reaction mix	0,5 µl
PCR product	1 µl

Table 10. Sequencing program

Sequencing program	
96 ℃	5 min
96 ℃	10 s
50 ℃	5 s
60 ℃	4 min

Statistical analysis

To study the breast cancer risk of Serpina3 918-1G>C variant the genotype frequencies from cases and controls were compared. Cohorts were divided into subgroups based on family history and cellular markers (ER+/- and TN). Using Fisher's exact test for each subgroup odds ratio (OR) with 95% confidence interval and two-sided p-value were calculated. Statistically a p-value of < 0,05 was considered significant.

Power analysis was conducted to study if the sample size in this study is large enough to reach a sufficient power (80 %). As a guideline the following values seen in literature (Koivuluoma et al., 2020) were used: OR=3 among all breast cancer patients when minor allele frequency (MAF) among the population controls is 0.006. In the power analysis the *"pwr"* package 1.3-0 in RStudio was utilized. For the analysis a set of MAFs (0.0005 – 0.03) and a set of OR (0.14 – 7.4) values were created. As a sample size the number of genotyped samples were used (cases = 3399 and controls = 3168). In the analysis the power curves for each MAF were calculated by using the chi square-test in the "pwr" package (pwr.chisq.test). A power versus OR graph was created. The significant level was set to be 0.05 and degree of freedom 2.

Results

In this study, association between breast cancer and splice site acceptor variant c.918-1G>C (rs199710314) in Serpina3, previously found in Northern Finland, was investigated in a case-control study. In total over 6000 samples were genotyped by TaqMan real-time PCR. Here the results are presented as graphs and tables covering power analysis, Fisher exact test for association, tumor histology, and cellular tumor markers.

Power analysis

Below in figure 8 is presented the power analysis. In the study done by Koivuluoma et al. (2020) the general results indicated an increased breast cancer risk with OR=3.0 among breast cancer cases when MAF=0.006 among controls. The computed plot shows an OR in the x-axis and power in the y-axis and is composed of several graphs each representing one MAF value from 0.0005 to 0.03 with significance value of 0.05. This analysis shows that the sample size (cases = 3399 and controls =3168) in this study is large enough to repeat the results of Koivuluoma et al. with power markedly above the threshold value of 80%. This can be concluded by interpreting the graph with MAF=0.004, when OR=3.0 the graph reaches a power above 90%.

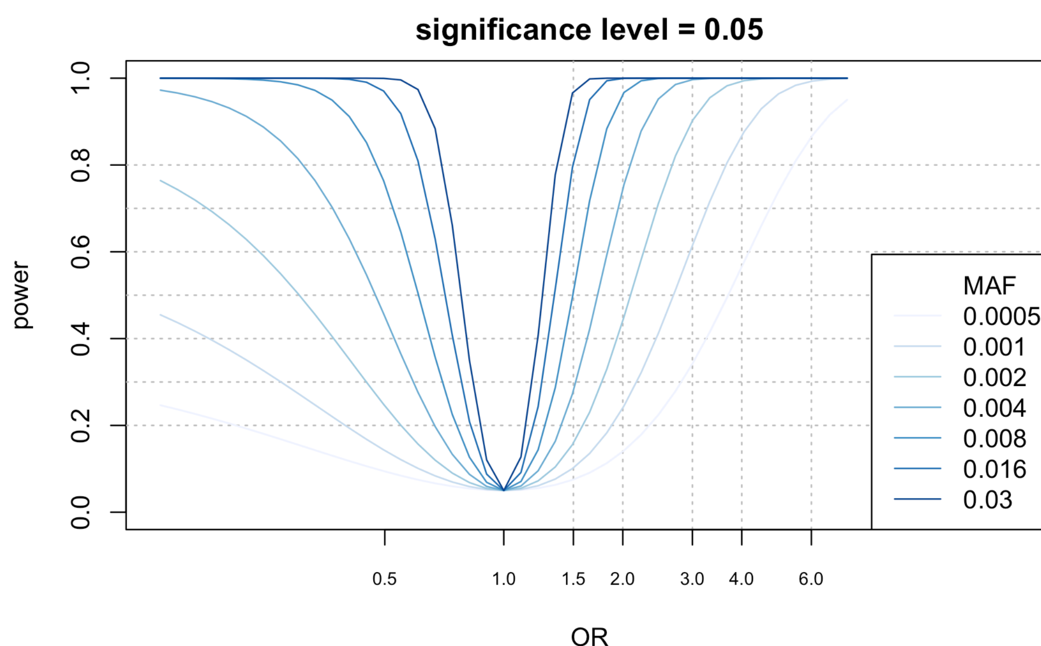

Figure 8. Power curves to detect if sample size is big enough to match the OR and frequency of Serpina3 c.918-1G>C variant carriers seen in literature. X-axis shows OR and y-axis shows power, and each curve represents one MAF. Sample sizes used are: 3399 BC patients and 3168 controls. OR = odds ratio, MAF = minor allele frequency

Serpina3 c.918-1G>C association with breast cancer risk

Fisher exact test was performed to statistically examine the association between *Serpina3* c.918-1G>C and breast cancer in cases and controls. Overall, the frequencies of c.918-1G>C variant between cases and controls do not differ significantly and in all subgroups $p>0.05$. This indicates that heterozygote splice site acceptor variant *Serpina3* c.918-1G>C does not increase the breast cancer risk in Southern Finland population.

In this analysis odds ratio is calculated for several patient subgroups separately as well as all patients together. Subgroups include familial patients and unselected patients. Amongst all breast cancer patient, the frequency of c.918-1G>C (0.32%) was almost equal to the control frequency (0.28%) with no elevated risk (OR=1.1, 95% CI=0.4-3.1, P=0.83). In familial patients the two subgroups, patients with strong familial background with at least tree breast cancer cases in first degree relatives and patients with one first degree relative, are pooled together resulting roughly the same but a little bit decreased mutation frequency (0.24%) as seen in all breast cancer cases.

Table 11. Frequencies of the Serpina3 c.918-1G<C variant in all breast cancer cases as well as in the subgroups. ER = estrogen receptor, TN= triple negative.

Study	Carriers/total	%	OR	95% CI	p
Helsinki					
All-BC	11/3394	0,32	1,1	0,4 – 3,1	0.83
Familial	3/1259	0,24	0,8	0,1 – 3,4	1
Unselected	9/2513	0,36	1,3	0,4 – 3,6	0,64
ER-positive	9/2677	0,34	1,2	0,4 – 3,5	0,81
ER-negative	1/595	0,17	0,6	0,01 – 4,3	1
TN	1/246	0,4	1,4	0,03 – 10,4	0,53
Control	9/3168	0,28			

Patients were further divided based on their breast cancer subtype to the following groups: ER+, ER- and triple negatives. There was no significant difference in the variant frequency between cases and controls.

Clinical characteristics of Serpina3 c.918-1G>C mutation carriers

The tumor histology and cellular markers of the mutation carriers in this study are shown in Table 12 and coupled with first diagnose age. The mean age of the first breast cancer diagnosis for c.918-1G>C carriers was 53.8 years when in unselected non-carriers it was 56.6. Eight of the tumors were ductal (73%) and only one carrier had a lobular tumor (9%). In addition, two of the tumors diagnosed as mix of ductal and lobular carcinoma are category as other (18%). The cellular markers of the tumors are divided into ER+/-, PR+/-, HER2+/- and Ki67+/. Tumors of all carriers (with data available i.e., not non applicable) were HER2 negative and nine both ER positive and PR positive (90%). Only one breast tumor was ER- and PR negative (10%) making it also the only triple negative case. In the Serpina3 variant carriers, the Ki67 marker status was almost equally divided, four carriers were negative (40%) for the marker and six were positive (60%). The one triple negative carrier was positive for Ki67 marker.

Table 12. Clinical and histopathological characteristics of Serpina3 c.918-1G>C variant among BC patients. Dg-age = age at diagnosis, pos = positive, neg = negative, NA = not applicable.

Dg-age	Tumor Histology	ER	PR	HER2	Ki67
68,2	Ductal	pos	pos	NA	neg
46,1	Ductal	pos	pos	neg	pos
55,9	Ductal	NA	NA	NA	NA
43,7	Ductal	neg	neg	neg	pos
43,3	Ductal	pos	pos	neg	neg
64,1	Lobular	pos	pos	NA	neg
49,4	Ductal	pos	pos	neg	neg
53,5	Other	pos	pos	NA	pos
54	Ductal	pos	pos	neg	pos
68	Ductal	pos	pos	neg	pos
46	Other	pos	pos	neg	pos

Discussion

Breast cancer is the leading cancer worldwide in women. The identified susceptibility genes that predispose to hereditary breast cancer explain less than half of the familial relative risk. To solve the remaining fraction of breast cancer the identification and study of novel low- and moderate risk genes is ongoing.

In this study, *Serpina3* c.918-1G>C variant was studied as a possible susceptibility allele for breast cancer in the population of Southern Finland. A case-control study was conducted by genotyping DNA from breast cancer patients (N=3399) as well as from population controls (N=3168). Control DNA came from either unknown blood donors in Helsinki region or Helsinki biobank. With the combination of Taqman real-time PCR and Sanger sequencing the results from this study suggest that the c.918-1G>C variant is not a breast cancer susceptibility allele at least in Southern Finland. Previously, a study by Koivuluoma et al. (2020) suggested that the same variant was significantly enriched amongst the Northern Finland breast cancer patients (N=1770), more specifically in the Northern Ostrobothnia region. The suggested association of *Serpina3* c.918-1G>C with breast cancer risk could not be validated in the present study including larger patient and controls series.

In this study, the frequency of variant carriers amongst all breast cancer patients was 0.28 % when in control population it was 0.27 %. In addition, in none of the subgroups were the results significant, as all p-values >0.05, with an OR of 0.8 for the familial and 1.3 for the unselected cases. Meanwhile, in the study by Koivuluoma et al., (2020) the variant showed a tree-fold increase in breast cancer risk in all breast cancer patients (1.8 % of cases) and a five-fold increase in familial cases (3.0 %) compared to the controls (0.6 %).

The control population frequency in Sequence Initiative Suomi project (SISu) (http://www.sisuproject.fi, SISu v.4.1) for the variant in the Helsinki region is 0.67 %. The frequency in SISu project is approximately two and a half time higher than in this study which might be explained by the variation in sample size. In this study 3168 controls were genotyped when in SISu the number is slightly above 500.

The difference in the c.918-1G>C variant carrier frequencies between Northern and Southern Finland population could be partially explained by the different genetic population structure. Due to the several bottlenecks, relocation of evacuees and migration movements the genetics of different regions

in Finland vary. The fine scale studies of the population structure have proven a clear clustering of regions with genetical similarities (Martin et al., 2018 and Kerminen et al., 2021). However, Helsinki region and Northern Ostrobothnia are geographically far from each other and populated at different times. Additionally, Helsinki region has higher migration levels and so more mixing in the population compared to Northern Ostrobothnia (Kerminen et al., 2017). Thus, regional differences in the enrichment of rare SNP variants like Serpina3 c.918-1G>C are possible due to the genetic differences in ancestry profiles (Locke et al., 2019). Similar regional enrichment can be seen in PALB2 and CHEK2 variants (rs180177102 and rs555607708) which were both increased in Eastern Finland up to 2.8 % when compared to the frequency very close to zero (0.0-0.2 %) in Western Finland in the FinnGen study (Mars et al., 2020).

In this study also the clinical characteristics i.e., the tumor histology and cellular markers were examined. Most of the tumors (73%) were ductal which follows the normal trend as majority of the diagnosed breast cancer cases are ductal carcinomas (Feng et al., 2018). Only one c.918-1G>C variant carrier showed a lobular tumor histology. In addition, two of the tumors were labelled as "other" due to the reason that they were diagnosed with a mixture of ductal- and lobular carcinomas. Interestingly, Koivuluoma et al., 2020 study showed a clear enrichment in medullary breast cancer cases in the *Serpina3* c.918-1G>C variant carriers when compared to the non-carriers among unselected cases (15,4 %; 4/26). However, similar correlation was not seen in this study since none of the variant carriers in unselected breast cancer cohort have been diagnosed with a medullary breast cancer (0 %; 0/9). To investigate an association between medullary breast cancer and c.918-1G>C variant, a study with greater number of medullar breast cancer cases should be conducted. In general, medullary breast cancer is very rear subtype and comprises approximately 3-5 % of all the diagnosed breast cancer cases (Dai et al., 2020). In the Helsinki series, 20 cases are found in the unselected cohort in total.

Based on the GnomAD database (https://gnomad.broadinstitute.org, gnomAD v2.1.1) and SISu project the c.918-1G>C variant is only enriched in Finnish population with an overall population frequency of 0.35-0.37 %. Moreover, in the study where more than 400 000 UK biobank participants were exome sequenced for rare variants the existence of the c.918-1G>C variant was not found. Several other rare SNPs including missense variants and loss-of-function (LoF) variants (14 in total) in *Serpina3* gene were discovered. Four missense variants had an OR above 2 but were not significantly associated with breast cancer risk. None of the four variants existed in the SISu project or in GnomAD for Finnish population. A gene burden test of *Serpina3* LoF variants did not indicate

an association to the breast cancer. (Backman et al., 2021 and NHGRI-EBI GWAS-catalog: https://www.ebi.ac.uk/gwas/home). Studies so far have only linked the variant in *Serpina3* to the breast cancer cases in Northern Finland, but this association could not be confirmed in the larger data set from Southern Finland. Together the results from this study and Backman et al., suggest that *Serpina3* is not a breast cancer susceptibility gene. However, if for instance subtype spesific association to breast cancer is discovered in further studies, it could open up new pathways to study breast cancer genetics and formation since *Serpina3* is not part of the DNA repair mechanism. Hence, it could offer new information of breast cancer prediction as well as its causal mechanism.

Future prospects

To further characterize the biological effect of *Serpina3* c.918-1G>C variant a transcriptome analysis would be the next step. *Serpina3* c.918-1G>C is proposed to be a heterozygous splice site acceptor which induces a stop codon into the beginning of exon 4 (https://www.ncbi.nlm.nih.gov/gene/, GeneID = 12: human Serpina3). Transcriptome analysis would tell whether the suspected truncated protein is produced or whether the mRNA is a target of a degradation and hence never translated into a protein. If the truncated protein is produced it would probably either lack the whole exon 4 or the rest of the protein: exons 4 and 5. In both cases (truncated protein or mRNA decay), one should study the effect of decreased expression level of *Serpina3* on a cell and tissue level.

The results from this study and the UKBB exome sequencing do not support SERPINA3 as a breast cancer susceptibility gene. However, further studies would be needed to investigate possible association of SERPINA3 variants with specific rare histological types of breast cancer.

Conclusion

Breast cancer cases keep climbing worldwide and the genetic reasoning in roughly half of the familial cases is unknown. Novel breast cancer variants are being sought actively and genes beyond DNA repair machine studied.

The first aim of this case-control study was to identify if *Serpina3* c.918-1G>C variant is associated with breast cancer in the Helsinki region. Secondly, the tumor histology and clinical characteristics of variant carriers were examined. Analysis of the c.918-1G>C variant in cases and control population revealed that the variant did not contribute as a novel breast cancer susceptibility gene in the Southern Finland population. In addition, tumor histology of the variant carriers was not associated with any spesific tumor histology but followed a normal trend as most of the carriers had a ductal carcinoma.

This study fulfilled well the set aims and was planned and carried out carefully. No results affecting errors were made. Altogether, this thesis offers theoretical background knowledge of breast cancer, comprehensive section on Serpina3, its structure, cellular function and disease linkage as well as a practical part. The carried-out experiments and results shine light to *Serpina3* c.918-1G<C variant prevalence among breast cancer cases and population controls in Helsinki region.

Acknowledgement

This master's thesis was conducted at the Department of Obstetrics and Gynaecology, University of Helsinki between June 2021 and February 2022.

I want to express my gratitude to docent Heli Nevanlinna for giving me this project and for guiding and advising me through it. I also want to thank my supervisor Ph.D. Maija Suvanto for the day-to-day help in the lab as well as with data handling and endless tips, comments and patience during the writing process. Additionally, I also thank MSc Anna Nurmi for her expertise in the lab and with data handling.

Lastly, I warmly thank my family and friends for supporting me from the start of this project till the end.

Helsinki, March 2022

Niina Aho

References

Articles

1. Abu-Helalah, M., Azab, B., Mubaidin, R., Ali, D., Jafar, H., Alshraideh, H., ... & Awidi, A. (2020). BRCA1 and BRCA2 genes mutations among highrisk breast cancer patients in Jordan. *Scientific reports*, *10*(1), 1-9.

2. Allinen, M., Launonen, V., Laake, K., Jansen, L., Huusko, P., Kääriäinen, H., ... & Winqvist, R. (2002). ATM mutations in Finnish breast cancer patients. *Journal of medical genetics*, *39*(3), 192-196.

3. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J., & Clarke, M. F. (2003). Prospective identification of tumorigenic breast cancer cells. *Proceedings of the National Academy of Sciences*, *100*(7), 3983-3988.

4. Antoniou, A. C., Casadei, S., Heikkinen, T., Barrowdale, D., Pylkäs, K., Roberts, J., ... & Tischkowitz, M. (2014). Breast-cancer risk in families with mutations in PALB2. *New England Journal of Medicine*, *371*(6), 497-506.

5. Baker, C., Belbin, O., Kalsheker, N., & Morgan, K. (2007). SERPINA3 (aka alpha-1-antichymotrypsin). *Front Biosci*, *12*(2821), 35.

6. Backman, J. D., Li, A. H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M. D., ... & Ferreira, M. A. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*, *599*(7886), 628-634.

7. Bombonati, A., & Sgroi, D. C. (2011). The molecular pathology of breast cancer progression. *The Journal of pathology*, *223*(2), 308–318.

8. Breast Cancer Association Consortium. (2021). Breast cancer risk genes—association analysis in more than 113,000 women. *New England Journal of Medicine*, *384*(5), 428-439.

9. Bubien, V., Bonnet, F., Brouste, V., Hoppe, S., Barouk-Simonet, E., David, A., ... & French Cowden Disease Network. (2013). High cumulative risks of cancer in patients with PTEN hamartoma tumour syndrome. *Journal of medical genetics*, *50*(4), 255-263.

10. Cao, L. L., Pei, X. F., Qiao, X., Yu, J., Ye, H., Xi, C. L., ... & Gong, Z. L. (2018). SERPINA3 silencing inhibits the migration, invasion, and liver metastasis of colon cancer cells. *Digestive diseases and sciences*, *63*(9), 2309–2319.

11. Chelbi, S. T., Wilson, M. L., Veillard, A. C., Ingles, S. A., Zhang, J., Mondon, F., ... & Vaiman, D. (2012). Genetic and epigenetic mechanisms collaborate to control SERPINA3 expression and its association with placental diseases. *Human molecular genetics*, *21*(9), 1968–1978.

12. Coignard, J., Lush, M., Beesley, J., O’mara, T. A., Dennis, J., Tyrer, J. P., ... & Peshkin, B. (2021). A case-only study to identify genetic modifiers of breast cancer risk for BRCA1/BRCA2 mutation carriers. *Nature communications*, *12*(1), 1–22.

13. Colomer, R., Aranda-López, I., Albanell, J., García-Caballero, T., Ciruelos, E., López-García, M. Á., ... & Palacios-Calvo, J. (2018). Biomarkers in breast cancer: A consensus statement by the Spanish Society of Medical Oncology and the Spanish Society of Pathology. *Clinical and Translational Oncology*, *20*(7), 815–826.

14. Dai, D., Shi, R., Wang, Z., Zhong, Y., Shin, V. Y., Jin, H., & Wang, X. (2020). Competing risk analyses of medullary carcinoma of breast in comparison to infiltrating ductal carcinoma. *Scientific reports*, *10*(1), 1–11.

15. Easton, D. F., Pharoah, P. D., Antoniou, A. C., Tischkowitz, M., Tavtigian, S. V., Nathanson, K. L., ... & Foulkes, W. D. (2015). Gene-panel sequencing and the prediction of breast-cancer risk. *New England Journal of Medicine*, *372*(23), 2243–2257.
16. Eerola, H., Blomqvist, C., Pukkala, E., Pyrhönen, S., & Nevanlinna, H. (2000). Familial breast cancer in southern Finland: how prevalent are breast cancer families and can we trust the family history reported by patients?. *European journal of cancer*, *36*(9), 1143-1148.

17. Eerola, H., Pukkala, E., Pyrhönen, S., Blomqvist, C., Sankila, R., & Nevanlinna, H. (2001). Risk of cancer in BRCA1 and BRCA2 mutation-positive and-negative breast cancer families (Finland). *Cancer Causes & Control*, *12*(8), 739–746.

18. Erkko, H., Xia, B., Nikkilä, J., Schleutker, J., Syrjäkoski, K., Mannermaa, A., ... & Winqvist, R. (2007). A recurrent mutation in PALB2 in Finnish cancer families. *Nature*, *446*(7133), 316–319.

19. Eroles, P., Bosch, A., Pérez-Fidalgo, J. A., & Lluch, A. (2012). Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. *Cancer treatment reviews*, *38*(6), 698–707.

20. Fagerholm, R., Hofstetter, B., Tommiska, J., Aaltonen, K., Vrtel, R., Syrjäkoski, K., ... & Nevanlinna, H. (2008). NAD (P) H: quinone oxidoreductase 1 NQO1* 2 genotype (P187S) is a strong prognostic and predictive factor in breast cancer. *Nature genetics*, *40*(7), 844-853.

21. Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., ... & Ren, G. (2018). Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & diseases*, *5*(2), 77–106.

22. Frey, S., Sticht, H., Wilsmann-Theis, D., Gerschütz, A., Wolf, K., Löhr, S., ... & Hüffmeier, U. (2020). Rare loss-of-function mutation in SERPINA3 in generalized pustular psoriasis. *The Journal of investigative dermatology*, *140*(7), 1451–1455.

23. Ghoussaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M. K., Dicks, E., ... & Durda, K. (2012). Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nature genetics*, *44*(3), 312–318.

24. Hallamies, S., Pelttari, L. M., Poikonen-Saksela, P., Jekunen, A., Jukkola-Vuorinen, A., Auvinen, P., ... & Nevanlinna, H. (2017). CHEK2 c. 1100delC mutation is associated with an increased risk for male breast cancer in Finnish patient population. *BMC cancer*, *17*(1), 1–5.

25. Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, *144*(5), 646-674.

26. Honrado, E., Benítez, J., & Palacios, J. (2006). Histopathology of BRCA1-and BRCA2-associated breast cancer. *Critical reviews in oncology/hematology*, *59*(1), 27–39.

27. Huebner, R. J., & Ewald, A. J. (2014, July). Cellular foundations of mammary tubulogenesis. In *Seminars in cell & developmental biology* (Vol. 31, pp. 124-131). Academic Press.

28. Huszno, J., & Grzybowska, E. (2018). TP53 mutations and SNPs as prognostic and predictive factors in patients with breast cancer. *Oncology letters*, *16*(1), 34–40.

29. Jonsson, P., Bandlamudi, C., Cheng, M. L., Srinivasan, P., Chavan, S. S., Friedman, N. D., ... & Taylor, B. S. (2019). Tumour lineage shapes BRCA-mediated phenotypes. *Nature*, *571*(7766), 576-579.

30. Kääriäinen, H., Muilu, J., Perola, M., & Kristiansson, K. (2017). Genetics in an isolated population like Finland: a different basis for genomic medicine?. *Journal of community genetics*, *8*(4), 319–326.

31. Kalli, S., Semine, A., Cohen, S., Naber, S. P., Makim, S. S., & Bahl, M. (2018). American joint committee on cancer's staging system for breast cancer: what the radiologist needs to know. *Radiographics*, *38*(7), 1921–1933.

32. Kantaputra, P. N., Chuamanochan, M., Kiratikanon, S., Chiewchanvit, S., Chaiwarith, R., Intachai, W., ... & Ngamphiw, C. (2021). A truncating variant in SERPINA3, skin pustules and adult-onset immunodeficiency. *Journal of Dermatology*.

33. Katayama, H., Tsou, P., Kobayashi, M., Capello, M., Wang, H., Esteva, F., ... & Hanash, S. (2019). A plasma protein derived TGFβ signature is a prognostic indicator in triple negative breast cancer. *NPJ precision oncology*, *3*(1), 1–8.

34. Kaunisto, M. A., Jokela, R., Tallgren, M., Kambur, O., Tikkanen, E., Tasmuth, T., ... & Kalso, E. A. (2013). Pain in 1,000 women treated for breast cancer: a prospective study of pain sensitivity and postoperative pain. *Anesthesiology*, *119*(6), 1410-1421.

35. Kelly-Robinson, G. A., Reihill, J. A., Lundy, F. T., McGarvey, L. P., Lockhart, J. C., Litherland, G. J., ... & Martin, S. L. (2021). The Serpin Superfamily and Their Role in the Regulation and Dysfunction of Serine Protease Activity in COPD and Other Chronic Lung Diseases. *International Journal of Molecular Sciences*, *22*(12), 6351.

36. Kerminen, S., Cerioli, N., Pacauskas, D., Havulinna, A. S., Perola, M., Jousilahti, P., ... & Pirinen, M. (2021). Changes in the fine-scale genetic structure of Finland through the 20th century. *PLoS genetics*, *17*(3), e1009347.

37. Kerminen, S., Havulinna, A. S., Hellenthal, G., Martin, A. R., Sarin, A. P., Perola, M., ... & Pirinen, M. (2017). Fine-scale genetic structure in Finland. *G3: Genes, Genomes, Genetics*, *7*(10), 3459-3468.

38. Kiiski, J. I., Pelttari, L. M., Khan, S., Freysteinsdottir, E. S., Reynisdottir, I., Hart, S. N., ... & Nevanlinna, H. (2014). Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. *Proceedings of the National Academy of Sciences*, *111*(42), 15172-15177.

39. Kiiski, J. I., Pelttari, L. M., Khan, S., Freysteinsdottir, E. S., Reynisdottir, I., Hart, S. N., ... & Nevanlinna, H. (2014). Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. *Proceedings of the National Academy of Sciences*, *111*(42), 15172-15177.

40. Kinzler, Kenneth W., and Bert Vogelstein. "Gatekeepers and caretakers." *Nature* 386.6627 (1997): 761-763.

41. Knudson, Alfred G. "Two genetic hits (more or less) to cancer." *Nature Reviews Cancer* 1.2 (2001): 157-162.

42. Koivuluoma, S., Tervasmäki, A., Kauppila, S., Winqvist, R., Kumpula, T., Kuismin, O., ... & Pylkäs, K. (2021). Exome sequencing identifies a recurrent variant in SERPINA3 associating with hereditary susceptibility to breast cancer. *European Journal of Cancer*, *143*, 46–51

43. Levitt, Nicola C., and Ian D. Hickson. "Caretaker tumour suppressor genes that defend genome integrity." *Trends in molecular medicine* 8.4 (2002): 179-186.

44. Li, S., Shen, Y., Wang, M., Yang, J., Lv, M., Li, P., ... & Yang, J. (2017). Loss of PTEN expression in breast cancer: association with clinicopathological characteristics and prognosis. *Oncotarget*, *8*(19), 32043.

45. Locke, A. E., Steinberg, K. M., Chiang, C. W., Service, S. K., Havulinna, A. S., Stell, L., ... & Freimer, N. B. (2019). Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature*, *572*(7769), 323–328.

46. Mars, N., Widén, E., Kerminen, S., Meretoja, T., Pirinen, M., della Briotta Parolo, P., ... & Ripatti, S. (2020). The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nature communications*, *11*(1), 1–9.

47. Martin, A. R., Karczewski, K. J., Kerminen, S., Kurki, M. I., Sarin, A. P., Artomov, M., ... & Daly, M. J. (2018). Haplotype sharing provides insights into fine-scale population history and disease in Finland. *The American Journal of Human Genetics*, *102*(5), 760–775.

48. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., ... & Ishiguro, J. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, *551*(7678), 92–94.

49. Moore-Smith, L. D., Isayeva, T., Lee, J. H., Frost, A., & Ponnazhagan, S. (2017). Silencing of TGF-β1 in tumor cells impacts MMP-9 in tumor microenvironment. *Scientific reports*, *7*(1), 1–10.

50. Negrini, Simona, Vassilis G. Gorgoulis, and Thanos D. Halazonetis. "Genomic instability—an evolving hallmark of cancer." *Nature reviews Molecular cell biology* 11.3 (2010): 220–228.

51. Nishida, Naoyo, et al. "Angiogenesis in cancer." *Vascular health and risk management* 2.3 (2006): 213.

52. Nieuwenhuis, M. H., Kets, C. M., Murphy-Ryan, M., Yntema, H. G., Evans, D. G., Colas, C., ... & Vasen, H. F. (2014). Cancer risk and genotype–phenotype correlations in PTEN hamartoma tumor syndrome. *Familial cancer*, *13*(1), 57-63.

53. Nurmi, A., Muranen, T. A., Pelttari, L. M., Kiiski, J. I., Heikkinen, T., Lehto, S., ... & Nevanlinna, H. (2019). Recurrent moderate-risk mutations in Finnish breast and ovarian cancer patients. *International journal of cancer*, *145*(10), 2692–2700.

54. Orvieto, E., Maiorano, E., Bottiglieri, L., Maisonneuve, P., Rotmensz, N., Galimberti, V., ... & Viale, G. (2008). Clinicopathologic characteristics of invasive lobular carcinoma of the breast: results of an analysis of 530 cases from a single institution. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *113*(7), 1511–1520.

55. Prat, A., Pineda, E., Adamo, B., Galván, P., Fernández, A., Gaba, L., ... & Muñoz, M. (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast*, *24*, S26-S35.

56. Rebbeck, T. R., Friebel, T. M., Friedman, E., Hamann, U., Huo, D., Kwong, A., ... & Meindl, A. (2018). Mutational spectrum in a worldwide study of 29,700 families with BRCA1 or BRCA2 mutations. *Human mutation*, *39*(5), 593–620.

57. Renwick, A., Thompson, D., Seal, S., Kelly, P., Chagtai, T., Ahmed, M., ... & Rahman, N. (2006). ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nature genetics*, *38*(8), 873-875.

58. Reya, T., Morrison, S. J., Clarke, M. F., & Weissman, I. L. (2001). Stem cells, cancer, and cancer stem cells. *nature*, *414*(6859), 105-111.

59. Sánchez-Navarro, A., González-Soria, I., Caldiño-Bohn, R., & Bobadilla, N. A. (2021). An integrative view of serpins in health and disease: The contribution of SerpinA3. *American Journal of Physiology-Cell Physiology*, *320*(1), C106-C118.

60. Sanrattana, W., Maas, C., & de Maat, S. (2019). SERPINs—from trap to treatment. *Frontiers in medicine*, *6*, 25.

61. Sarantaus, L., Huusko, P., Eerola, H., Launonen, V., Vehmanen, P., Rapakko, K., ... & Nevanlinna, H. (2000). Multiple founder effects and geographical clustering of BRCA1 and BRCA2 families in Finland. *European Journal of Human Genetics*, *8*(10), 757–763.

62. Schon, K., & Tischkowitz, M. (2018). Clinical implications of germline mutations in breast cancer: TP53. *Breast cancer research and treatment*, *167*(2), 417–423.

63. Siming, G. A. O., Honglin, Z. H. U., Xiaoxia, Z. U. O., & Hui, L. U. O. (2018). Cathepsin G and its role in inflammation and autoimmune diseases. *Archives of rheumatology*, *33*(4), 498.

64. Sin, W. C., & Lim, C. L. (2017). Breast cancer stem cells—from origins to targeted therapy. *Stem cell investigation*, *4*.

65. Smalley, M., & Ashworth, A. (2003). Stem cells and breast cancer: a field in transit. *Nature Reviews Cancer*, *3*(11), 832–844.

66. Spurdle, A. B., Couch, F. J., Parsons, M. T., McGuffog, L., Barrowdale, D., Bolla, M. K., ... & Henderson, B. E. (2014). Refined histopathological predictors of BRCA1 and BRCA2 mutation status: a large-scale analysis of breast cancer characteristics from the BCAC, CIMBA, and ENIGMA consortia. *Breast Cancer Research*, *16*(6), 1–16.

67. Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. "The cancer genome." *Nature* 458.7239 (2009): 719-724.

68. Sud, A., Kinnersley, B., & Houlston, R. S. (2017). Genome-wide association studies of cancer: current insights and future perspectives. *Nature Reviews Cancer*, *17*(11), 692–704.

69. Syrjäkoski, K., Kuukasjärvi, T., Waltering, K., Haraldsson, K., Auvinen, A., Borg, Å., ... & Koivisto, P. A. (2004). BRCA2 mutations in 154 Finnish male breast cancer patients. *Neoplasia*, *6*(5), 541–545.

70. Syrjakoski, K., Vahteristo, P., Eerola, H., Tamminen, A., Kivinummi, K., Sarantaus, L., ... & Nevanlinna, H. (2000). Population-based study of BRCA1 and BRCA2 mutations in 1035 unselected Finnish breast cancer patients. *Journal of the National Cancer Institute*, *92*(18), 1529-1531.

71. Toss, A., Tenedini, E., Piombino, C., Venturelli, M., Marchi, I., Gasparini, E., ... & Cortesi, L. (2021). Clinicopathologic Profile of Breast Cancer in Germline ATM and CHEK2 Mutation Carriers. *Genes*, *12*(5), 616.

72. Vahteristo, P., Bartkova, J., Eerola, H., Syrjäkoski, K., Ojala, S., Kilpivaara, O., ... & Nevanlinna, H. (2002). A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer. *The American Journal of Human Genetics*, *71*(2), 432–438.

73. Vahteristo, P., Eerola, H., Tamminen, A., Blomqvist, C., & Nevanlinna, H. (2001). A probability model for predicting BRCA1 and BRCA2 mutations in breast and breast-ovarian cancer families. *British journal of cancer*, *84*(5), 704-708.

74. Wang, Q., Dhindsa, R. S., Carss, K., Harper, A. R., Nag, A., Tachmazidou, I., ... & Petrovski, S. (2021). Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*, *597*(7877), 527-532.

75. Weinberg, R. A., & Hanahan, D. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57-70.

76. Wendt, C., & Margolin, S. (2019). Identifying breast cancer susceptibility genes–a review of the genetic background in familial breast cancer. *Acta Oncologica*, *58*(2), 135–146.

77. Wendt, C., Muranen, T. A., Mielikäinen, L., Thutkawkorapin, J., Blomqvist, C., Jiao, X., ... & Lindblom, A. (2021). A search for modifying genetic factors in CHEK2: c. 1100delC breast cancer patients. *Scientific Reports*, *11*(1), 1–9.

78. Wiedow, O., & MEYER-HOFFERT, U. (2005). Neutrophil serine proteases: potential key regulators of cell signalling during inflammation. *Journal of internal medicine*, *257*(4), 319–328.

79. Wilson, T. J., Nannuru, K. C., & Singh, R. K. (2009). Cathepsin G–mediated activation of pro–matrix metalloproteinase 9 at the tumor-bone Interface promotes transforming growth factor-β signaling and bone destruction. *Molecular Cancer Research*, *7*(8), 1224–1233

80. Wiseman, Bryony S., and Zena Werb. "Stromal effects on mammary gland development and breast cancer." *Science*296.5570 (2002): 1046–1049.

81. Yamamura, J., Miyoshi, Y., Tamaki, Y., Taguchi, T., Iwao, K., Monden, M., ... & Noguchi, S. (2004). mRNA expression level of estrogen-inducible gene, α1-antichymotrypsin, is a predictor of early tumor recurrence in patients with invasive breast cancers. *Cancer science*, *95*(11), 887–892.

82. Zhou, M. L., Chen, F. S., & Mao, H. (2019). Clinical significance and role of up-regulation of SERPINA3 expression in endometrial cancer. *World journal of clinical cases*, *7*(15), 1996.

Books

1. Aittomäki, K., Moilanen, M., Perola, M. (2016). *Lääketieteellinen genetiikka.* Duodecim.
2. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular biology of the cell*. New York: Garland Science.
3. Hartwell, L. H., Goldberg, M. L., Fischer, J. A., Hood, L. (2015). *Genetics from genes and genomes.* New York: Mc Graw-Hill Education.
4. Schatten, H. (2013). *Cell and molecular biology of breast cancer.* New York: Humana Press.
5. Strachan, T., Goodship, J., Chinnery, P. (2015). *Genetics and genomics in medicine.* New York: garland Science
6. Weinberg, R. A. (2014). *The biology of cancer.* New York: Garland Science.
7. Woodward, J. (2014). Bi-allelic SNP genotyping using the TaqMan® assay. In *Crop breeding* (pp. 67-74). Humana Press, New York, NY.

Websites

1. AstraZeneca PheWAS Portal. *Gene: Serpina3*. https://azphewas.com/geneView/7e2a7fab-97f0-45f7-9297-f976f7e667c8/SERPINA3/glr/binary
2. Global cancer observatory. *Cancer today.* https://gco.iarc.fr/today/fact-sheets-cancers
3. GWAS catalog. *Study: GCST90079606.* https://www.ebi.ac.uk/gwas/studies/GCST90079606
4. Finnish cancer registry. *Cancer statistic.* https://cancerregistry.fi/statistics/cancer-statistics/
5. Cancer Society of Finland, All about cancer. *Stages of cancer, differentiation and staging of cancer.* https://www.allaboutcancer.fi/facts-about-cancer/stages/#2413a64f
6. GnomAD. V.2.1.1. https://gnomad.broadinstitute.org/variant/14-95088677-G-C?dataset=gnomad_r2_1
7. Sequencing Initiative Suomi. (01.09.2016). SISu v.4.1 http://search.sisuproject.fi/#/variant/14:95088677
8. Breastcancer. (7.4.2021). *Molecular subtypes of breast cancer.* https://www.breastcancer.org/symptoms/types/molecular-subtypes
9. WHO. (21.09.2021). *Cancer.* https://www.who.int/news-room/fact-sheets/detail/cancer