Genetic variation and allelic imbalance in a selection of genes in breast cancer patients.

Christine Haugen

Norwegian University of Life Sciences
Department of Animal and Aquacultural Sciences
Master Thesis 60 credits 2010

Acknowledgements

The work on this thesis was performed between July 2009 and May 2010 as the final part of my Master's Degree in Biotechnology at the University of Life Sciences (UMB) in Ås, Norway. The SNP genotyping with MassArray was done by the staff at Centre for Integrative Genetics (CiGene) at UMB. The rest of the work was performed by me at the Departement of Genetics, Institute for Cancer Research, Norwegian Radium Hospital Rikshospitalet, Oslo University Hospital.

I would like to show my appreciation to my supervisor Professor Vessela N. Kristensen for accepting me into her Cancer Genome Variation group and giving me much encouragement and the opportunity to work with this exiting subject. Great thanks to my main supervisor PhD Silje Nordgard for being patient and caring, and for dedicating many hours of guidance and advice for me, and a push when needed.

To everyone at the Departement of Genetics: thank you for making me feel welcome, and providing me with a lot of help and advice through the year. A special thanks to research technician Grethe I. Grenaker Alnæs and bioengineer Hilde Johnsen for all the help given with the practical in the lab, and to PhD Hege Edvardsen for helping me with the MassArray part of the thesis. Thank you to research associate Daniel Nebdal and PhD-student Kristin Valborg Reinertsen for friendship, support and many good discussions.

Thanks to lab technician Arne Roseth and SNP Platform Manager Paul Berg at CiGene for letting me observe the MassArray SNP genotyping process and answering all my questions.

My gratitude to my parents for being patient, helpful and supportive in my five years at UMB, and to my brother and his family for reminding me that there are other important things in life than work.

And to my friends, Sean and Jenny; thank you for being there for me.

Abstract

Genetic variation, such as Single Nucleotide Polymorphisms (SNPs), are naturally occurring characteristics of the genome that differs between individuals of a species, and in some cases affect the risk of developing a disease. When the phenotype is affected by the genotype it happens through expression, and the level of expression itself can be considered a phenotype. When two alleles have different expression levels it is known as Allelic Imbalance (AI). Breast cancer (BC) is a complex disease which is influenced by genetic variation and level of expression of certain genes, along with other risk factors, e.g. Mendelian inherited gene variants (like *BRCA1* and *BRCA2*) and hormone replacement therapy (HRT). This thesis examines the variation in germline DNA and tumour expression level in BC patients. SNPs in 9 haplotypes associated with Reactive Oxygen Species (ROS) pathways, and previously shown to have significantly different genotype frequencies in BC cases and controls, were genotyped with MassArray in a larger number of BC cases and healthy controls, and the frequency distribution of the two groups was compared. This validation showed that all 9 haplotypes was significantly associated with BC risk. In addition, 20 SNPs in 19 genes were genotyped in tumour RNA with the TaqMan SNP Genotyping assays to measure the level of expression of each allele relative to each other, and 50 % was shown to have significant AI.

Contents

Abbreviations

AA:	Amino Acid
AI:	Allelic Imbalance
AR:	Allelic Ratio
BC:	Breast Cancer
cDNA:	complementary Deoxyribonucleic Acid
CI:	Confidense Interval
CiGene:	Centre of Integrative Genetics
CIN:	Cervical Intraepithelial Neoplasia
Cis:	In close proximity to the gene
CNA:	Copy Number Alterations
CNV:	Copy Number Variation
CT:	Cyclic Threshold
cSNP:	coding Single Nucleotide Polymorphism
DAE:	Differential Allelic Expression
DCIS:	Ductal Carcinoma In Situ
Df:	Degrees of freedom
DNA:	DeoxyriboNucleic Acid
dNTP:	deoxyriboNucleotide Tri-Phosphate
e.g.:	example given
EMSA:	Electrophoretic Mobility Shift Assay
eQTL:	expression Quantiative Trait Loci
HRT:	Hormone Replacement Therapy
HWE:	Hardy-Weinberg Equilibrium
i.e.:	id est (that is)
IGFs:	Insulin-like Growth Factors
iSNP:	intron SNP
KM:	Kaplan-Meier estimator
LD:	Linkage Disequilibrium
LOD:	Log Odds
MD:	Mammographic Density
nrSNP:	non-regulatory Single Nucleotide Polymorphism
nsSNP:	non-synonymous Single Nucleotide Polymorphism
NTC:	Non Template Control
PCR:	Polymerase Chain Reaction
QTL:	Quantitative Trait Loci
Rn:	Normalized reporter value
RNA:	RiboNucleic Acid
ROS:	Reactive Oxygen Species
rSNP:	regulatory Single Nucleotide Polymorphism
SD:	Standard Deviation
SNP:	Single Nucleotide Polymorphism
sSNP:	synonymous Single Nucleotide Polymorphism
TNM:	Tumour, Node, Metastasis
Trans:	Distant from the gene
UMB:	Universitetet for Miljø- of Biovitenskap (Norwegian University of Life Sciences)
VNTR:	Variable number tandem repeats
vs:	versus

1 Introduction

1.1 Genetic variation and expression

1.1.1 Genetic variation

Naturally occurring characteristics in the genome that differ between different individuals in a species are called genetic variation. At any given position two or more versions of the sequence may have emerged during the evolution giving more than one allele. This variation may have arisen due to mutation, unequal recombination, duplication, inversion, or insertion or deletion of a sequence (indels) (Futuyma D J, 2005).

The vast majority of the variation in DeoxyriboNucleic Acid (DNA) sequences is likely to be neutral, with no or little effect on a trait, including susceptibility to disease (Halliburton R, 2004). Some variants may have a large role in the development of a disease, commonly referred to as monogenic, or Mendelian, disease. For instance, cystic fibrosis and Huntington's, are both caused by a mutation in a single gene (Halliburton R, 2004). However, the majority of variation has only a limited impact on disease risk, where increased disease susceptibility is the combination of multiple genetic variants and environmental factors. This type of complex disease could be viewed a a sum of quantitative traits, and the variations affecting it are known as Quantitative Trait Loci (QTL) (Halliburton R, 2004). Different types of variation includes Copy Number Variation (CNV), a common denominator for deletions, insertions, inversions and duplications above 1 kilobase (kb) in size (Redon R et al., 2006), Variable Number Tandem Repeats (VNTR), i. e. different types of short sequence repeats (Halliburton R, 2004), and Single Nucleotide Polymorphisms (SNP), variation in a single base above 1 % in frequency. SNPs are composing 90 % of human genetic variation, with frequency of one per 300 bases in the genome (The International HapMap Consortium, 2003). Though theoretically a SNP may harbour more than two variants they are usually biallelic (Vignal A et al., 2002).

The impact of a SNP may depend on its location. A SNP located in regulatory regions is known as a regulatory SNP (rSNP) and may reside up or down-stream of the gene. A SNP in the coding region is called a coding SNP (cSNP), and in the intronic space they are referred to as intronic SNPs (iSNP) (figure 1). A SNP in the intergenic region may have no effect on regulation of the gene and is then known as a non-regulatory SNP (nrSNP).

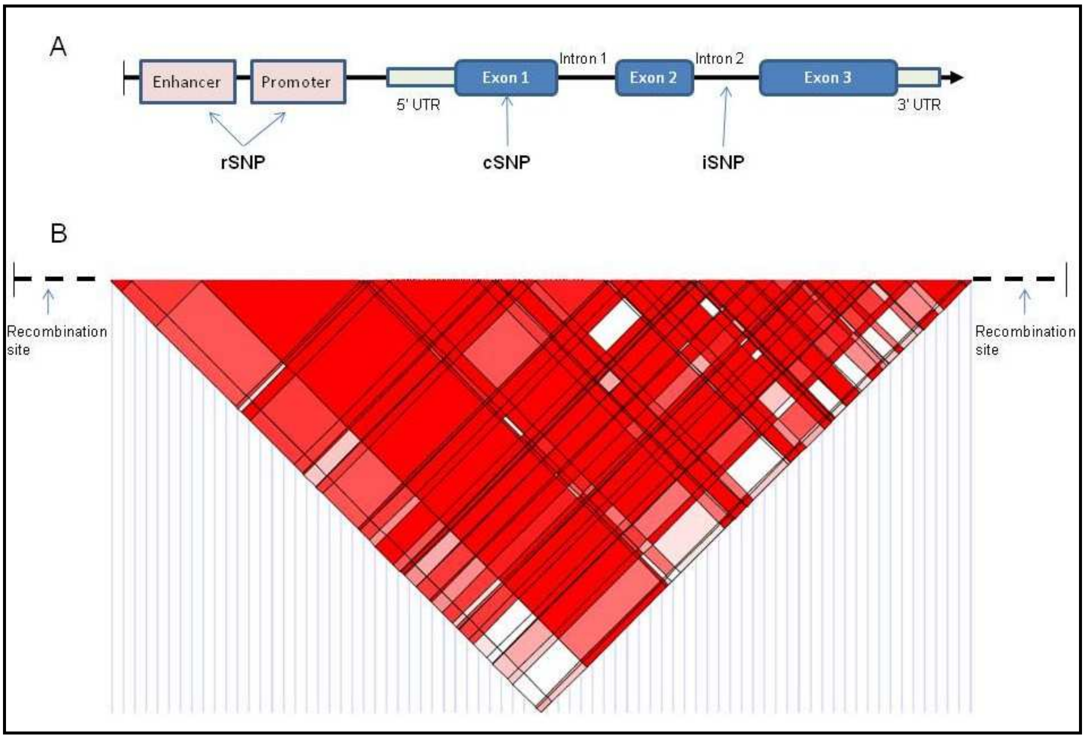

Figure 1. The different positions of SNPs relative to a gene. A. The rSNP is located in the regulatory region, cSNPs in exons, and the iSNP is in the intron of a gene. A SNP in the intergenic region may be an nrSNP, and an rSNP may be located in the coding region as well. **B.** The linkage disequilibrium block marks the SNPs as being linked and composing a haplotype block. The red marks where there is linkage (D' = 1), while the white squares show where recombination occur. The pink squares are areas where D'<0,5 but the log odds (LOD) score is high (*NQO2*, genome.ucsc.edu).

A SNP in the coding region may have an impact on the protein depending on the position in the triplet that makes up one codon. The codon translates to one amino acid (AA), and due to the degenerate nature of the genetic code, a SNP may not give rise to a different AA, referred to as silent or synonymous (sSNP), SNPs resulting in an AA change are called non-synonymous (nsSNP) (Halliburton R, 2004). However, sSNP may still exert a regulatory function and have an effect on the expression. Each codon matches a different transfer RiboNucleic Acids (tRNA), and they are present in different concentrations. This may cause different transcription rates according to the different alleles of an sSNP. The iSNPs may also have an effect on the processing of the transcript if located at specific sites, such as splicing sites or protein binding boxes.

SNPs located in close proximity to each other may be in linkage disequilibrium (LD), which is when two loci are inherited together more often than by chance. Theoretically two loci are considered in LD if the frequency of recombination between them is less than 50 % (Halliburton R, 2004), however, in practice the cut off used is usually lower. Linked SNPs are located between recombination hotspots (figure 1), and the alleles are inherited together as a haplotype.

One great advantage with haplotypes is that if one genotype only a few selected SNPs in the LD block, one may theoretically genotype them all. These SNPs are referred to as haplotype tagSNPs (htSNPs) and they may be identified using the HapMap database (hapmap.ncbi.nlm.nih.gov). About 1 % of all SNPs in the human genome cannot be captured by tagSNPs (Frazer K A et al., 2007), and this is mainly due to their location in recombination hotspots.

1.1.2 Variation in gene expression

The phenotype is affected by the genotype through expression, and expression can itself be considered a phenotype (Rockman M V and Kruglyak L, 2006). Expression of a gene may depend on multiple factors including, in addition to DNA polymorphisms, that are studied here, also micro RNAs and methylation of regulatory site in close proximity to the gene (cis), as well as trans-acting factors such as transcription factors, which may also have regulatory variants, such as DNA polymorphisms. Though the trans-acting regulatory mechanisms are more important for gene expression, 25-35 % of the differences in gene expression level between individuals may be explained by variation in cis-acting regulating sequences (Pastinen T and Hudson T J, 2004). In fact, most known regulatory polymorphisms are located in the promoter regions and the effect these variants have on expression may be important for development and prognosis of diseases (Stranger B E et al., 2005).

Considering expression as a phenotype and the amount of expression as a quantitative trait, it may, like other quantitative traits, be affected by several loci. The variation in these expression Quantitative Trait Loci (eQTL) determines the amount of transcript produced. An eQTL can reside in the regulatory sequence or in the coding sequence of a gene and to identify and determine its effect on the expression is as difficult as with any other quantitative trait. Unless the study involves a very large number of individuals, only those loci harbouring a strong effect on transcript level can be detected. These loci can exert their effect either in cis or in trans or both. The LD between the functional and nearby non-functional loci may complicate their identification, and those loci exerting their effect through haplotypes rather than single polymorphisms may further hamper the detection (Rockman M V and Kruglyak L, 2006).

When alleles have different expression levels at a single locus it is referred to as allelic imbalance (AI), differential allelic expression (DAE) (Maia A T et al., 2009) or allelic-specific expression (Pastinen T and Hudson T J, 2004). The imbalance may be complete, effectively making heterozygotes monozygotically expressed. An example is genes whose expression patterns depend on whether the allele is paternally or maternally imprinted. Imprinting is associated with methylation or histone modification, and interindividual variability in transcription levels of the imprinted genes have been observed (Pastinen T and Hudson T J, 2004). The amount of transcript produced for each allele is affected by functional polymorphisms as well as environmental factors, implying that gene expression may vary between tissues as these harbour different environments (Rockman M V and Kruglyak L, 2006). A recent study suggests that the AI of human blood and healthy breast tissue are similar in a selection of genes with possible association to breast cancer susceptibility (Maia A T et al., 2009).

Allelic imbalance is a common phenomenon in humans (Lo H S et al., 2003) and it may be used to identify the SNPs with an impact on expression and potentially more complex phenotypes. Given the effect of genetic variation on transcription, and the impact of variation on risk and prognosis of complex diseases, such as breast cancer (Chang H Y et al., 2005; Liu R et al., 2007; Naderi A et al., 2007; Sorlie T et al., 2006; van ', V et al., 2002; van d, V et al., 2002; Wang Y et al., 2005), identifying these variants may be a step towards better prediction of risk and outcome. Assuming LD between cSNP and rSNP, measuring AI is a simple and adequate initial screen to identify the candidates for functional validation.

1.2 Breast cancer

Cancer is a collection of diseases recognized by abnormal and rapid growth of cells. Breast cancer (BC) is the most common type of cancer among women worldwide (WHO fact sheet no. 297). Breast carcinomas developed from epithelial cells lining the ducts and lobules are the most common form of breast cancer tumours, but non-epithelial tumours do exist as well (Lee J H et al., 2010).

Several risk factors increase the possibility of developing BC. Being a woman is the most noticeable risk, as less than 1 % of all breast cancer patients are male (Ottini L et al., 2010). Having breast cancer in the family also increases the chances of developing the disease as several genes, including *BRCA1* and *BRCA2*, harbour variants that increase the odds (Antoniou A et al., 2003). Life-history traits, such as late first pregnancy or number of children, also influence the risk (Althuis M D et al., 2004), and environmental factors may play a role (Lof M and Weiderpass E, 2009). Oral contraceptives and hormone therapy may also increase the risk (Althuis M D et al., 2004).

The transformation of a healthy breast into an advanced tumour is a multistage process. Increased density in the breast, as determined by mammogram, is associated with elevated risk of developing breast cancer, and this may be regarded as the first step (McCormack V A and dos S S, I, 2006).

When the tumour has appeared the next steps are the different stages of breast cancer, Ductal carcinoma in situ (DCIS) being the first stage (called Tis in the Tumour Node Metastasis (TNM) staging system). These are tumours of the ducts or lobules without invasion to the nearby tissue. Stages T1-T3 depends on the size of the tumours; T1 being less than 2 cm in diameter, T2 carcinomas between 2 and 5 cm, and T3 being everything above 5 cm. T4 is advanced carcinoma of any size, and are either inflammatory or have extensions either to the chest wall or skin. The TNM classification do in addition take spreading to nearby nodes (N0-3), and metastasis (M0 = no metastasis, M1 = present) into account (Brystkreft. Diagnostikk og behandling. En veiledning., 5th edition)

1.3 Reactive Oxygen Species

Reactive Oxygen Species (ROS) are molecules or ions formed by the incomplete reduction of one electron of an oxygen atom. ROS are important in humans for several reasons, including being part of the phagocytes' arsenal when destroying microbial agents, aiding the regulation of signal transduction and playing a part in the regulation of gene expression. However, they may also cause oxidative damage to nucleic acids, proteins and lipids, and ROS are known to cause mutations in the *TP53* gene, a known tumour suppressor. Factors that create and maintain ROS may therefore contribute to the development of tumours, and antioxidants that destroy ROS may help inhibit tumour development (Pan J S et al., 2009).

1.4 Background

Our department has previously reported a study with genotyping of 1030 SNPs in DNA from blood of 193 female breast cancer patients. The 213 genes selected were involved in ROS metabolism and signalling, DNA repair and apoptosis. (Edvardsen H et al., 2006). Furthermore the patients' germline genotype data were also compared to their tumour's genome wide gene expression data in 50 of the cases. The expression of multiple transcripts showed a significantly higher correlation than expected by chance with SNPs *in cis* (Kristensen V N et al., 2006). By comparing genotype frequencies of breast cancer patients with healthy individuals, a number of SNPs associated with breast cancer risk were revealed (unpublished). To validate the results the SNPs needed to be genotyped in large cohorts of BC patients and healthy individuals (study 1).

A later study investigating the role of functional SNPs in response to certain treatments, found several SNPs with significant association with gene expression (Nordgard S H et al., 2008a). The SNPs of these studies were selected for being associated with the expression of relevant genes, and, in the latter study, for being functional. In addition, an investigation correlating genome wide SNPs and gene expression data, i.e. with no known association with breast cancer was performed, and discovered novel players in the initiation and development of the disease, which are validated here (study 2).

1.5 Aim

The aim of this thesis was to examine genetic variation in germline DNA and variation in gene expression level in breast cancer for a selection of SNPs. This was approached from two different angles:

1. SNPs in ROS pathways with significantly different genotype frequency distribution in breast cancer patients and controls were genotyped in a larger cohort of patients and controls, and the genotype frequency distributions were compared.
2. SNPs previously associated with significantly different gene expression levels were genotyped in heterozygotic RiboNucleic Acids (RNA) with real-time polymerase chain reaction (PCR) to measure the relative amount of each allele. The data was used to calculate the level of AI. In addition, a case control analysis was performed on the germline genotype frequency data.

2 Materials and methods

2.1 Materials

The materials included in this thesis are collected previously for other studies. Informed consent from the donors and approval by the regional ethics board were acquired prior to this study. The materials are listed in table 1, including number of samples per study, and a more detailed explanation of each material follows;

Table 1. Genotyped materials. Column 2, 3 and 4: number of individuals genotyped with MassArray, DNA and RNA samples genotyped with TaqMan for the study of AI, respectively.

Material	MasseArray	TaqMan		Description	RNA isolation*
	N DNA	N DNA	N RNA		
DCIS	-	-	89	Ductal carcinoma in situ tumors.	Column purification
LB	45	22	7	Blood and tumour specimen from females with stage 3 and 4 BC.	Trizol extraction
FU	24	30	30	Blood and tumour specimen from females with stage 3 and 4 BC.	Trizol extraction
LN	105	24	-	Blood from healthy women.	-
MAM04	412	-	-	Blood of patients with stage 2 BC.	-
MB	120	-	-	Blood from BC patients.	-
MDG	185	187	59	Blood and breast biopsies from patients with dense MD** and newly diagnosed breast cancer.	Column purification
Micma	699	132	-	Blood from patients with mainly early stage BC.	-
NOWAC	525	-	-	Blood from healthy women.	-
SIFFK	210	-	-	Blood from healthy women.	-
TMBC	1019	-	-	Blood from healthy women.	-
ULL	119	44	41	Blood and primary tumor from patients with mainly early stage BC.	Trizol extraction
XRAT	273	-	-	Blood from BC patients who received radiotherapy.	-

* Method of RNA isolation used for this cohort
** Mammographic density

DCIS

A material collected between 1986 and 2004 for the study of *TP53* mutations in early stage breast cancer. The 118 tumour specimen were sampled from women with pure Ductal Carcinoma in Situ (DCIS) (N=32), invasive breast carcinoma (N=38) or a mixture of the two (N=48) (Zhou W et al., 2009). RNA was isolated by column purificatio, and 89 of the 118 RNA samples were included here, representing DCIS and early stages of BC.

LB

Blood and tumour specimen collected between1993 and 2001 for a study examining the effect of certain *TP53* mutations on resistance to Doxorubicin treatment and relapse of breast cancer. Patients were between 32 and 88 years of age with locally advanced breast cancer (stage 3 and 4). They were treated with Doxorubicin in an adjuvant setting, and tumour biopsies were taken both before (N=51) and after (N=37) treatment (Aas T et al., 1996). RNA was isolated by the Trizol extraction method. While 45 blood DNA samples were genotyped with MassArray, 22 blood DNA samples and 7 tumour RNA samples of the before treatment batch, were genotyped with TaqMan (representing stage 3 and 4 BC).

FU

This cohort consisted of specimen from 35 patients (37-82 years of age) with stage 3 and 4 breast cancer that received neoadjuvant treatment of 5-fluorouracil and mitomycin. The study examined the role of specific *TP53* mutations in response to a non-anthracycline treatment. Tumour specimen were collected both before and after treatment in the period 1993-2001 (Geisler S et al., 2003). This thesis included 24 blood DNA samples in study 1, and 30 DNA samples from blood and 30 RNA samples from tumour prior to treatment (representing stage 3 and 4 BC) in study 2. RNA was isolated with the Trizol extraction method.

LN

A collection of 109 blood samples from normal postmenopausal women (55-72 years of age), with at least two consecutive negative mammograms over a period of two years, and who were not on Hormone Replacement Therapy (HRT) (Helle S I et al., 2002). LN is geographically matched to LB and FU. In this thesis 24 DNA samples were genotyped with TaqMan and 10 used as controls for the study of AI (study 2), and 105 were genotyped with MassArray and used as control for the case control analysis of SNPs in the ROS pathways (study 1).

Mam04

A cohort of 464 patients (stage 2 and 3) treated with adjuvant radiotherapy between 1998 and 2002 and designed to examine late clinical and biochemical effects of the treatment. The study participants had to be 75 years or younger in 2004 and with no recurrence or other cancers (Landmark-Hoyvik H et al., 2009). In this thesis 412 DNA samples from blood was genotyped with MassArray in study 1.

MB

Blood and tumour DNA collected from 360 breast cancer patients between 1972 and 1991 (31 to 85 years of age), with primary tumour or breast cancer metastasis. The study examined the prognostic significance of selected mutations (Andersen T I et al., 1993). Here 120 blood DNA samples were genotyped with MassArray.

MDG

Biopsies collected from dense breast and small cancers from women aged between 22 and 87 years. The study is ongoing and has currently collected blood DNA and tissue specimen from 121 women without breast cancer and 65 with. The goal is 200 with in 100 in each subgroup. The study is designed to analyse density variation in healthy breast and BC (unpublished). RNA was isolated with the column purification method, and in this thesis 185/187 blood DNA samples and 59 tumour RNA samples from the group with breast cancer were used, representing early stage BC.

Micma

Blood, tumour and bone marrow specimen from patients (32-93 years of age) mainly with stage 1 and 2 breast cancer. The material was used in a study to examine the importance of isolated tumour cells in bone marrow of breast cancer patients (Wiedswang G et al., 2003). In this thesis, 699 blood DNA samples were genotyped for SNPs in ROS pathways with MassArray (study 1) and 132 blood DNA samples were genotyped with TaqMan (study 1).

Nowac

Blood samples collected from healthy women, in the age range of 30-70 years, living in the Tromsø area and with no history of BC, determined by cross-reference to the Norwegian Cancer register. The samples have been collected since 1991 and includes at present more than 100000 individuals. Information was collected through extensive questionnaires, including details about parity, lifestyle, diet and use of HRT. Follow-up ensure that participants who later develop breast cancer are reassigned to the case group. The aim is to create a databank of women representative for the entire female population in their respective age-groups (Lund E et al., 2003). In this thesis 525 DNA blood samples were genotyped with MassArray and used as control for the case control analysis in study 1.

SIFFK

Blood samples from 220 randomly selected apparently healthy women between 20 and 40 years of age collected in 1991/1992, and ensured to show no sign of Cervical Intraepithelial Neoplasia (CIN). These were to be control samples in a study estimating the association between CIN grade II-III and presence of DNA from the human papillomavirus (Helland A et al., 1998). 210 blood DNA samples were genotyped with MassArray and used as controls in this thesis.

TMBC

Blood samples from 1041 healthy women above 50 years of age with a negative mammogram, collected in 2001 and 2002. Females with breast cancer were excluded. The participants were interviewed by a trained nurse concerning their current and previous postmenopausal HRT use, reproductive and menstrual factors, previous history of cancer and smoking status. The participants completed questionnaires in both 2001 and 2002. The study aimed to classify mammograms and examine their relationship to selected risk factors for breast cancer development (Gram I T et al., 2005). In study 1, 1019 DNA blood samples were included in the control cohort of case control analysis of SNPs in ROS pathways.

Ull

Primary tumour samples obtained from 212 breast cancer patients (28-91 years of age) between 1987 and 1994. Blood samples were collected in 1996 from 130 of the patients. The material was used in a study that examined the relationship between abnormal P53 protein and no expression *of P21* in human BC tumours (Bukholm I K et al., 1997). The tumours were stage 1 to 3, and RNA was isolated by the Trizol extraction method. Used in this thesis were 119 samples of blood DNA for the genotyping of SNPs in the ROS pathways (study 1), and 44 samples of blood DNA and 41 samples of tumour RNA for study 2, representing the early to middle stages of BC.

XRAT

Blood DNA from 275 breast cancer patients who received radiotherapy, grouped by the dosage they received. The treatment was performed between 1975 and 1986 and blood samples were collected in 1996. The purpose was to evaluate these patients for adverse sideeffects caused by the radiotherapy (Edvardsen H et al., 2007). In this thesis, 273 blood DNA samples were genotyped with MassArray.

2.2 Methods

2.2.1 RNA isolation

RNA were previously isolated by either of two methods; Guanidinium thiocyanate-phenol-chloroform extraction (TRIzol® extraction) by Invitrogen (do not include a removal of residual DNA step), or column purification with DNAse treatment.

2.2.2 Genotyping of SNPs in study 1 with MassArray

SNPs in genes with connection to the ROS pathways were genotyped in DNA on a MassArray® Platform with the iPLEX® Gold assays. The system is produced by Sequenom, Inc., and its outline is illustrated in figure 2. The SNP of interest and surrounding sequence are amplified by Polymerase Chain Reaction (PCR), and remaining nucleotides (dNTPs) deactivated by a dephosphorylating SAP treatment. Next step is the annealing of primers to the DNA and subsequent extension. The primers are complementary to the sequence adjacent to the SNP and elongated with the SNP. Detection is then performed by a Sequenom MALDI-TOF mass spectrometer. The different alleles of the SNP are differentiated by their different mass.

Briefly, samples were prepared by dilution to 20 ng/µl and transferred to 96-well plates with a volume of 30 µl per sample. Each plate contained 94 samples and two blanks. When the concentration was not previously known, the samples were measured with a Saveen Biotech Nanodrop 1000.

Genotyping was performed at Center for Integrative Genetics (CiGene) at the Norwegian University of Life Sciences (UMB) in Ås, according to the recommended protocol from Sequenom (www.sequenom.com). Assays were designed with the MassARRAY® Assay Design v.3.1 software and ordered from Sequenom, Inc. Sequences are listed in supplementary table 1. Data analysis was performed with MassARRAY® Typer v.4.0 software.

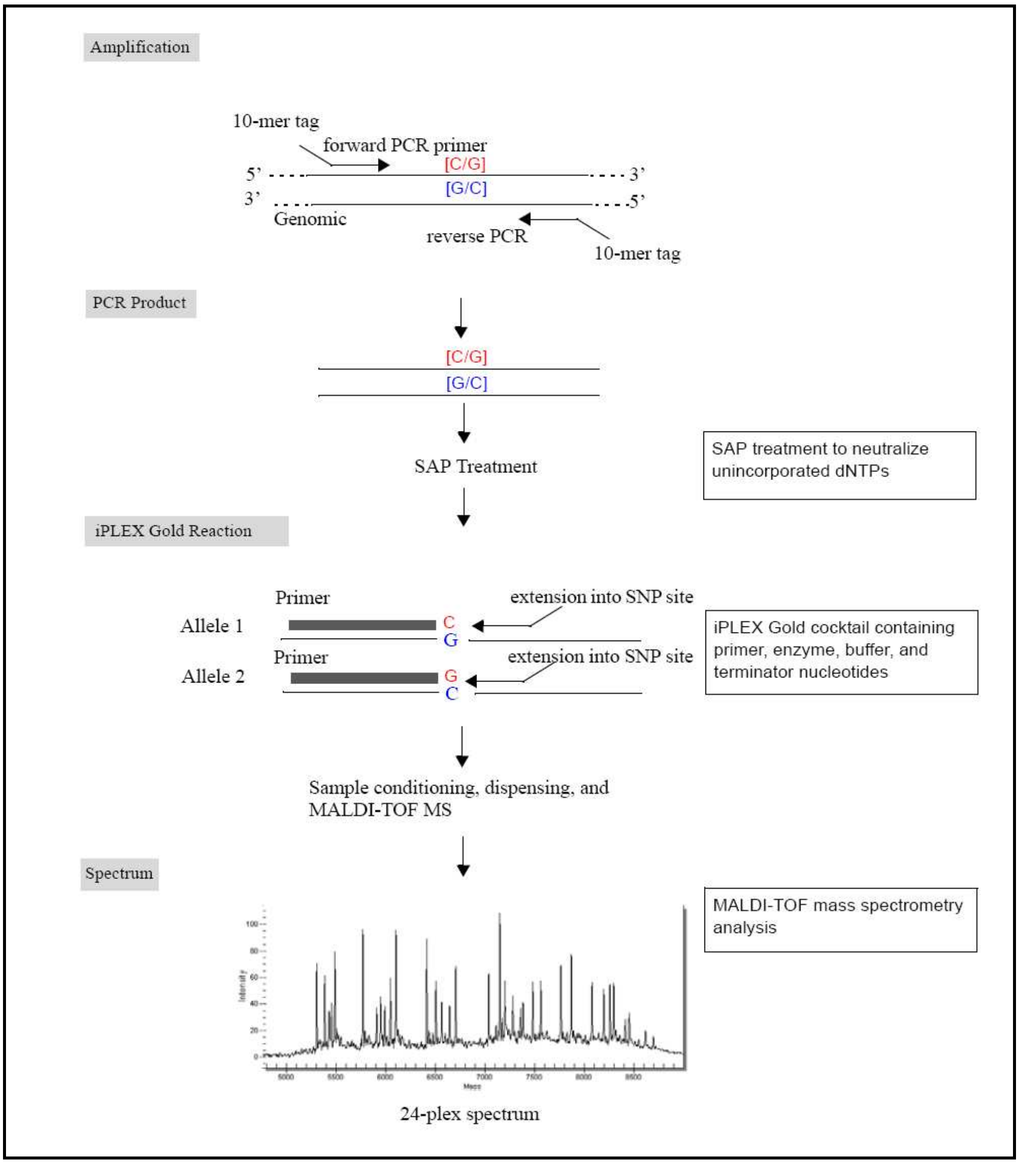

Figure 2. Genotyping with the MassArray System. Flowchart of genotyping (from the top down); PCR amplification of target sequence including the SNP to be genotyped, SAP treatment for removal of unincorporated dNTPs, annealing of the primers to the target sequence and subsequent extension of the SNP, and the measure of nucleotide size with the MALDI-TOF mass spectrometry (figure from www.sequenom.com).

2.2.3 Genotyping of SNPs in study 2 with TaqMan

The 20 selected SNPs were genotyped in DNA and complementary DNA (cDNA), created from the RNA specimens, with TaqMan® SNP Genotyping Assays to determine the allelic ratios (AR) of each gene. TaqMan, outlined in figure 3, is real-time PCR, where the amplification product is measured when produced. TaqMan probes have fluorescent dye attached along with a quencher. The probe attaches to the SNP and the surrounding sequence, and the dye does not fluoresce as long as both dye and quencher are attached to the probe. During PCR, the polymerase destroys the probe, releasing fluorescent dye from quencher. This causes the dye to fluoresce, signalling that the SNP has been polymerized. The probe has to fit perfectly, and the two alleles have a probe with a different dye. The probe with the right allele has the highest affinity for the sequence, and genotype can thereby be read by the emitted signal. The volumes of each reactant for the protocols of DNA and cDNA genotyping are given in table 2.

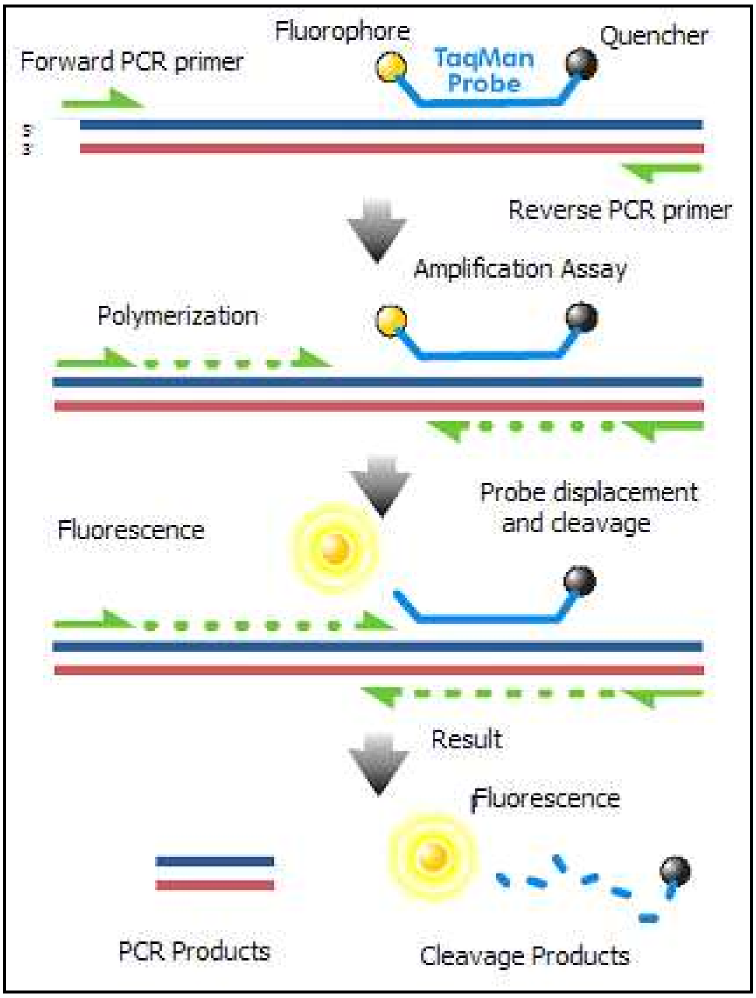

Figure 3. Genotyping with TaqMan®. The probe and primers attach to target sequences (top), followed by polymerization and degradation of the probe, causing dye to be released (middle) and fluoresce. The result is an equal amount of fluorescing dyes and PCR products (bottom), and the amount of PCR product can be measured by how much signal is present (figure from en.wikipedia.org/wiki/TaqMan).

Table 2. The reagents for the TaqMan protocols. Highlighting the differences between the DNA and cDNA SNP genotyping protocols. The volumes are in µl.

Reactant	DNA	cDNA
Mastermix	2,5	5
Primers/probes	0,0625	0,5
H_2O	0	3,83
Template	2,44	0,67
Total	5	10

2.2.3.1 DNA genotyping

Genotyping was performed according to the SNP DNA genotyping protocol (www3.appliedsystems.com), and the volumes are listed in table 2. SNP assays were ordered from Applied Biosystems; 15 pre-designed and 5 custom made. Sequences for the custom assays were retrieved from the SNPper database (SNPper.chip.org), and confirmed by Blat search in the UCSC Genome Browser (genome.ucsc.edu). Sequences for all the SNPs are listed in supplementary tables 2 (pre-designed) and 3 (custom made).

Real-time PCR was performed on a 7900HT Fast Real-Time PCR System with the software SDS 2.3, under Allelic Quantification (AQ) settings, which reads the fluorescence level during the PCR. The Allelic Discrimination post-read process is performed after the PCR and reads the total level of fluorescence present. The post-read predicts the genotypes based on the total amount of signals The DNA template concentration was 5 ng/µl, and there was one Non Template Control (NTC) per SNP per plate, containing only master mix, primer/probes and water.

2.2.3.2 cDNA genotyping

Genotyping of cDNA was performed to determine the level of alleles expressed relative to each other, the allelic ratio, for a given SNP. A random selection of RNA specimen were controlled for quality, and all RNAs were DNAse treated if needed and reverse transcribed into cDNA prior to the genotyping.

2.2.3.2.1 RNA quality control

The purpose of the quality control was to determine whether the RNA samples were degraded, as well as investigating whether the DNAse treatment may have an affect on the RNA. Quality control was performed with Agilent 2100 Bioanalyzer for a subset of random selected RNA specimen prior to cDNA synthesis. Five random FU samples; where 3 were done both before and after DNAse treatment, and 6 Ull samples were chosen.

The Bioanalyzer utilizes a set of micro channels to separate nucleic acid fragments by electrophoresis according to size. When the fragments reach the detection point the bioanalyzer detects and records the fluorescence added to the nucleic acid prior to electrophoresis. The results can then be analyzed by use of the software, which returns estimated concentration, a plot of the fluorescence level versus fragments size, i.e. the time before the fragments reach the detector (figures 7 and 8), and a RNA Integrity Number (RIN) which gives the level of degraded RNA. The preparation of the chip and the analysis of results were performed according to the manufacture recommended protocol (www.chem.agilent.com).

2.2.3.2.2 DNAse treatment and cDNA synthesis

Before cDNA synthesis, removal of residual DNA was performed for all samples isolated by the TRIzol® extraction protocol, due to the lack of DNAse treatment in this protocol. This was accomplished with the DNA-free™ DNAse Treatment and Removal Reagents kit, purchased from Applied Biosystems, according to the producer recommended protocol (www3.appliedbiosystems.com). The kit remove all residual DNA with nuclease free DNAse I, and then degrades the DNAse. The materials FU, LB and Ull were DNAse treated.

cDNA synthesis was performed for all tumour RNA specimen of the 5 cohorts FU, LB, Ull, MDG and DCIS (5 ng of RNA in a 20 µl reaction), with the High Capacity cDNA Reverse Transcription Kit, purchased from Applied Biosystems, according to the manufacture recommended protocol (www3.appliedbiosystems.com).

2.2.3.2.3 TaqMan genotyping

Genotyping with TaqMan® SNP Genotyping assays were performed according to protocol for genotyping of cDNA (see table 2) for all 20 SNPs on the FU, LB and Ull materials, and for 15 of the 20 SNPs for MDG and DCIS (see table 6 for details). In addition, all SNPs, except three, were genotyped with a slightly modified protocol for Ull and/or FU/LB. Table 5 lists the details. The variation in the protocols is as follows:

- DNA protocol with the volume of cDNA recommended by the cDNA protocol (0,67 µl)
- DNA protocol with ½ the volume of cDNA recommended by the cDNA protocol (0,34 µl)
- DNA protocol with ¼ the volume of cDNA recommended by the cDNA protocol (0,17 µl)
- cDNA protocol with ½ the volume of cDNA recommended by the cDNA protocol (0,34 µl)

The cDNA genotyping was performed with the same method, SNP assays, instrument and software as DNA genotyping (see 2.2.2.1). All samples were genotyped in triplets, and included for each SNP on each plate were triplets of three control (LN) DNA samples heterozygote for that SNP (for 50:50 ratio), and triplets of NTC and reference Ambion® RNA control.

2.3 Statistics

2.3.1 Haplotypes and selection of tagSNPs

Haplotypes and htSNPs were determined using Haploview 4.1 (Barrett J C et al., 2005). The haplotype blocks were defined by the four gamete rule. The SNPs are paired and the population frequencies are calculated for all 4 possible haplotypes. Recombination events are assumed to have occurred if all 4 haplotypes are seen with a frequency of 1 % or more. The blocks are formed where only 3 gametes are observed. htSNPs were picked by pair wise tagging only, using the standard r^2-threshold (0,8).

2.3.2 Survival analysis

Survival analysis was performed for the 20 SNPs genotyped with TaqMan with the Kaplan-Meier estimator (KM) and the Cox Proportional Hazards models.

Kaplan-Meier measure the effect a variable may have on survival of each study participant and the risk of reaching the endpoint, e.g. failure or death, at any given time point. The number of individuals (e.g. patients or machinery) in the study are counted at specific times, and as the individuals reach the endpoint they are not counted further. The advantage with the Kaplan-Meier is that it takes into account participants that are removed from the study before the endpoint. These are censored rather than registered as fail, and hence, included in the survival analysis (Kaplan E L and Meier P, 1958).

The Cox Proportional Hazards, like KM, measures the correlation between variables and survival, and the risk of reaching the endpoint at any given time. But, unlike KM, the Cox Model allow for the analysis of the effect of several variables on the survival risk at the same time and is also more useful than KM when one or more of the covariates are continuous (Cox D R, 1972).

For the SNPs in this study, the KM was used to estimate the correlation between genotype and survival, and the Cox model utilized to assay the correlation between expression and survival. The genotypes and survival data was extracted from a previous study at our department on the Micma material (Nordgard S H et al., 2008b). The cohort expression data is currently unpublished. Both survival analyses were performed in SPSS version 16.0.1 (SPSS Inc.), with a p-value significance cut off less than 0,05.

2.3.3 Test for Hardy-Weinberg Equilibrium

The control samples for all 65 SNPs were tested for Hardy-Weinberg Equilibrium (HWE). A population is said to be in HWE when both allele and genotype frequencies remain constant from generation to generation. This indicates that the locus is not influenced by evolution in this population, i.e., no non-random mating, mutation, selection or gene flow influencing this locus.

The test for Hardy-Weinberg was performed with the observed genotype frequencies and the expected genotype frequencies calculated from the former. The observed frequencies are the basis for the allele frequencies (p and q). The frequencies expected for a locus in HWE for the homozygotes is the allele frequency for that allele raised to the power of 2 (p^2 and q^2), while for the heterozygote it is the product of the two allele frequencies and number of alleles in the genotype (2*p*q). It is then possible to compare the two populations, the observed and the expected, with a statistical test. A significant difference would mean that the population is not in HWE. The comparison was performed with a Pearson’s chi-square goodness-of-fit test in Excel 2007 (Microsoft Office). The chi-square takes the difference between the observed and the expected for each of the genotype frequencies, raised to the power of 2, and divides it with the expected frequencies. The sum of the result for each of the genotypes is the test statistic. The p-value (probability of similarity) can then be found with the help of a chi-square distribution table and Degrees of freedom (Df=1 for a HWE test with 3 genotypes) (Halliburton R, 2004). A p-value below 0,001 was considered as a significant deviation from HWE (Haploview 4.1 standard significance threshold, (Barrett J C et al., 2005)).

2.3.4 Case control analysis

To determine whether there is a possible association between the variants genotyped in this thesis and risk of developing breast cancer, a Pearson’s chi-square goodness-of-fit test was performed for all 65 SNPs. In this test the control samples served as the theoretical frequency distribution that the breast cancer cases were tested against. This test was performed in SPSS 16.0.1 (SPSS Inc), and the correlation was considered significant when the p-value was below 0,05.

2.3.5 Calculation of allelic ratios and test for AI

Raw data from the RNA genotyping was taken from the SDS 2.3 software Allelic Quantification setting. This is the point (i.e. in number of cycles) where the increase in fluorescence is at its highest, i.e. the log phase when the reaction has maximum amplification. This is known as the cyclic threshold (CT) and one value is returned for each allele for each well. The CT gives an approximation of amount of mRNA fragment present with the correct genotype for each sample. By dividing the CT for one allele on the other, one can obtain a ratio that show the expression level of one allele compared with the other. A ratio of 1 (0 when log2 transformed) is equal to a 50:50 expression of the two.

For every SNP, raw data was extracted from the SDS 2.3 software, and ratios calculated. The percentage of samples with no CT or a CT higher than 35 (i.e. no calls) was estimated and the ratios removed (figure 5, step 1). An outlier was defined as any sample outside 1,5 times the interquartile range. This range is the upper quartile (75 % of the samples are below this point) minus the lower quartile (25 % of the samples are below this point), the top and bottom lines of the box in a box plot, and 1,5 times this is the distance from the end of the box to a point 1 and a half times the length of the box. Triplets with only one value left after removal of no calls and outliers were excluded (figure 5, step 2). This procedure was performed for each material separately (FU/LB, Ull, DCIS and MDG) and the controls. FU and LB was considered as one material due to their study similarity and small population sizes.

The allelic ratios of the control samples were pooled for each SNP and an average ratio was estimated. For each RNA specimen the ratio was calculated as the average allelic ratio of the triplets or duplets (figure 5, step 2). This ratio was adjusted with control to remove differences in the values caused by the chemical and physical properties of the probes. As the control is DNA from blood, i.e. 50:50 ratio of each allele, this would pull the ratio for equal expression of the alleles down to 1 (0 when log2 transformed) for the samples (figure 4). The adjustment was accomplished by dividing the allelic ratios for each sample with the average allelic ratio for the controls (figure 5, step 4). The samples were then Log2 transformed to generate akin to a normal distribution (figure 5, step 5).

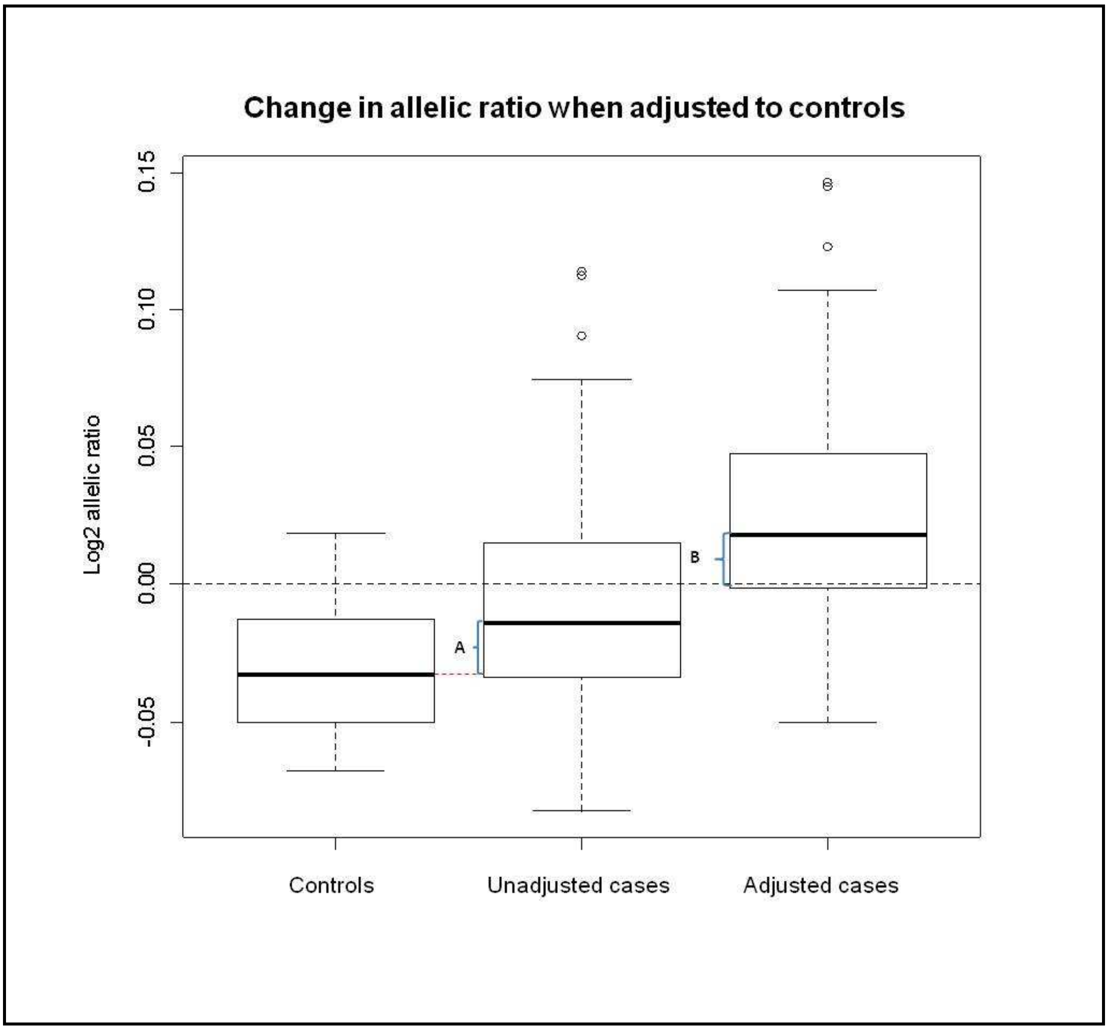

Figure 4. Adjusting the case values to the reference. The Log2 transformed allelic ratio is calculated on the basis of the cyclic threshold given by the SDS software during real-time PCR. The cases are adjusted for the difference in signal caused by chemical and physical properties of the probes, by dividing on the average allelic ratio of the control samples. In the box plot, distance from the average of the unadjusted cases to the average of the controls (marked with **A**) is approximately equal to the distance between the average of the adjusted cases and 0 (marked with **B**), showing that after adjustment the 50:50 ratio of the alleles in the samples would lie at 0 (plot made with R version 2.9.1 (R Foundation)).

Average Log2 adjusted allelic ratios were estimated for each material separately and combined (figure 5, step 5). The case samples were tested for normal distribution (prior to Log2 transformation). This was performed with the Lillifors Significance Correction and Shapiro-Wilk test in SPSS. These tests compare the values with the expected values of a normal distributed population. A p-value below 0,05 for at least one of these tests were considered not normally distributed. A two-tailed Welch T-test was performed for each material and for the combined set for each SNP, if normally distributed. This test returns the probability of the two cohorts being equal, by comparing the mean of the cases to the controls. The two cohorts have different sample size and, presumably, different variances, and therefore a Welch T-test was performed rather than a student's t-test. If the material was not normally distributed a Mann-Whitney U test was performed instead. This non-parametric test serves the same purpose as the t-test, but does not require a normal distribution as it compares the distribution of the samples rather than the mean. The tests were performed on the unadjusted average ratios of the triplets/duplets (figure 5, step 3). Figure 5 displays a schematic overview of the calculation of AR and p-values.

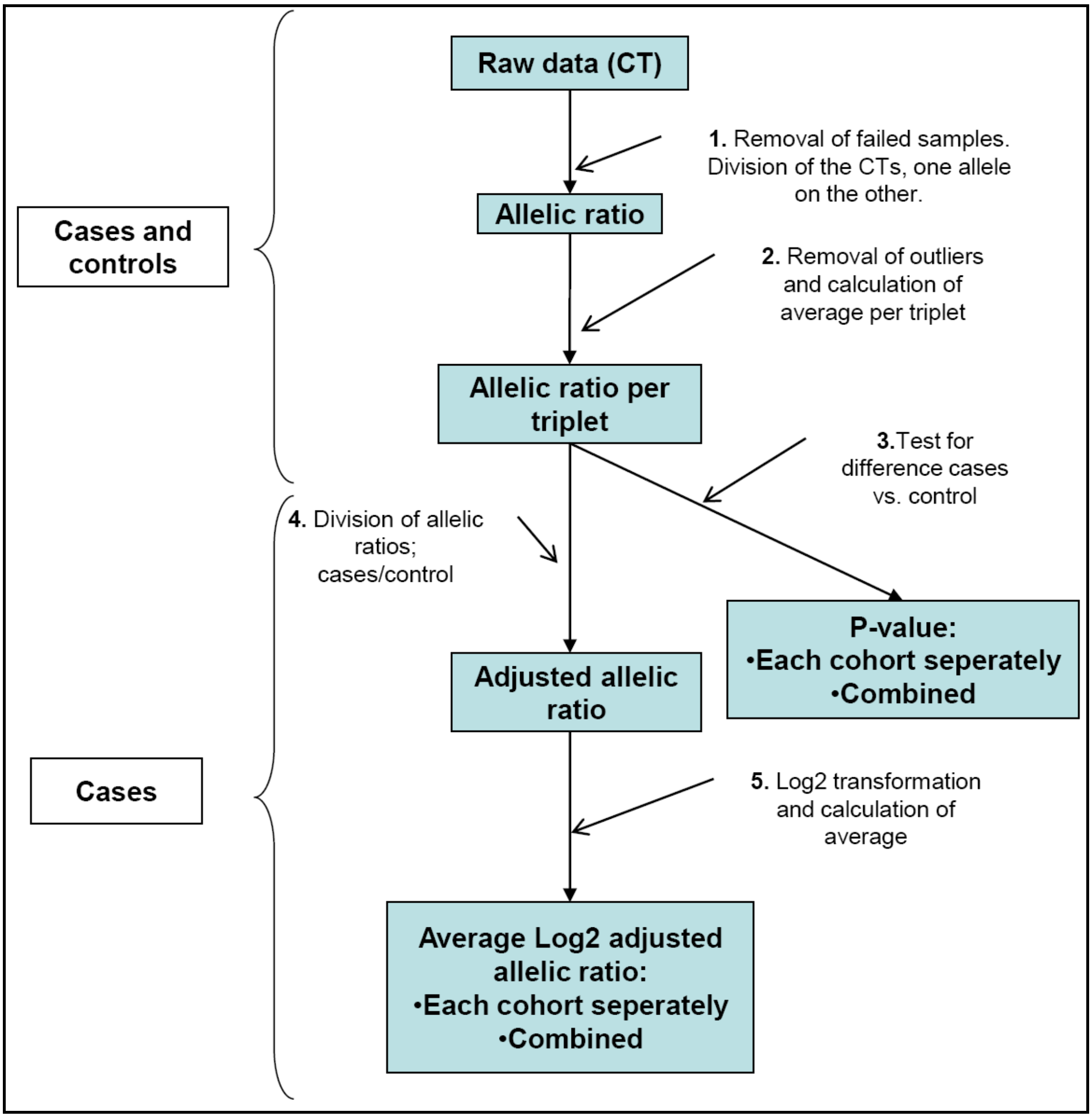
Cases and controls
Raw data (CT)
1. Removal of failed samples. Division of the CTs, one allele on the other.
Allelic ratio
2. Removal of outliers and calculation of average per triplet
Allelic ratio per triplet
3.Test for difference cases vs. control
4. Division of allelic ratios; cases/control
P-value:
•Each cohort seperately
•Combined
Adjusted allelic ratio
Cases
5. Log2 transformation and calculation of average
Average Log2 adjusted allelic ratio:
•Each cohort seperately
•Combined

Figure 5. A step by step outline of the calculation of allelic ratios and tests for differences. 1. The failed samples are removed and the CT for one allele divided by the other. **2.** Outliers are removed and an average allelic ratio per triplet is estimated. **3.** Testing for differences between cases and control. **4.** Adjusting by division with average allelic ratio of the control. **5.** Log2 transformation and calculation of average adjusted allelic ratio.

2.3.6 Aberration detection in the breast carcinomas

The tumour specimens for DCIS, MDG and UII cohorts were inspected for Copy Number Aberrations (CNAs) for each gene genotyped in study 2. The data was extracted from an ongoing study in our department performed with Agilent 244K CGH Microarrays on tumour DNA (unpublished).

3 Results

3.1 Selection of SNPs

3.1.1 Selection of SNPs for case control analysis (study 1)

A previous study genotyped SNPs located in genes associated with the Reactive Oxygen Species (ROS) pathway with SNP-IT™ (Edvardsen H et al., 2006). Using the genotype frequencies from this study, some haplotypes were found to significantly differ between breast cancer cases (N=169) and controls (N=86), indicating a connection between the associated genes and breast cancer risk (unpublished). Furthermore, these SNPs have been previously shown to have an association with tumour expression (Kristensen V N et al., 2006). The 45 SNPs genotyped on the MassArray platform in this thesis were selected for the validation of the result in 1757 cases and 1859 controls. These SNPs represent the htSNPs from all 9 haplotypes that had significant frequency difference between controls and cases in the pilot study, and were associated with the expression level of multiple transcripts. In this thesis each haplotype is named by the gene it is associated with. Table 3 lists all SNPs and genes/haplotypes.

Table 3. The SNPs genotyped in this study. For each SNP the p-values are listed for the Hardy-Weinberg test (controls) and case control analysis (bold font marks the SNPs with significant p-values) In addition, a 95 % confidence interval is given for case control analysis. The SNPs from study 1 are listed alphabetically by gene (haplotype) and the SNPs from study 2 are ordered according to priority.

SNP ID	Gene	Location	GT	Frequency		P-value HWE	P-value case control analysis [95 % CI]
				Cases	Controls		
Study 1							
rs215094	*ABCC1*	Intron	CC	1,8	3,6	0,75	**0,000** [0,000-0,001]
			CT	26,5	29,9		
			TT	71,7	66,5		
rs215067	*ABCC1*	Intron	CC	0,5	1,1	0,005	0,082 [0,074-0,092]
			CT	12,4	13,2		
			TT	87,2	85,8		
rs2062541	*ABCC1*	Intron	AA	37,6	33,8	0,9	**0,035** [0,034-0,046]
			AG	47,4	48,9		
			GG	15,0	17,3		
rs903880	*ABCC1*	Intron	AA	5,2	5,3	0,75	0,963 [0,962-0,973]
			AC	34,2	34,6		
			CC	60,6	60,1		
rs212083_a	*ABCC1*	Intron	CC	67,8	70,8	0,9	0,137 [0,124-0,146]
			CT	28,3	26,1		
			TT	3,9	3,1		
rs212083_b	*ABCC1*	Intron	AA	4,9	4,6	–[1]	–[1]
			AG	28,4	26,1		
			GG	66,7	69,3		
rs1381548	*BCL2*	Intron	CC	40,8	40,7	0,5	0,696 [0,685-0,714]
			CT	45,9	46,8		
			TT	13,4	12,5		
rs1481031	*BCL2*	Intron	AA	9,8	12,4	0,9	**0,016** [0,014-0,022]
			AG	44,9	45,8		
			GG	45,3	41,8		
rs1982673_a	*BCL2*	Intron	GG	3,6	2,8	–[2]	–[2]
			GT	24,6	27,0		
			TT	71,8	70,2		
rs1982673_b	*BCL2*	Intron	GG	3,8	3,0	–[2]	–[2]
			GT	20,6	21,9		
			TT	75,6	75,1		
rs1016860	*BCL2*	3' UTR	AA	0,9	1,4	0,75	0,181 [0,165-0,190]
			AG	18,3	19,9		
			GG	80,8	78,7		
rs2062011	*BCL2*	Intron	AA	54,5	51,4	0,5	0,08 [0,073-0,09]
			AT	39,5	41,0		
			TT	6,0	7,5		
rs1481030	*BCL2*	Intron	AA	100,0	100,0	–[3]	–[3]
			AG	0,0	0,0		
			GG	0,0	0,0		
rs2715438	*IGF1R*	Intron	CC	4,6	3,5	0,5	0,219 [0,206-0,232]
			CT	30,9	32,1		
			TT	64,6	64,4		
rs2137680	*IGF1R*	Intron	AA	7,6	8,5	<0,0001[4]	0,411 [0,404-0,436]
			AG	35,0	33,1		
			GG	57,4	58,4		

rs907807	*IGF1R*	Intron	AA	83,0	82,4	<0,0001[5]	**0,002** [0,001-0,004]
			AG	16,2	15,4		
			GG	0,8	2,2		
rs871335	*IGF1R*	Intron	GG	63,2	53,8	0,75	**0,000** [0,000-0,001]
			GT	32,5	38,8		
			TT	4,3	7,4		
rs1567811	*IGF1R*	Intron	CC	9,5	11,1	0,95	**0,036** [0,035-0,048]
			CG	41,8	44,3		
			GG	48,7	44,6		
rs1568502	*IGF1R*	Intron	AA	59,0	61,3	0,5	0,087 [0,081-0,099]
			AG	35,2	34,3		
			GG	5,9	4,4		
rs2160227	*IL1R1*	Intron	GG	55,7	49,4	0,5	**0,001** [0,000-0,003]
			GT	37,3	42,5		
			TT	7,0	8,0		
rs997049	*IL1R1*	Intron	AA	20,0	16,4	0,95	**0,006** [0,004-0,009]
			AT	48,5	48,1		
			TT	31,5	35,4		
rs1805386	*LIG4*	Coding	CC	3,0	3,0	0,9	0,525 [0,518-0,55]
			CT	27,2	28,9		
			TT	69,8	68,1		
rs1805388	*LIG4*	Coding	CC	70,6	63,9	0,95	**0,000** [0,000-0,001]
			CT	25,2	32,1		
			TT	4,2	4,0		
rs2232640	*LIG4*	Coding	AA	100,0	100,0	–[3]	–[3]
			AG	0,0	0,0		
			GG	0,0	0,0		
rs1805389	*LIG4*	Coding	CC	97,9	98,3	<0,0001	0,669 [0,708-0,737]
			CT	1,8	1,4		
			TT	0,3	0,2		
rs230525	*NFKB1*	Intron	AA	45,7	43,3	0,5	0,241 [0,23-0,257]
			AG	43,1	44,2		
			GG	11,1	12,5		
rs1609798	*NFKB1*	Intron	CC	47,9	45,3	0,1	**0,022** [0,02-0,03]
			CT	42,6	42,5		
			TT	9,4	12,2		
rs230505	*NFKB1*	Intron	AA	0,0	0,0	–[6]	–[6]
			AC	0,0	0,0		
			CC	0,0	0,0		
rs1585214	*NFKB1*	Intron	CC	32,9	35,9	0,25	0,158 [0,144-0,167]
			CT	49,2	46,6		
			TT	17,9	17,5		
rs1801	*NFKB1*	Intron	CC	13,6	15,2	0,25	0,355 [0,342-0,372]
			CG	46,1	45,9		
			GG	40,3	38,9		
rs230531	*NFKB1*	Intron	AA	45,5	43,2	0,5	0,223 [0,206-0,232]
			AG	43,5	44,2		
			GG	11,0	12,6		
rs230498	*NFKB1*	Intron	AA	12,0	14,0	0,75	0,229 [0,217-0,244]
			AG	46,8	45,9		
			GG	41,2	40,1		
rs1598857	*NFKB1*	Intron	CC	36,5	33,8	0,1	**0,05** [0,048-0,062]
			CT	47,1	46,9		
			TT	16,4	19,4		
rs1020760	*NFKB1*	Intron	CC	33,2	30,9	0,25	0,083 [0,076-0,094]
			CG	48,5	48,0		
			GG	18,2	21,0		

rs854539	*PPP1R9A*	Intron	AA	45,6	45,9	0,75	0,968 [0,97-0,98]
			AG	43,8	43,4		
			GG	10,5	10,7		
rs854523	*PPP1R9A*	Intron	CC	29,0	28,6	0,95	**0,011** [0,009-0,016]
			CT	45,3	49,7		
			TT	25,7	21,7		
rs854524	*PPP1R9A*	Intron Coding	AA	27,7	29,4	0,9	0,291 [0,282-0,311]
			AG	49,5	49,8		
			GG	22,7	20,8		
rs854518	*PPP1R9A*	Intron	AA	17,9	17,1	0,5	0,824 [0,817-0,841]
			AT	49,1	49,7		
			TT	32,9	33,2		
rs705377	*PPP1R9A*	Intron	CC	0,0	0,0	_[6]	_[6]
			CT	0,0	0,0		
			TT	0,0	0,0		
rs958379	*PPP3CA*	Intron	CC	64,2	58,4	0,1	**0,002** [0,001-0,004]
			CT	30,6	35,0		
			TT	5,2	6,6		
rs920559	*PPP3CA*	Intron	CC	1,3	2,5	0,1	**0,004** [0,002-0,006]
			CG	20,4	23,2		
			GG	78,3	74,3		
rs1021965	*PPP3CA*	Intron	AA	84,9	83,8	0,5	0,235 [0,23-0,257]
			AG	14,2	15,7		
			GG	0,8	0,5		
rs13340	*TXNIP*	3’ UTR	CC	100,0	100,0	_[3]	_[3]
			CT	0,0	0,0		
			TT	0,0	0,0		
rs7212	*TXNIP*	3’ UTR	CC	93,0	90,8	0,75	**0,048** [0,043-0,057]
			CG	6,8	9,0		
			GG	0,2	0,3		
rs7211	*TXNIP*	3’ UTR	CC	92,6	90,7	0,75	0,093 [0,077-0,095]
			CT	7,1	9,1		
			TT	0,3	0,3		
rs2791749	*TXNIP*	Intron	AA	100,0	100,0	_[3]	_[3]
			AG	0,0	0,0		
			GG	0,0	0,0		
rs2791750	*TXNIP*	Intron	CC	0,0	0,0	_[3]	_[3]
			CT	0,0	0,0		
			TT	100,0	100,0		
Study 2							
rs801719	*CERK*	Coding	CC	47,8	45,1	0,5	0,141 [0,096-0,155]
			CT	40,4	49,0		
			TT	11,8	5,9		
rs1801200	*ERBB2*	Coding	AA	47,4	62,5	0,25	0,068 [0,028-0,065]
			AG	44,0	20,8		
			GG	8,6	16,7		
rs1064608	*MTCH2*	Coding	CC	46,3	40,0	0,25	0,142 [0,105-0,166]
			CG	41,2	52,0		
			GG	12,5	8,0		
rs10409364	*RAB8A*	Coding	AA	10,2	9,2	0,9	0,455 [0,446-0,534]
			AG	47,2	40,8		
			GG	42,6	50,0		
rs12347	*MTRR*	Coding	AA	1,5	1,1	0,75	0,92 [0,942-0,977]
			AG	21,6	23,1		
			GG	76,8	75,8		
rs2015205	*QRSL1*	Coding	AA	32,2	38,2	0,75	0,543 [0,525-0,612]
			AG	48,7	45,1		
			GG	19,0	16,7		

rs3192149	*TOPBP1*	Coding	GG	58,1	62,6	0,25	0,315 [0,294-0,378]
			GT	37,5	30,3		
			TT	4,4	7,1		
rs4129190	*FLJ10916*	Coding	AA	74,0	57,8	0,1	**0,002** [0,000-0,006]
			AG	22,3	40,2		
			GG	3,7	2,0		
rs3088040	*USP36*	Coding	AA	20,2	22,5	0,5	0,681 [0,69-0,768]
			AG	42,6	45,1		
			GG	37,1	32,4		
rs7562391	*PPIL3*	Coding	AA	79,9	82,4	0,05	0,71 [0,728-0,803]
			AC	17,9	14,7		
			CC	2,2	2,9		
rs2243603	*SIRPB1*	Coding	CC	3,7	10,8	0,05	**0,028** [0,015-0,045]
			CG	33,3	31,4		
			GG	63,0	57,8		
rs1143684	*NQO2*	Coding	CC	4,8	4,9	0,9	0,896 [0,925-0,965]
			CT	30,9	33,3		
			TT	64,3	61,8		
rs1494961	*HEL308*	Coding	CC	25,5	16,7	0,25	0,115 [0,085-0,141]
			CT	46,1	56,9		
			TT	28,4	26,5		
rs973730	*ESCO1*	Coding	CC	3,9	2,2	0,75	0,539 [0,531-0,618]
			CT	33,7	29,7		
			TT	62,4	68,1		
rs2863095	*MRPL43*	Coding	CC	57,7	59,4	0,9	0,857 [0,819-0,882]
			CT	38,2	35,6		
			TT	4,0	5,0		
rs2636	*MCTP1*	Coding	AA	24,5	33,7	0,75	0,057 [0,033-0,072]
			AT	44,3	46,5		
			TT	31,1	19,8		
rs2255546	*LRAP*	Coding	CC	23,3	18,8	0,75	0,647 [0,647-0,729]
			CT	50,6	51,8		
			TT	26,1	29,4		
rs2290911	*SH3YL1*	Coding	AA	33,7	31,7	0,9	0,913 [0,904-0,95]
			AG	46,2	48,5		
			GG	20,1	19,8		
rs2294008	*PSCA*	Coding	CC	30,1	36,3	0,5	0,516 [0,48-0,568]
			CT	50,4	45,1		
			TT	19,5	18,6		
rs10380	*MTRR*	Coding	CC	81,1	83,2	0,9	0,853 [0,830-0,891]
			CT	18,1	15,8		
			TT	0,7	1,0		

GT: Genotype

CI: Confidence interval

[1] Not calculated due to problems with the genotyping.

[2] Not calculated due to the presence of two heterozygotic clusters. The accurate frequencies could not be determined.

[3] Not calculated since the SNP is monomorphic

[4] P-value for HW test is 0,005 after removal of the control cohorts that were not in HWE.

[5] P-value for HW test is 0,9 after removal of the control cohorts that were not in HWE.

[6] Not calculated due to genotype frequencies not being available

3.1.2 Selection and prioritising for study 2

3.1.2.1 Selection of SNPs

19 of the 20 SNPs were selected for showing a Bonferroni (BF) corrected significant association between tumour expression level and germline genotypes in an initial analysis of 103 early stage BC patients (Nordgard et al., unpublished). The genotyping of SNPs in transcripts (cDNA) would validate the presence of allelic imbalance in these genes. BF corrects for the higher likelihood of rejecting the null hypothesis when true, when many tests are performed on the data (Bonferroni C E, 1935; Bonferroni C E, 1936).

The SNPs were further selected for being coding SNPs, in order to be expressed, and *in cis*. In addition, the gene had to be outside of known CNV, to prevent this type of variation from interfering with the results. Other selection criteria were htSNPs with a high level of heterozygosity as these would give a higher number of samples to test, and genes with multiple eQTL hits and with relevance to breast cancer. The selected panel consisted of 19 cSNPs located in 18 different genes.

The last SNP, rs1801200, is located in an exon of *ERBB2* (*Her2*), a gene known for its elevated expression (Perou C M et al., 2000) and loss of heterozygosity (Nordgard et al., unpublished) in a subset of breast carcinomas. Further, this SNP was shown to have allelic imbalance in a recent study ((Milani L et al., 2007)), and it may have therapeutic relevance. See table 3 for a list of all the SNPs.

3.1.2.2 SNP prioritisations

In the eventuality that some SNPs had to be excluded from the study due to limited material availability, the 20 SNPs in study 2 were prioritised according to the following characteristics. *ERBB2* (rs1801200) was given a high priority due to its clinical relevance. The other SNPs were prioritised first according to the germline frequency of heterozygosity (a high frequency gives increased statistical power due to higher number of samples), and secondly after the results of the survival analysis. One SNP was significantly associated with survival, *CERK* (rs801719), with a p-value of 0,01 (figure 6). None of the genes were significant for the expression vs. survival test with the Cox model, but the transcript associated with rs801719, had a p-value of 0,111. Seen together, the models give an indication of a correlation between the gene, *CERK*, and survival, giving this gene a higher priority.

After the initiation of cDNA genotyping, the failure rates became another priority variable. The SNPs with the highest failure rates had lower priority and those with a failure rate above 90 % were excluded from further genotyping. Table 3, 4 and 6 are ordered according to this combined priority.

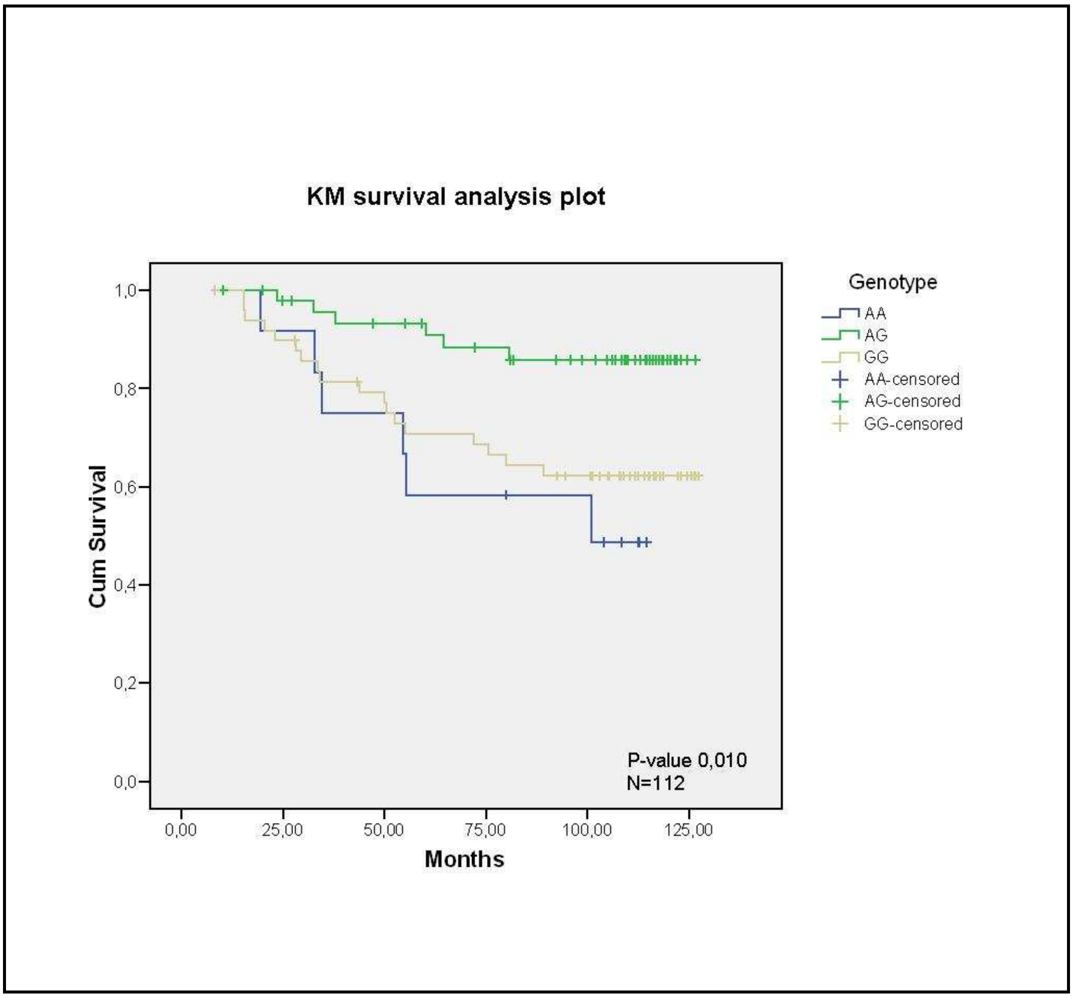

Figure 6. Correlation between a variant in *CERK* and BC specific survival. The SNP, rs801719, showed a significant association with survival, p-value of 0,01, in a cohort of 112 early stage BC. The heterozygote is associated with higher probability of survival than either of the homozygotes (plot extracted from SPSS 16.0 (SPSS, Inc.)).

3.1.2.3 Selection of cases for genotyping in tumour cDNA in study 2

All SNPs were genotyped in germline DNA to determine the heterozygotic cases, which would be genotyped in tumour cDNA. The results are listed in table 4. The LN, Micma and partially MDG cohorts had been genotyped in a previous study at our department for 19 of the SNPs (Nordgard S H et al., 2008b), and was not genotyped again in this study. The MDG cohort was only genotyped with TaqMan for those individuals that did not have previous genotype data. The SNP in *ERBB2* (rs1801200) did not have any genotype data for either of the cohorts. In addition to the cohorts listed in table 4, 132 of the Micma cohort and 24 of the LN were genotyped in germline DNA, and the number of heterozygotes was 57 and 5 respectively. Neither cohort was genotyped in tumour cDNA, but 3 heterozygotic LN specimens were used as control in the AI study. The DCIS cohort did not have any germline DNA available, and was genotyped in 89 tumour cDNA specimens for all SNPs.

Table 4. Number of specimen in the genotypings of germline DNA and tumour cDNA. Also included is the number of heterozygotic individuals revealed in the DNA genotyping. For some SNPs there is a higher amount of specimens in the cDNA genotyping than in the DNA genotyping for the MDG cohort, as genotype data was extracted from a previous study.

	FU/LB			**UlI**			**MDG**		
SNP	**N DNA**	**N Het.**	**N cDNA**	**N DNA**	**N Het.**	**N cDNA**	**N DNA**	**N Het.**	**N cDNA**
rs801719	52	18	13	44	19	19	61	27	22
rs1801200	52	25	21	44	17	16	185	64	22
rs1064608	52	20	13	44	16	16	61	30	26
rs10409364	52	28	22	44	17	16	61	25	28
rs12347	52	9	7	44	11	11	61	8	13
rs2015205	52	26	17	44	21	19	61	29	27
rs3192149	52	25	16	44	17	18	61	27	19
rs4129190	52	9	6	44	9	9	61	19	16
rs3088040	52	22	17	44	14	13	61	34	32
rs7562391	52	6	6	44	7	6	61	13	15
rs2243603	52	15	10	44	16	15	61	23	23
rs1143684	52	16	12	44	11	10	61	24	21
rs1494961	52	22	13	44	25	23	61	33	32
rs973730	52	19	14	44	16	15	61	24	17
rs2863095	52	16	11	44	13	13	61	27	28
rs2636	52	26	18	44	19	19	61	32	0
rs2255546	52	30	21	44	18	15	61	27	0
rs2290911	52	23	15	44	13	12	61	33	0
rs2294008	52	20	16	44	19	16	61	34	0
rs10380	52	8	6	44	8	8	61	7	0

N DNA: Number of germline DNA samples genotyped.
N Het.: Number of heterozygotic samples found in the genotyping of germline DNA.
N cDNA: Number of tumour cDNA samples genotyped.

3.2 Genotyping of germline DNA

3.2.1 Success rates and reproducibility

3.2.1.1 Success rates and reproducibility in study 1

Success rates were above 98 % for all 45 SNPs with the exception of two, rs705377 and rs230505, which failed completely. Five SNPs were monomorphic and they were monomorphic, or almost monomorphic, in the HapMap database, but not in the initial study population genotyped with SNP-IT™. In addition, 51 samples failed completely for all SNPs.

One SNP, rs1982673, had two heterozygotic clusters making the results unreliable (table 3; for rs1982673_a all heterozygotes are included, and for rs1982673_b only the largest cluster is included), and one SNP, rs212083, was genotyped twice (rs212083_a and rs212083_b in table 3) due to erroneous primers caused by the primer design software. Of these two, rs212083_b is not particularly good (success rate 92 %), however this round showed 98 % similarity with rs212083_a. In addition three random SNPs were controlled against the genotypes from the SNP-IT study for samples that were genotyped both times, and all three had less than 3 % mismatch between the two.

3.2.1.2 Success rates and reproducibility in study 2

Success rates for genotyping DNA on TaqMan were above 97 % for all 20 SNPs and 50 % showed a 100 % success rate. Random samples were regenotyped for four random SNPs and found to match 100 % with the previous results.

The control samples in the cDNA genotyping rounds had a general high level of success for the 15 SNPs genotyped in all materials. Only *USP36* (rs3088040) had a high failure for the controls. For all 15 SNPs at least 27 allelic ratios for 3 samples were obtained from the software (3 samples x triplet x 3 separate rounds of genotyping = 27 allelic ratios), but for rs3088040, 7 allelic ratios in 2 different samples passed processing criteria (all failed samples and outliers removed). The remaining 14 of the 15 SNPs had more than 20 allelic ratios left for all three control samples. For the 5 SNPs, with >90 % failure rates for the tumour cDNA specimen, there were 0-9 allelic ratios left of the controls depending on each SNP. These were only genotyped once, and so there were only 9 allelic ratios possible (3 samples x triplets).

The same three control samples were genotyped 3-6 times depending on the SNP. The raw data (CT for each allele) displayed in general very few discrepancies between the different genotyping rounds. And as any sizable difference that might exist for some SNPs were presumed removed as outliers, these differences were therefore believed to have no effect on the calculation of allelic imbalance.

3.2.2 Hardy-Weinberg Equilibrium and case control analysis

The threshold for significant p-values are set at 0,001 for the test for HWE in the control samples, using the Haploview 4.1 standard, and 0,05 for the case control analysis. For a full list of all p-values for all SNPs see table 3.

3.2.2.1 The SNPs genotyped in study 1

All SNPs were in Hardy-Weinberg equilibrium with the exception of three. Two of these, rs2137680 and rs907807, were out of equilibrium due to specific control cohorts, and hence these populations were excluded from all further analysis on these SNPs, both located in *IGFR1*. When these groups were removed the control samples were in HWE with p-values of 0,005 for rs2137680 and 0,9 for rs907807. The purpose of the removal of these populations was to bring the controls into HWE so the case control analysis could be performed. The last SNP, rs1805389 (*LIG4*), deviated from the HWE for all control groups.

All significant p-values from the case control analysis are marked in bold in table 3. All genes had at least one significant SNP. *IGFR1* and *IL1R1* had the highest amount of significant SNPs, with 3 out of 6 and 2 out of 2, respectively. The two SNPs deviating from HWE in some of the control cohorts (rs2137680 and rs907807), had p-values of 0,411 and 0,002 respectively. After removal of the control groups that deviated from HW, the p-values for cases vs. controls were 0,104 for rs2137680 and 0,259 for rs907807, effectively removing the significance of the latter.

3.2.2.2 The SNPs genotyped in DNA in study 2

To increase the statistical weight of the case control analysis, germline genotypes from more individuals were acquired from an ongoing study at our departement (Nordgard S H et al., 2008b), where the cohorts extracted were 102 controls and 112 cases. There was no previous genotype data available for *ERBB2* (rs1801200), and the case control analysis is therefore performed on data from individuals genotyped in this study (266 cases and 24 controls). Two of the 20 SNPs in the TaqMan genotyping, *FLJ10916* (rs4129190) and *SIRPB1* (rs2243603), showed significant difference in frequency between case and control (bold, table 3), with p-values of 0,002 and 0,028, respectively. In addition, 2 SNPs, *ERBB2* (rs1801200) and *MCTP1* (rs2636), were borderline insignificant, with p-values of 0,068 and 0,057, respectively. None of the SNPs showed a significant deviation from the HWE for the controls.

3.3 Genotyping of tumour cDNA in study 2

3.3.1 Quality control and reproducibility

3.3.1.1 Quality control

A quality control was performed to determine if the RNA were degraded and whether the DNAse treatment would have a negative effect on the RNA. The DNAse treatment was performed to ensure that all RNA samples genotyped was completely free of genomic DNA as presence of this could interfere with the results. Only FU, LB and Ull were treated with DNAse. These materials had been isolated with the Trizol extraction method, a protocol that do not include DNA removal step. DCIS and MDG were isolated by the column purification method, which included DNAse treatment, and therefore, there was no need to perform another DNA removal.

The results of the quality control with the bioanalyzer are shown in representative plots (figure 7 and 8). All controlled samples were similar to these and displayed no or limited degradation of RNA. All expected peaks were present, and peaks indicating DNA was not observed. The RNA showed an equally fine quality after DNAse treatment (bottom figure 7) as prior (top figure 7), indicating no or limited ill effect of the treatment. The plots showed no indication of DNA present in the RNA samples, however the bioanalyzer plots may not detect small amounts of DNA, and so the DNAse treatment was performed on all samples isolated with the Trizol extraction.

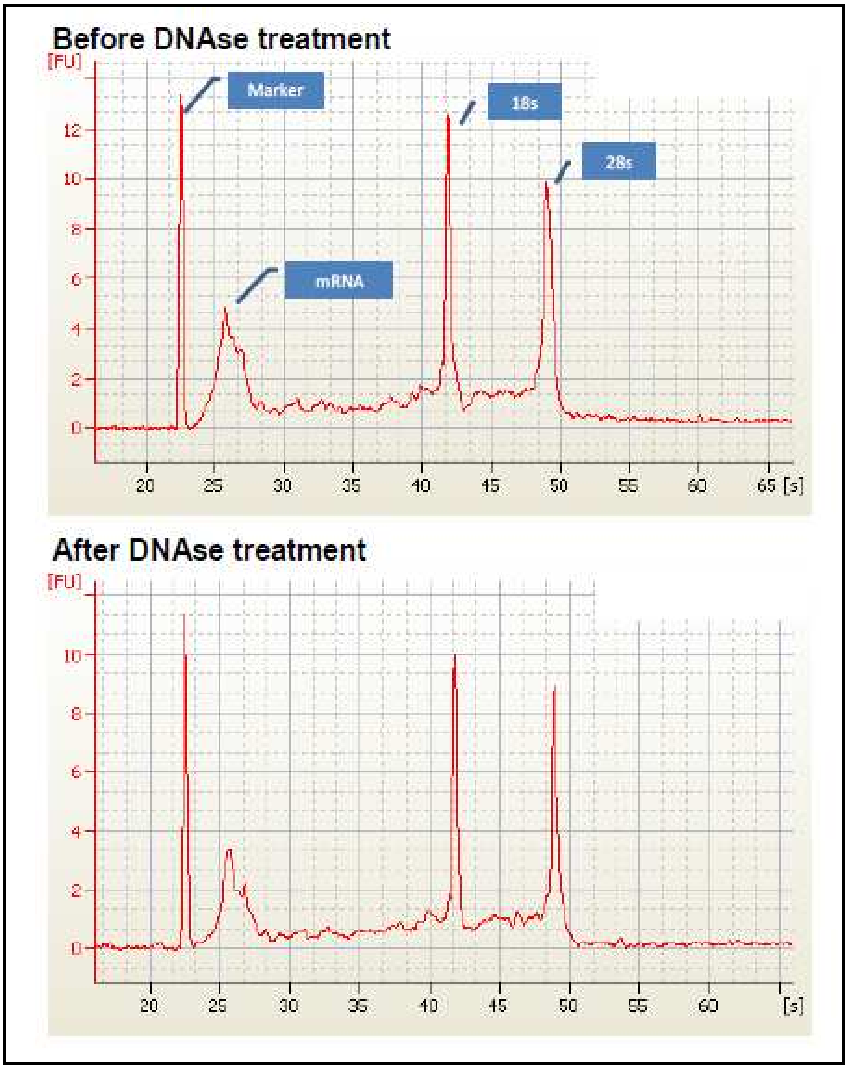

Figure 7. RNA quality before (top) and after (below) DNAse treatment. The y-axis is the fluorescence level measured in fluorescence units (FU), and the x-axis is the time (in seconds (s)) the sample runs through the electrophoresis. The upper figure have markers explaining all visible peaks; the first and second are the marker and mRNA, the third and fourth the 18s and 28s ribosomal RNAs. If there was DNA present another clear peak would be visible (plot from the Bioanalyzer software).

The UII cohorts displayed higher failure rates than FU/LB (see table 6). Yet, the Bioanalyzer showed that the RNA samples for UII have generally good quality (Figure 8). RIN is a number that indicates the level of degradation in the RNA. Seven and up shows good quality with limited degradation. RIN value of 5 indicates some partial degradation (agilent.com). All samples tested (both FU and UII) had RNA Integrity Number (RIN) above 7, with the exception of one UII sample with a RIN of 5,1.

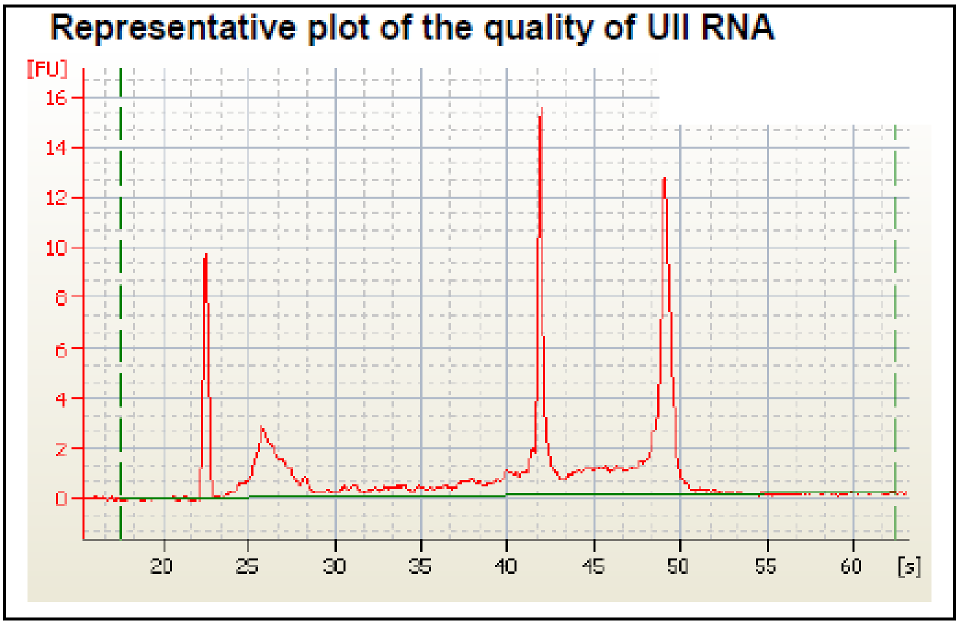

Figure 8. The general RNA of the UII cohort. As in figure 7, the expected peaks are present and no DNA peak was visible (plot from the Bioanalyzer software).

3.3.1.2 Failure rates and reproducibility

Failure rates are listed in table 6, in the column marked % no call. As the SNPs were prioritized partly according to failure rates, there is a clear pattern in the table. The percentage was generally low (< 50 %) for the first 10 SNPs in table 6. The rest showed a relative high amount of failure, though the extent depended on the cohort. The last 5 SNPs had more than 90 % failure, and of these 3 had 100 %. As a consequence, these SNPs were not cDNA genotyped for DCIS and MDG cohorts, and AI was not estimated for these SNPs due to low power in the remaining samples. A BLAT search in the UCSC genome browser was performed for the probe sequences of the 10 SNPs with high failure rates, showing that 60 % were directly on the border to an intron, and for the SNPs with >90 % failure rates the frequency was 4 out of 5.

When examining failure rates per material, it is higher for Ull than the others. One extreme example is *MRPL43* (rs2863095) with 100 % failure for Ull and 3 % for DCIS. However, the differences in failure rates are limited when considering SNPs with generally low failure rates. It is the SNPs with elevated failure rates that show a clear difference between the materials. After Ull, FU/LB has the highest failure rates, with MDG and DCIS having generally low failure rates for almost all 15 SNPs.

Genotyping of the FU cohort was repeated 3 times for *TOPBP1* (rs3192149). In all three independent rounds the allelic ratio and failure rates remained approximately the same.

The post-read (Allelic Discrimination) gives a normalized reporter (Rn) value for each allele for each well. This value is the signal from the reporter divided by the signal of the passive reference dye in the master mix, which is added to measure the background signal. This implies that the Rn value is the signal value where the background signal variations between wells are removed. The Non Template Controls (NTC) had relatively high Rn values for many of the SNPs. Initially the elevated NTC values were assumed to be caused by contamination, however, the Rn values for the NTCs for all SNPs were approximately 4 times higher than the corresponding value in the DNA genotyping. This coincides with the fact that the percentage of primers and probes of the total volume is 4 times higher for the cDNA protocol compared with the DNA protocol (table 2). The amount of primers and probes compared to the total volume is 1,25 % (0,0625 µl primers and probes of 5 µl total volume) for the DNA protocol, while it is 5 % (0,5 µl primers and probes of 10 µl total volume) for the cDNA protocol. In addition, the multicomponent plots given by the SDS 2.3 software showed no amplification (see figure 10 for a comparison of the fluorescence signals of NTC and amplified samples).

3.3.1.3 Optimalisation of cDNA genotyping

For selected SNPs several different genotyping protocols were run on Ull and FU. This optimalisation was performed to lower the general failure rates, and was based on two protocols; the regular cDNA protocol with a full volume of template (0,67 µl) and DNA protocol with half the volume (according to the cDNA protocol) of template (0,34 µl). For a list of the differences in the two protocols (DNA and cDNA protocol) see table 2. In short, the cDNA protocol has a lower template volume and a higher volume of primers/probes than the DNA protocol, plus double the total volume. When the template volume is decreased the amount of water added is increased so the total volume remains the same and the template concentration is diluted. The DNA protocol with template volume from the DNA protocol (2,44 µl) was also attempted, but no amplification was seen, presumably due to the high template concentration. Other protocols attempted was DNA protocol with full (0,67 µl) and one quarter (0,17 µl) volume of template and cDNA protocol with half (0,34 µl) volume of template (of the cDNA protocol). Results are listed in table 5. Those SNPs (*FLJ10916* (rs4129190), *USP36* (rs3088040) and *ESCO1* (rs973730)) only genotyped according to the cDNA protocol with full template volume are not listed.

Generally, the allelic ratio remained approximately the same independent of the genotyping protocol, being within the natural variation of each round as represented by the standard deviation (SD). Only *RAB8A* (rs10409364) had a higher difference between the two allelic ratios than the SD. However, the failure rates (% no call) is generally higher for the DNA protocol with half the volume of cDNA than it is for the regular cDNA protocol. The SNPs rs3129149 and rs1494961, were tested with different template volumes as well as different protocols and show a higher level of failure rates when halving the template volume. In addition, they show the same difference in failure rates when only the protocols differ, and the template volumes are the same. Four SNPs, *QRSL1* (rs2015205), *MRPL43* (rs2863095), *LRAP* (rs2255546) and *SH3YL1* (rs2290911) showed a lower failure rate for one or both materials with the DNA protocol compared to the cDNA protocol. However, this difference was either very low or, in the case of *MRPL43*, Ull showed a lower failure rate while FU was higher for the DNA protocol.

The final calculations of allelic ratios were performed on the results from the genotyping with the cDNA protocol (full volume of template).

Table 5. The different protocols and amounts of cDNA. All SNPs are listed, with the exception of *FLJ10916*, *USP36* and *ESCO1*.

		UII			FU		
SNP	**Protocol**	**% no call**	**Allelic ratio**	**SD**	**% no call**	**Allelic ratio**	**SD**
CERK	cDNA	7	1,03	0,04	10	1,00	0,01
	DNA	33	0,99	0,03	23	1,02	0,02
ERBB2	cDNA	58	1,01	0,03			
	DNA	88	1,00	0,01			
MTCH2	cDNA	0	0,99	0,01			
	DNA	13	1,02	0,02			
RAB8A	cDNA	13	1,04	0,03			
	DNA	29	1,00	0,03			
MTRR (rs12347)	cDNA	12	1,03	0,02	10	1,07	0,01
	DNA	36	1,02	0,01	81	-	-
QRSL1	cDNA	28	1,02	0,01			
	DNA	26	1,02	0,02			
TOPBP1	cDNA				30	0,99	0,04
	DNA*				46	0,99	0,01
	DNA				21	1,00	0,02
	DNA***				58	1,03	0,05
PPIL3	cDNA	72	-	-	61	-	-
	DNA	100	-	-	100	-	-
SIRPB1	cDNA	96	-	-			
	DNA	98	-	-			
NQO2	cDNA	90	-	-	42	1,01	0,02
	DNA	93	-	-	78	1,01	0,02
HEL308	cDNA	97	-	-	56	0,98	0,02
	cDNA**				80	-	-
	DNA	100	-	-	85	1,00	0,01
MRPL43	cDNA	100	-	-	55	0,96	0,01
	DNA	79	0,97	0,01	100	-	-
MCTP1	cDNA	100	-	-	89	-	-
	DNA	100	-	-	100	-	-
MTRR (rs10380)	cDNA	100	-	-	100	-	-
	DNA	100	-	-	100	-	-
LRAP	cDNA	100	-	-	95	-	-
	DNA	91	-	-	97	-	-
PSCA	cDNA				100	-	-
	DNA				100	-	-
SH3YL1	cDNA				100	-	-
	DNA				89	-	-

SD: Standard deviation
-: Not enough material for calculation
Blank: SNP was not run for both protocols with that material.
cDNA protocol is in general performed with 0,67 µl cDNA, and DNA protocol with 0,34 µl, with a few exceptions:
* 0,67 µl cDNA
**0,34 µl cDNA
*** 0,17 µl cDNA

3.3.2 Allelic Imbalance

Allelic ratio is based on the CTs of each allele for each sample. The fluorescence from the genotyped allele with the VIC dye was divided by the signal from the allele with the FAM dye. Figure 9 displays the raw data for one SNP grouped by cohort. The y-axis is the allele marked with the VIC dye, and the x-axis is the FAM dye. The CT value is reversely proportional to the amount of target cDNA it takes to reach the amplification maximum. In other words, the higher the CT value the lower the amount of RNA present originally.

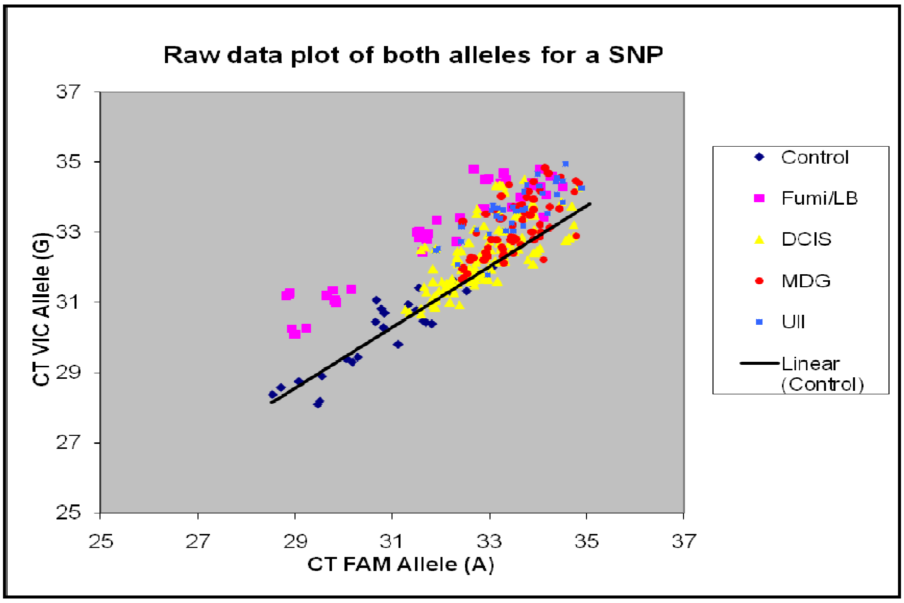

Figure 9. Raw data of all samples for a SNP (rs2015205). The SNP was chosen for being representatively skewed without being amongst the most skewed SNPs. Coloured according to cohort and with a ratio reference line based on the control samples. The y axis is the cycle threshold (CT) for allele G (VIC dye), and the x axis for allele A (FAM dye). The 50:50 ratio would be a line going from origin (0,0) and in a 45° diagonal up and right. The fact that the controls do not cluster along that line suggests a strong influence by the chemical and physical properties of the different probes on the raw data values. The samples are clustered above the ratio reference line, showing a tendency for a higher cycle threshold for the G allele, which, if true, would mean that this allele is underrepresented compared to the A allele (plot made in Excel 2003 (Microsoft Office)).

Figure 9 demonstrates the importance of adjusting the samples for differences between the allelic probes. The plot indicates that the FAM probe has a higher signal intensity than VIC for this SNP. The inequality of the probes is also shown for a different SNP in figure 10. These plots are made on the basis of the fluorescence signals per cycle. These plots show that the sample with the high allelic ratio (B) has less difference between VIC and FAM than the one with the low sample (A). However, the controls (C) do also show a large deviation from the 50:50 reference line, indicating that the majority of the difference is caused by the probes rather than the AI. Adjusting the samples for the control nullifies this difference, and B has in fact the highest AI. When there is no visible amplification (NTC (D)), the two signal levels are almost the same, indicating that the dyes themselves are not very different from each other, but rather that the probes' abilities to attach to their corresponding sequence differ.

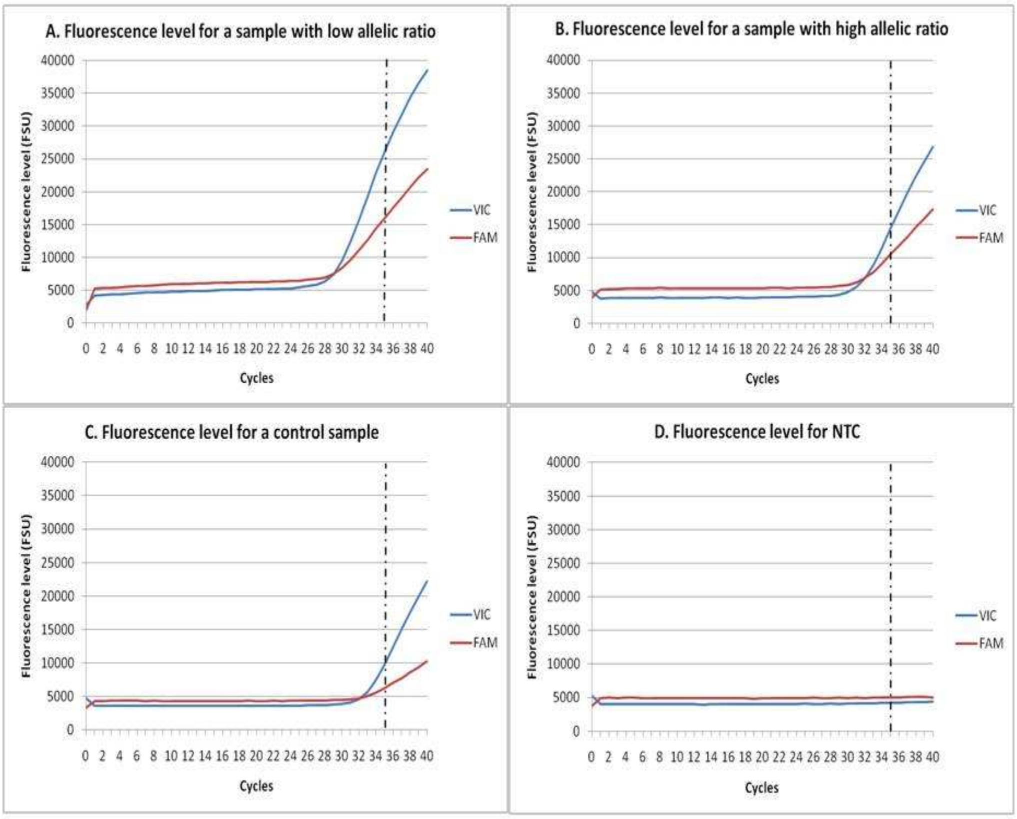

Figure 10. Fluorescence levels per cycle for a selection of samples for one SNP (rs1064608). The fluorescence level is given in fluorescence spectra units (FSU). **A**. Sample with low allelic ratio (0,017 log2 adjusted allelic ratio), **B**. High allelic ratio (0,087 log2 adjusted allelic ratio), **C**. Control sample and **D**. None Template Control (NTC). The vertical line shows the lower limit to what is considered expressed (35 cycles) (plots made with Excel 2007 (Microsoft Office)).

The threshold for what is considered expressed was set to be 35 cycles, which is the same used by Maia et al. (Maia A T et al., 2009). Every allele above this threshold was removed as failed or not expressed, and the corresponding allele was removed as well regardless of CT.

There was no prior knowledge of the genotypes for patients in the DCIS cohort. No blood samples were available; hence, the germline genotype could not be determined prior to cDNA genotyping. The homozygotes was removed if the SDS 2.3 post-read (Allelic Discrimination) gave a clear calling of the genotypes or if there were three clear groups to infer the genotypes from. For the SNPs with undistinguishable genotypes, all CTs below 35 were used. For these the actual percentage of failure could not be inferred due to failed heterozygotic samples being indistinguishable from homozygotes. These are marked as N/A in the % no call column of table 6 for DCIS, and the combined failure rate are calculated based on no call numbers for MDG, Ull and FU/LB cohorts. The SNPs without calling show generally a lower average allelic ratio and a wider spread of the individual allelic ratios than the SNPs that did not get any genotype calling (table 6 and figure 11).

Table 6. Degree of AI in breast carsinomas. Materials are ordered by stage, with the combined set first. SNPs are listed by priority. Blank cell; the material was not genotyped for this SNP, or not enough samples were available for calculation.

		All				MDG				DCIS				UII				Fumi/LB			
SNP	Gene	% no call	N	AR	P-value	% no call	N	AR	P-value	% no call	N	AR	P-value	% no call	N	AR	P-value	% no call	N	AR	P-value
rs801719	*CERK*	3	84	-0,01	>0,1*	2	22	-0,02	0,01	0	39	-0,02	0,01*	7	18	0,04	0,001	10	10	-0,01	>0,1
rs1801200	*ERBB2*	43	48	0,02	>0,1	20	17	0,02	>0,1	N/A	16	0,01	>0,1	58	6	0,01	>0,1	56	9	0,02	0,1
rs1064608	*MTCH2*	1	97	0,04	>0,1*	0	25	-0,02	0,01*	2	46	0,08	<0,0001	0	16	-0,01	>0,1*	0	10	0,02	0,01
rs10409364	*RAB8A*	2	91	0,00	>0,1*	0	23	0,02	0,1	0	34	-0,08	0,001	13	13	0,06	0,001	0	21	0,08	<0,0001
rs12347	*MTRR*	4	47	0,02	>0,1*	0	13	-0,01	>0,1	0	19	0,00	>0,1	12	10	0,04	<0,0001	10	6	0,09	<0,0001
rs2015205	*QRSL1*	20	89	0,02	>0,1*	11	23	0,02	>0,1*	N/A	41	0,01	>0,1*	28	13	0,03	<0,0001	24	13	0,08	<0,0001
rs3192149	*TOPBP1*	23	69	0,02	0,001	32	13	0,00	>0,1	2	31	0,03	<0,0001	30	13	0,01	>0,1*	30	12	-0,01	>0,1*
rs4129190	*FLJ10916*	29	26	0,07	<0,0001	15	12	0,06	<0,0001	7	8	0,09	0,0001	67	3	0,10	0,01	44	3	0,03	>0,1
rs3088040	*USP36*	44	67	0,04	<0,0001	28	23	0,03	<0,0001	N/A	38	0,04	>0,1*	69	3	0,01	>0,1	55	7	0,04	0,0001
rs7562391	*PPIL3*	43	24	-0,03	0,0001	42	9	-0,04	0,01*	25	12	-0,03	0,0001	72	1			61	2		
rs2243603	*SIRPB1*	76	16	0,00	>0,1	67	4	0,00	>0,1	N/A	8	0,02	>0,1*	96	1			67	3	-0,05	0,001
rs1143684	*NQO2*	53	19	-0,01	>0,1	43	12	-0,02	0,01	N/A	0			90	0			42	7	0,01	>0,1
rs1494961	*HEL308*	62	24	0,00	>0,1	41	19	0,00	>0,1	N/A	0			97	0			56	5	-0,03	0,01
rs973730	*ESCO1*	79	15	0,01	0,01	78	3	0,02	0,001	73	7	0,00	>0,1	100	0			67	5	0,01	>0,1
rs2863095	*MRPL43*	42	41	-0,04	<0,0001	40	16	-0,04	<0,0001	3	20	-0,04	<0,0001	100	0			55	5	-0,06	0,0001
rs2636	*MCTP1*	95	2											100	0			89	2		
rs2255546	*LRAP*	97	1											100	0			95	1		
rs2290911	*SH3YL1*	100	0											100	0			100	0		
rs2294008	*PSCA*	100	0											100	0			100	0		
rs10380	*MTRR*	100	0											100	0			100	0		

% no call: Percentage of failed in the real-time PCR. Total number of failed divided with total number of heterozygotes. Amount of outliers removed is not included in this percentage. For the SNPs without available failure rates for the DCIS, the combined percentages were calculated on the basis of MDG, UII and FU/LB.
N/A: Homozygotes could not be differentiated from no-called heterozygotes. For these, N is based on all called specimen.
N: Number of specimen after the removal of no calls and outliers. Allelic ratio and P-values are calculated on the basis of these.
AR: Log2 transformed average allelic ratio, adjusted for control. 0 = 50:50 ratio for the two alleles.
P-value: Welch's t-test based on the unadjusted ratios.
* The samples were not normally distributed, and so the p-value is from a Mann-Whitney U test.

After the removal of the failed and outlying CT values, some samples had only one allelic ratio left of the triplet, and these were also removed. The samples with 2 or 3 allelic ratios per triplet were kept. For each sample, triplet or duplet, an average allelic ratio was calculated. If the number of average allelic ratios (N) for each SNP and material were below 3, no further calculations were performed. FU and LB were regarded combined. Three samples were considered adequate for calculations due to them being an average of 6-9 allelic ratios after the removal of all outliers. However, consideration for the low number should be made when interpreting the results.

After adjusting the case ratios by the control ratios, and log2 transformation, the average was estimated for all the samples for each of the 4 materials, MDG, DCIS, Ull and FU/LB, and combined. The samples that were outliers for the combined selection, but not for the cohort were omitted from the overall average. All average ratios are listed in table 6 (columns marked AR) and displayed in figures 11-13.

The AR of the cohorts was tested for normal distribution, one per material and for the combined set. A t-test was performed for the normally distributed sample groups and a Mann-Whitney U test for the ones that did not fit into a normal distribution. The p-values below 0,05 were considered significant.

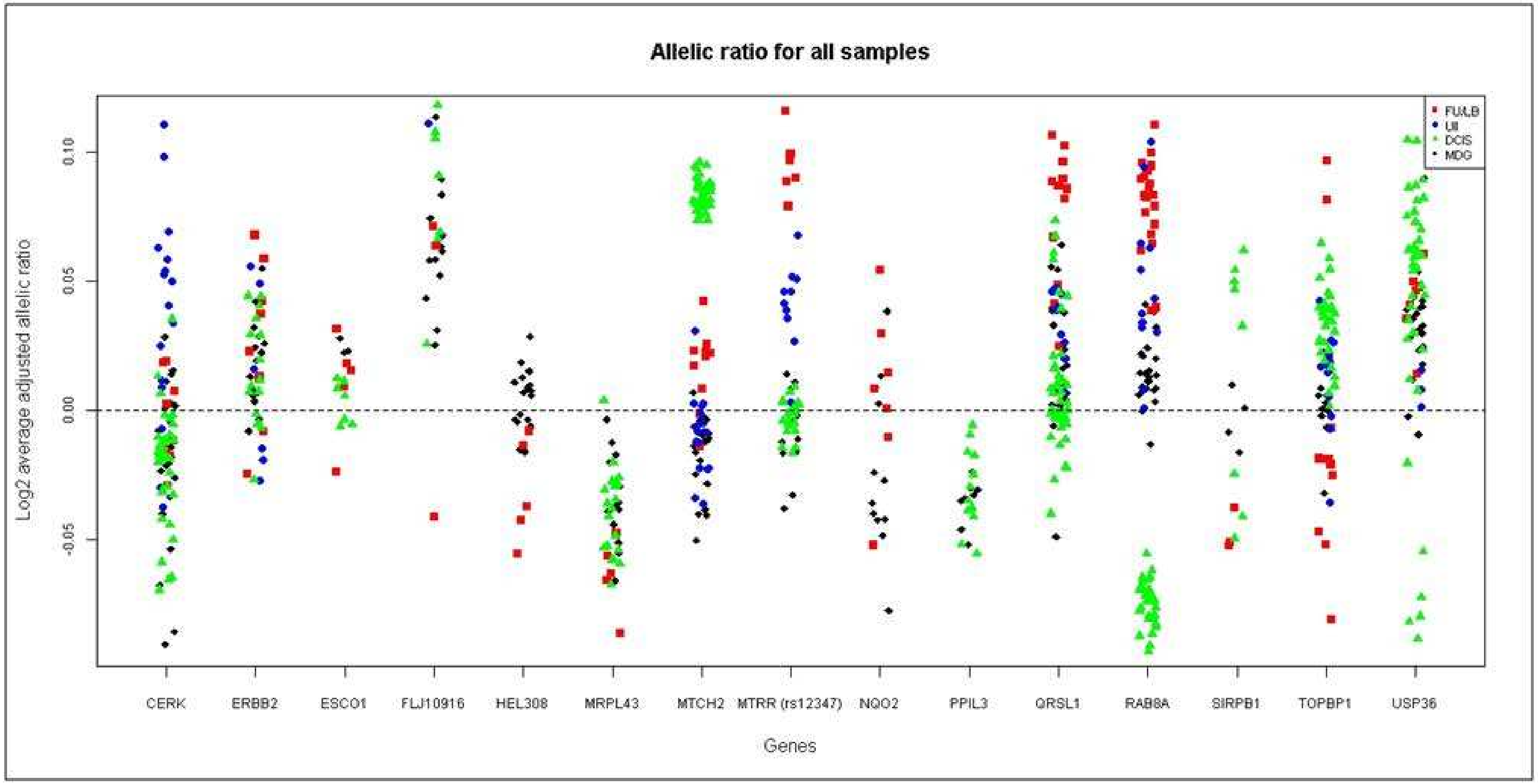

Figure 11. The allelic ratio for all samples grouped by gene and colour coded by material. The ratios are adjusted with controls and log2 transformed. The 50:50 ratio lies at 0 (dotted line) and the genes are listed alphabetically. The materials display limited differences, except for *MTCH2*, *RAB8A* and *MTRR* (rs12347), where DCIS behave differently then all other cohorts (plot made in R version 2.9.1 (R Foundation)).

The MDG cohort showed generally low allelic ratio. The other cohorts were highly spread with great variance. Four SNPs, *MRPL43* (rs2863095), *USP36* (rs3088040), *FLJ10916* (rs4129190) and *PPIL3* (rs7562391), were consistently displaying higher expression of one allele in almost all samples. *ERBB2* (rs1801200) was situated around the 50:50 ratio. Some SNPs indicate a difference in allelic ratios for the different materials (figure 11 and table 6). Especially notable were *RAB8A* (rs10409364) with a negative AR for DCIS (-0,08) and positive for the rest (FU/LB with AR at 0,08), and *MTCH2* (rs1064608) with high AR in DCIS (0,08) and the rest around 0. For *MTRR* (rs12347) DCIS and MDG were around 0 and the rest high above, and *CERK* (rs801719), with Ull spread out in the other direction compared to the other materials.

The p-values for the t- and Mann-Whitney U tests had a general pattern of high AR = low p and low AR = high p, however this pattern was not consistent for all SNPs and cohorts. The tests were performed to set a threshold for what can be considered significant AI. By this threshold, 6 SNPs were considered significant AI for the combined material. For each material seen separately, there were 9 significant SNPs for MDG and FU/LB, 7 for DCIS and 5 for Ull. All of these had allelic ratio above 0,01 in either direction, however, the average allelic ratios were generally above 0,01 independent of the p-value. SNPs with generally low AI and high p-values included *ERBB2* (rs1801200), *SIRPB1* (rs2243603), *NQO2* (rs1143684), *HEL308* (rs1494961) and *ESCO1* (rs973730).

There was no or limited general pattern between allelic ratios and RNA isolation method (figure 12). The plot does not just separate on the account of RNA isolation method, but also by other differences between these two groups of materials. One SNP, *MTRR* (rs12347), show a differential pattern between the two.

Given that the different cohorts included in this thesis represent to a certain extent the different stages of BC, the allelic ratios were examined by the samples' breast cancer stage. Samples with unknown stage were omitted, and stage 3 and 4 are grouped together, representing advanced tumours. The combined stage 3 and 4 displayed a different pattern than the early stages. A relatively high allelic ratio was found solely in the late stage carcinomas in 40 % of the examined SNPs (*RAB8A*, *MTRR*, *QRSL1*, *HEL308*, *SIRPB1* and *MRPL43*).

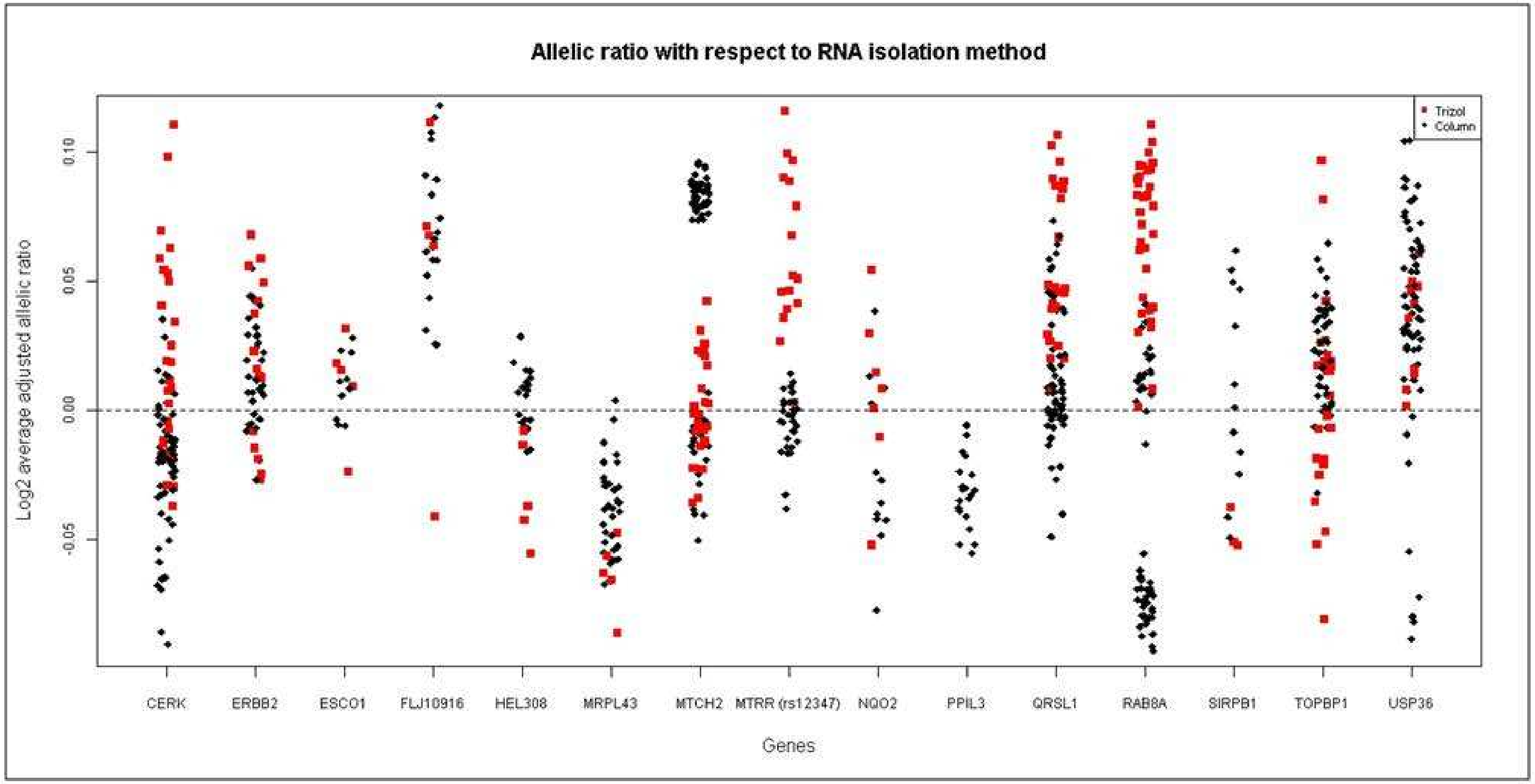

Figure 12. The allelic ratio for all samples grouped by gene and colour labeled by RNA isolation method. The ratios are adjusted with controls and log2 transformed. The 50:50 ratio lies at 0 (dotted line) and the genes are listed alphabetically. The red spots are samples isolated by Trizol extraction (FU, LB and UII cohorts) and the black spots are samples isolated by column purification (MDG and DCIS cohorts). There is generally very limited difference between the two isolation methods, however, *MTRR* (rs12347) do indicate a discrepancy between the two methods (plot made in R version 2.9.1 (R Foundation)).

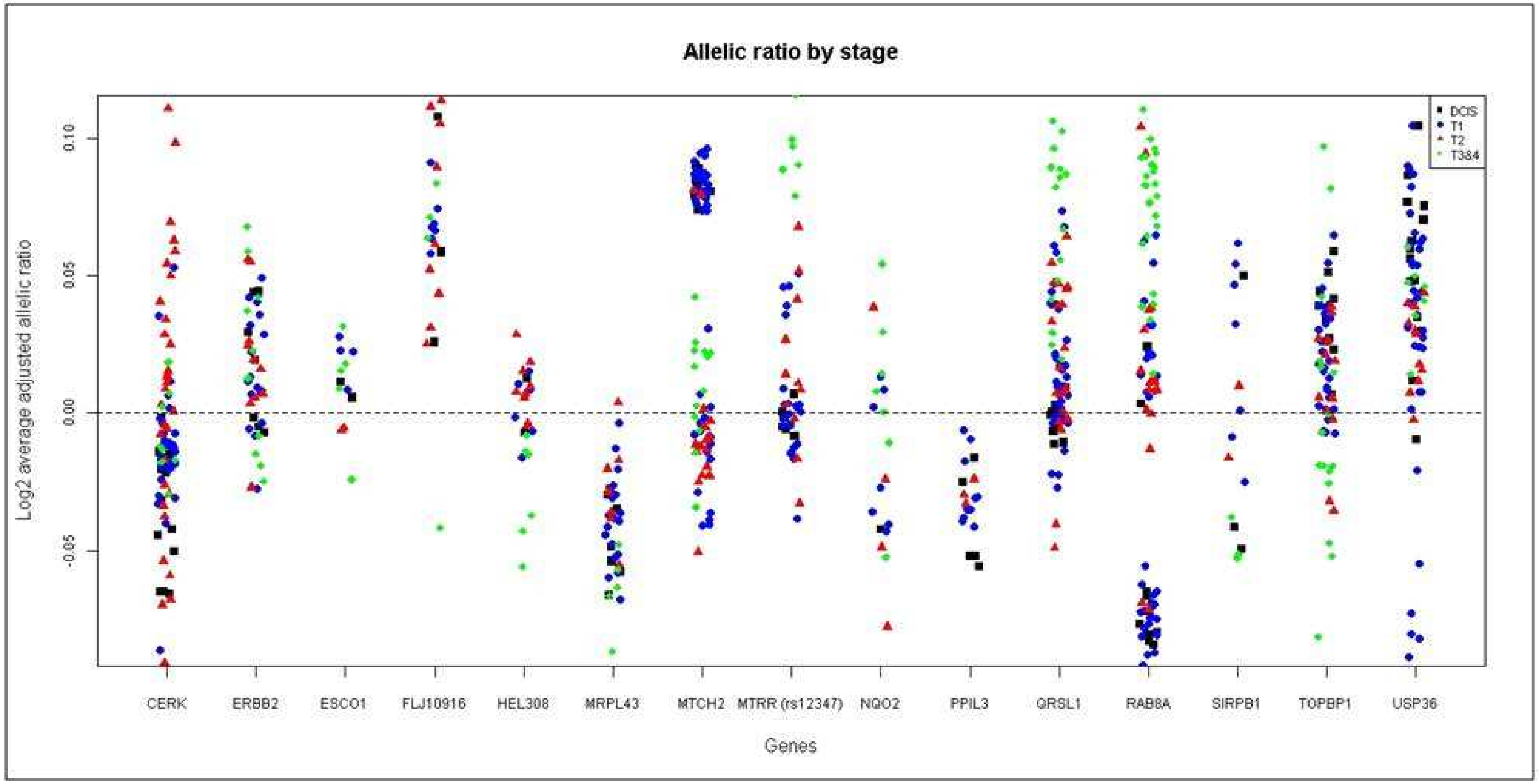

Figure 12. The allelic ratio for all samples grouped by gene and colour labeled by stage. The ratios are adjusted with controls and log2 transformed. The 50:50 ratio lies at 0 (dotted line) and the genes are listed alphabetically. Samples with unknown stage are omitted, and stage 3 and 4 combined to represent advanced tumours (plot made in R version 2.9.1 (R Foundation)).

3.3.3 Copy Number Alterations in tumour

Prior to selection of candidate SNPs for cDNA genotyping the SNPs were ensured to be outside all known CNVs. The genes were examined for CNAs in the specimens that were successfully cDNA genotyped, to investigate whether any pattern of aberrations coincided with failure rates and level of AI. *CERK* (rs801719), *NQO2* (rs1143684), *QRSL1* (rs2015205) and *USP36* (rs3088040) had a high number of aberrations among the samples successfully genotyped in cDNA. The rest of the genes displayed only a limited number of CNAs. However, 40 % (8 of 20) of the genes (*ERBB2*, *CERK*, *QRSL1*, *USP36*, *RAB8A*, *MCTP1*, *LRAP* and *PSCA*) had a general high number of aberrations. Of these, 3 were not successfully genotyped in cDNA (*MCTP1*, *LRAP* and *PSCA*). Ull showed a general high level of aberrations for many of the genes.

4 Discussion

4.1 About the methods

The genotyping platforms utilized in this thesis were chosen according to the study design. The deciding matter is usually cost efficiency, which is dependent on both number of samples and SNPs, and availability. The 45 SNPs in ROS pathways were to be genotyped in large number of samples (N=3749), and the number of SNPs were moderately high. MassArray can genotype a large number of samples fast and with relatively low cost. Affymetrix Molecular Inversion Probes (affymetrix.com) or Illumina Goldengate (illumina.com) platforms were appropriate alternatives, however, for the number of SNPs and samples genotyped in this study MassArray was the most cost efficient platform available.

For the 20 SNPs in the AI study, the aim was to develop a novel method of genotyping for the determination of differences in the expression for the two alleles of each SNP. Other possible platforms available were, in addition to MassArray, e.g. Illumina Goldengate and pyrosequencing. However, once again the cost efficiency come into play as the number of SNPs were too low to make Illumina or MassArray cost efficient. Pyrosequencing, on the other hand, would have been equally applicable for this study.

As documented by this thesis, there may be some differences with regard to success rates and accuracy between the two genotyping platforms. However, these differences are small enough to become irrelevant. Both platforms show a high level of reproducibility and success rates for genotyping germline DNA.

For the genotyping with MassArray (study 1), two SNPs failed completely and 5 were found to be monomorphic. The monomorphic SNPs were also monomorphic or almost monomorphic in the HapMap database, but not in our previous study using SNP-It™ on a Norwegian cohort (Edvardsen H et al., 2006). This is why these SNPs were included in the MassArray study. The discrepancies between the two runs of genotyping could be due to error in the calling process of either of the two platforms. The MassArray did have a 3 % mismatch rate compared with SNP-IT. There is a possibility that the error occurred in the genotyping with SNP-IT. As the SNPs were monomorphic in the HapMap database, it may be that these SNPs simply are monomorphic.

Several factors needed to be taken into consideration when the TaqMan genotyping in study 2 were designed. Firstly, the possible presence of DNA in the RNA extracted material. This method of measuring allelic imbalance is very sensitive to the 50:50 ratio provided by DNA fragments with the same sequence. As these would register as expression in the final analysis, being virtually indistinguishable from the RNA fragments, it would affect the results. One could argue that any germline DNA present in the sample would give a 50:50 ratio, as the patient was heterozygote for the SNP, and would therefore not affect the result. The actual amount of RNA present was not measured, just the relative amounts of each allele compared with each other, and if some heterozygote DNA was present this would not cause an unnatural skewing of the results as the allelic ratio for the DNA would be 0. However, as the genotyping was performed in tumour tissue, not blood, it is uncertain whether only one DNA copy of each allele exists. And hence, DNAse treatment was performed for all RNA samples that had not previously been through a DNA removal step. The quality control run on the bioanalyzer showed that the DNAse treatment itself had limited or no effect on RNA quality. DNAse treatment of the samples that had gone through such a treatment during the isolation procedure was considered redundant and not performed.

Another issue to consider is the problem with the unknown germline genotypes in the DCIS material. SNPs with clear calling of the RNA samples could be underrepresented due to removal of heterozygotes along with the homozygotes, while the SNPs that did not have a clear genotyping of the material could possibly have homozygotes in the calculations. This could theoretically cause a lower allelic ratio for the former and a higher allelic ratio for the latter than the true allelic ratio of the material. The opposite was seen in this thesis and the reason is unclear. However, the no-called SNPs failed to be called because the samples spread wildly making it difficult to detect the clusters of the three genotypes. The ARs are highly spread for the SNPs without calling, and so the low average AI is presumed to be due to the AR of the individual samples cancelling each other out. This does not solve the question of whether the heterozygotes' under- or overrepresentation affect the results, and without germline DNA available there is no effective way of detecting the possible discrepancies. This is one reason why it is important to identify the heterozygotes beforehand.

Initially the large number of elevated NTC values in the cDNA genotyping was worrisome as it implied a large scale contamination, either of the assays themselves or of the plates during preparation. The investigation to determine the source of the contamination gave no results and the raw data from each genotyping was then reviewed more closely. It led to the conclusion that the cause was not contamination, but rather that some assays have a higher level of non-specific fluorescence than others, and this was confirmed by Applied Biosystems. This high level of fluorescence does not influence the results as the background signals are removed from the CT by the software. In addition, the adjustment of the samples to the controls would remove any other unspecified fluorescence and crosstalk of the probes.

There was a high level of failure for 10 of the assays for the cDNA genotyping. This could simply be that the genes were limited or unexpressed thereby exceeding the CT threshold. These genes could also be highly expressed, as seen when the genotyping with the DNA protocol and 2,44 µl cDNA failed. There is also a possibility that it is due to chemical or physical problems with the probe, for instance, that the probe does not have as high affinity to the cDNA compared to the DNA target sequence. These assays were designed for genomic DNA, and may therefore not be as ideal to use on cDNA. This is not uncommon as Milani et al reported that 19 of 79 genes failed when genotyped in cDNA, but were successful in DNA (Milani L et al., 2007). The reason for the lower affinity may be that the cDNA have no introns. If the probes attach partly to the sequence of the intron, it would have low affinity for the cDNA. Of the SNPs with high failure rates genotyped in this study, more than half were on the border of an intron, making this a highly possible reason for failure. Another possibility is that aberrations in tumour cause one of the alleles to be extremely high or unexpressed, i.e. amplification or deletion. Either of these would cause the TaqMan genotyping to fail. If this occurs the SNP would in fact be more in AI than indicated by the results in this thesis.

Ull had much higher failure rates compared with other materials. This could be biological, i.e. that the genes simply were too low or highly expressed in this material compared to other materials. Another possibility is that the carcinomas in the Ull cohort are more degraded, though the quality control indicates that this was not the case. However, the bioanalyzer plots show the quality of the ribosomal RNA more clearly than mRNA, and so there is a possibility that the mRNA in the samples were generally of poorer quality than indicated by the ribosomal RNA. The Ull, FU and LB cohorts all had high failure rates, and they were also the oldest materials of the 5 included in this study of AI. Repeated freezing and thawing of the material could to some extent degrade the RNA without it being detected on the bioanalyzer. In addition, Ull, FU and LB cohorts were all isolated with the Trizol extraction protocol, while MDG and DCIS were isolated with the column purification method. Though both are high quality methods their differences may account for the observed variation in failure rates. Alternatively the DNAse treatment may have caused the increased failure rates. Ull, FU and LB were all DNAse treated prior to cDNA synthesis and this may have caused some degradation of RNA. Finally, the cause may be due to aberrations in the genes, as Ull had a high level of CNAs. This may have contributed to making the Ull material more failure prone than the others. However, the high stage BC tumours have generally more aberrations than the lower (Gao Y et al., 2009), thereby implying that FU/LB should have more failure than the Ull cohort. Either way, it is highly unlikely that the cause of the increased failure rate would have an effect on the allelic ratios, an assumption strengthen by the lack of correlation between allelic ratio and RNA isolation method seen in this thesis.

Optimisation of the method in study 2 was performed in an attempt to lower the failure rates. Several different protocols were attempted, and the allelic ratios remained approximately the same regardless of protocol for all SNPs, but one, indicating that the results themselves are independent of the protocol used.

The DNA protocol with half amount of cDNA showed consistently higher failure rates than the regular cDNA protocol. Four SNPs showed a lower failure rate for the DNA protocol; however the differences for three of these were small enough to be considered insignificant. The 4th SNP, *MRPL43* (rs2863095), had a higher failure rate for FU/LB and lower for Ull with the DNA protocol. Why exactly is uncertain, but it may be related to the general differences in failure rates between Ull and the other materials, as suggested by 100 % failure for Ull, 3 % for DCIS and 50 % for MDG and FU/LB. There were generally few samples with aberrations in this gene, with the exception of the Ull cohort, which had many.

The cause of the difference in failure rates may be due to the protocols or the template volumes. Two SNPs were genotyped with the same template volume and different protocols, and the same protocols with different template volumes. The failure rates were higher for the DNA protocol and for the halving of template. Presumably, the cDNA protocol, having higher primer/probe concentration and total volume, increase the chance of each target sequence attaching to a probe. The DNA protocol, having a lower amount of primers and probes, will give a weaker signal due to fewer probes attaching to the target sequences. If the template is diluted (when the volume is halved), less signal is emitted due to the lower concentration of the target sequence.

In the end it was decided to use the regular cDNA protocol for all SNPs as this had generally lower failure rate, giving a higher amount of samples to calculate the AR. The 5-10 SNPs with high failure rates may yield better results with a change in the protocols, however, the changes would have to be more profound than the ones attempted here. Also, choosing a SNP deeper into an exon would probably yield better results for the assays where the targeted sequences were partially in an intron.

4.2 Genetic variation and allelic imbalance

4.2.1 Genetic variation in ROS pathways (study 1)

Three of the SNPs were not in Hardy-Weinberg equilibrium in the controls. For one of these, rs1805389 in *LIG4*, the T allele is very rare. As the genotype frequencies don't always fit the asymptotic chi-square distribution, a goodness-of-fit test is not always applicable for measuring HWE, especially when one of the alleles is rare or N is low (Halliburton R, 2004). The test may give a significant p-value even when the population is in HWE. However, for all three of these it may simply be that the SNPs deviate from HWE in these populations. Two of the SNPs, rs2137680 and rs907807, reside within the same gene (*IGF1R*), strengthening the possibility that evolutionary forces are at work. In addition, both genes in question are important components in the metabolism and actions of ROS. And so, it is not too far fetched to believe that some natural selection or other evolutionary forces may influence the gene. The HWE test is not adequate proof that evolution is affecting the locus, nor does it reveal the factors throwing the population off the equilibrium. In addition to evolutionary forces like natural selection, gene flow and non-random mating, the deviations can also be due to random chance or unknown subpopulation structures in the control cohorts.

Deviation from the HWE has in the past been synonymous with genotyping errors. However, current genotyping methods are accurate enough to not cause the deviation from HW. Fu et al. investigated the effect of missing call bias on HWE and allele frequency distributions, and its occurrence in current genotyping platforms. No calls are often clustered between one of the homozygote groups and the heterozygotes, and may result in either an overrepresentation or underrepresentation of heterozygotes compared to the homozygotes. Though they did not investigate MassArray specifically they do show that missing call bias occurs in modern genotyping platforms, and that it may be beneficial to lower the number of no calls, despite the increase in genotyping error rate that would follow (Fu W et al., 2009). So the deviations from HWE in this thesis may be caused by missing call bias, even though the success rate is high. As deviations from the HWE may affect the result of association studies ((Trikalinos T A et al., 2006)), the control populations causing the deviation from HWE were removed from the combined control group for the two relevant SNPs. The removal brought the controls into HWE and changed the p-values of the similarity test between cases and control, leaving rs907807 insignificant, however; this does not change the overall result for *IGF1R* as it still has two SNPs with significant p-values.

In the case control analysis of study 1, the 9 haplotypes had at least one SNP each with significant difference between cases and controls. This supports our previous finding of these genes being associated with breast cancer risk, and indicates that these significant SNPs may possibly be in strong LD with the causative variant, though further investigation, including functional studies, would be needed for this to be confirmed. These haplotypes, in addition to having previously shown association with breast cancer risk and tumour expression (Kristensen V N et al., 2006; unpublished study), are located in genes that are involved in the metabolism and function of Reactive Oxygen Species, DNA-repair and apoptosis, and these processes have a relevance to tumourigenesis. ROS induce damage to nucleic acids, proteins and lipids, which can cause abnormal activities in the cell. In addition, ROS is involved with apoptosis, and tumourigenesis and suppression through the RAS-RAF-MEK-ERK pathway (Pan J S et al., 2009), a pathway relaying signals from the membrane resulting in gene regulation through the manipulation of chromatin structure and the activation of various transcription factors (Orton R J et al., 2005). Damage to DNA may cause tumourigenesis to occur if left unattended, due to loss or damage to genes involved with the normal activity and mitosis of the cell. Genes involved in DNA-repair is needed to repair the DNA before further mitosis, while genes involved in apoptosis recognizes the damages and cause the cell to self destruct to prevent further replication and proliferation (Plotkin J B and Nowak M A, 2002). Variants in specific genes involved in these processes being revealed to have an association with breast cancer risk sheds more light on the process of tumourigenesis in the breast.

4.2.2 Genetic variation and allelic imbalance (study 2)

The purpose of genotyping the 20 coding SNPs in germline DNA with TaqMan was to determine the heterozygotic samples, but the results could also be used for a case control analysis. However, the total N is low for both cases and control (273 and 102 respectively), and so the test have less statistical weight than the ones performed for the SNPs genotyped in study 1 (1757 cases and 1859 control samples).

The 20 SNPs genotyped with TaqMan were all in HWE (p>0,001). Two SNPs showed significant differences in frequency between cases and controls, indicating an association between these variants and breast cancer risk. Two other SNPs were borderline insignificant, one of these being in *ERBB2* with known relevance to breast cancer. However *ERBB2* had only 24 individuals in the control cohort, which gives the test low power for this SNP.

The survival analysis gave one significant SNP, rs801719 in *CERK*. That the other SNPs were insignificant is not unexpected, as none of these 19 genes were selected for having an association with breast cancer specific survival. The one gene that was selected for its clinical significance, rs1801200 (*ERBB2*) was not tested for correlation with survival. The SNP in *CERK* showed a higher survival rate for the heterozygotes, implying overdominance. That is, the genotype of the heterozygote has a higher fitness than either of the homozygotes. That would mean that either allele may have less advantage than the sum. This test alone does not prove any real connection between this SNP and survival, but a recent study show an association between the expression of *CERK* and survival which support the result found in this thesis (Ruckhaberle E et al., 2009). In addition, the test does not prove that any of the other SNPs don't have a connection with survival, and further studies would be required. However, the results of these tests were adequate for the use in this thesis, as their sole purpose was to prioritise the SNPs.

All SNPs show some level of AI, however the threshold for what is significant can be difficult to set, especially with the high level of variance between the individual tumour samples. For instance, *RAB8A*, *USP36*, *TOPBP1* and *CERK* are spread over the entire spectrum from -0,1 through 0 to 0,1. For this reason, a t-test or Mann-Whitney U test was performed for each material and combined for each SNP to give an idea of the significance of the AI seen. The AR and p-values were calculated for each separate material for two reasons; firstly, to establish if there were any differences between the different patient cohorts (representing different stages), and secondly, to see if the difference affected the overall average AR. For instance, *RAB8A* (rs10409364) had an average AR of -0,08 for DCIS and +0,08 for FU/LB, but and overall average of 0,002, indicating no AI. But this overall average is due to the different cohorts showing AI for different alleles, not because the AI is low. The p-values show an equal pattern.

Besides the spread patterns of the allelic ratios, there is a tendency for MDG to be close to 0. Otherwise there is no consistent pattern for the various cohorts. MDG represent the low stages of BC, and there is a tendency for late stage to have a higher AI. This may indicate a connection between tumour expression and these genes, however, it could also be due to chance, tumour aberrations or differences in the materials as most of the samples with stage 3 and 4 are from FU/LB, and the pattern was not consistent for all SNPs.

DCIS has a different AI than the other cohorts for 3 SNPs. The reason for this is uncertain. DCIS represent early stage breast carcinoma, and according to the pattern set by MDG and stage 3 and 4 DCIS should be clustered close to 0. This is true for one of the SNPs (rs12347 in *MTRR*), however the other two have a high AI and, in addition, the SNP in *RAB8A* is in AI for a different allele in DCIS compared to the other cohorts. It may be that the ductal carcinoma in situ have a different expression pattern for these genes, however that does not explain why the pure invasive samples of the DCIS material show the same pattern. If this was a general tendency for the low stage carcinomas in these genes, then MDG, and a subset of the Ull cohort, should have displayed the same pattern.

The high variance in AI between individuals is common, for instance, Lo et al. reported a high level of interindividual differences in allelic ratios (Lo H S et al., 2003). The study by Lo et al., as well as the studies performed by Maia et al. and Milani et al., used a standard curve of heterozygotic DNA to estimate the accurate transcript level of each allele (Lo H S et al., 2003; Maia A T et al., 2009; Milani L et al., 2007). In this thesis, the ratio was found by using DNA with equal concentrations from three different heterozygotic, healthy individuals, and the accurate transcript levels can therefore not be inferred. Yet, one can see from the clustering of individual allelic ratios around 0 and the high p-values, that 5 of the genes have low AI.

This is especially noticeable for *ERBB2*, which displayed a high AI in the study by Milani et al., but low in this thesis, along with insignificant p-values. However, the failed specimen in the genotyping had a tendency to fail for either one or the other allele, rarely both. And since both alleles were needed to calculate the AI these specimen had to be removed. This may imply that the AI is in fact higher than estimated, because a highly skewed AR would cause one allele to fail completely in the genotyping. The gene has generally a very high level of aberrations, but there were few among the samples successfully genotyped. It is possible that the samples with high level of aberrations failed during genotyping because the level of AI is too high. If this is the case with *ERBB2*, then several of the other genes may suffer from the same problem.

The 4 remaining genes with low AI all have high failure rates. Only one of them has probes overlapping with an intron (*SIRPB1*), and though they have a high level of aberrations in the Ull cohort (which failed completely for all of these SNPs) they do not generally have many in DCIS and MDG. Whether aberrations have contributed to the high failure rates and low AI of these genes is unknown, however it can be investigated. Measuring the total level of expression with TaqMan Gene Expression assays is an option to discover whether low expression caused the pattern seen, and measuring AI in a cohort where the aberrations for each sample is known can show whether the CNAs is a factor. Both of these can be performed with the TaqMan platform; however, another option is to cDNA genotype the same samples on a different platform, e.g. pyrosequencing or Illumina Goldengate. Neither Milani et al. nor Maia et al. reported a problem with aberrations, however the former used a different genotyping platform (Tag-microarray minisequencing) and the latter did not measure AI in tumour tissue (Maia A T et al., 2009; Milani L et al., 2007). The problem with TaqMan is that if the AI is too high, giving either too much or too low expression of one allele, the reaction will simply fail. Due to the general high level of aberrations in DNA from tumours, this platform may not be ideal for detection of AI in tumour expression unless the protocol is optimised further.

Four SNPs showed AI for one allele for almost all individuals. Of these, *USP36* has a few members of the DCIS cohort spread into the opposite allele, but the rest of this material, along with the other cohorts, are in AI for one of the alleles. These SNPs have generally low p-values, strengthening the evidence of AI. One of the SNPs, rs4129190 in *FLJ10916,* had, in addition, a significant p-value in the case control analysis.

The remaining six SNPs have a highly spread pattern, however, for all of them the p-values are significant for at least some of the cohorts. Most notably are *RAB8A*, *MTRR* and *QRSL1*, which have highly significant p-values and high AI for the Ull and FU/LB cohorts. All six SNPs show significant AI with a great deal of interindividual variance.

This shows that variants in 10 out of 15 genes have a differential expression in breast carcinomas, with 4 being preferential for one allele. These genes are involved in processes, such as, proliferation, transport, translation and apoptosis, all important for the cell, and changes to these processes may help cause tumourigenesis. A low expression level of a transport gene may cause waste products to remain in the cell or organelle, or a deficiency of an important molecule, because there is not enough of the transport protein to move them across the membranes at the proper pace. Variants in genes involved in translation may affect the level of proteins produced, which could have an influence on many processes inside and outside of the cell depending on the affected proteins. And the importance of the genes involved with apoptosis and proliferation have been mentioned earlier. The genes involved in these processes are therefore possible locations for variants that influence expression level in a way that may affect the risk of tumourigenesis.

4.3 Conclusions and future research

At least one SNP per haplotype was validated for having a correlation with risk of breast cancer in study 1, and the possible causative agents may have been found in the SNPs that were significant for each gene, or in strong LD with them. From here the next step would be to estimate the haplotypes for the 9 genes and see if the difference in haplotype distribution previously identified can be validated. This can be done by estimating the haplotypes using software, e.g. PHASE, and then performing a case control analysis for all haplotypes. For many of the cases and controls there are available expression data and these can be used to see if the SNPs/haplotypes found to be associated to BC risk have an effect on the expression of the gene in which the SNPs resides.

Of the SNPs successfully genotyped in cDNA in study 2, 10 out of 15 showed evidence of AI, and this marks these genes as possible candidates for harbouring tumourigenic variants. The remaining 5 may have had low AI because of aberrations in the gene, and this can be discovered by further genotyping with TaqMan or other genotyping platforms, such as Illumina or pyrosequencing. This thesis raises the question of whether TaqMan is a suitable genotyping platform for detection of AI in tumour tissue with the current producer recommended protocol. The high level of aberrations in tumour DNA may limit the use of TaqMan for this type of study, though more investigation is needed.

A functional study, e.g. Electrophoretic Mobility Shift Assay (EMSA), of candidate rSNPs in the promoter of the genes that showed a high level of AI, may yield a possible cause for the differential expression.

5 References

Aas T, Borresen A L, Geisler S, Smith-Sorensen B, Johnsen H, Varhaug J E, Akslen L A and Lonning P E. 1996. Specific P53 mutations are associated with de novo resistance to doxorubicin in breast cancer patients. **Nat.Med.** 2 (7): 811-814.

Althuis M D, Fergenbaum J H, Garcia-Closas M, Brinton L A, Madigan M P and Sherman M E. 2004. Etiology of hormone receptor-defined breast cancer: a systematic review of the literature. **Cancer Epidemiol.Biomarkers Prev.** 13 (10): 1558-1568.

Andersen T I, Holm R, Nesland J M, Heimdal K R, Ottestad L and Borresen A L. 1993. Prognostic significance of TP53 alterations in breast carcinoma. **Br.J.Cancer.** 68 (3): 540-548.

Antoniou A, Pharoah P D, Narod S, Risch H A, Eyfjord J E et al. 2003. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. **Am.J.Hum.Genet.** 72 (5): 1117-1130.

Barrett J C, Fry B, Maller J and Daly M J. 2005. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics.** 21 (2): 263-265.

Bonferroni C E. 1935. Il calcolo delle assicurazioni su gruppi di teste. **Studi in Onore del Professore Salvatore Ortu Carboni**. 13-60.

Bonferroni C E. 1936. Teoria statistica delle classi e calcolo delle probabilità. **Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze**. 8 (3-62.

Bukholm I K, Nesland J M, Karesen R, Jacobsen U and Borresen A L. 1997. Relationship between abnormal p53 protein and failure to express p21 protein in human breast carcinomas. **J.Pathol.** 181 (2): 140-145.

Chang H Y, Nuyten D S, Sneddon J B, Hastie T, Tibshirani R et al. 2005. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. **Proc.Natl.Acad.Sci.U.S.A.** 102 (10): 3738-3743.

Cox D R. 1972. Regression Models and Life-Tables. **Journal of the Royal Statistical Society, Series B (Methodological)**. 34 (2): 187-220.

Edvardsen H, Irene Grenaker A G, Tsalenko A, Mulcahy T, Yuryev A et al. 2006. Experimental validation of data mined single nucleotide polymorphisms from several databases and consecutive dbSNP builds. **Pharmacogenet.Genomics.** 16 (3): 207-217.

Edvardsen H, Kristensen V N, Grenaker Alnaes G I, Bohn M, Erikstein B, Helland A, Borresen-Dale A L and Fossa S D. 2007. Germline glutathione S-transferase variants in breast cancer: relation to diagnosis and cutaneous long-term adverse effects after two fractionation patterns of radiotherapy. **Int.J.Radiat.Oncol.Biol.Phys.** 67 (4): 1163-1171.

Frazer K A, Ballinger D G, Cox D R, Hinds D A, Stuve L L et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. **Nature.** 449 (7164): 851-861.

Fu W, Wang Y, Wang Y, Li R, Lin R and Jin L. 2009. Missing call bias in high-throughput genotyping. **BMC.Genomics.** 10:106. (106-

Futuyma D J. 2005. Evolution. **Sinauer Associates, Inc., USA**. 165-169. ISBN 978-0-87893-187-3.

Gao Y, Niu Y, Wang X, Wei L and Lu S. 2009. Genetic changes at specific stages of breast cancer progression detected by comparative genomic hybridization. **J.Mol.Med.** 87 (2): 145-152.

Geisler S, Borresen-Dale A L, Johnsen H, Aas T, Geisler J, Akslen L A, Anker G and Lonning P E. 2003. TP53 gene mutations predict the response to neoadjuvant treatment with 5-fluorouracil and mitomycin in locally advanced breast cancer. **Clin.Cancer Res.** 9 (15): 5582-5588.

Gram I T, Bremnes Y, Ursin G, Maskarinec G, Bjurstam N and Lund E. 2005. Percentage density, Wolfe's and Tabar's mammographic patterns: agreement and association with risk factors for breast cancer. **Breast Cancer Res.** 7 (5): R854-R861.

Halliburton R. 2004. Introduction to Population Genetics. **Pearson Prentice Hall, USA**. 59-65, 74, 91-94, 121, 187, 197-200, 525. ISBN 0-13-016380-5.

Helland A, Olsen A O, Gjoen K, Akselsen H E, Sauer T, Magnus P, Borresen-Dale A L and Ronningen K S. 1998. An increased risk of cervical intra-epithelial neoplasia grade II-III among human papillomavirus positive patients with the HLA-DQA1*0102-DQB1*0602 haplotype: a population-based case-control study of Norwegian women. **Int.J.Cancer.** 76 (1): 19-24.

Helle S I, Ekse D, Holly J M and Lonning P E. 2002. The IGF-system in healthy pre- and postmenopausal women: relations to demographic variables and sex-steroids. **J.Steroid Biochem.Mol.Biol.** 81 (1): 95-102.

Kaplan E L and Meier P. 1958. Nonparametric Estimation from Incomplete Observations. **Journal of the American Statistical Association**. 53 (282): 457-481.

Kristensen V N, Edvardsen H, Tsalenko A, Nordgard S H, Sorlie T et al. 2006. Genetic variation in putative regulatory loci controlling gene expression in breast cancer. **Proc.Natl.Acad.Sci.U.S.A.** 103 (20): 7735-7740.

Landmark-Hoyvik H, Reinertsen K V, Loge J H, Fossa S D, Borresen-Dale A L and Dumeaux V. 2009. Alterations of gene expression in blood cells associated with chronic fatigue in breast cancer survivors. **Pharmacogenomics.J.** 9 (5): 333-340.

Lee J H, Park S, Park H S and Park B W. 2010. Clinicopathological features of infiltrating lobular carcinomas comparing with infiltrating ductal carcinomas : a case control study. **World J.Surg.Oncol.** 8 (1): 34-

Liu R, Wang X, Chen G Y, Dalerba P, Gurney A et al. 2007. The prognostic role of a gene signature from tumorigenic breast-cancer cells. **N.Engl.J.Med.** 356 (3): 217-226.

Lo H S, Wang Z, Hu Y, Yang H H, Gere S, Buetow K H and Lee M P. 2003. Allelic variation in gene expression is common in the human genome. **Genome Res.** 13 (8): 1855-1862.

Lof M and Weiderpass E. 2009. Impact of diet on breast cancer risk. **Curr.Opin.Obstet.Gynecol.** 21 (1): 80-85.

Lund E, Kumle M, Braaten T, Hjartaker A, Bakken K, Eggen E and Gram T I. 2003. External validity in a population-based national prospective study--the Norwegian Women and Cancer Study (NOWAC). **Cancer Causes Control.** 14 (10): 1001-1008.

Maia A T, Spiteri I, Lee A J, O'Reilly M, Jones L, Caldas C and Ponder B A. 2009. Extent of differential allelic expression of candidate breast cancer genes is similar in blood and breast. **Breast Cancer Res.** 11 (6): R88-

McCormack V A and dos S S, I. 2006. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. **Cancer Epidemiol.Biomarkers Prev.** 15 (6): 1159-1169.

Milani L, Gupta M, Andersen M, Dhar S, Fryknas M, Isaksson A, Larsson R and Syvanen A C. 2007. Allelic imbalance in gene expression as a guide to cis-acting regulatory single nucleotide polymorphisms in cancer cells. **Nucleic Acids Res.** 35 (5): e34-

Naderi A, Teschendorff A E, Barbosa-Morais N L, Pinder S E, Green A R et al. 2007. A gene-expression signature to predict survival in breast cancer across independent data sets. **Oncogene.** 26 (10): 1507-1516.

Nordgard S H, Alnaes G I, Hihn B, Lingjaerde O C, Liestol K et al. 2008a. Pathway based analysis of SNPs with relevance to 5-FU therapy: relation to intratumoral mRNA expression and survival. **Int.J.Cancer.** 123 (3): 577-585.

Nordgard S H, Johansen F E, Alnaes G I, Bucher E, Syvanen A C, Naume B, Borresen-Dale A L and Kristensen V N. 2008b. Genome-wide analysis identifies 16q deletion associated with survival, molecular subtypes, mRNA expression, and germline haplotypes in breast cancer patients. **Genes Chromosomes.Cancer.** 47 (8): 680-696.

Norsk Bryst Cancer Gruppe (NBCG). 1998. Brystkreft. Diagnostikk og behandling. En veiledning. 5th edition.

Orton R J, Sturm O E, Vyshemirsky V, Calder M, Gilbert D R and Kolch W. 2005. Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway. **Biochem.J.** 392 (Pt 2): 249-261.

Ottini L, Palli D, Rizzo S, Federico M, Bazan V and Russo A. 2010. Male breast cancer. **Crit Rev.Oncol.Hematol.** 73 (2): 141-155.

Pan J S, Hong M Z and Ren J L. 2009. Reactive oxygen species: a double-edged sword in oncogenesis. **World J.Gastroenterol.** 15 (14): 1702-1707.

Pastinen T and Hudson T J. 2004. Cis-acting regulatory variation in the human genome. **Science.** 306 (5696): 647-650.

Perou C M, Sorlie T, Eisen M B, van de R M, Jeffrey S S et al. 2000. Molecular portraits of human breast tumours. **Nature.** 406 (6797): 747-752.

Plotkin J B and Nowak M A. 2002. The different effects of apoptosis and DNA repair on tumorigenesis. **J.Theor.Biol.** 214 (3): 453-467.

Redon R, Ishikawa S, Fitch K R, Feuk L, Perry G H et al. 2006. Global variation in copy number in the human genome. **Nature.** 444 (7118): 444-454.

Rockman M V and Kruglyak L. 2006. Genetics of global gene expression. **Nat.Rev.Genet.** 7 (11): 862-872.

Ruckhaberle E, Karn T, Rody A, Hanker L, Gatje R, Metzler D, Holtrich U and Kaufmann M. 2009. Gene expression of ceramide kinase, galactosyl ceramide synthase and ganglioside GD3 synthase is associated with prognosis in breast cancer. **J.Cancer Res.Clin.Oncol.** 135 (8): 1005-1013.

Sorlie T, Wang Y, Xiao C, Johnsen H, Naume B, Samaha R R and Borresen-Dale A L. 2006. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. **BMC.Genomics.** 7:127. (127-

Stranger B E, Forrest M S, Clark A G, Minichiello M J, Deutsch S et al. 2005. Genome-wide associations of gene expression variation in humans. **PLoS.Genet.** 1 (6): e78-

The International HapMap Consortium. 2003. The International HapMap Project. **Nature.** 426 (6968): 789-796.

Trikalinos T A, Salanti G, Khoury M J and Ioannidis J P. 2006. Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated gene-disease associations. **Am.J.Epidemiol.** 163 (4): 300-309.

van ', V, Dai H, van d, V, He Y D, Hart A A et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. **Nature.** 415 (6871): 530-536.

van d, V, He Y D, van't Veer L J, Dai H, Hart A A et al. 2002. A gene-expression signature as a predictor of survival in breast cancer. **N.Engl.J.Med.** 347 (25): 1999-2009.

Vignal A, Milan D, SanCristobal M and Eggen A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. **Genet.Sel Evol.** 34 (3): 275-305.

Wang Y, Klijn J G, Zhang Y, Sieuwerts A M, Look M P et al. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. **Lancet.** 365 (9460): 671-679.

Wiedswang G, Borgen E, Karesen R, Kvalheim G, Nesland J M et al. 2003. Detection of isolated tumor cells in bone marrow is an independent prognostic factor in breast cancer. **J.Clin.Oncol.** 21 (18): 3469-3478.

Zhou W, Muggerud A A, Vu P, Due E U, Sorlie T, Borresen-Dale A L, Warnberg F and Langerod A. 2009. Full sequencing of TP53 identifies identical mutations within in situ and invasive components in breast cancer suggesting clonal evolution. **Mol.Oncol.**

6 Supplementary

Table 1. Primer sequences for all SNPs genotyped with MassArray.

SNP	Gene	Sequence
rs13340	*TXNIP*	ACGTTGGATGATCCCTATCTCCTAACACAG
rs2715438	*IGF1R*	ACGTTGGATGTGGGACACTGATGCTGTATG
rs1805386	*LIG4*	ACGTTGGATGATGCCACTCCTTGTCATCTC
rs1381548	*BCL2*	ACGTTGGATGGGGTGATCGGAATAGTGATG
rs854539	*PPP1R9A*	ACGTTGGATGTCAGAGACCTCACCCTAATC
rs1805388	*LIG4*	ACGTTGGATGTTGATGGCTGCCTCACAAAC
rs7212	*TXNIP*	ACGTTGGATGCCTCTAGTTTCTCATGGCAG
rs230525	*NFKB1*	ACGTTGGATGGTAAGATTACGGGAAAAGTG
rs7211	*TXNIP*	ACGTTGGATGCCTTTTCCCAAAGTTTTGGC
rs215094	*ABCC1*	ACGTTGGATGCTGAGATCCGTGGAGTGAG
rs215067	*ABCC1*	ACGTTGGATGTGTTGGCTGCTTTCTGTAAC
rs854523	*PPP1R9A*	ACGTTGGATGAGCTGAATATCAAAAGCACC
rs2062541	*ABCC1*	ACGTTGGATGGTTCCGTGAACTTGAATGTG
rs2791749	*TXNIP*	ACGTTGGATGTGCCTCGGGTAGTTAAAGTC
rs2137680	*IGF1R*	ACGTTGGATGAGTGCTACAGGTGAGGAAAG
rs1481031	*BCL2*	ACGTTGGATGAGGGTCGTTTCTGAGTCTAC
rs854524	*PPP1R9A*	ACGTTGGATGAGCTGAGGTGTTATGAAGTC
rs1609798	*NFKB1*	ACGTTGGATGTCACTGTCATGACTGCTCAC
rs907807	*IGF1R*	ACGTTGGATGGCATTCTGCATGAGGCATTG
rs958379	*PPP3CA*	ACGTTGGATGGTCAATCTTAAGGATGACTGC
rs2160227	*IL1R1*	ACGTTGGATGGGTTAACGCAGAATTGAAAG
rs230505	*NFKB1*	ACGTTGGATGTAGGCCATCCAAACGTAAAG
rs871335	*IGF1R*	ACGTTGGATGAGTTCCAAACACCTGTTCAC
rs1585214	*NFKB1*	ACGTTGGATGCCCTGCAAATCTGCATGAAC
rs1801	*NFKB1*	ACGTTGGATGCTGCGGTATGAGTCTGTATC
rs1982673	*BCL2*	ACGTTGGATGGTGCCATACTTTAAAAAATTC
rs920559	*PPP3CA*	ACGTTGGATGTAGTTTGACCATGCAGAGGG
rs1016860	*BCL2*	ACGTTGGATGAGAGCCAGTATTGGGAGTTG
rs2232640	*LIG4*	ACGTTGGATGACAAAAGAGGTGAAGGGTGG
rs854518	*PPP1R9A*	ACGTTGGATGGATTTTAGCAGCTGTTATG
rs903880	*ABCC1*	ACGTTGGATGCAGGGCCCCATCCTGGATT
rs1567811	*IGF1R*	ACGTTGGATGCCACACAAATCCTAAATGGG
rs2062011	*BCL2*	ACGTTGGATGACAAGCCTCCAGGAATCCAC
rs2791750	*TXNIP*	ACGTTGGATGAAGTTCGGCTTTGAGCTTCC
rs230531	*NFKB1*	ACGTTGGATGTCAGTTTCCTAGCATAACAC
rs1805389	*LIG4*	ACGTTGGATGAGGAACGTGAGATGCAACAG
rs230498	*NFKB1*	ACGTTGGATGCGTGTCTCCTGTTGTATGTC
rs705377	*PPP1R9A*	ACGTTGGATGACAGCACAGACACAGGTTTC
rs1568502	*IGF1R*	ACGTTGGATGGGATGACCGCATAGAGGAAC
rs997049	*IL1R1*	ACGTTGGATGAACTGTTTCCAAAAAGCCAG
rs1021965	*PPP3CA*	ACGTTGGATGCTTCTGTGCTATTTTCTGCTC
rs1598857	*NFKB1*	ACGTTGGATGGCCAAAACACTGTGGTGTAT
rs1481030	*BCL2*	ACGTTGGATGGTTGTCTAACCTAGTGGTTC
rs1020760	*NFKB1*	ACGTTGGATGATAGAAAGCACTCAAAGAGG
rs212083_a	*ABCC1*	ACGTTGGATGCCATCATGGACTACACAAGG
rs212083_b	*ABCC1*	ACGTTGGATGTGTTGGAAATTCCTTCTGCC

Table 2. Context sequences for all pre-designed assays for TaqMan genotyping. This is the sequence that surrounds the SNP. The target specific primer sequences are not listed.

SNP	Gene	Context Sequence
rs801719	*CERK*	CTGCCAGCTTCCAGAGCTTATCATC[C/T]GTGCTCCCAGAACAGACGCCAAGGT
rs12347	*MTRR*	AGGAGGAAGCCCCAGCAAAGTATGT[A/G]CAAGACAACATCCAGCTTCATGGCC
rs3192149	*TOPBP1*	AAAAGAGCTGCTTTACTTTCAGGCT[G/T]ATGTGAAACTGGAATTTCCACTGGC
rs4129190	*FLJ10916*	GAAGAGCTCCCACAGTTGGACAGAG[G/A]GACCCTGTGCCAGTGGAGCACACTC
rs7562391	*PPIL3*	GAATAGTTATGTCCTTAATGTGTAC[A/C]TCATTAAGAGGTCGGTATGTCTTCT
rs2243603	*SIRPB1*	AGTGGAGCAGTAGGAGCCAGCGCTG[C/G]TTCTGGAAATCAGGGAAGAGGAGGA
rs1143684	*NQO2*	AGTGTCTGATTTGTATGCCATGAAC[C/T]TTGAGCCGAGGGCCACAGACAAAGA
rs1494961	*HEL308*	CCAAGGTCATTTGATGATGACTCAA[C/T]TGTTTTCTTAGCAACATTAATTTCC
rs973730	*ESCO1*	TATCATGTAGCTTTGTCCCTAAAAA[C/T]GTACTTTGAGTCACTGGATGATGCA
rs2863095	*MRPL43*	GTGGGGCAACTCTTCACTGTGCTTG[C/T]ACCTGGGCTGGGGCAGGATCCTGAA
rs2636	*MCTP1*	CACAGGTCACCGTCACTCACTTCTT[A/T]ATCCTTTCGCCAAAGGAAGCCACTT
rs2255546	*LRAP*	TGCCAGACGTCCAAAGGGGCAGCAA[C/T]TAGCATGGGATTTTGTAAGAGAAAA
rs2290911	*SH3YL1*	GATAGAGCTTATATTCATTTCTGTT[A/G]CCTGCCAAAAAAGAGGAGAGTGGTG
rs2294008	*PSCA*	CTCCACCACAGCCCACCAGTGACCA[C/T]GAAGGCTGTGCTGCTTGCCCTGTTG
rs10380	*MTRR*	ATTATATTTCAGAAAAGAGCTCAGA[C/T]ATTTCCTTAAGCATGGGATCTTAAC

Table 3. Primer and probe sequences for all custom designed arrays for TaqMan genotyping.

SNP	Gene	Forward Primer	Reverse Primer	VIC reporter	FAM reporter
rs1801200	*ERBB2*	CCTGACCCTGGCTTCCG	ACCAGCAGAATGCCAACCA	ACGTCCATCATCTCTG	CGTCCATCGTCTCTG
rs1064608	*MTCH2*	TCAGGTCACAACAATAAGTCTTCCC	GGAATATGAGCCGAGGAAATAGCTT	CGGAAGGTCGCCTTT	CGGAAGGTCCCCTTT
rs10409364	*RAB8A*	CCTCTGCAGACGTCGAAAAGA	TTCCTTGGAAACTTGTCTCTTGTCA	ACTTGTTCCCGAGTATCA	CTTGTTCCCAAGTATCA
rs2015205	*QRSL1*	GAAGCTGCTGGTCACAAAACG	CCCTGTTGCACTCTCAAACCAA	ATAGGACTGCAGTTTAT	CAATAGGACTACAGTTTAT
rs3088040	*USP36*	GTTGCCAGAGGCCAGTGA	GGCTCTCCCACAAAGGTCTTTT	CCCCCAGAGCCC	CCCCCGGAGCCC