8-2014

Genetic and mechanistic analysis of rat mammary cancer susceptibility.

Jennifer Sanders
University of Louisville

GENETIC AND MECHANISTIC ANALYSIS OF RAT MAMMARY CANCER SUSCEPTIBILITY

By

Jennifer Sanders
B.S. University of Louisville, 2009

A Dissertation
Submitted to the Faculty of the
School of Medicine
of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

Department of Biochemistry and Molecular Biology
University of Louisville
Louisville, KY

August 2014

GENETIC AND MECHANISTIC ANALYSIS OF RAT MAMMARY CANCER SUSCEPTIBILITY

By

Jennifer Sanders
B.S. University of Louisville, 2009

A Dissertation Approved on

June 9th, 2014

by the following Dissertation Committee:

_____

Dr. David Samuelson- Dissertation Director

_____

Dr. James Shull

_____

Dr. Carolyn Klinge

_____

Dr. Ronald Gregg

_____

Dr. Ted Kalbfleisch

ACKNOWLEDGEMENTS

I would like to thank my mentor, Dr. David Samuelson, for his support and guidance. I would like to thank my committee members, Drs Ronald Gregg, Ted Kalbfleisch, Jim Shull and Carolyn Klinge for their guidance. I also need to thank Xin Xu for his help.

Most importantly, I would like to thank my husband, Michael, without whom this would never be possible. I also would like to thank Charlie and Winston for being the best distraction possible.

ABSTRACT

GENETIC AND MECHANISTIC ANALYSIS OF RAT MAMMARY CANCER SUSCEPTIBILITY

Jennifer Sanders

June 9th, 2014

Breast cancer is a complex disease, which is influenced by genetic, epigenetic and environmental components. Genetic susceptibility to breast cancer is made up of high, moderate and low penetrance alleles. High and moderate penetrance alleles are rare and constitute only a small percentage of the genetic susceptibility. Most variation in genetic susceptibility is controlled by low- penetrance, common polymorphisms. Comparative genetics uses model organisms to study human disease. Rat strains exhibit different susceptibility phenotypes to chemical induced carcinogenesis. The Wistar-Furth (WF) rat strain is susceptible to chemically induced mammary carcinogenesis, while the Wistar-Kyoto (WKy) and Copenhagen (COP) rat strains are resistant. Selective breeding and linkage analyses of these rat strains after treatment with 7,12-dimethylbenz[a]anthracene (DMBA) have been used to identify eight rat mammary cancer quantitative trait loci (QTLs) in the rat. This dissertation focuses on two of these QTLs, *mammary carcinoma susceptibility loci 1b and 6* (*Mcs1b* and *Mcs6)*. *Mcs6* has been identified and physically confirmed using WF.WKy congenic animals and maps to a 33Mb region. This locus will have to be mapped to a narrower interval in order for functional studies to be practical. I

was able to map the *Mcs6* locus to a region of 8.5Mb on rat chromosome *7.* The *Mcs1b* locus maps to a region of 1.8Mb on r at chromosome *2*. *Mcs1b* contains the rat orthologous region to a breast cancer risk associated region marked by SNP *rs889312*. This makes the *Mcs1b* congenic rat an ideal model for studying the mechanism of *rs889312*. The goal of my project is to identify all *Mcs1b* sequence variants between the two rat strains and test for gene regulatory functions. I was able to identify 70 SNPs and 2 INDELs using next- generation sequencing. Three rat SNPs have gene regulatory function differences between the two rat alleles. Out of the seven human SNPs that tag SNP *rs889312*, four exhibit gene regulatory differences between the major and minor alleles, and therefore, may be functional orthologs to the rat *Mcs1b* candidate SNPs. Overall, I was able to fine map the *Mcs6* region and identify several candidate rat and human *Mcs1b/ MCS1B* SNPs.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER I

INTRODUCTION

Breast Cancer Statistics

In 2013, approximately 232,340 new breast cancer cases was expected to be diagnosed, making breast cancer the most commonly diagnosed cancer among women (excluding skin cancer). A woman's lifetime chance of developing breast cancer in 2013 was 1 in 8. The lifetime chance of developing breast cancer has risen since the 1970's when a woman's lifetime chance of being diagnosed with breast cancer was 1 in 11. This rise in breast cancer diagnoses is attributed to an increase in life expectancy and breast cancer incidence [1].

Breast cancer is the second leading causing of cancer related deaths in women after lung cancer. However, for women between 20-59 years of age, breast cancer is the leading cause of cancer related deaths [2]. Due to improvements in awareness, screening and breast cancer treatments, the breast cancer death rate has dropped 34% between the years 1990-2010 [1]. However, approximately 39,620 women will have died of breast cancer in 2013, highlighting the need for even better early detection and prevention methods.

Breast Cancer Risk Factors

Breast cancer is a complex disease, made up of environmental, epigenetic and genetic factors. A 12.5% lifetime chance of developing breast cancer is a population

based estimate and a woman's individual risk may vary depending on breast cancer risk factors [1]. These breast cancer risk factors include age, gender, reproductive history, genetic and non- heritable factors [3]. Age and gender are the most important breast cancer risk factors and breast cancer risk increases for females with increase in age [1]. However, there are several environmental breast cancer risk factors. These include but are not limited to: radiation, estrogen exposure, smoking, high alcohol intake, unhealthy diets and environmental pollutants such as polycyclic aromatic hydrocarbons [4-12].

A. Radiation as a Risk Factor for Breast Cancer

Radiation is one of the most potent inducers of breast cancer, since the mammary gland is sensitive to radiation induced carcinogenesis [13]. Timing of the radiation exposure is essential, with exposed women under the age of 20 having a higher risk of developing breast cancer than women exposed at an older age. This is likely due to the breast tissue being relatively undifferentiated before the age of 20 [13, 14]. Women are generally not exposed to high dosages of radiation; the few exceptions include survivors of the atomic bomb detonations in Japan during World War II and the nuclear disaster at Chernobyl. Breast cancer incidence data from these disasters suggests an increase in breast cancer risk after high radiation exposure that is more pronounced in women who were younger at the time of exposure and those who received higher the dose of radiation was [14, 15]. Women can also be exposed to low dosages of radiation, usually due to medical procedures. These include: 1) women being monitored for tuberculosis infection by X-rays; 2) women being treated for benign breast disease or acute post-partum mastitis; 3) childhood cancer survivors; 4) adult cancer survivors; 5) women treated for

benign disorders as children using radiation and 6) breast cancer screening through mammograms [7, 13]. Low dosage radiation exposure is considered carcinogenic but the benefits from radiation treatments outweigh the risks [7]. The breast cancer risk after low dose exposure increases if women are positive for a genetic mutation that increase breast cancer susceptibility [6].

B. Estrogens as a Risk Factor For Breast Cancer

The US National Toxicology Program of the Department of Health and Human Services categorizes estrogens as carcinogens [9, 16]. The involvement of estrogens in breast carcinogenesis was established through several key observations: 1) hormonal replacement therapies, which increase circulating estrogens, also increase breast cancer risk; 2) bilateral oophorectomy (reduction in circulating estrogens) in both animals and in humans protects from breast cancer; 3) parity decreases breast cancer risk and 4) treatment with anti-estrogen drugs such as tamoxifen decreases chances of developing breast cancer [17-21].

The main estrogen studied in breast cancer carcinogenesis is 17β-estradiol (E2); however, other estrogens include estrone (E1) and estriol (E3). E2 is secreted by the ovaries in pre-menopausal women and is synthesized by the aromatase enzyme in pre-menopausal and post-menopausal women [9]. E2's role in breast carcinogenesis is complex and involves many different mechanisms and cellular pathways. E2 can bind and activate estrogen receptors (ER) α and β, which in turn act as transcriptional regulators of several genes involved in cell proliferation and cell cycle progression [9]. One example of this is the role of E2 in upregulating anti-apoptotic Bcl-2 and Bcl-Xl

genes in breast cancer cell lines and therefore preventing apoptosis [22]. Another example of a mechanism of E2's involvement in breast cancer is the observation that E2 increases the secretion of interleukin 8 (IL-8) and vascular endothelial growth factor (VEGF), both needed for the formation of blood vessels to growing tumors [23]. A more controversial mechanism by which estrogens may be involved in breast cancer is through initiating DNA damage. Some evidence suggests that estrogens and their metabolites can cause direct and indirect DNA damage through the formation of free-radicals [18, 24].

The role of estrogens in breast cancer has been used to develop treatments specific to estrogen-responsive breast cancers that express ERα, since this makes up the majority of breast cancer cases (about 70%). Currently, there are three different types of treatments for ER-positive cancers that use the role of estrogens as their basis. The first type of treatment is selective estrogen receptor modulators (SERMs). These are synthetic chemical compounds that compete with estrogens for ERα binding, but inhibit transcriptional activity in a cell- type and promoter context. There are currently three SERMs on the market: raloxifene, toremifene and tamoxifen [9, 25]. Another type of treatment includes compounds that increase ER turnover in the cells. Fulvestrant is an example of a drug approved for this type of treatment [26]. A third type of breast cancer treatment that takes advantage of the role of estrogens in breast cancer is aromatase inhibitors. Aromatase is an enzyme needed for the synthesis of estrogen and is extremely important in post-menopausal women, since the synthesis of estrogen in the adipose tissue becomes the primary source of estrogen for obese women [9].

C. Alcohol as a Risk Factor for Breast Cancer

The link between alcohol intake and breast cancer risk has been established. A meta-analysis of 100 epidemiological studies identified a correlation between alcohol intake and breast cancer risk at high alcohol consumption (<45g/day) with RR of 1.46 (95%CI= 1.33-1.61) compared to nondrinkers. This results in a 7.1% increase in breast cancer risk with every 10g/ day of alcohol consumption [11, 27]. Interestingly, low alcohol intake (1drink/day or 5-14g/day) is also associated with an increase in breast cancer risk in several meta-analyses (relative risk (RR) = 1.05, 95% CI 1.02-1.08 in one study) [11, 28]. Several mechanisms have been identified as to how alcohol is involved in breast carcinogenesis. These mechanisms include changes in hormone and hormone receptor levels, increased cell proliferation, DNA adduct formation, increased cyclic adenosine monophosphate (cAMP), changes in potassium channels and modulation of gene expression all due to alcohol or its metabolites [29]. Of important consideration is the alcohol metabolite acetaldehyde, which is a known carcinogen. Acetaldehyde promotes inflammation, can cause DNA damage and can inhibit DNA repair [30]. It is recommended for females to limit alcohol intake to ≤1 drink/day [11].

D. Smoking as a Risk Factor for Breast Cancer

Cigarette smoke contains thousands of chemicals, 69 of which are known carcinogens, and 20 of them are mammary carcinogens [10, 31]. Many of these carcinogens can be activated through enzymes and eventually lead to DNA adduct formation. Some of these carcinogens found in cigarette smoke reach the breast tissue [32].

The relationship between smoking and breast cancer remains complex. Recent studies have indicated that active smoking is associated with an increase in breast cancer risk, particular when age of smoking initiation was very early (before the age of 20) [10, 31]. In one particular study, the risk of breast cancer associated with smoking was determined to have an odds ratio of 1.28 (95% CI= 1.17-1.39) [33]. The relationship of passive smoking and breast cancer remains even more elusive with several studies showing no r elationship while some studies suggest that even passive smoking can increase breast cancer risk [10]. A positive correlation between smoking and breast cancer risk has been established in a subset of people, who carry the NAT2 slow acetylation genotype, especially among post-menopausal women. In a meta-analysis of breast cancer risk, the relative risk (RR) was 1.27 (95%CI= 1.16-1.39) in people with the NAT2 slow acetylation genotype compared to a RR o f 1.05 (95%CI= 0.95-1.17) in people with the rapid NAT2 acetylation genotype [31, 34]. NAT2 is involved in detoxifying several carcinogens in tobacco smoke. A NAT2 slow acetylation genotype person is homozygous for a NAT2 variant, while a rapid NAT2 genotype is composed of a homozygous wildtype genotype [34]. This indicates that there is interplay between genetics and the environment when it comes to smoking and increases in breast cancer risk.

E. Environmental Pollutants as a Risk Factor for Breast Cancer

The two environmental pollutants most often associated with breast cancer risk are polycyclic aromatic hydrocarbons (PAHs) and polychlorinated biphenyls (PCBs). PAHs are generated during combustion. Exposure to PAHs can occur through several

ways. PAHs are present in smoked and grilled foods, air pollution, vehicular exhausts and in cigarette smoke [10, 12]. Several PAHs are known mammary carcinogens, including benzo(a)pyrene [12, 35]. PAHs form DNA adduct and subsequently cause DNA damage [12]. Polychlorinated biphenyls (PCBs) were used in electrical equipment until banned in the U.S. The main route of exposure to PCBs is through eating contaminated fish from contaminated rivers near industrial areas. High PCB levels can be found in breast milk post exposure since that PCBs accumulate in fat [12, 36]. PCBs can activate several hormone receptors and PCB metabolites can form DNA adducts leading to carcinogenesis [37].

F. Diet and Breast Cancer

Diet plays an important factor in influencing breast cancer risk. Increased breast cancer risk is associated with obesity, high fat intake in post-menopausal women and high total energy intake in both pre- and post-menopausal women [4, 38, 39]. Breast cancer risk is influenced by intake of different types of fats. High intake of saturated fats is associated with increase in breast cancer risk in post-menopausal women, while polyunsaturated fats increase breast cancer risk in both pre- and post-menopausal women. Monounsaturated fats seem to have no influence on breast cancer risk [4, 40]. In addition, a diet low in bread and fruits, as well as high in meat, fish, butter, other animal fats and margarine is associated with in increase in breast cancer risk (hazard ratio= 2.00, 95% CI 1.30-3.09) [41]. It has been postulated that fats may increase E2 levels as a possible mechanism for the increase in breast cancer risk [42]. There are also several foods which have a protective effect against breast cancer. A diet rich in fiber reduces chances of

developing ERα+ breast cancer [43]. Also, a diet rich in soy is associated with a reduction in breast cancer risk, possibly due to some of the components of soy having anti-oxidant and anti-inflammatory properties [44]. The American Cancer Society suggests to reduce weight/ weight gain in overweight and obese people, to adopt a physically active lifestyle, to eat foods high in vegetables and fiber, and to reduce intake of fats to reduce the chances of developing breast cancer [45].

G. Epigenetic Factors and Breast Cancer Risk

There are several pieces of evidence that suggest epigenetic factors are involved in breast cancer risk. One example is that in a meta-analysis of the effects of genistein, a compound in soy, on breast cancer risk suggests that a diet rich in genistein before puberty protects from breast cancer later [46]. Genistein's breast cancer protective effect is thought to possibly act through the epigenetic effects of genistein [46]. Genistein is known to affect DNA methylation by both increasing methylation and by inhibiting DNA methyltransferase, resulting in a decrease in DNA methylation [47, 48]. Another study that identified a link between epigenetics and breast cancer risk showed that lower levels of DNA methylation of repetitive elements in white blood cells correlated with an increase risk of developing breast cancer. A decrease in DNA methylation in repetitive elements is associated with genomic instability. Some repetitive elements have the ability to integrate themselves into different genomic regions. This can potentially disrupt gene function, therefore repetitive elements are often highly methylated and therefore silenced [49-51]. Another piece of evidence for the link between epigenetics and breast cancer comes from the observation that the daughters of women given diethylstilbestrol (DES)

(a synthetic estrogen) during their pregnancy are more likely to develop breast cancer than women who haven't been exposed. The generational jump in breast cancer risk suggests that there may be epigenetic factors at play, possibly through the disregulation of DNA methyltransferases expression levels [52, 53].

Genetic Factors of Breast Cancer Risk

A. General Introduction to Genetic Factors of Breast Cancer Risk

The genetic contribution towards breast cancer has been approximated by comparing the concordance of cancer between monozygotic and dizygotic pairs of twins. The idea being that if there is a higher concordance of cancer between pairs of monozygotic twins than there is between dizygotic twins, breast cancer susceptibility has a genetic component. Overall, a monozygotic twin is more likely to be diagnosed with breast cancer if there is an affected twin, suggesting that there is a genetic component to breast cancer. About 25-32% of breast cancer susceptibility can be attributed to genetic factors [54-56].

Breast cancer heritability is made up of three different classes of alleles. The first class is made up of high penetrance risk alleles, such as *BRCA1, BRCA2, STK11* and *TP53*. These alleles have a strong effect on breast cancer risk, with relative risk >8. However, mutations in these genes are very rare and therefore, only a small percentage of the population is affected [57-59]. It is estimated that only 20-25% of genetic heritability of breast cancer can be explained by high penetrance risk alleles [60]. Family-based linkage analysis were used to identify these high penetrance breast cancer risk genes [59].

Genome-wide association studies have not mapped any additional high penetrance risk alleles, suggesting that no further high-penetrance risk alleles exist [60].

A second class of breast cancer risk alleles is made up of moderate risk alleles. These risk alleles have a relative risk of 2-8. Some of these breast cancer risk alleles were identified through targeted sequencing, usually of genes known to interact with BRCA1 and BRCA2 or genes that are active in DNA repair pathways. Other breast cancer risk alleles in this class were identified through linkage analysis of families suffering from rare syndromes that include breast malignancies as a symptom [59]. Genes in this class of breast cancer risk alleles include *PTEN, PALB2, CDH1, ATM, BRIP1*, and *CHECK2* [57-60]. These moderate risk alleles are rare in the population and therefore, the contribution of this class of breast cancer risk alleles to the heritability of breast cancer is estimated to be only 3% [60].

Since the majority of breast cancer heritability cannot be explained by high and moderate penetrance alleles, the majority of breast cancer heritability is thought to be made up of common, low penetrance risk alleles. These alleles make up the third class of breast cancer risk alleles. The risk allele frequencies of common, low penetrance risk alleles are >5% in a human population. This means this class of breast cancer risk alleles affects a large part of the respective population. The relative risk of this class of risk alleles is low and as of yet no risk allele in this group with a relative risk higher than 1.5 has been identified [59]. Initially, this class of breast cancer risk alleles was identified through case- control, association studies using candidate genes. A more common and more fruitful method of identifying these risk alleles was developed in 2007 when the first genome-wide association study (GWAS) of breast cancer was published [59, 61].

Since 2007, 72 common, low penetrance risk alleles have been identified and the most recent study suggested that up to 1000 common, low penetrance risk alleles may exist [59, 62]. The 72 common, low penetrance risk alleles identified so far only make up about 14% of the heritability of breast cancer. The remaining heritability may be made up of low penetrance risk alleles that have very low effect sizes that cannot be easily identified through GWA studies due to limits of detection. Alternatively, risk alleles that interact with each other and therefore cannot be identified through genome-wide association studies may make up the remaining heritability [59, 62]. GWA studies are designed to test a limited number of SNPs that tag enough SNPs to cover the whole genome. Tagged SNPs are found in areas of low recombination and are therefore inherited together with the GWA study SNP. Tagged SNPs are found at linkage disequilibrium blocks or haplotypes. The boundaries of LD blocks or haplotype blocks are delimited by recombination hot-spots. This means that a GWAS identified SNP may tag the actual SNP, but the causative variant(s) may not be the SNP tested in the GWA study [59]. Genes making up the three classes of genetic susceptibility and their minor allele frequency can be found in Figure 1. Discussion of breast cancer susceptibility genes will focus on germ line mutations in these genes.

Surprisingly, high and moderate penetrance risk mutations are found in coding regions of their respective genes. This makes identifying the causative gene relatively easy. However, low penetrance risk alleles are generally found in non-coding or intergenic regions. This makes identifying the causative gene difficult, since the causative gene can be long distances away and may affect gene regulation through complex chromatin arrangements. Low penetrance risk alleles located in non-coding or intergenic

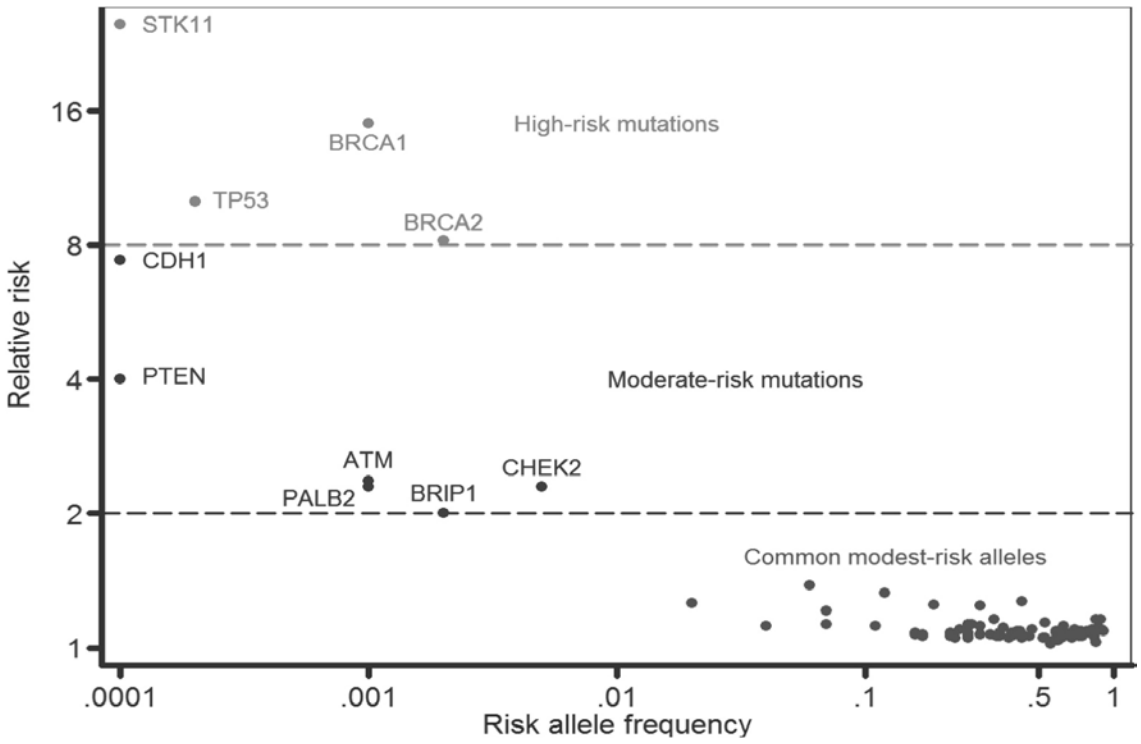


Figure 1: Architecture of genetic susceptibility to breast cancer. Adapted from Ghoussaini et al. (2013) [59]. Figure shows the genetic architecture of breast cancer risk. Relative risk and risk allele frequencies for breast cancer risk genes are shown. Breast cancer risk genes are subdivided into classes of level of risk.

regions are thought to be enhancers that affect the expression of genes long distances away.

B. *BRCA1* and *BRCA2*

The first breast cancer gene was identified in 1990 and cloned in 1994. Breast cancer susceptibility 1 or *BRCA1* was identified through a linkage analysis of 23 families with 146 cases of breast cancer. *BRCA1* was mapped to chromosome 17q21 [63, 64]. The BRCA1 protein contains 24 exons and is 1863 amino acids long. BRCA1 is an extremely versatile protein, which interacts with several different proteins to form distinct complexes. Its most known function is DNA repair, however, the protein is also involved in cell cycle control and transcriptional regulation and generally acts as a tumor suppressor [65, 66]. Upon DNA damage, the DNA damage sensors ataxia telangiectasia mutated (ATM) and ataxia telangiectasia mutated rad3-related (ATR) will phosphorylate BRCA1 leading to the recruitment of BRCA1 to DNA damage foci that locate to sites of DNA damage. BRCA1 is involved in homologous recombination, an error-free pathway to repair double stranded breaks [67].

BRCA1 is also known to be involved in cell cycle control. The protein interacts with cell cycle proteins E2F, CDC2 and cyclins. BRCA1 levels increase in late G1 phase. High levels of BRCA1 results in the upregulation of p21 and G1-S cell cycle arrest [65].

BRCA1 is a known transcriptional regulator. The C-terminus of the BRCA1 protein is known to interact with transcriptional activators and repressors. One protein BRCA1 is known to interact with is the RNA polymerase II holoenzyme and functions as

a transcriptional activator. BRCA1 is also known to interact with TP53 at the p21 gene, resulting in the upregulation of p21 [65].

The second major breast cancer susceptibility gene (Breast cancer 2, early onset or *BRCA2*) was identified in 1994 through a linkage analysis of 15 families with multiple cases of early onset breast cancer. *BRCA2* was localized to a region on chromosome *13* [68]. Like BRCA1, BRCA2 is involved in DNA damage repair, in particular, BRCA2 is involved in homologous recombination. BRCA2 interacts with several DNA damage repair proteins including RAD51, a protein known to cover single stranded DNA strands formed during homologous recombination [69].

The risk of developing breast cancer by the age of 70 is as high as 85% in *BRCA1* and *BRCA2* mutation carriers [67]. *BRCA1* and *BRCA2* mutations also increase the risk of being diagnosed with higher grade/ stage and ER negative tumors, and increase the risk of metastasis [67, 70]. In fact, the majority of breast cancers in *BRCA1* mutation carriers are ER negative (70-90%), while the majority of breast cancers in *BRCA2* mutation carriers are ER positive (60-75%) [71]. More than 2000 mutations have been identified in the *BRCA1/2* genes [70]. The vast majority of *BRCA1/2* mutations result in a truncated protein with only a small minority resulting in amino acids substitutions [67]. Male breast cancer is a rare disease with a ratio of 1:175 men to women developing breast cancer. *BRCA1* and *BRCA2* are known risk factors for male breast cancer with *BRCA2* being more important in the genetic predisposition to male breast cancer [72]. The lifetime chance of men developing breast cancer is 1.8% for *BRCA1* carriers and 8.3% for *BRCA2* carriers [71]. *BRCA1/2* mutation carriers are often advised to take prophylactic measures to prevent developing breast cancer or ensuring early detection. These measures include

annual mammography or magnetic resonance imaging (MRI) starting at age 30. Other prophylactic measures include bilateral mastectomy, salpingo-oophorectomy and chemoprevention using the anti-estrogen tamoxifen [71, 73].

C. *TP53* and *PTEN*

Another breast cancer predisposition gene is *TP53*. *TP53* is implicated in Li-Fraumeni syndrome, which is a rare autosomal dominant syndrome that is associated with an increase in childhood and adult cancers including breast cancer. TP53 is a tumor suppressor protein involved in cell cycle control and apoptosis [74]. Breast cancer is the most commonly found cancer in female *TP53* mutation carriers. However, *TP53* mutations are rare and make up only 0.1% of all breast cancer cases [73].

Mutations in *PTEN* are associated with another rare autosomal dominant cancer syndrome called Cowden's syndrome. Cowden's syndrome is associated with an increase risk in several cancers including skin, bowel, thyroid and breast and the presence of pathognomonic physical features, including facial trichemmoma, acral keratoses and oral papillomatous papules. PTEN is a tumor suppressor protein and female *PTEN* mutation carriers have a lifetime chance of developing breast cancer that is 75% [74].

D. FGFR2

The breast cancer susceptibility genes discussed so far are high penetrance risk alleles and mutations in these genes increase a woman's chance of developing breast cancer significantly. However, another set of risk alleles are associated with a much lower increase in breast cancer risk. Low penetrance risk alleles are found at a much

higher frequency in the population than high penetrance risk alleles and therefore are thought to make up the majority of genetic breast cancer risk. One of the earliest low penetrance risk alleles identified through a genome- wide association study is located within the fibroblast growth factor receptor 2 (*FGFR2*) gene and it has been confirmed through several other studies in multiple populations [61, 75-78]. It is also the most studied low penetrance breast cancer risk allele to date. SNPs associated with breast cancer risk are located on a 7.5kb linkage disequilibrium block (LD block) within intron 2 of *FGFR2* [61]. Several risk associated SNPs map to this locus and it is not known if there is a single causative variant, multiple independent or multiple interacting variants in this region. *FGFR2* is expressed higher in homozygotes of the minor allele than the major allele, suggesting that these SNPs are involved in gene regulation of *FGFR2* possibly through altering transcription factor binding sites [79]. Through electrophoretic mobility shift assays (EMSAs), several transcritiption factors including OCT-1, RUNX2, FOXA1 and E2F1 were shown to bind to the minor and major alleles of the risk SNPs differentially. Furthermore, chromosome confirmation capture (3C) of the *FGFR2* locus suggests that the *FGFR2* risk SNPs are brought to close proximity of the *FGFR2* promoter through chromosomal looping [79, 80]. The *FGFR2* locus provides the first mechanism into how low penetrance risk alleles that are located in intergenic or intronic regions can affect gene regulation of candidate susceptibility genes.

E. *MAP3K1* and *rs889312*

The same study that identified a breast cancer risk SNP in the intron of *FGFR2* also found several other SNPs that associate with breast cancer risk in an European

population, one of them being *rs889312* [61]. This SNP will be important for the majority of the studies reported here. The minor allele frequency (MAF) of *rs889312* is about 30% in the European population. *rs889312* increases breast cancer risk slightly with a per allele odds ratio of 1.13 (95%CI= 1.10-1.16) [61]. GWA studies using a wide variety of different populations confirmed previous findings for *rs889312*, indicating that this SNP is associated with breast cancer risk in several different populations. This included studies with populations of European, Korean, South American, Chinese and Tunesians [78, 81-84]. The SNP is also found to be associated with several breast cancer subtypes including ER+, ER- and triple negative [85, 86].

rs889312 is located in an intergenic region which is located within a linkage disequilibrium block (LD block) of 280kb. This LD block contains three genes, namely *MAP3K1*, *SETD9* and *MIER3*. *rs889312* is located closed to *MAP3K1*; therefore, it is referred to by that name in the literature. Initially, *MAP3K1* was considered the most likely causative gene in this region due to its role in cell signaling. However, this does not mean that *MAP3K1* is the causative gene, since *rs889312* is located on a haplotype block with *MAP3K1*, *SETD9* and *MIER3*, any of these genes could potentially be the causative gene. *SETD9* is not expressed in human breast tissue and there is little conservation between human and rat *SETD9*. Therefore, *SETD9* is an unlikely candidate for conferring the mammary carcinogenesis phenotype of the *rs889312* region. However, *MIER3* is expressed higher in breast tumors compared to normal breast tissue, suggesting that *MIER3* might have a role in mammary carcinogenesis [87]. *rs889312* is in linkage disequilibrium with at least six other SNPs at an r^2 value of 0.8. All seven of these SNPs are located on the 280kb LD block. This means that any of these seven SNPs could be the

causative SNP. Therefore, *rs889312* is not necessarily the causative SNP in this region but tags the causative SNP. It is currently not known which of the *rs889312* correlated SNP is causative, or if multiple causative SNPs are present. Interestingly, with more and more genomes being sequenced, more SNPs that tag *rs889312* may be identified. This also means that the causative SNPs for this region may not be in public databases yet. Figure 2 shows the map of the LD block containing *rs889312*.

Animal Models of Breast Cancer

Studying human polymorphisms that associate with breast cancer risk comes with a set of challenges. Studying these polymorphisms in humans directly is limiting. Crucial experiments cannot be performed in humans directly due to ethical and feasibility constrains.

A. Human Breast Cancer Cell Lines

Human cell lines have been used to circumvent the challenge of studying diseases in a human model. Human cells lines have several advantages. They can be immortalized and easily grown. They can be transfected with foreign DNA and exposed to chemicals. Genomic manipulation is also easily accomplished in human cell lines. However, there are disadvantages to using human cell lines. There is a broad range of human breast cancer cell lines that differ in their genomes and cell environments drastically. It is important to select the cell line that is most appropriate for the experiment. Human breast

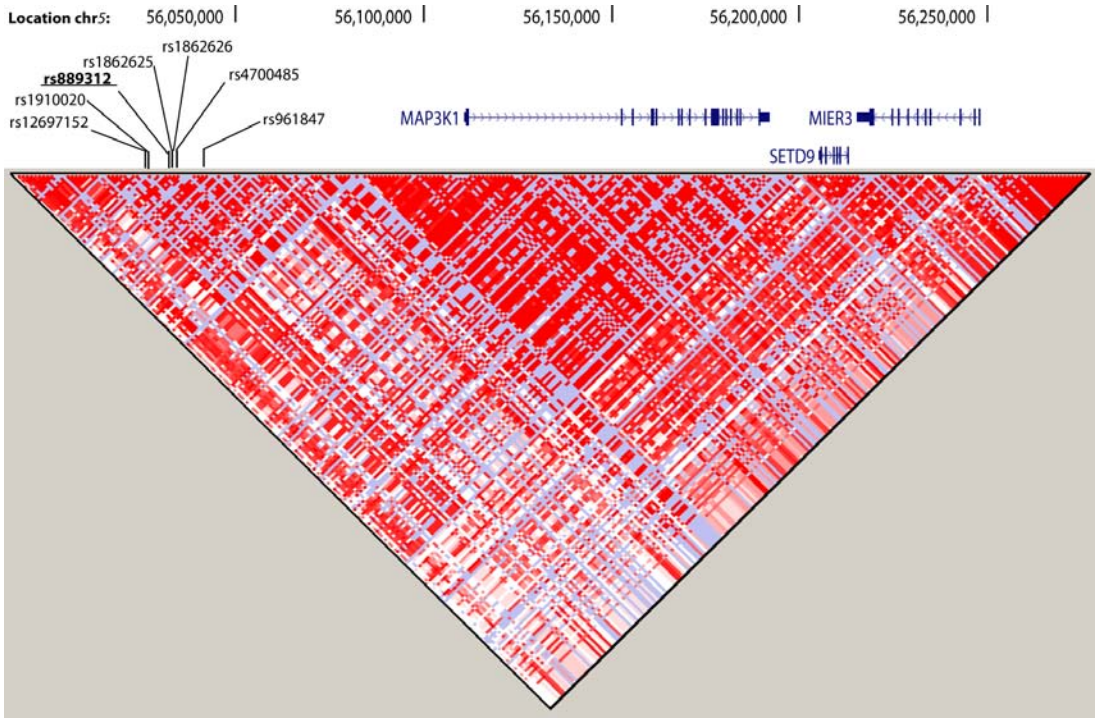


Figure 2. LD block containing *rs889312* and tagged SNPs. Blue lines indicate gene transcripts in this region. The LD block is 280kb in size.

cancer cell lines are prone to change genotypically and phenotypically with each passage, giving rise to subpopulations. This is especially the case if these cells lines have been used for long periods of time. Another problem with human cells lines is a high rate of false misidentification and contamination with different cells lines [88, 89]. These disadvantages have led to the identification of several non- human breast cancer animal models that can be used instead of or to enhance cell line studies.

B. Canine Models of Breast Cancer

The canine has been used as a b reast cancer animal model. Canines develop mammary cancer spontaneously. Canine mammary tumors are similar to human breast cancer biologically and clinically. In particular, several genes that are deregulated in mammary cancer are deregulated in both humans and canines. Problems with canine models include the relatively high costs of housing canines compared to rodents and ethical considerations of using animals commonly viewed as pets [90].

C. Mouse Models of Breast Cancer

Rodent breast cancer models have become increasing popular. The mouse model of breast cancer has been popular because of the ability to perform genetic manipulations in the mouse. There are several advantages of mouse models of breast cancer. The biology and development of the mouse mammary gland is well known and characterized. Mouse mammary glands develop mammary cancer spontaneously but mammary cancer can also be induced with carcinogens such as 7,12-dimethylbenz[a]anthracene (DMBA). With the identification of mammary gland specific promoters, many transgenic mouse

models of breast cancer have become available. It is also possible to transplant a mouse mammary gland [91]. Another advantage is that the mouse genome is very well characterized and mapped, especially compared to other rodent models such as the rat [92]. However, there are also disadvantages to mouse models of breast cancer. The biology of mammary cancer in mouse differs in some aspects from the biology of human breast cancer. While the majority of human breast tumors are responsive to estrogen (ER+), mouse mammary tumors are generally independent of estrogen. Also, the majority of human breast cancer metastases locate to the bone, while mouse mammary cancer metastasis locate to the lung [91].

Several transgenic mouse models of breast cancer exist. These include mice that overexpress breast cancer genes such as *C-myc* or *Ras*. These genes are often under the control of the mouse mammary tumor virus long terminal repeat promoter (MMTV). This results in a strong, gland-wide expression of the oncogene resulting in multifocal tumorigenesis. This is in contrast to human, single focal point tumorigenesis [91, 93]. Also, mice carrying gene knock-outs have been used to study the function of several mammary gland tumor suppressors. *Brca1* and *Brca2* knock- out mice all have been established [94, 95]. In general, these mice develop mammary cancer with less multifocality than transgenic animals. Conditional knock-out mice, that exhibit gene knock-out only in targeted tissues, have been used to overcome negative effects from whole- organism knock-outs of genes such as embryonic lethality [91].

Mice have also been used in experimental metastasis models of breast cancer. Here, human breast cancer cells are injected into the heart of immune- compromised host mice. This model has been used to study bone metastasis of human breast cancer cells in

a mouse, however, this model skips important steps in the metastasis pathway, such as cell-cell detachment, invasion of local tissue and intravasation. Several other experimental metastasis models exist that differ in injection site and site of metastasis [93].

D. Rat Models of Breast Cancer

Another popular rodent model of breast cancer is the laboratory rat. While mouse models of breast cancer have been very popular due to the fact that transgenic and knock-out mice are readily available, the rat model has some advantages over the mouse models of breast cancer. Rats develop spontaneous mammary tumors, but chemical or oncogene induction is also possible [96]. Rat mammary tumors are more similar to human breast tumors than mouse mammary tumors are in etiology and biology. Mice often develop mammary tumors that are associated with a viral etiology unlike rat and human mammary tumors. [92]. Rat mammary tumors are hormone sensitive, which is the same with the majority of human breast tumors [97-99]. DMBA treatment of both rats and mice can be used for carcinogenesis. However, rats require only a single dose, develop more tumors and exhibit a shorter mean latency, which makes them easier to use in an experiment. DMBA induced tumors in the rat do not often metastasize and are localized to the mammary gland. However, in mice leukemias, skin, lung, ovarian and stomach cancers are common, resulting in early termination of experiments [99]. Rats develop spontaneous mammary tumors, but the tumor incidence rate, tumor grade and age of tumor onset vary with different inbred rat strains [100]. Also, inbred rat strains exhibit differential susceptibility to chemical, radiation and hormone induced carcinogenesis [96,

100]. A study using DMBA as a carcinogen showed that a single dose of DMBA is sufficient to induce multiple mammary tumors in the outbred Sprague- Dawley strain [101]. Subsequent studies with different inbred and outbred rat strains showed that rat strains differ in their susceptibility to DMBA carcinogenesis. The Sprague-Dawley and Wistar-Furth (WF) rat strains are highly susceptible to DMBA carcinogenesis, while the Long-Evans and F344 are resistant [102]. Other studies revealed that the Copenhagen (COP) and Wistar-Kyoto (WKy) are resistant to DMBA induced carcinogenesis [100, 103, 104].

Differences in susceptibility to tumorogenesis after DMBA treatment are not due to a strain difference in the ability to metabolize DMBA [100]. This indicates that there is a genetic component to mammary cancer susceptibility in different inbred rat strains. These rat strains can be used to study genetic elements that determine genetic susceptibility to mammary cancer. This can then be translated back to humans, making the study of genetic elements of breast cancer possible.

Mammary Cancer Quantitative Trait Loci (QTLs) in the Rat

Several rat QTLs that associate with mammary cancer risk have been identified. The first study to identify mammary cancer QTLs in the rat used the susceptible WF and the resistant COP rat strains. Progeny from a (COP x WF)$_{F1}$ x WF backcross were treated with DMBA and tumors were counted. A genetic linkage analysis using informative microsatellite markers between the two rat strains was performed. This resulted in the identification of one mammary cancer QTL, named *M*ammary *C*ancer *S*usceptibility 1 (*Mcs1*) on rat chromosome *2*, with a LOD score of 3.8 [105]. Several more *Mcs* loci were

identified in an extension of the original study. Here, two independent (COP x WF)$_{F1}$ x WF backcrosses and a F2 generation were treated with DMBA and a linkage analysis was performed. This resulted in the confirmation of *Mcs1* and the identification of *Mcs2-4* [106]. *Mcs2-4* map to rat chromosomes *7*, *1* and *8*, respectively. While *Mcs1-3* are associated with a decrease in tumor number, the *Mcs4* locus is associated with an increase in tumor number in animals that carry the COP allele in the *Mcs4* region [106]. In an additional study, a genetic linkage analysis using the susceptible WF and the resistant WKy strain was performed to identify additional *Mcs* loci. In this study, a (WFxWKy)$_{F1}$ x WF backcross was used for the linkage analysis post DMBA treatment. This resulted in the identification of four more loci, named *Mcs5-8*. These loci are located on rat chromosomes *5*, *7*, *10* and *14* respectively. The WKy allele for *Mcs5*, *Mcs6* and *Mcs8* all reduce the susceptibility to DMBA induced carcinogenesis, while the WKy allele for *Mcs7* increases susceptibility. In addition, a locus was identified that interacts with *Mcs8*, named *M*odifier of *Mcs* or *Mcsm1* [107]. *Mcs2* and *Mcs6* overlap extensively, however it is currently not known if both QTLs map to the same locus [108].

The most studied *Mcs* locus is the *Mcs5* locus. The *Mcs5* locus was initially confirmed and mapped to a region of 115Mb on rat chromosome *5* using congenic lines [109]. Congenic lines are generated by mating two rat inbred strains to generate a F1 generation. One of the rat strains is considered the donor (in this case the COP or WKy resistant rat strains) and the other rat strain is the recipient (in this case the WF or susceptible rat strain). The F1 generation is then backcrossed to the recipient (in this case the susceptible WF rat strain) for up to ten generations. At each generation the genotype of the animals is determined and only animals that maintain the donor allele in the region

of interest are selected. This results in the introgression of the donor allele in to a recipient genetic background only in the region of interest. The phenotype of these congenic animals can then be compared to the inbred parent strain to identify if the genomic region of interest is involved in modulating the phenotype. A diagram of the technique is shown in figure 3. DMBA treatment of several more congenic lines for the *Mcs5* locus resulted in the identification of three *Mcs5* subloci, named *Mcs5a*, *Mcs5b* and *Mcs5c*. WKy alleles in the *Mcs5a* and *Mcs5c* regions decrease susceptibility, while a WKy allele in the *Mcs5b* region increases susceptibility [110]. Further fine mapping using congenic resulted in the identification of a synthetic QTL within *Mcs5a*, where at least one WKy allele has to be present on the same chromosome at two distinct loci within the *Mcs5a* locus. These distinct loci are named *Mcs5a1* and *Mcs5a2*. *Mcs5a1* and *Mcs5a2* are located in close proximity to the genes *Fbxo10* and *Frmpd1*. *Mcs5a* congenic animals show differential expression levels of *Fbxo10* and *Frmpd1* in thymus tissue compared to WF homozygous animals. Also, *Mcs5a1* and *Mcs5a2* contain the rat orthologous region to human genomic loci that associate with breast cancer risk, making this a great model to study the function of these human risk loci [111]. *Mcs5a* acts through the immune system and the WKy allele at the *Mcs5a* is associated with an increase of γδTCR+ T-cells in the mammary glands compared to WF homozygous rats [112]. Furthermore, chromatin looping of the *Mcs5a* region appears to be a mechanism by which the *Mcs5a* alleles affect *Fbxo10* expression levels in T-cells and this mechanism appears to be conserved between the rat and human [113]. Overall, the *Mcs5a* locus exemplifies how the rat can be used as a model to study the mechanisms of breast

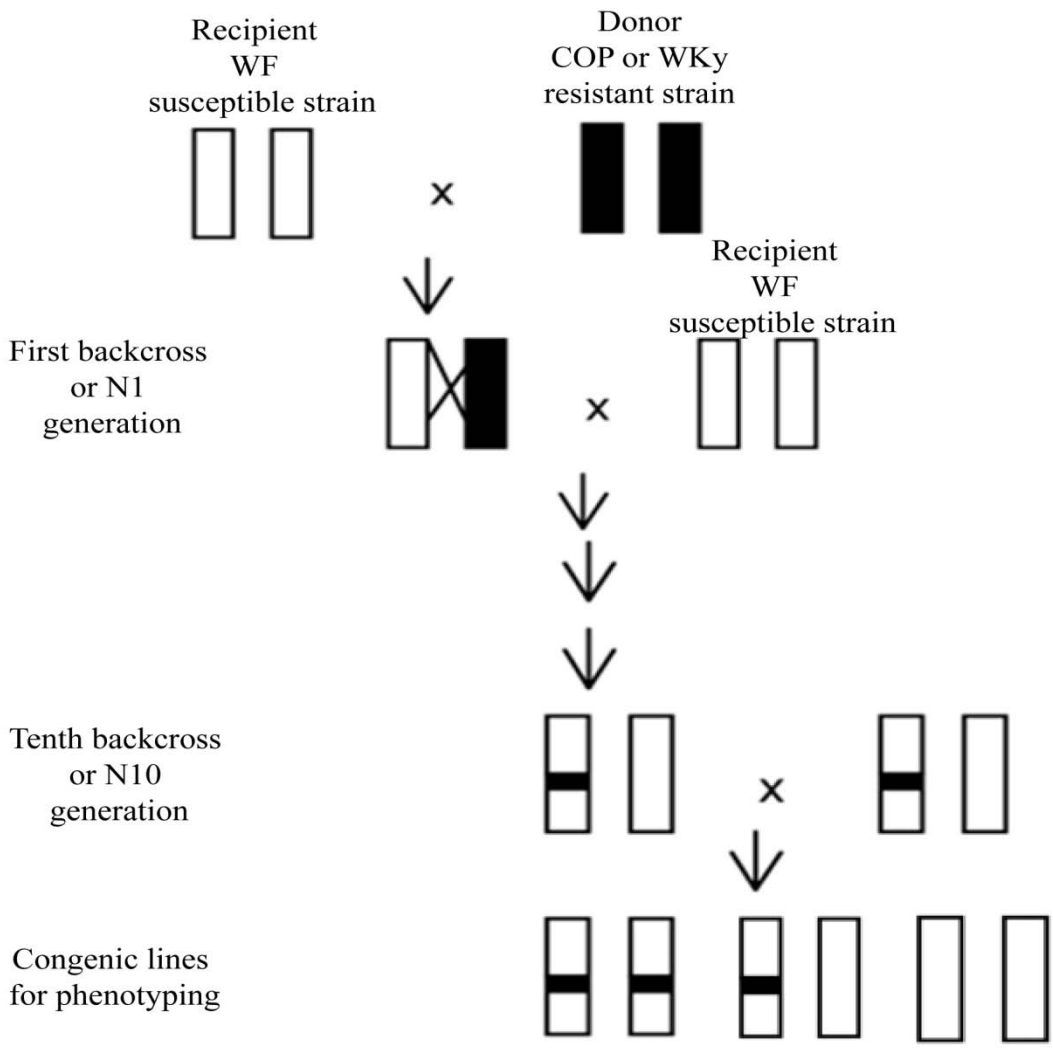


Figure 3. Generation of congenic animals. Adapted from Kim et al. [114]. Two inbred rat strains are mated to generate a heterozygous F1 generation. The F1 generation is backcrossed to the recipient strain for ten generations. At each generation, the genotype of the animals is test to ensure donor DNA is present in region of interest. This method will introgress donor DNA into recipient genome only in regions of interest. Black bars indicate DNA from donor strain, while white bars indicate DNA from recipient strain.

cancer susceptibility in the human. Another set of rat mammary cancer QTLs were identified using DMBA treatment and a linkage analysis of crosses between the susceptible (SPRD-Cu3) and the resistant WKy rat strain. This resulted in the identification of *Mcstm1* and *Mcstm2* on rat chromosomes *5* and *18*and two loci involved in modifying tumor growth rate, named *Mcsta1* and *Mcsta2* on rat chromosomes *10* and *18* [115, 116].

The mammary cancer loci discussed so far have all been identified through linkage analysis using DMBA as a method of carcinogenesis induction. However, several rat mammary cancer QTLs have been identified through using E2 as a method of inducing carcinogenesis. These studies used the August-Copenhagen-Irish (ACI) rat strain, which is susceptible to E2- induced carcinogenesis and the COP or Brown-Norway (BN) rat strains, which are resistant to E2- induced carcinogenesis. This resulted in the identification of several rat mammary cancer QTLs, named *E*strogen *i*nduced *m*ammary *c*ancer loci or *Emca1-2*, *Emca4-8* [117, 118].

Several rat mammary cancer QTLs overlap; however, at this point it is not known if they map to the same loci. Further fine mapping studies using congenics are needed to determine if overlapping mammary cancer QTLs map to the same region.

<u>*Mcs 6*</u>

The work presented in this dissertation will involve two different mammary cancer susceptibility loci, *Mcs6* and *Mcs1b*.

Mcs6 was identified in a linkage analysis using the susceptible WF and the resistant WKy rat strain post DMBA treatment [107]. *Mcs6* was confirmed using several WF.WKy

congenic lines spanning potential regions of the locus as shown in Figure 4 and Table 1. Congenic lines were generated in the same manner as depicted in Figure 3. WF.WKy congenic lines A, B and F all had a phenotype of fewer tumors compared to WF homozygous animals (3.5, 3.9, 3.4 a nd 7.3 tumors per rat respectively). This indicates that there is a genetic element in these lines that is modifying mammary cancer susceptibility and that these lines contain the *Mcs6* locus. Lines C and E showed a similar tumor multiplicity when compared to WF homozygous animals (8.0, 7.2 and 7.3 tumors per rat respectively). This indicates that these lines do not contain the *Mcs6* locus. This maps the *Mcs6* locus to a region of 33Mb on rat chromosome *7*, between bp 22,382,725 and 55,364,398. Overall, the WKy allele at the *Mcs6* locus results in a 55% reduction in tumor multiplicity. Interestingly, the *Mcs6* locus overlaps the *Mcs2* locus, which was confirmed and physically mapped using WF.COP congenic lines [108]. It is currently not known if both loci map to the same location and the phenotype results from the same genetic element(s) in both QTLs. There are 111 genes annotated in the *Mcs6* region using the UCSC Rat Nov. 2004 (Baylor3.4/rn4) Genome Browser [119]. None of these annotated genes are known breast cancer susceptibility genes. Figure 5 contains all 111 genes annotated in the *Mcs6* locus.

The human orthologous region of the *Mcs6* locus maps to a contiguous region on human chromosome *12* between base positions 71,299,117 to 105,502,699. The human region is inverted with respect to the rat region. All genes annotated in the rat *Mcs6* region are also annotated in the human orthologous *MCS6* region [108]. Therefore, the *Mcs6* rat model can be used to study the human breast cancer susceptibility gene(s) in this region. Several potential breast cancer susceptibility SNPs that map to the *MCS6*

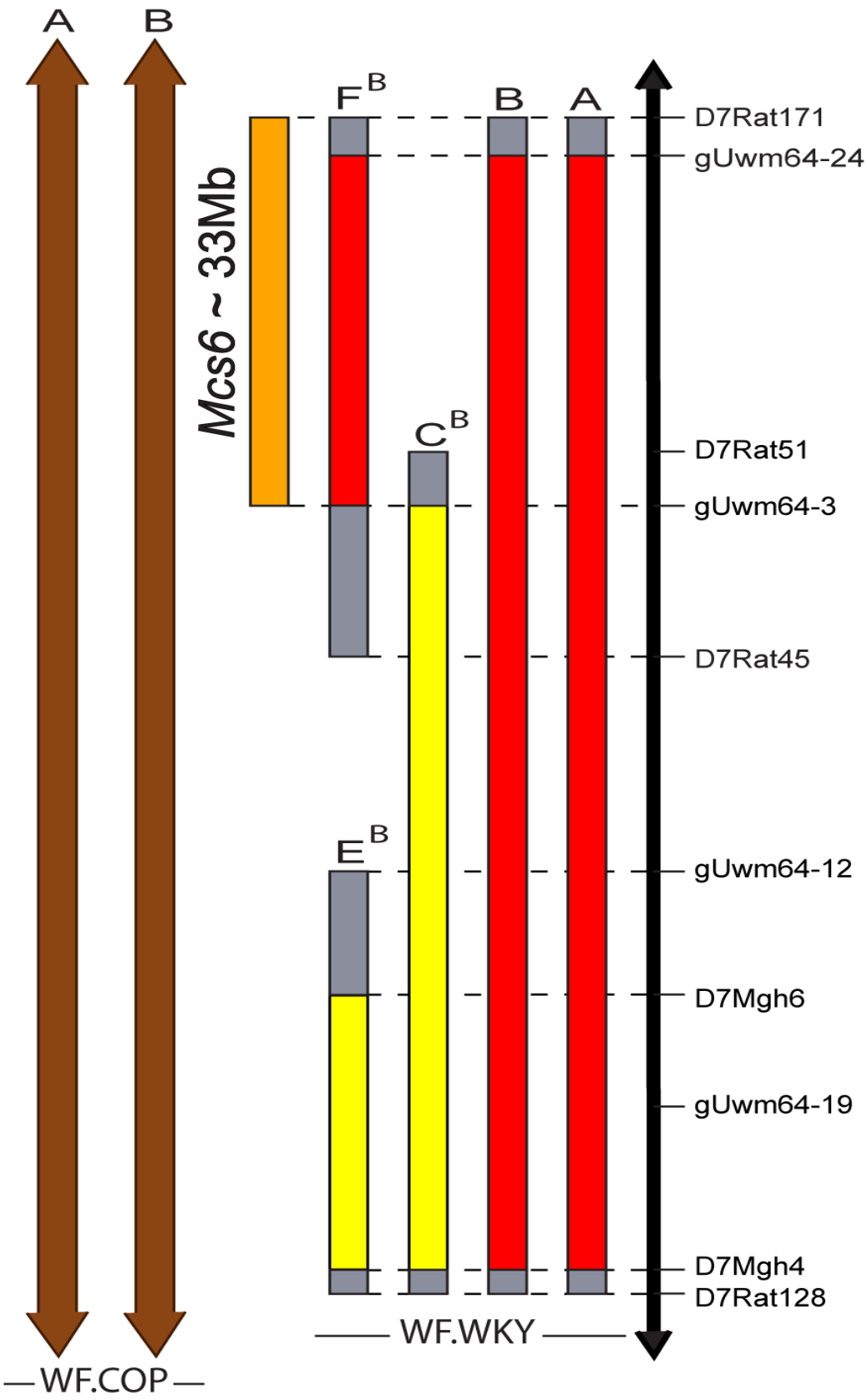

A
B
F B
B
A
Mcs6 ~ 33Mb
C B
D7Rat171
gUwm64-24
D7Rat51
gUwm64-3
D7Rat45
E B
gUwm64-12
D7Mgh6
gUwm64-19
D7Mgh4
D7Rat128
WF.WKY
WF.COP

Figure 4. WF.WKy congenic lines that define the *Mcs6* locus. Adapted from Sanders et al. (2011) [108]. Figure shows segment of rat chromosome *7* where the *Mcs6* locus is thought to be located and congenic lines that led to the mapping of the locus. Informative markers between the WF and WKy rat strains used for genotyping are shown on right. Red bars indicate WF.WKy segments that resulted in a resistant phenotype of fewer tumors compared to WF homozygous animals. These congenic lines are thought to contain the *Mcs6* locus. Yellow bars indicate WF.WKy segments that resulted in a susceptible phenotype of the same number of tumors compared to WF homozygous animals. These congenic lines are thought not to contain the *Mcs6* locus. Grey bars indicate regions of recombination for which informative markers are missing. Brown bars indicate WF.WKy segments for the *Mcs2* locus These congenic lines are thought to contain the *Mcs2* locus. The *Mcs2* and *Mcs6* locus overlap extensively, however it is not known if the two loci map to the same region. Superscript letters next to congenic line name indicate congenic line of origin. Overall, using these congenic lines, the *Mcs6* locus was mapped to a region of 33Mb on rat chromosome *7*, as indicated by orange bar.

Table 1. Summary of mammary carcinoma multiplicity phenotypes from WF.WKy and WF.COP rat chromosome *7* congenic lines used to map *Mcs6* and *Mcs2*. Adapted from Sanders et al. (2011) [108].

	WF.WKy Congenic Line					**WF.COP Congenic Line**		
	A	**B**	**C**	**E**	**F**	**A**	**B**	**WF**
Congenic Region[1] *Marker/ Marker*	*D7Rat171/ D7Rat128*	*D7Rat171/ D7Rat128*	*D7Rat51/ D7Rat128*	*gUwm64-12/ D7Rat128*	*D7Rat171/ D7Rat45*	*D7rat39/ D7Uwm12*	*D7rat39/ D7Uwm12*	-
Mean (SD) Mammary Carcinomas *per* Rat	3.5 (2.5)	3.9 (2.6)	8.0 (4.1)	7.2 (3.0)	3.4 (2.1)	2.4 (1.6)	2.0 (1.5)	7.3 (3.6)
n	17	32	23	43	28	16	16	19
p-value[2]	0.0012	0.0007	0.7521	0.7893	0.0001	<0.0001	<0.0001	-

[1]Markers spanning the maximal WKy or COP Chr *7* segment that was introgressed onto a susceptible WF genetic background are given.
[2]p-values are from Mann-Whitney nonparametric *post hoc* tests comparing each congenic line to the WF phenotype after a statistically significant Kruskal-Wallis test with a p-value < 0.0001.

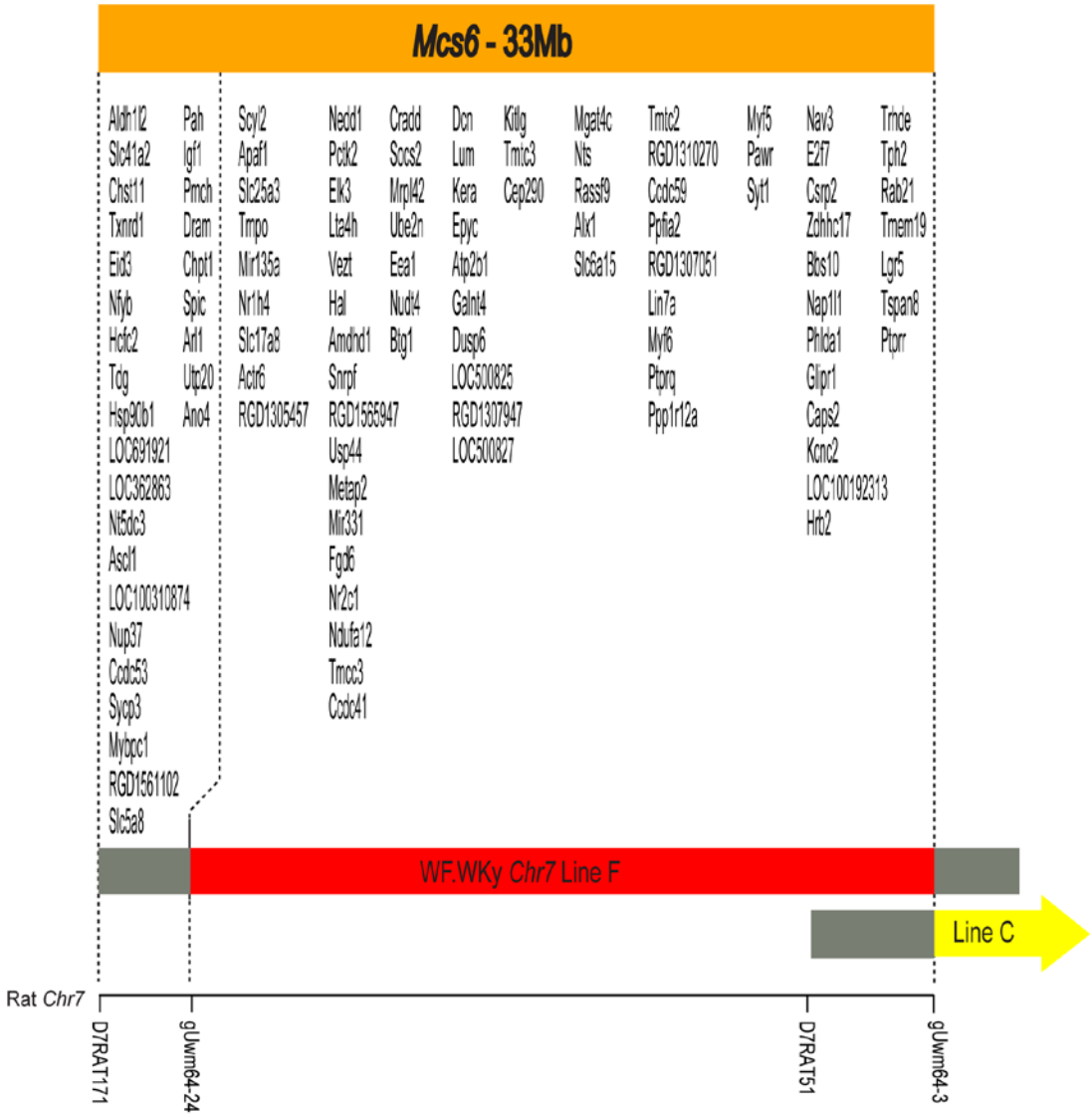


Figure 5. Annotated genes in the *Mcs6* locus. Figure adapted from Sanders et al. (2011) [108]. Known an predicted transcripts annotated in the UCSC Rat Nov. 2004 (Baylor3.4/rn4) Genome Browser. X- axis represents the region of rat chromosome *7* that contains the *Mcs6* locus. Dashed lines mark ends of the *Mcs6* locus. Color filled bars indicate congenic animals that have been tested for the *Mcs6* locus. Grey bars indicate regions of recombination with no known genetic markers. Genomic markers between the WF and WKy rat strains are shown on the bottom.

locus have been studied in GWA studies of breast cancer. Five of these SNPs reached the final validation step of their respective GWA study, but did not reach genome-wide significance after the validation step. These are *rs4146372*, *rs1154865*, *rs17740709*, *rs7310517* and *rs10507088* [120-122]. A list of all GWAS identified SNPs for the MCS6 locus can be found in Table 2. *rs1154865* had a p-value that was closest to reaching genome-wide significance at $6.6x10^{-7}$, with a p-value of $1x10^{-7}$ required for significance [120]. In a recent GWA study *rs17356907* was identified to associate with breast cancer risk. This SNP is located with the human orthologous region to *Mcs6*. The p-value for the SNP is $1.8x10^{-22}$ and reached genome-wide significance. It is located in close proximity to the gene NTN4 [62].

The *Mcs6* locus is currently too large for a practical functional study. Fine mapping of this locus is needed to identify a smaller region and to identify candidate genes. However, the *Mcs6* model is ideal for studying breast cancer susceptibility in the *MCS6* region. The human orthologous region to the *Mcs6* locus is located in a contiguous region on human chromosome *12*. This results in less complexity when studying the locus. Also, the *Mcs6* locus maps to a human orthologous region that contains several GWA study identified polymorphisms. Fine mapping of this locus will be the main goal discussed in this dissertation for the *Mcs6* locus.

Mcs1b

The second part of this dissertation will discuss the *Mcs1b* locus. The *Mcs1* locus was identified through a linkage analysis using the DMBA carcinogenesis susceptible

Table 2. Human SNPs in the *Mcs6* Orthologous Region that Have Been Reported in Genome-Wide Association Studies as Potentially Associating with Breast Cancer Susceptibility. Adapted from Sanders et al. (2011) [108]

SNP ID	Human Chr Location	Rat Chr Location	OR	P-value
rs4146372	*12q72131594*	*7q54588907*	nr	7.0×10^{-5}
rs1154865	*12q 72276104*	*7q54476579*	nr	6.6×10^{-7}
rs17740709	*12q83423340*	*7q43765270*	0.91	0.002
rs7310517	*12q89149235*	*7q37548015*	nr	1.04×10^{-3}
rs10507088	*12q97879744*	*7q29148172*	nr	5.12×10^{-4}
rs17356907	*12q 96027759*	*7q30853605*	0.91	1.8×10^{-22}

Chr, chromosome; OR, odds ratio; nr, not reported

WF rat strain and the resistant COP rat strain [106]. WF.COP congenic lines spanning different intervals of the *Mcs1* locus revealed three subloci, named *Mcs1a-c* [123]. Subsequently, several WF.COP congenic lines for the *Mcs1b* locus were developed, that spanned different regions of the potential *Mcs1b* locus as shown in Figure 6. Five WF.COP congenic lines showed the same tumor multiplicity post DMBA treatment as WF homozygous animals and these lines have a susceptible phenotype. These congenic lines are K, F3, U2, W2 and I4. This indicates that these congenic lines do not contain the *Mcs1b* locus. Two WF.COP congenic lines resulted in fewer tumors post DMBA treatment as compared to WF homozygous animals and therefore have a r esistant phenotype. These congenic lines are T and N3. Line T and N3 resulted in a similar tumor multiplicity. This indicates that the *Mcs1b* is located within these congenic lines. This delimits the *Mcs1b* locus to a region of 1.8Mb on rat chromosome *2*. The exact location for the *Mcs1b* locus is rat chr*2*:42,364,155-44,195,382. Animals homozygous for the COP allele in the *Mcs1b* locus showed a 56% reduction in tumor multiplicity compared to WF homozygous animal. An ectopic mammary gland transplant assay and subsequent treatment with DMBA revealed that the *Mcs1b* locus is acting in a mammary gland autonomous manner [87].

Interestingly, the *Mcs1b* locus has a human ortholog. The human orthologous region to the *Mcs1b* locus is located on human chromosome *5*. Its exact location is chr*5*: 54,816,178-57,003,049. The human orthologous region to the *Mcs1b* locus maps to the GWAS identified risk locus defined by the SNP *rs889312* at 5q11.2 that associates with an increase in breast cancer risk [61]. This risk locus is often referred to by the name *MAP3K1* and has been previously discussed in this dissertation (see section E of Genetic

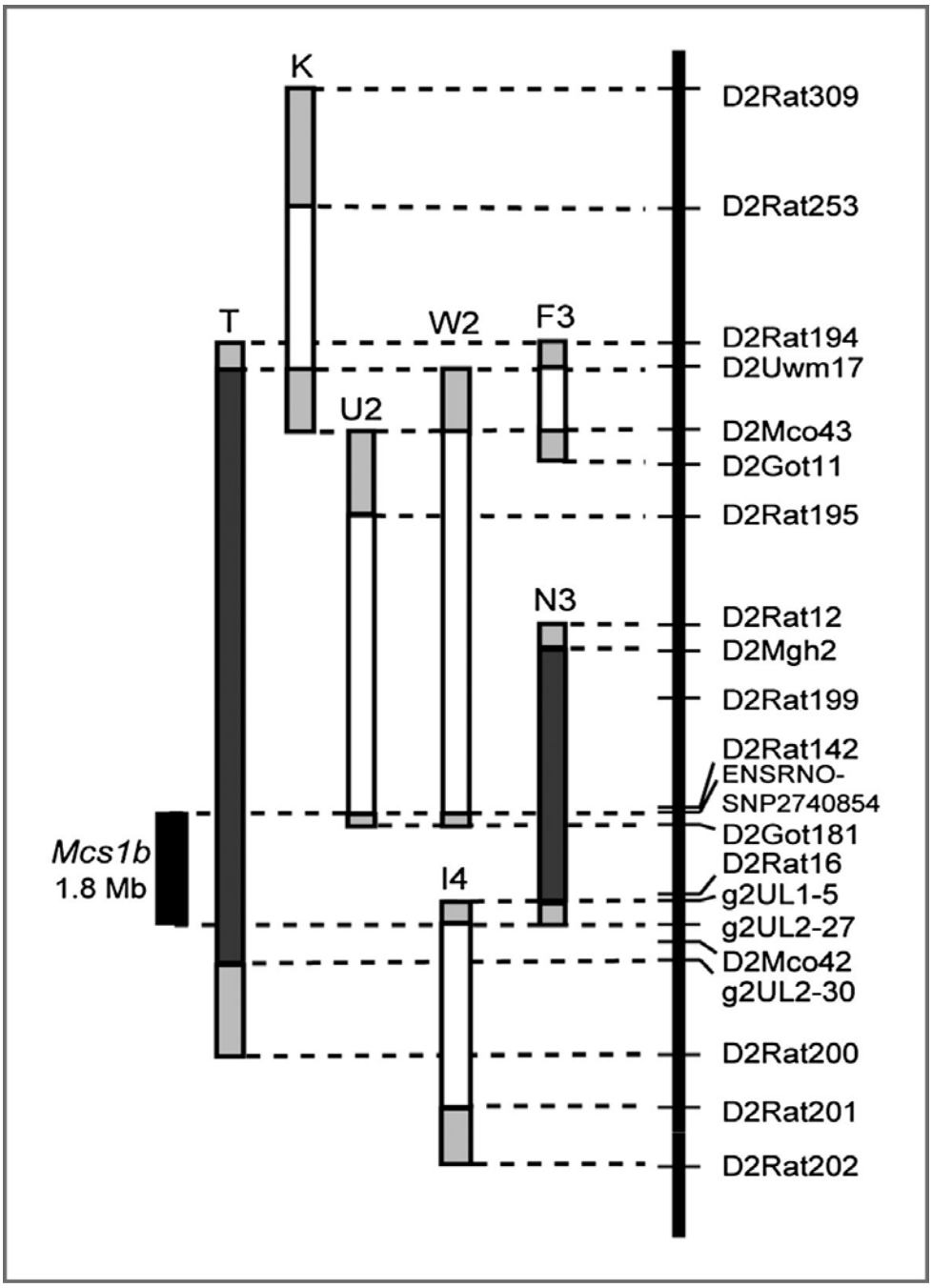

K
T
W2
F3
U2
N3
I4
Mcs1b
1.8 Mb
D2Rat309
D2Rat253
D2Rat194
D2Uwm17
D2Mco43
D2Got11
D2Rat195
D2Rat12
D2Mgh2
D2Rat199
D2Rat142
ENSRNO-
SNP2740854
D2Got181
D2Rat16
g2UL1-5
g2UL2-27
D2Mco42
g2UL2-30
D2Rat200
D2Rat201
D2Rat202

Figure 6. Map of congenic lines that define the *Mcs1b* locus. Adapted from denDekker et al. (2012) [87]. Map shows potential location for the *Mcs1b* locus on rat chromosome *2*. Dark grey bars are WF.COP congenic lines that resulted in fewer tumors compared to WF homozygous animals post DMBA treatment. These lines are thought to contain the *Mcs1b* locus. White bars are WF.COP congenic lines that have a phenotype similar to WF homozygous animals post DMBA treatment. These congenic lines are thought to not contain the *Mcs1b* locus. Informative markers between the two rat strains are shown on the right. Black bar is the location of the delineated *Mcs1b* locus. The *Mcs1b* is located on rat chromosome *2* and is about 1.8Mb in size.

factors of breast cancer risk and Figure 2). There are currently nine transcripts located within the *Mcs1b* locus that are expressed within the rat mammary gland. These are *Gpbp1*, *Map3k1*, *MIER3*, *Ankrd55*, *Il6st*, *Il31ra*, *Ddx4*, *Slc38a9*, and *Ppap2a*. None of these transcripts have coding sequence variants between the COP and WF rat strains; therefore the causative genetic variant(s) is likely to be regulatory in nature [87].

The *Mcs1b* locus is an ideal model for the 5q11.2 identified human breast cancer risk locus. As previously discussed, there are seven potential risk SNPs in the human *MCS1B* locus (see section E of Genetic factors of breast cancer risk and Figure 2). Since there are seven potential risk SNPs in the human *rs889312* marked breast cancer risk locus, it is important to determine how many potential risk associated sequence variants there are in the rat *Mcs1b* locus. Therefore, identifying all sequence variants between the two rat strains in the *Mcs1b* locus and identifying potential mechanisms of their action will be my goal for the *Mcs1b* locus.

Overall Goal

The overall goal is to identify genetic elements within the *Mcs6* and *Mcs1b* loci that modify mammary cancer susceptibility. The goal is then to determine the underlying mechanism and identify orthologous genetic elements within human breast cancer risk loci.

Hypothesis and Aims

The hypothesis is that both the *Mcs6* and *Mcs1b* contain genetic elements that control mammary cancer susceptibility. These genetic elements can be identified and act through regulation of gene expression within the *Mcs6* and *Mcs1b* loci.

Aim 1: Fine map the *Mcs6* locus to a smaller chr *7* defined chromosomal segment using congenic WF.WKy lines and DMBA- induced mammary carcinoma multiplicity phenotyping.

Aim 2: Map and annotate *Mcs1b* sequence differences between susceptible WF and resistant COP alleles using sequence capture, next-generation sequencing and *in silico* approaches.

Aim 3: Perform functional analysis on *Mcs1b* sequence variants and human tagged *rs889312* SNPs.

Aim 4: Perform analysis of the overlap between rat genetic loci that associate with mammary cancer risk and human GWAS identified breast cancer risk SNPs.

CHAPTER II

FINE MAPPING OF THE MCS6 LOCUS

Introduction

The *Mcs6* locus was initially identified in a linkage analysis. The DMBA carcinogenesis susceptible WF rat strain and the resistant WKy were bred to generate a F1 generation. Subsequently, the F1 generation was backcrossed to the susceptible WF rat strain, generating an $(WFxWKy)_{F1}$ x WF backcross. This was repeated to generate the F2 generation. Rats were then phenotyped for tumor multiplicity post DMBA treatment. This resulted in the identification of four loci controlling mammary cancer susceptibility in the WKy rat. These loci are *Mcs5, Mcs6, Mcs7* and *Mcs8* [107]. The *Mcs6* locus was physically confirmed and mapped using congenic lines as shown in Figure 4. The locus currently maps to a region of 33Mb on rat chromosome *7.* The exact location of the locus is between genetic markers *D7Rat171* and *gUwm64-3* (chr*7*:22,382,725-55,364,398). There are 111 transcripts annotated in this region using the UCSC genome browser and none of them are known breast cancer susceptibility genes. Importantly, all of the rat transcripts annotated in this region are also found in the human orthologous region to the *Mcs6* locus [108]. Some of the *MCS6* genes are expressed differentially between normal breast tissue and ductal breast cancer tissue using an Oncomine search. Out of the 111 annotated genes, 14 genes show increased expression levels in breast cancer tissue compared to normal breast cancer tissue. Also, 9 genes are expressed lower in breast cancer tissue compared to normal breast tissue.

One gene was both expressed higher and lower in breast cancer tissue, depending on which study was considered in the analysis. While there are 111 annotated genes in the *Mcs6* region according to the UCSC genome browser, there were far more annotated genes in the Rat Genome Database (137 genes) and the Ensemble Genome Browser (215 genes) [108].

The human orthologous region to the *Mcs6* locus is found contiguously on human chromosome *12*. The human orthologous region is inverted with respect to the rat *Mcs6* region and is found at human chromosome *12*: 71,299,117 to 105,502,699 [108]. The rat *Mcs6* locus and human *MCS6* orthologous region share 37.7% of the bases and the span 99.9% of the size according to the UCSC genome browser. The human *MCS6* region contains several SNPs that have been tested in GWA studies of breast cancer. Five of these SNPs entered the final validation step of their respective study but were found not to be genome-wide significant [120-122]. One SNP, *rs17356907,* was identified in a large breast cancer GWA study in 2013. The minor allele odds ratio is 0.91 (0.89–0.93), suggesting that the SNP has a protective effect on breast cancer [62]. The SNP was identified in a GWA study using a population of European decent. The SNP was then tested in a GWA study using a population of East Asian decent and similar results were seen [124]. It is located in close proximity to the NTN4 gene. This makes the *Mcs6* rat model an ideal model for studying the mechanisms of breast cancer susceptibility that are present in the human *MCS6* region.

The *Mcs6* locus as it is currently mapped is very large. There are over a hundred genes present in this region, making identification of a candidate gene difficult. Also, it is

difficult to target this region for DNA sequencing in both rat strains due to its size to identify candidate genetic variants between the rat strains. The goal for this locus is to fine map it to a region that is small enough in size to warrant functional analysis of this locus.

To fine map the *Mcs 6* locus, several WF.WKy congenic lines for the locus were developed that span different intervals of current *Mcs6* locus. Developing congenic lines is a laborious process that involves the selective breeding of rats over a long period of time. It is currently the preferred method of identifying smaller intervals for rat QTLs. It would also be possible to perform RNA-seq for the genes in the targeted region. However, analysis of the expression levels of so many genes can be complex. It is also not known if the *Mcs6* locus is autonomous to the rat mammary gland, and therefore selecting the right cell type would be a challenge. It is also possible to identify all genetic variation between the two rat strains in this region using whole genome sequencing. This would result in the identification of a lot of sequence variants and parsing out a potential candidate would be difficult. Another reason why fine mapping the *Mcs6* locus is essential is because there is a possibility of several distinct QTLs may be present in this region, which will only be identified through fine mapping of this locus. Several *Mcs* loci have been mapped to distinct subloci upon fine mapping using congenic lines, including the *Mcs1* and *Mcs5* loci [110, 123]. This is particularly important since the *Mcs6* locus currently spans a large portion of the chromosome. The *Mcs6* locus would therefore benefit from fine mapping to a smaller region. None of the annotated transcripts in the *Mcs6* region are known breast cancer susceptibility genes. Therefore, studying this locus can result in the identification of a novel breast cancer susceptibility gene.

The goal for the *Mcs6* locus is to fine map the locus to a smaller region on rat chromosome *7*. The hypothesis is that the *Mcs6* locus is located within on of three independent WF.WKy congenic lines that span the 33Mb of the locus.

Methods

A. Generating WF.WKy congenic animals for the *Mcs6* locus

All animals used for this study were housed by the University of Louisville Research Resources Center Animal Facility. All protocols were approved by the University of Louisville IACUC committee.

Three WF.WKy congenic lines were developed to fine map the *Mcs6* locus. These are lines D, H and I. WF.WKy congenic line D was developed through a backcross of the resistant *Mcs6* congenic line B (Shown in Figure 4) to the WF rat strain (obtained from Harlan). The resulting pups were genotyped for informative markers in the *Mcs6* region and recombinants were determined. A recombinant would be any animal that showed a shorter WKy allele than the original line B. Recombinants were crossed again to WF animals to expand the line. Brothers and sisters containing the same recombinant WKy allele in the regions were mated to fix the new line. WF.WKy congenic lines H and I were generated through the same breeding scheme as shown in Figure 3. In short, WF and WKy inbred rats were obtained from Harlan and bred with each other. The resulting heterozygous F1 males were backcrossed to WF inbred females to generate a $(WFxWKy)_{F1}$ x WF backcross. The resulting pups were genotyped to ensure a WKy present at the *Mcs6* locus. The rats were then backcrossed seven more times until the N8 generation. At each backcrossing the animals were genotyped to ensure the desired WKy

allele was still present. At the N8 generation, brothers and sisters that contained the same WKy allele were mated. The brothers and sisters used for this mating were heterozygous for the WKy allele of interest. Their pups will be 25% WF homozygous, 50% heterozygous and 25% WKy homozygous for the area of interest. The WF homozygous and WKy homozygous females were selected for phenotyping. Line I was also tested at the N9 generation. To get the N9 generation, the N8 generation was backcrossed to WF females. Brothers and sisters were mated and resulting homozygous offspring were tested.

B. Genotyping of Animals

Animals had to be 12 weeks of age for breeding. Pups were tail clipped at 5-8 days of age and tattooed for identification. Tails were digested in genomic lysis solution supplemented with proteinase K at 15mg/ml. Tails were digested at 55°C overnight and extracted using protein precipitation solution (Qiagen) and an isopropanol- ethanol DNA extraction. DNA was amplified using GeneAmp Fast PCR Master Mix (Life Technologies) and primers for microsatellite markers between the two rat strains. PCR reactions were run on a 3% high resolution agarose gel (GeneMate) alongside PCR reactions of control DNA samples. SNPs between the two rat strains were also used for genotyping. DNA was amplified using a TaqMan Genotyping Master Mix (Life Technologies) and primers and probes specific for the SNP marker. Analysis was done on a StepOne Plus QPCR machine (ABI) using the StepOne software for genotyping analysis. Informative markers used to generate congenic lines for the *Mcs6* locus are shown in Table 3.

Table 3. Informative markers used to genotype congenic lines for *Mcs6*.

ID	Position[#]	Forward sequence	Reverse sequence	Probe sequence (rat strain this probe matches)	Used for congenic lines
rs13459010	25,172,732	AATTACCTTTCTACCATTAGCTTTGACA	CCACCCCAGATAAAAAGGAATTAA	AGGACAGTTCAATTC (WKy) TAAAAGGACAGTCCAATTC (WF)	D, H
rs65052669	46,915,037	TGCACATCTCCATCTTAAGAGCTT	ATACAATGGACTTTTTCAGCCATAAA	TTATGTTTCACAGGAGC (WKy) TTATGTTTCACAGGAACCC (WF)	D, I
g7UL1-39	25,498,127-25,498,190	CATTGCTGTCCGCTCAACAC	TGGTGCTTGTCCAGCTCTCTC		D, H
D7ARB16	47,145,688-47,145,774	ACATACATACACACCAGAACGC	ATCACTGAACGTAACAGGTGAC		D, I
D7Mit28	29,204,226-29,204,458	AGTCCGAAGCCATATGTTGG	TAAACCTATGAATTGCCCGC		D, H
D7Rat182	33,195,152-33,195,408	CAAATGTTCAGACCACTAGATGAGA	TACCCACCTCCACCCAGATA		D, H
D7Rat83	37,616,505-37,616,890	CAGGTGAGGAGTGAATGAGGA	CCCAGGCAGAGATCATCAAT		D, I
D7Rat103	44,462,790-44,463,229	CTGGTGCTTTTGGGTCTGTT	GTGTCAAACTGTGGGGATCC		D, I
gUwmn64-3	55,364,143-55,364,398	AACAGTTCTCCTTTCCCTTC	TCTCTTTGCCAGTCTGTTTT		I
D7Rat45	68,056,925-68,057,251	GAGGAGGAAACAGAGGTGGTC	CACTCCCTGAGGCTCCTATG		I
D7Rat51	50,883,678-50,883,793	TTTGATGGACTTGGGACTTTC	GTGGGCAAGGAATGAGTAGG		I
g7UL3-17	53,034,818-53,034,865	CACTGTTTGGTCTCCCTGTGG	ACCAGAAGGAACAGGCAGCA		I
D7Mgh6	96,648,984-96,649,133	ACCCCAAAGACTTAAAAAATTAGC	TGGCTTGTTAATCGTGACTACTG		I
D7Rat171	22,382,725-22,382,935	CGATGTTGTTCGAGGTGCTA	CCTTCATTCCACCTTTGGTC		H

[#]Position on rat chromosome *7* using RGSC Genome Assembly 3.4, Wf is Wistar- Furth and WKy is Wistar- Kyoto rat strain

C. Phenotyping of animals

For each congenic lines, 18-65 female rats were tested. Female rats used in phenotyping experiments were housed in an all-female rat room to ensure no exposure to male rat hormones. The animals were administered DMBA at 50-55 days of age with a single dose of 65mg/Kg body weight through oral gavage. DMBA was prepared by suspension with sesame oil at 20mg/ml and then heated in a boiling waterbath for 20 minutes. The experiment was stopped 15 w eeks post DMBA treatment and tumor multiplicity was determined by counting all mammary tumors $\geq 3 \times 3$mm. The spleens were removed for phenotyping using the method described above. The data was analyzed using Systat 13™. A nonparametric Mann- Whitney U test was performed, comparing each congenic lines to the WF homozygous animals tested, after a significant Kruskal-Wallis test for congenic lines H and I. A nonparametric Mann-Whitney U test was performed comparing line D animals to WF animals tested alongside line D animals.

D. Functional analysis of *Mcs6* congenic lines

Location and annotated genes of the possible *Mcs6* locus were determined using the UCSC genome browser with the rat genome assembly Nov. 2004 (Baylor 3.4/ rn4) [119]. The human orthologous region to the *Mcs6* locus was determined using the "In other genomes (convert)" function in the UCSC Genome Browser. To determine informative markers between the two rat strains, sequences for regions of interest were downloaded from the UCSC genome browser and manually scanned for regions of dinucleotide repeats. Primers against the repeats were designed and tested using WF and WKy control DNA as described in the genotyping methods section. SNPs between the

two rat strains were identified using the "SNPlotyper" function from the Rat Genome Database [125]. Primers were designed against potential SNP regions using Primer 3. A PCR reaction for WF and WKy control DNA was performed using Accuprime Taq (Life Technologies) and the resulting samples were run on a 1% agarose gel. PCR products were purified using a PCR Purification Kit (Qiagen). PCR products were then sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies). Sequenced products were cleaned by adding Agencourt AMPure XL beads and 80% ethanol. Beads were washed with 80% ethanol and DNA was eluted using molecular grade water. Sequences were submitted to the University of Louisville DNA Core for analysis. To identify SNPs tagged by GWAS identified polymorphisms, the software Haploview version 2 was used. All SNPs with an r^2 of 0.8 were considered tagged [126].

Results

A. Phenotyping of WF.WKy congenic line D

The first WF.WKy congenic line tested was line D. Line D was initially tested at the University of Wisconsin- Madison in the Dr. Michael Gould lab. In total, 35 congenic line D females were tested in Wisconsin. These animals developed on average 5.6 tumors per rat with a standard deviation of 4.0. Nineteen WF homozygous animals tested alongside the line D congenic animals showed a tumor multiplicity of 7.3 tumors per rat with a standard deviation of 3.6. A statistical analysis resulted in a p-value of 0.049. Since this p-value is only marginally significant and there may be complexity to the *Mcs6* locus and more animals needed to be tested. Thirty more line D congenic animals and 22 WF homozygous animals were tested at the University of Louisville. The line D congenic

females developed 5.1 tumors per rat with a standard deviation of 2.6, while the WF homozygous females developed 5.0 tumors pre rat with a standard deviation of 2.8. The data were then pooled: A total of 65 congenic females and 41 WF homozygous females were tested. Line D congenic females showed a tumor multiplicity of 5.4 tumors per rat with a standard deviation of 3.4. Results are shown in Table 4. The WF homozygous females showed a tumor multiplicity of 6.0 rats per animals with a standard deviation of 3.3. The tumor multiplicity data for the WF females was lower than previously reported [107]. This may be due to differences in the environments at different animal care facilities. The tumor multiplicity between the line D congenic animals and the WF homozygous animals is not statistically significant (p-value 0.18). This suggests that the *Mcs6* locus is not located within line D. The original mapping of the *Mcs6* locus is shown in Figure 4. The *Mcs6* locus was mapped to a region of 33Mb on rat chromosome *7*. An updated congenic line map that includes line D is shown in Figure 7. Note, some of the congenic lines that are not needed to further fine map the *Mcs6* locus have been removed from Figure 7. Line F as shown in Figure 7 spans the entirety of the originally mapped *Mcs6* locus. Congenic line D overlaps line F extensively. Since the WF.WKy congenic line D has a susceptible phenotype and does not contain the *Mcs6* locus, this splits the locus into two possible locations. The first possible location is the region of recombination for line D and F. There are no known genetic markers between the WF and WKy rat strains in this region. This region is 325Kb in size. The second possible location is the region on non-overlap between the congenic lines F and D. This region is 8.5Mb in size. To test if the *Mcs6* locus is located in one of these two possible locations,

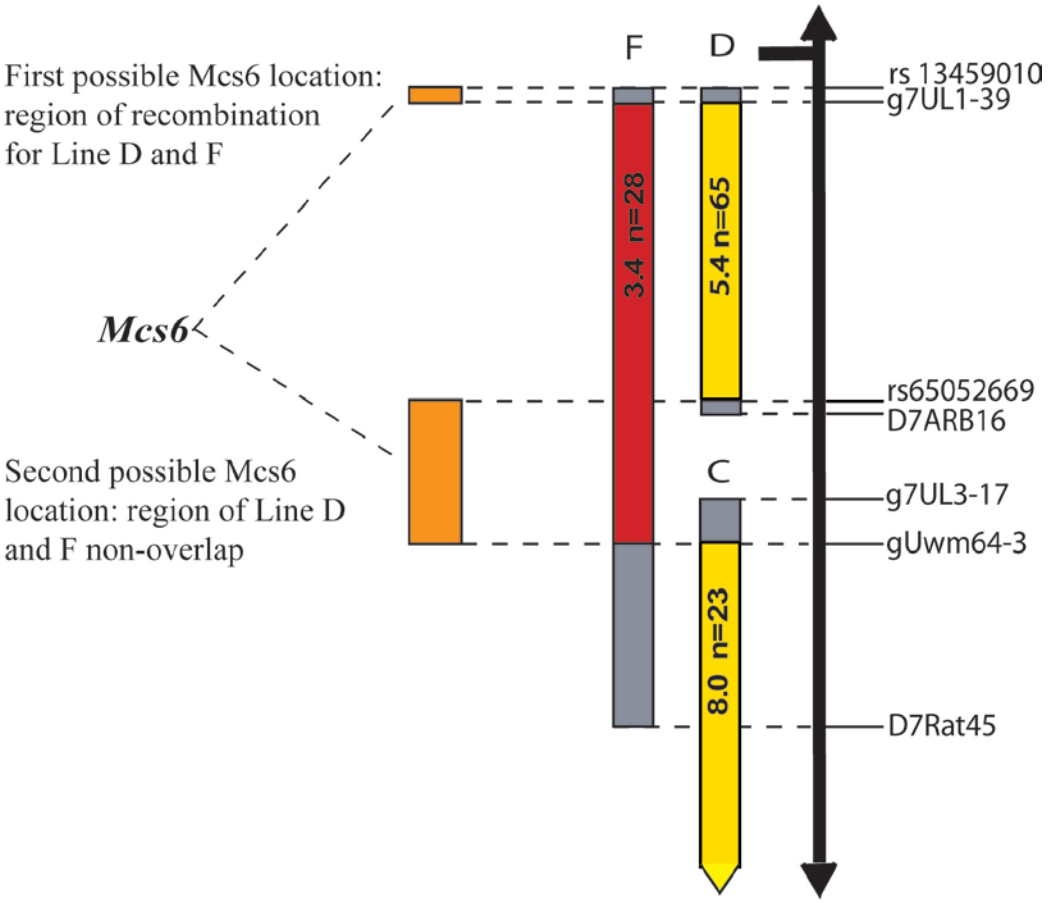

First possible Mcs6 location: region of recombination for Line D and F
Mcs6
Second possible Mcs6 location: region of Line D and F non-overlap
F
D
C
3.4 n=28
5.4 n=65
8.0 n=23
rs 13459010
g7UL1-39
rs65052669
D7ARB16
g7UL3-17
gUwm64-3
D7Rat45

Figure 7. *Mcs6* congenic line map with the addition of WF.WKy congenic line D. Yellow bars indicated WF. WKy congenic lines have a susceptible phenotype. This means these animals developed the same amount of tumors as WF homozygous animals and the *Mcs6* locus is not located within these genomic regions. The red bar (line F) indicates a WF.WKy congenic line that has a resistant phenotype. This means these animals developed fewer tumors compared to WF homozygous animals and the *Mcs6* locus is found in this genomic region. Tumor multiplicity and number of animals tested is shown inside congenic line. Line F spans the entirety of the previously mapped *Mcs6* locus. Grey bars indicate regions of recombination. Informative markers between the two rat strains are missing in these regions. Congenic line C spans a larger genomic region as shown in this figure. The entirety of line C can be seen in Figure 4. Congenic line D splits to *Mcs6* locus into two possible locations indicated by orange bars.

two independent congenic lines spanning the regions of interest were developed and phenotyped.

B. Phenotyping of WF.WKy congenic lines H and I

Two separate WF.WKy congenic lines spanning the two possible location of *Mcs6* were developed as described in Figure 3. The two independent congenic lines were named H and I and their genomic location are shown in Figure 8. Eighteen WF.WKy congenic line H animals were treated with DMBA and tumors were counted. Line H animals developed on average 6.6 tumors per animal with a standard deviation of 2.3 tumors. Eighteen animals were also tested for line I. Line I animals developed on average 3.4 tumors per rat with a standard deviation of 2.2 tumors. The tumor multiplicity of these two congenic lines was compared to WF homozygous animals that were treated at the same time as congenic lines H and I. Overall, 20 WF homozygous animals were tested. The WF homozygous animals developed on average 7.7 tumors per rat with a standard deviation of 3.9 tumors. Results are shown in Table 4. Line H is not statistically different from WF homozygous animals with a p-value of 0.27. Line I is statistically significant with a p-value of 0.0005. A WKy allele in the line I genomic region results in a 56% reduction in tumor multiplicity compared to WF homozygous animals. These results map the *Mcs6* locus to part of the line I genomic region. An updated congenic line map for the *Mcs6* locus is shown in Figure 9. The map shows the new location of the *Mcs6* locus. It is interesting that the tumor multiplicity phenotype for congenic lines I and F are identical, suggesting that the same genetic element may be acting in both congenic lines. The *Mcs6* locus was reduced in size from 33Mb to 8.5Mb on rat chromosome *7* using congenic

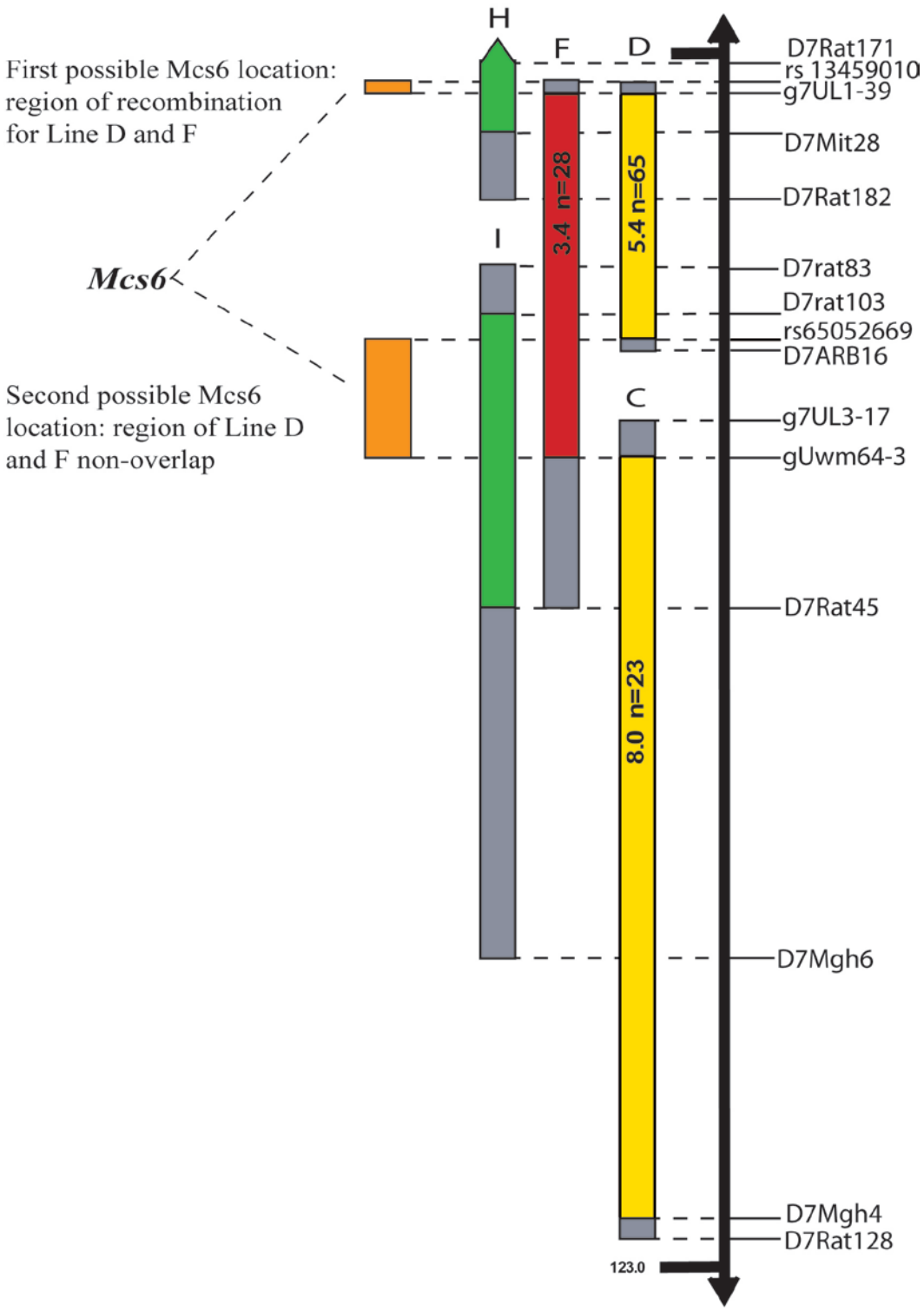

First possible Mcs6 location: region of recombination for Line D and F
Mcs6
Second possible Mcs6 location: region of Line D and F non-overlap
H
F
D
I
C
3.4 n=28
5.4 n=65
8.0 n=23
D7Rat171
rs 13459010
g7UL1-39
D7Mit28
D7Rat182
D7rat83
D7rat103
rs65052669
D7ARB16
g7UL3-17
gUwm64-3
D7Rat45
D7Mgh6
D7Mgh4
D7Rat128
123.0

Figure 8. *Mcs6* congenic line map showing the location of independent congenic lines H and I. Yellow bars indicated WF. WKy congenic lines have a susceptible phenotype. This means these animals developed the same amount of tumors as WF homozygous animals and the *Mcs6* locus is not located within these genomic regions. The red bar (line F) indicates a WF.WKy congenic line that has a resistant phenotype. This means these animals developed fewer tumors compared to WF homozygous animals and the *Mcs6* locus is found in this genomic region. Tumor multiplicity and number of animals tested is shown inside congenic line. Line F spans the entirety of the previously mapped *Mcs6* locus. Grey bars indicate regions of recombination. Informative markers between the two rat strains are missing in these regions. Green bars indicate the location of the two independent WF.WKy congenic lines H and I. These lines were generated to map the *Mcs6* locus to two possible locations indicated by orange bars.

Table 4. Summary of mammary carcinoma multiplicity phenotypes from WF.WKy congenic lines D, H and I used to map *Mcs6*

	WF.WKy congenic line			**WF.WKy congenic line**	
	H	**I**	**WF**	**D**	**WF**
Congenic region *Marker/Marker*[1]	*~D7Rat171/ D7Rat182*	*D7rat83/ D7Mgh6*	-	*rs13459010/ D7ARB16*	-
Mean (SD[4]) mammary carcinomas per rat	6.6 (2.3)	3.4 (2.2)	7.7 (3.9)	5.4 (3.4)	6.0 (3.3)
N	18	18	20	65	41
p-value	0.27[2]	0.0005[2]	-	0.18[3]	-

[1]Markers spanning the maximal WKy or COP Chr *7* segment that was introgressed onto a susceptible WF genetic background are given. Note proximal end of line H is not known.
[2]p-values are from Mann-Whitney nonparametric *post hoc* tests comparing each congenic line to the WF phenotype after a statistically significant Kruskal-Wallis test with a p-value 0.0002
[3]p-value is from Mann-Whitney nonparametric test comparing line D to the WF phenotype
[4]Standard deviation

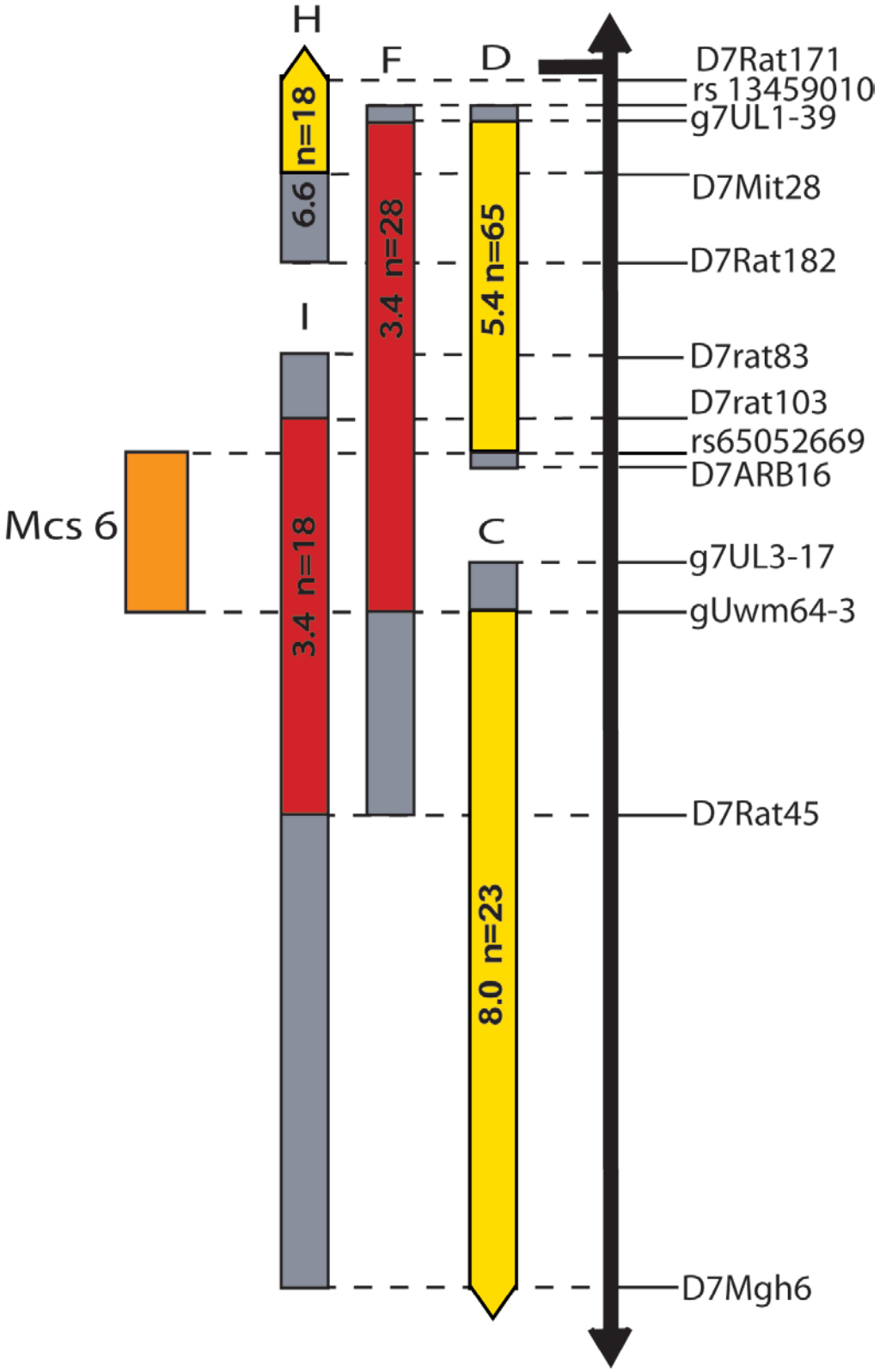

H
F
D
I
C
n=18
6.6
3.4 n=28
5.4 n=65
3.4 n=18
8.0 n=23
Mcs 6
D7Rat171
rs 13459010
g7UL1-39
D7Mit28
D7Rat182
D7rat83
D7rat103
rs65052669
D7ARB16
g7UL3-17
gUwm64-3
D7Rat45
D7Mgh6

Figure 9. *Mcs6* congenic line map of fine mapped *Mcs6* locus. Yellow bars indicated WF.WKy congenic lines have a susceptible phenotype. This means these animals developed the same amount of tumors as WF homozygous animals and the *Mcs6* locus is not located within these genomic regions. The red bars (line F and I) indicate WF.WKy congenic lines that have a resistant phenotype. This means these animals developed fewer tumors compared to WF homozygous animals and the *Mcs6* locus is found in these genomic regions. Tumor multiplicity and number of animals tested is shown inside congenic line. Grey bars indicate regions of recombination. Informative markers between the two rat strains are missing in these regions. Phenotyping results from WF.WKy congenic lines map the *Mcs6* locus to the region shown in orange. The locus was reduced in size from 33Mb to 8.5Mb using congenic lines D, H and I.

lines D, H and I. The location of the *Mcs6* locus is now on r at chr*7*: 46,915,037-55,364,398.

Discussion

The goal of aim 1 was to fine map the *Mcs6* locus to a smaller genomic region. The *Mcs6* locus was initially mapped to a region of 33Mb on rat chromosome *7* using WF.WKy congenic animals. The *Mcs6* locus was fine mapped to a region of 8.5Mb using three WF.WKy congenic lines and DMBA- induced mammary carcinoma multiplicity phenotyping. The new location for the *Mcs6* locus is chr*7*: 46,915,037-55,364,398. Fine mapping of the *Mcs6* locus resulted in a 75% reduction in size of the locus. Previously, 111 transcripts were annotated in the *Mcs6* locus, making a functional analysis of this locus difficult [108]. The fine mapped *Mcs6* locus contains 22 t ranscripts. These are shown in Figure 10. None of these genes are known breast cancer susceptibility genes, but all transcripts found in the rat *Mcs6* locus are also found in the human orthologous region. Therefore, studying the *Mcs6* locus can result in the identification of a novel breast cancer susceptibility gene. It is possible to perform RNA seq or design primers for RT-QPCR for the 22 genes annotated in this region. This could reveal any differences in the expression levels between the two rat strains and could help identify a candidate gene for the *Mcs6* locus. However, it is necessary to determine if the *Mcs6* locus acts in a mammary gland autonomous manner in order to determine which tissue type to use for RNA expression analysis.

An Oncomine search revealed that several of the genes found in the *Mcs6* locus are differentially expressed between normal human breast tissue and ductal breast cancer tissue. These genes are PPP1R12A, ZDHHC17, NAV3 and TSPAN [108]. PPP1R12A or

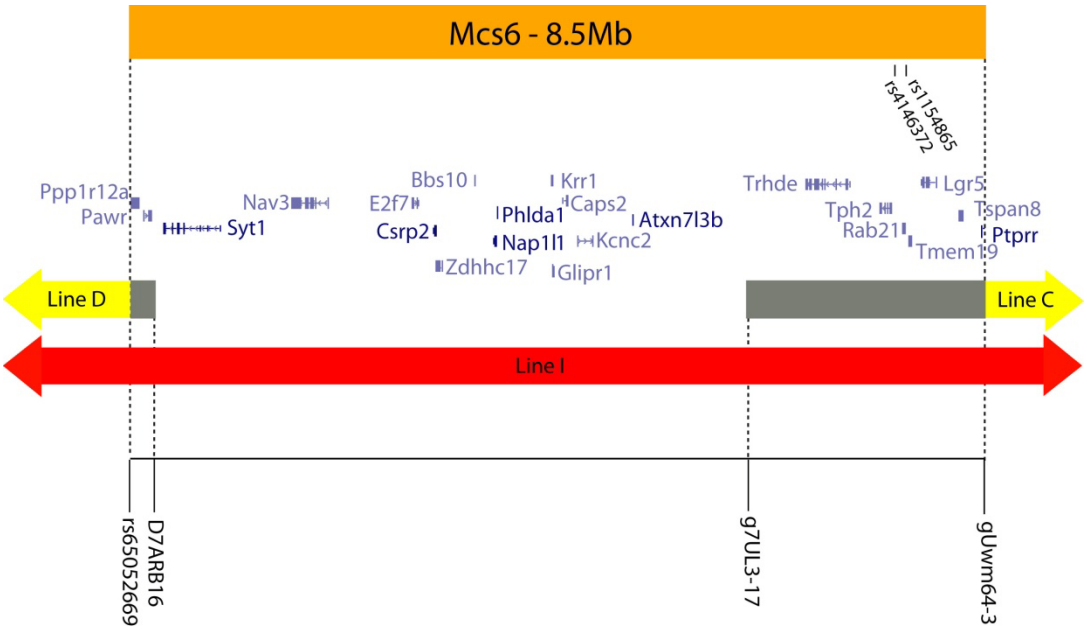


Figure 10. T ranscript map of the fine mapped *Mcs6* locus. Yellow bars indicated WF. WKy congenic lines that resulted in a susceptible phenotype. This means these animals developed the same amount of tumors as WF homozygous animals and the *Mcs6* locus is not located within these genomic regions. The red bar (line I) indicates a WF.WKy congenic line that has a r esistant phenotype. This means these animals developed fewer tumors compared to WF homozygous animals and the *Mcs6* locus is found in this genomic region. Grey bars indicate regions of recombination. Informative markers between the two rat strains are missing in these regions. Congenic lines D, C and I map the *Mcs6* locus to a 8.5Mb region shown in orange. Informative markers used to delineate congenic lines are shown at the bottom. Rat orthologous region to human breast cancer GWAS identified SNPs are shown on t op right. All annotated transcripts according to the UCSC genome browser Nov. 2004 (Baylor 3.4/rn4) assembly are shown in blue. The fine mapped *Mcs6* locus contains 22 annotated transcripts.

Myosin Phosphatase- Targeting Subunit 1 (MYPT1) is a subunit of Myosin Phosphatase and is involved in smooth muscle contraction and possibly hypoxia [127]. MYPT1 is also involved in phosphorylation of RB1 leading to cell cycle progression [128]. ZDHHC17 or Huntingtin- Interacting Protein 14 (HIP-14) is a protein involved in endocytosis and is implicated in Huntington's disease [129]. NAV3 or Neuron Navigator 3 is a protein involved in axonal guidance. Chromosomal aberrations of chromosome 12 in several cancers have resulted in the loss of the NAV3 gene. This implicates this gene in several types of cancers including colorectal cancer, T-cell lymphomas, neuroblastomas and squamous cell carcinomas [130-133]. TSPAN or Tetraspanin is part of a family of transmembrane proteins involved in cell signaling. Proteins in this family are involved in regulation of cell growth and motility. Members of the Tetraspanin family are implicated with a variety of cancers including ovarian carcinomas [134].

Six human GWAS identified polymorphisms map to the rat orthologous region of the *Mcs6* locus. A list of the SNPs can be found in Table 2. One of these GWAS identified polymorphisms reached genome-wide significance in its respective study. This SNP is *rs17356907*. The rat orthologous region of this SNP maps to the proximal end of WF.WKy congenic line D. The SNP was found to reduce breast cancer risk with an OR of 0.91 [62]. It is interesting that this polymorphism maps to a congenic line that has a susceptible phenotype and therefore does not contain the *Mcs6* locus, indicating that *rs17356907* may not be the human ortholog to the *Mcs6* locus. It is possible that the *Mcs6* locus is complex and that there are multiple genetic elements controlling mammary cancer susceptibility. It appears that there is at least one genetic element in congenic line I that modifies mammary cancer susceptibility. However, congenic line D is very large

and may contain phenotypically opposing genetic elements that mask the phenotype. To identify if there are opposing genetic elements located in congenic line D, recombinant congenic lines that span shorter intervals than line D need to be phenotyped.

Two GWAS identified polymorphisms map to the rat orthologous region contained in congenic line I. These are *rs1154865* and *rs4146372*. Both SNPs failed the last validation step in their respective study. r*s1154865* was the closest to reaching genome-wide significance with a p -value of $6.6x10^{-7}$ when $1x10^{-7}$ is required for genome-wide significance. *rs1154865* tags at least six other SNPs. There is no tagging information for SNP *rs4146372* in the Haploview database. It is possible that one of these two identified SNPs or the SNPs they tag are the human ortholog of *Mcs6*. However, it is also possible that the *MCS6* SNP has not been identified in a GWAS study or has not been made available in public databases yet.

There are currently seventeen known SNPs between the WF and WKy rat strains in the *Mcs6* region. These were identified using the SNPlotyper software at the Rat Genome Database [125]. The identified SNPs in the *Mcs6* region are shown in Table 5. There are likely to be many more sequence variants between the two rat strains in this region. It is necessary to sequence both rat strains in the *Mcs6* region to identify all sequence variants. However, sequencing technique for targeted sequencing currently available require smaller genomic regions for sequencing. Further fine mapping of the *Mcs6* locus using recombinant congenic lines that originate from line I may be necessary to feasibly identify candidate *Mcs6* sequence variants.

Table 5. *Mcs6* SNPs between the WF and WKy rat strain.

ID	Position[#]	Forward Primer	Reverse Primer
rs1349010	22613971	TGCTGAAACTGCATTCAAAGA	TTGCATCTCTAACTCCTGGGTA
rs13457291	29833956	TCAACCTTTGCCCTTTCATT	GGGTTGCAGAGGGATATACTGA
SNP2793282	40318750	TCCTTTTGCCCATGTTTCTC	TCTTGATGGCTTCATGGACA
rs63992414	41254704	AGGACAAAAGACATCCCCAGT	CATGAATCTCAAAAGGAGCTGTT
SNP2793293	41716640	TGTCATTGCTTCCCCTCAAT	AAGAGGCAGCGTTTAAGGTG
SNP2793296	42001299	AAGAAAAGAAAATGTGGGACCTT	CGGGCATGAAAATGTCAATC
rs64625218	44306171	TTAGAAAAGGAAGCGGGTCA	CAGCATTGAAAAGGAGATGGA
rs64542424	44669020	GGATAGGTCTAATCGTGGAGGA	AGCGTCGCTGGTAGTGGTAG
rs1348617	46124020	CCCACCCACTCTACCTCATC	ACTGCCATGAATGGAAGGTC
rs66070275	46166066	ATTGCATCAGTTCGCACAAG	ATTCAGTGGCCTGGTTCATC
rs65272077	46420580	ATGCTTGCGGTCTTTGTACC	AATATGCCTGAGCCGTTTTG
rs65052669	46915037	TGTGACTTGCACATCTCCATC	TGGATACTGGCACCTCAATG
rs13453157	47762262	CTGAAGGAACTCGTGGAGGA	GGCCACAGGGGTACAGAGT
rs66140753	51524725	TCTCAGTCTGGGACACCTCA	CAACACCCCACAAGGAGACT
rs63864648	51704903	AGAGCCCAGACTTCGTCCTT	GCAATGACAGGGGTCTCAGT
rs64230269	52093888	GCTTGCGTGCTGGTTTTACT	TTTCTCGTGAATGGGGAAAG
SNP2793378	53123731	GCTTTGAGCACTGATGCTTTC	TGCGTGTACAATCCCAACAT

[#]positions are on rat chromosome *7* using UCSC Genome Browser Nov.2004 (Baylor 3.4/rn4)

In conclusion, the *Mcs6* locus was fine mapped from a region of 33Mb to a region of 8.5Mb. Previously, the locus contained 111 annotated genes, while the fine mapped *Mcs6* locus contains 22 genes. The fine mapped *Mcs6* locus can be used to identify candidate genes by testing the differences in expression levels of the *Mcs6* genes between the two rat strains.

CHAPTER III

IDENTIFICATION OF MCS1B SEQUENCE VARIANTS

Introduction

The second part of this dissertation is focused on the rat mammary cancer susceptibility locus *Mcs1b.* The *Mcs1* locus was identified through a linkage analysis of the DMBA mammary carcinogenesis resistant COP rat strain and the susceptible WF rat strain [106]. The *Mcs1* locus contains three subloci that were identified using several WF.COP congenic lines. These subloci are *Mcs1a*, *Mcs1b* and *Mcs1c* [123]. The *Mcs1b* locus was further fine mapped using congenic animals. It maps to a region of 1.8Mb on rat chromosome *2* [87]. A map showing the congenic lines defining the *Mcs1b* locus is shown in Figure 6.

The *Mcs1b* locus has a human ortholog. The SNP *rs889312* was identified in a breast cancer genome-wide association study in 2007. The SNP has an OR of 1.13 (1.10-1.16), meaning that the minor allele of this SNP increases the chances of developing breast cancer [61]. *rs889312* is in linkage disequilibrium with at least six other SNPs, meaning that any of these SNPs could be the actual causative SNP. *rs889312* and the six SNPs it tags are located on a haplotype block that is 280kb in size. The haplotype block is shown in Figure 2. There are three transcripts located within the *rs889312* haplotype block. These are *MAP3K1*, *SETD9* and *MIER3*. However, *Setd9/ SETD9* is not expressed

in the rat mammary gland and human breast tissue [87]. There is the potential of more SNPs tagged by *rs889312* being identified through ongoing sequencing studies such as the 1000 Genomes Project. This means that the causative SNP may not be located in public databases yet. This makes studying the genetic elements controlling genetic susceptibility to breast cancer in the human difficult. It is easier to study the genetic elements in the rat, since only two inbred rat strains need to be sequenced to identify all the genetic variation that is present between the two rat strains in this region. Since there are multiple candidate SNPs in the human, it is likely that there will be multiple genetic variants present in the rat. Note, only SNPs are used in human breast cancer genome-wide association studies. It is possible that *rs889312* tags an INDEL that is the actual causative genetic element. Therefore, SNPs and INDELs will be identified between the two rat strains in the *Mcs1b* region. It is not possible to identify copy number variation (CNV) with the sequence capture technique that was used for this aim. Sequence capture depends on a microarray chip and complementary primers, which introduces bias when it comes to sequence frequency. Identification of copy-number variation depends on the analysis of sequence frequency across a region.

The goal for this aim is to identify all the genetic variants between the COP and WF rat strains in the *Mcs1b* region to generate a list of candidates. The hypothesis is that there are multiple genetic variants between the two rat strains in this region.

Methods

A. Sanger sequencing of rat orthologous region to *rs889312* risk SNP bin

Figure 2 shows the human haplotype block that associates with breast cancer risk. The *rs889312* correlated SNP are found in a SNP bin cluster within this haplotype block. The rat orthologous region to the *rs889312* risk SNP bin was sequenced in the WF and COP rat strain using standard Sanger sequencing. This region is on rat chr *2*: 43,168,806-43,185,894. For this, the splenic DNA from a homozygous WF and homozygous COP animal was used. The region to be sequenced was divided into 14 fragments and overlapping primers for each fragment were designed using Primer 3. DNA was amplified using Accuprime Taq (Life Technologies) and the resulting samples were run on a 1% agarose gel (Gene*Mate*) and stained with SybrGold (Life technologies). PCR products were purified using a QiaQuick PCR Purification Kit (Qiagen) or if multiple bands were present, the right size band was extracted using a QiaQuick Gel Purification Kit (Qiagen). PCR products were sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies). Sequenced products were cleaned by adding Agencourt AMPure XL beads and 80% ethanol. Beads were washed with 80% ethanol and DNA was eluted using molecular grade water. Sequences were submitted to the University of Louisville DNA Core for analysis. Sequences were analyzed using the DNASTAR Lasergene 8 SeqMan program.

B. Library preparation for sequence capture of the *Mcs1b* locus

The targeted region for sequencing the *Mcs1b* locus is on rat chromosome *2:* 42,200,000- 44,500,000. The DNA sequence for this targeted region was identified by using the UCSC Genome Browser DNA function. The Nov. 2004 (Baylor 3.4/rn4) assembly was used and this sequence is based on the Brown Norway (BN) rat strain. The

targeted DNA sequence was then submitted to NimbleGen. The sequence capture arrays were custom NimbleGen Sequence Capture Developer 385K Arrays. The custom NimbleGen sequence capture microarrays covered 88.9% of the targeted bases in the *Mcs1b* region. As a control, a human sequence capture microarray was used. The human practice array used was NimbleGen Sequence Capture Practice 385K Array. DNA libraries for the sequence capture were prepared using the GS FLX Titanium General Library Preparation Method Kit (Roche). DNA used was from a WF homozygous animal and a WF.COP line T congenic animal. DNA was extracted using the Blood and Tissue DNAeasy kit (Qiagen). Human genomic DNA (Bioline) was used for practice arrays. The libraries were prepared according to the library preparation manual. In short, the libraries were prepared by fragmenting 10μg of gDNA using a nebulizer and nitrogen gas for 1 minute. The fragmented DNA was run on a 1% low melting agarose gel (Lonza) and stained with SybrGold (Life technologies). Fragments between 800-500bps were gel extracted using a Gel Purification Kit (Qiagen). DNA quality was assessed using a DNA Bioanalyzer DNA 7500 LabChip (Agilent). DNA fragments were polished and adapters were ligated. DNA fragments were then immobilized on magnetic streptavidin-coated beads, via the biotin moiety of one of the adaptors. A fill- in reaction was performed to remove any nicks in the DNA. As a final step the library was melted off the beads, resulting in single stranded DNA. The quality and quantity of the ssDNA was determined using a Bioanalyzer RNA Pico 6000 LabChip (Agilent).

C. Sequence capture of *Mcs1b* libraries

The prepared libraries were hybridized to the microarrays using the Titanium Optimized Sequence Capture Array Delivery Kit according the manufacturer's instructions. In short, the pre-captured DNA was amplified using the GC-RICH PCR system dNTP pack (Roche) according to manufacturer's instructions. PCR primers used are shown in Table 6. PCR products were purified using the QiaQuick PCR Purification Kit (Qiagen). DNA quality and yield were measured using a Bioanalyzer DNA 7500 LabChip (Agilent). The amplified libraries were then annealed to the microarray sequence capture chips using the NimbleGen Hybridization System 4. The hybridization system was heated to 42°C for three hours prior to hybridization. 3µg human COT DNA (C_0t) was added to the amplified libraries to get rid of repetitive DNA. Hybridization enhancing oligos were added to the amplified DNA and the DNA was denatured at 95°C for 10 minutes. The DNA was loaded onto the microarray sequence capture chip and allowed to hybridize in the hybridization machine for 72 hours at 42°C. Microarrays were washed and captured DNA was eluted using 125mM NaOH. Captured DNA was purified using the Qiagen MinElute PCR Purification Kit. Samples were amplified using the same conditions as for the pre-captured LM- PCR. Samples were purified using the Qiagen QiaQuick PCR Purification Kit. Quantity and Quality of the captured DNA was determined using a Bioanalyzer DNA 7500 LabChip (Agilent) and NanoDrop 2000 spectrophotometer.

D. RT- QPCR on captured versus pre-captured samples to determine enrichment of libraries in *Mcs1b* DNA

Table 6. Primers used for *Mcs1b* sequence capture.

Name	Sequence	Step used for
LM-PCR 454 Ti-A Oligo	CCA TCT CAT CCC TGC GTG TC	Pre- and post- capture LM-PCR
LM-PCR 454 Ti-B Oligo	CCT ATC CCC TGT GTG CCT TG	Pre- and post- capture LM-PCR
Hybridization Enhancing 454 Ti-A	CCA TCT CAT CCC TGC GTG TCC CGA CTC	Hybridization step
Hybridization Enhancing 454 Ti-B	CCT ATC CCC TGT GTG CCT TGG CAG TCT	Hybridization step
QPCR oligo NSC-0237, forward	CGC ATT CCT CAT CCC AGT ATG	Confirmation of *Mcs1b* DNA enrichment
QPCR oligo NSC-0237, reverse	AAA GGA CTT GGT GCA GAG TTC AG	Confirmation of *Mcs1b* DNA enrichment
QPCR oligo NSC-0247, forward	CCC ACC GCC TTC GAC AT	Confirmation of *Mcs1b* DNA enrichment
QPCR oligo NSC-0247, reverse	CCT GCT TAC TGT GGG CTC TTG	Confirmation of *Mcs1b* DNA enrichment
QPCR oligo NSC-0268, forward	CTC GCT TAA CCA GAC TCA TCT ACT GT	Confirmation of *Mcs1b* DNA enrichment
QPCR oligo NSC-0268, reverse	ACT TGG CTC AGC TGT ATG AAG GT	Confirmation of *Mcs1b* DNA enrichment
QPCR oligo NSC-0272, forward	CAG CCC CAG CTC AGG TAC AG	Confirmation of *Mcs1b* DNA enrichment
QPCR oligo NSC-0272, reverse	ATG ATG CGA GTG CTG ATG ATG	Confirmation of *Mcs1b* DNA enrichment
A50 control inside 1 forward	CAC TCT CGG GTG AGA CAA CA	Confirmation of *Mcs1b* DNA enrichment
A50 control inside 1 reverse	TGT CGA AGG CAC AAA GAC TG	Confirmation of *Mcs1b* DNA enrichment
A50 control inside 2 forward	GGA GTG CTT TTC GGA CAG AG	Confirmation of *Mcs1b* DNA enrichment
A50 control inside 2 reverse	CGG AAG GAA AGC AAG AGT TG	Confirmation of *Mcs1b* DNA enrichment
A50 control inside 3 forward	ACA GTA CCC ACT GGC TGG TC	Confirmation of *Mcs1b* DNA enrichment
A50 control inside 3 reverse	GCT GGT TGC TCA TTC TAG GG	Confirmation of *Mcs1b* DNA enrichment
A50 control inside 4 forward	CAC ATG GCA CCA TCT CAA GT	Confirmation of *Mcs1b* DNA enrichment
A50 control inside 4 reverse	CCT GGT TGC CAC TAC AGT CA	Confirmation of *Mcs1b* DNA enrichment
A50 control outside 1 forward	GAG GCT GTT GAT GAT GCA GA	Confirmation of *Mcs1b* DNA enrichment
A50 control outside 1 reverse	TGG GAA GTG CCA TGT TGT AA	Confirmation of *Mcs1b* DNA enrichment
A50 control outside 2 forward	CAT GCT TCT GGT TTT GGT GA	Confirmation of *Mcs1b* DNA enrichment
A50 control outside 2 reverse	TGG AGA AAG CAT GAC CAC AG	Confirmation of *Mcs1b* DNA enrichment

To determine if the sequence capture was successful in enriching the post-captured libraries in *Mcs1b* DNA, a RT- QPCR reaction can be performed on the pre-captured versus the post- captured samples. For the QPCR reaction, SyrbGreen (Life Technologies) and 1ng of ssDNA was used. As a control, gDNA was used. Primers for the QPCR reaction are shown in Table 6. The RT- QPCR reactions were run on an ABI PRISM 7900HT Sequence Detection System. The Ct values for the post-captured samples were then subtracted from the Ct value of the pre-captured sample. A positive delta Ct value indicates enrichment.

E. 454 next-generation sequencing, assembly of *Mcs1b* genomes for WF and COP rat strains and identification of genetic variants

Mcs1b enriched WF and COP libraries were sent to the University of Kentucky for 454 next-generation sequencing. Two lanes of an 8- well plate were used for the WF library and 3 lanes were used for the COP library. The resulting sequences were aligned separately against the Rattus norvegicus (Brown- Norway) genome build 4.1 using SSAHA2 to generate BAM files. A pileup and BCF file were made for each position with a coverage of≥1 on rat chromosome *2*: 42,200,000-44,500,000. If the allele count of the non reference allele made up less than 25% of the total coverage, that record was flagged. An SQL query was used to select those positions from the database where either sample 1 was different from the reference and had a total coverage of 20 or greater OR sample 2 was different from the reference and had a total coverage of 15 or greater AND that consensus call that had a sufficient enough coverage to pass was not flagged. After a position was selected, the calls for all samples at that position were retrieved. If the call

was heterozygous and one of the alleles had a count of 2 or less, it was converted into a homozygous call of the stronger allele. If both alleles had a count of 2 or less, it was converted to a N/N.

F. Confirmation of *Mcs1b* genetic variants

All 454 next generation sequencing data was visualized using IGV 2.0 (http://www.broadinstitute.org/igv/v2.0). The data were filtered using an SQL query to identify SNPs between the two rat strains. Also, the data was filtered to identify any INDELs between the two rat strains with a coverage of ≥5 for each sample. Primers were designed for each SNP and INDEL to be confirmed using Primer 3. In total, primers were designed for 130 potential SNPs and 13 INDELs. The WF and line T (COP) DNA used for confirmation of genetic variants was pooled from three homozygous animals each. DNA was amplified using Accuprime Taq (Life Technologies) and the resulting samples were run on a 1% agarose gel (Gene*Mate*) and stained with SybrGold (Life technologies). PCR products were purified using an ExcelaPure 96-well UF PCR Purification kit (EdgeBio) or QiaQuick PCR Purification kit (Qiagen). PCR products were sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies). Sequenced products were cleaned by adding Agencourt AMPure XL beads and 80% ethanol. Beads were washed with 80% ethanol and DNA was eluted using molecular grade water. Sequences were submitted to the University of Louisville DNA Core for analysis. Sequences were analyzed using the DNASTAR Lasergene 8 SeqMan program.

G. Sequencing of gaps using Sanger sequencing

There are several gaps in the *Mcs1b* sequencing data for the WF and COP rat strain. Sequencing gaps in the rat orthologous region to the *rs889312* haplotype block were sequenced using standard Sanger sequencing (RNO2: 42,974,029-43,219,434). To identify gaps, the program IGV 2.0 was used to visualize the WF and COP assemblies. The rat orthologous region to the *rs889312* was scanned manually for gaps between the two rat strains and location of gaps were noted. Primers surrounding the gaps were designed using Primer3. The DNA from three homozygous WF and three homozygous WF.COP line T animals were used for sequencing. DNA was sequenced as described above.

H. Bioinformatic analysis of *A46-SNP-A*, *A074-SNP-17* and *A074-SNP-18*

The program Alternative Splice Site Predictor (http://wangcomputing.com/assp/index.html) was used to determine changes in splicing sites for the *Map3k1* gene. The entire *Map3k1* sequence was added to the program (RNO2: 43,062,252-43,102,501) and differences in the predicted splicing sites when changing the SNP alleles were determined.

The program Variant Visualizer from the Rat Genome Database (http://rgd.mcw.edu/rgdweb/front/select.html) was used to identify the *Mcs1b* candidate SNP alleles in different rat strains. The program uses the RGSC Genome Assembly 3.4. The region of interest was entered into the program and all available rat strains were selected for analysis. Additionally, genotyping data for the WF/Nhsd and COP/NHsd rat strains were added since the program does not include the WF rat strain.

The UCSC Genome Browser "in other genomes (convert)" function was used to identify the human orthologous region to the identified sequence variants between the two rat strains. When there was no orthologous region identified, the LAGAN alignment tool with the settings VISTA for CNS - window: 30 bp, min width: 30bp, CNS identity: 60%, Min Y%: 50 was used. The UCSC Genome Browser uses a large window size to determine orthology resulting in no orthologous regions identified.

The human orthologous region for *A102-INDEL-2* was determined using the UCSC Genome Browser "in other genomes (convert)" function. Since the INDEL does not have a human orthologous region using the genome browser, additional surrounding sequence was added until an orthologous region was found. DNA was then downloaded from the genomes browser for the human orthologous region with an additional ±50kb of surrounding sequence. This was pasted into Microsoft Word and manually analyzed for repetitive elements.

Results

A. Identification of *A046-SNP-A* through Sanger sequencing of the WF and COP rat strain

We initially sequenced the rat orthologous region to the human *rs889312* SNP bin to identify any sequence variants between the WF and COP rat strain. The rat orthologous region to the *rs889312* SNP bin is located on rat chromosome *2*: 43,168,806-43,185,894. This 17kb region was divided into 14 fragments and overlapping primers for the 14 fragments were designed. Out of the 14 fragments, 11 were fully sequenced in both the WF and COP rat strain using Sanger sequencing. No PCR could be generated for fragments that did not yield sequence. Sequencing resulted in the identification of one

SNP named *A046-SNP-A*. This is located on rat chromosome *2:* 43,175,144. The WF and Brown-Norway (BN) reference share the same allele, while the COP allele is different.

B. Confirmation of *Mcs1b* enrichment after sequence capture

Sequencing of the *rs889312* SNP bin resulted in the identification of *A046-SNP-A*. It is possible that there are more genetic variants that are located outside the orthologous region to the *rs889312* SNP bin that are involved in the *Mcs1b* phenotype. Therefore, it is necessary to sequence the entire *Mcs1b* region. NimbleGen sequence capture arrays were used to sequence the 1.8Mb defining *Mcs1b* region, since it is not feasible to sequence the entire *Mcs1b* region using Sanger sequencing. This will result in the identification of every sequence variant between the two rat strains. The NimbleGen Sequence Capture mircoarrays can be used to enrich a DNA library in a targeted region. Compared to whole genome sequencing, sequence capture allows for less sequencing reactions to get a good coverage across the targeted region. A RT-QPCR reaction can be performed on the pre-captured versus the post-captured sample to ensure that a DNA library has been enriched in the targeted region after sequence capture. The idea is that the post-captured sample should have a lower Ct value than the pre-captured sample. The post-captured sample is then subtracted from the pre-captured sample to get a delta Ct value. The delta Ct value should be positive if there is enrichment of the DNA library in the targeted region. Primers used for the QPCR reactions are shown in Table 6.

The results of the RT- QPCR reactions are shown in Table 7. The NCS primers are recommended by NimbleGen. These primers are complementary to sequences that are found on every NimbleGen Sequence Capture array and contain sequences that are universal among different species. Several primer pairs used were specific to the *Mcs1b*

Table 7. RT-QPCR results for enrichment in targeted regions of sequence capture libraries. NCS primers are used for every NimbleGen Sequence Capture array and are universal for different species. A50 primers are specific to the *Mcs1b* rat region and can therefore not be used on human control DNA.

	Delta Ct value for sample		
Primer	**Human**	**WF (susceptible)**	**COP (resistant)**
NCS 0237	10.0	4.9	4.6
NCS 0247	9.3	4.3	5.2
NCS 0268	10.8	6.9	6.4
NCS 0272	10.1	6.3	5.2
A50 control inside 1		11.2	10.5
A50 control inside 2		10.2	10.0
A50 control inside 3		10.4	9.7
A50 control inside 4		7.0	8.4
A50 control outside 1		-1.7	-2.7
A50 control outside 2		-0.6	-0.9

targeted region. Two primer pairs were located outside the *Mcs1b* locus and four pairs were located inside the targeted region. The delta Ct values are positive for all samples when using the NCS primers indicating that these targeted regions were enriched. NCS regions are included on all sequence capture microarrays. However, the Ct values for the WF and COP libraries are lower than for the human library. This could indicate that the human library is of better quality. However, all three libraries were prepared at the same time and should be of equal quality. NimbleGen recommends the NCS primers as being universal for different species. It is possible that there is a difference in the affinity of the primers to human versus rat DNA, resulting in lower delta Ct values.

Primers that are located within the targeted *Mcs1b* region (A50 control inside primers) show a positive delta Ct value for both the WF and COP rat strains, indicating that the libraries were successfully enriched in the *Mcs1b* DNA. As a control, primers outside of the targeted region (A50 control outside primers) show no enrichment, indicated by a negative delta Ct value.

C. Identification of *Mcs1b* genetic variants between the WF and COP rat strain

The WF and COP libraries were sequenced using 454 next- generation sequencing. The goal was to get a coverage of at least 15X for both assemblies. The coverage for the WF rat strain was 20.0X and for the COP rat strain 15.7X. The average read length for the WF sequences was 328bp and for the COP sequences 317. Overall, the sequencing resulted in coverage for 91.7% of all the bases found in the targeted region. The WF and COP assemblies were filtered for sequence variants between the two rat strains and primers were designed to confirm potential sequence variants using Sanger

sequencing. Out of 130 potential SNPs, 67 SNPs were confirmed between the two rat strains. A list of the identified SNPs is shown in Table 8. No PCR product could be generated for nine potential *Mcs1b* SNPs and therefore these SNPs could not be confirmed. A list of potential *Mcs1b* SNPs that could not be confirmed is found in Table 9. Out of 13 INDELs that were tested between the two rat strains, two were confirmed using Sanger sequencing. A transcript map for the *Mcs1b* locus including the sequence variation in the locus is shown in Figure 11. Most of the sequence variation between the two rat strains is located at the extreme ends of the *Mcs1b* locus. WF.COP congenic animals that have a susceptible phenotype and do not contain the *Mcs1b* locus were genotyped using the new markers. These congenic lines extend to the ends of the markers located at the extreme ends of the *Mcs1b* locus. This means that all these sequence variants are ruled out as candidate causative sequence variants for the *Mcs1b* locus. The positions for the WF.COP susceptible congenic lines are shown in Figure 11. This leaves four potential sequence variants: *A046-SNP-A*, *A074-SNP-17*, *A074-SNP-18* and *A102-INDEL-2*. These potential genetic variants are marked in Figure 11 and bolded in Table 8. *A046-SNP-A*, *A074-SNP-17* and *A074-SNP-18* are located within the rat orthologous region to the human haplotype block containing the *rs889312* SNP bin. *A102-INDEL-2* is upstream of the *Gpbp1* gene.

D. Filling in gaps in the *Mcs1b* sequence using Sanger sequencing

The *Mcs1b* sequence capture assemblies for the WF and COP rat strain contain gaps. These gaps are due to three different issues: 1) there are gaps in the Brown Norway rat reference sequence used to design the sequence capture microarrays. 2) There are gaps

Table 8. SNPs and INDELs between the WF and COP rat strains in the *Mcs1b* region. The SNPs are organized according to genomic location. The table also includes *A046-SNP-A*. INDELs are organized according to genomic location. Bolded variants are *Mcs1b* candidate sequence variants.

Name	Location[#]	Reference (BN) allele	WF allele	COP allele
A074-SNP-1	42364375	A	G	A
A074-SNP-65	42364706	C	C	T
A074-SNP-2	42365768	G	A	G
A074-SNP-3	42366362	T	G	T
A074-SNP-4	42367078	A	T	A
A074-SNP-5	42367311	C	G	C
A074-SNP-6	42367553	G	A	G
A074-SNP-66	42374226	A	G	A
A074-SNP-7	42374481	T	C	T
A074-SNP-64	42375050	C	T	C
A074-SNP-8	42375207	A	T	A
A074-SNP-9	42375271	A	T	A
A074-SNP-10	42375321	C	T	C
A074-SNP-67	42375547	G	A	G
A074-SNP-11	42376800	A	G	A
A074-SNP-12	42377178	C	T	C
A074-SNP-13	42378143	C	T	C
A074-SNP-14	42378758	C	T	C
A074-SNP-15	42378793	C	T	C
A074-SNP-16	42378804	C	T	C
A074-SNP-17	43071787	C	C	A
A074-SNP-18	43090006	C	T	C
A046-SNP-A	43175144	C	C	T
A074-SNP-19	44147176	C	A	C
A074-SNP-20	44148139	G	A	G
A074-SNP-21	44149528	G	T	G
A074-SNP-22	44150206	T	T	G
A074-SNP-23	44152022	A	G	A
A074-SNP-24	44152126	G	A	G
A074-SNP-25	44152673	A	A	G
A074-SNP-26	44152745	A	G	A
A074-SNP-27	44152882	T	G	T
A074-SNP-28	44152995	T	C	T
A074-SNP-29	44153181	A	T	A
A074-SNP-30	44153277	G	A	G
A074-SNP-31	44153858	G	A	G
A074-SNP-32	44156135	G	G	A
A074-SNP-33	44157354	T	T	G
A074-SNP-34	44157459	C	T	C
A074-SNP-35	44157568	C	C	T
A074-SNP-36	44158255	G	C	G
A074-SNP-37	44158391	G	A	G
A074-SNP-38	44158578	G	G	A
A074-SNP-39	44158871	G	G	T
A074-SNP-40	44158875	C	C	T

Table 8 continued.				
A074-SNP-41	44161442	G	G	A
A074-SNP-42	44165615	A	T	A
A074-SNP-43	44167496	A	T	A
A074-SNP-44	44167791	G	C	G
A074-SNP-45	44170531	A	A	G
A074-SNP-46	44170619	T	T	C
A074-SNP-47	44170980	C	T	C
A074-SNP-48	44171160	T	C	T
A074-SNP-49	44172139	A	G	A
A074-SNP-50	44192478	T	T	G
A074-SNP-51	44193102	A	G	A
A074-SNP-52	44196643	T	C	T
A074-SNP-53	44199441	G	A	G
A074-SNP-54	44201507	C	G	C
A074-SNP-55	44201538	A	T	A
A074-SNP-56	44205934	G	A	G
A074-SNP-57	44206713	G	A	G
A074-SNP-58	44206843	G	A	G
A074-SNP-59	44207194	G	A	G
A074-SNP-60	44207705	G	A	G
A074-SNP-61	44208687	G	A	G
A074-SNP-62	44209120	A	G	A
A074-SNP-63	44209260	C	C	A
A102-INDEL 1	42378873	TTG		TTG
A102-INDEL 2	42709213	TAGA	TAGA	

[#]rat chromosome *2*. UCSC Genome Browser Nov. 2004 (Baylor 3.4/rn4)

Table 9. Potential *Mcs1b* SNPs that cannot be confirmed using Sanger sequencing.

Position	Reference (BN) allele	WF allele	COP allele	WF coverage	COP coverage
43,775,309	G	A/G	A/G	16	10
44,108,905	T	C/T	C/T	17	10
44,108,908	T	G/T	G/T	17	10
44,108,958	G	A/C/G	A/G	19	8
44,108,967	A	A/T	A/T	28	13
44,138,671	T	A/T	A/T	7	11
44,138,675	T	C/T	C/T	8	11
42,826,948	T	C/T	T	5	1
42,672,597	T	T	C	24	1

[#]rat chromosome *2*. UCSC Genome Browser Nov. 2004 (Baylor 3.4/rn4)

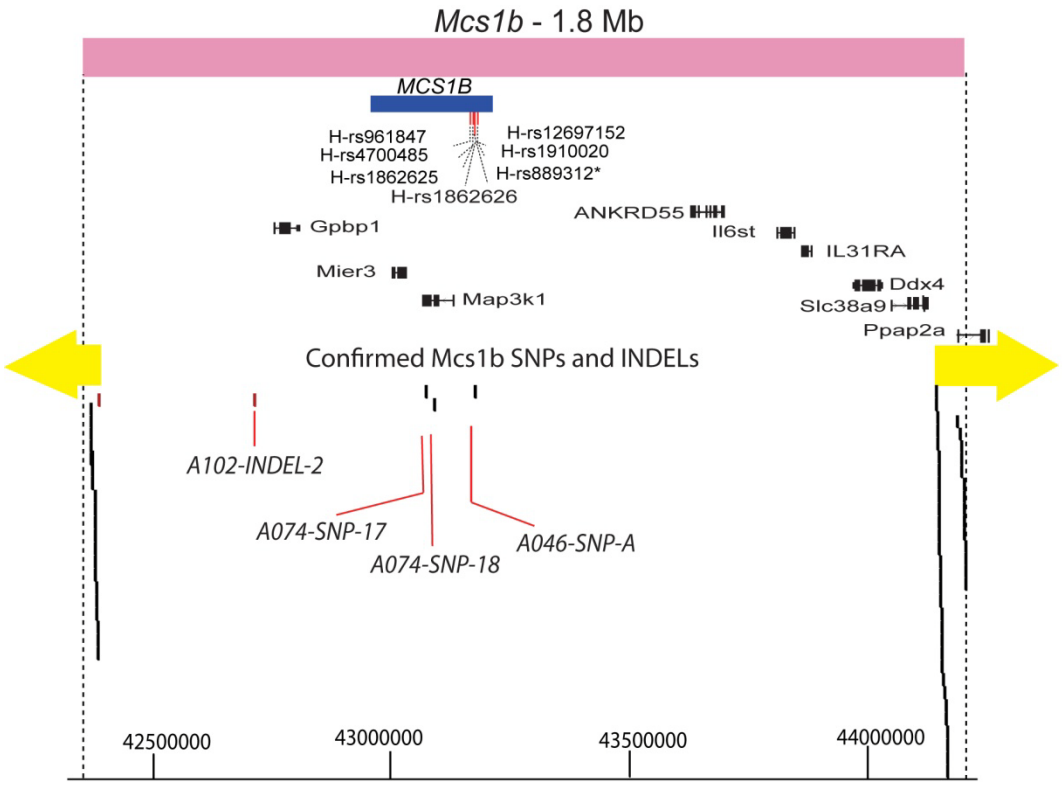


Figure 11. Transcript map for the *Mcs1b* locus showing all genetic variation between the two rat strains. Pink bar indicated the length of the *Mcs1b* locus. Blue bar indicates the rat orthologous region to the human haplotype block containing *rs889312* tagged SNPs. Yellow bars indicate WF.COP congenic lines that have a susceptible phenotype and do not contain the *Mcs1b* locus. SNPs are shown in black, while INDELs are shown in red. *Mcs1b* transcripts are shown in black. Also shown are transcripts located within this region.

due to repetitive elements in the sequence. When designing the sequence capture arrays, NimbleGen will remove all repetitive elements and these regions will not be included on the arrays. 3) There are gaps due to some regions having a low coverage either in one of the rat assemblies alone or in both. We focused on filling sequencing gaps in the rat orthologous region to the *rs889312* haplotype block. There are 75 gaps in this region. The size of the gaps varies from 1-1200bps. Standard Sanger sequencing was used to fill in these gaps. Out of the 75 gaps, 45 (60%) were fully sequenced between the WF and COP rat strain. There were no additional sequence variants found between the two rat strains. Three gaps were only partially sequenced, meaning that the entire gap could not be sequenced. Out of the 75 gaps, 27 (36%) were not sequenced at all. The 75 gaps make up a total of 17,008bps. The gaps that were not sequenced make up 10,883bps or 64% of the gap bases that we attempted to sequence.

E. Bioinformatic analysis of *Mcs1b* sequence variants

A074-SNP-17 and *A074-SNP-18* are located within introns of the *Map3k1* gene. *A074-SNP-17* is located in intron 11 and *A074-SNP18* is located in intron 2 of *Map3k1*. It is possible that one or both of these SNPs result in a change in the splicing of the *Map3k1* gene. To test this, the program Alternative Splice Site Predictor (http://wangcomputing.com/assp/index.html) was used. This program scans sequences for consensus splicing sites. There are no changes in the splicing pattern between the WF and COP *A074-SNP-17* and *A074-SNP-18* alleles. This indicates that *A074-SNP-17* and *A074-SNP-18* do not change splicing of the *Map3k1* gene. Previous QPCR experiments in the lab also revealed no additional *Map3k1* splicing variants.

To determine if the rat alleles for the *Mcs1b* candidate SNPs are common among different rat strains, the program Variant Visualizer from the rat genome database was used. This program allows for the identification of the sequence for all known SNPs among different rat strains. In total, there is information for 19 rat strains in this program. We added the genotyping information for the COP/NHsd and WF/NHsd rat strains, since these are not found using the program. Results from the analysis can be found in Table 10. The variant allele for *A046-SNP-A* is found both in the COP and ACI rat strains. Since the COP and ACI rat strain differ in their susceptibility, it may be unlikely that *A046-SNP-A* is an independently acting variant for the *Mcs1b* phenotype. However, genetic susceptibility to mammary cancer is a complex trait and not all information about the genetic make-up of these two rat strains is known. The variant allele for *A074-SNP-18* is found in the DMBA carcinogenesis susceptible WF and SS (Salt-Sensitive) rat strains. Since the *Mcs1b* phenotype is due to the presence of the COP allele in the *Mcs1b* region, it is likely that the variant allele is found in the COP rat strain. Therefore, *A074-SNP-18* is not likely to be an independently acting variant for the *Mcs1b* phenotype. However, *A074-SNP-17* is a rare polymorphism. The variant allele of *A074-SNP-17* is found only in the COP or resistant rat strain. Since the causative *Mcs1b* variant is likely to be found in the COP rat strain and not common to rat strains with susceptible DMBA phenotypes, *A074-SNP-17* is a strong candidate for the *Mcs1b* phenotype conferred. It is possible, however, that all three or a combination of the *Mcs1b* candidate SNPs are responsible for the *Mcs1b* phenotype.

The sequencing of the *Mcs1b* region between the two rat strains resulted in the identification of one INDEL called *A102-INDEL-2*. This INDEL is located outside of the

Table 10. Sequence results for candidate rat SNPs in different rat strains using Variant Visualizer. The Rat Genome Database program Variant Visualizer was used. Highlighted sequences show variant allele.

Sensitivity to DMBA carcinogenesis	**Rat strain**	***A074-SNP-17***	***A074-SNP-18***	***A046-SNP-A***
Resistant	BN/NHsdMcwi	C	C	C
	BN/SsN	C	C	C
	COP/Crl	A	C	T
	COP/NHsd	A	C	T
	SHR/Olalpcv	C	C	C
	Wky/N	C	C	C
Intermediate	ACI/Eur	C	C	T
	ACI/N	C	C	T
	F344/N	C	C	C
	Le/Stm	C	C	C
Susceptible	Buf/N	C	C	C
	SS/JrHsdMcwi	C	T	C
	WF/NHsd	C	T	C
Unknown	BN-Lx	C	C	C
	FHH/EurMcwi	C	C	C
	FHL/EurMcwi	C	C	C
	GH/OmrMcswi	C	C	C
	M520/N	C	C	C
	MR/N	C	C	C
	SHRSP/Gcrc	C	C	C
	SR/JrHsd	C	C	C

rat orthologous region to the *rs889312* haplotype block. *A102-INDEL-2* is located within a stretch of repetitive sequence. The INDEL repeat and surrounding sequences for *A102-INDEL-2* are found in Table 11. The WF allele has one additional TAGA repeat that is shown in Table 11. The three candidate *Mcs1b* SNPs have a human ortholog with the *rs889312* SNP bin. To identify any human orthologous INDELs, we manually scanned the human orthologous region to *A102-INDEL-2* and an additional ±50kb surrounding sequence for repetitive regions. To identify the immediate orthologous region for the INDEL, we had to go 500bp out on either side of the INDEL until a region that is conserved between humans and rats was found. No repetitive sequence was found in this region, suggesting that there is no human ortholog to *A102-INDEL-2* or it is located outside of the queried region.

Discussion

Sequence capture and 454 next-generation sequencing has proven to be a fruitful technique for the *Mcs1b* locus. The *Mcs1b* locus was initially mapped using WF.COP congenic lines to be 1.8Mb in size [87]. Using sequence capture, 69 variants between the two rat trains were discovered. *A046-SNP-A* was identified using standard Sanger sequencing. Another two SNPs in the *Mcs1b* region were previously known. These are A12-oo and A12-v. There are a total of 72 variants between the two rat strains in the *Mcs1b* region. There are only two confirmed INDELs and the rest of the variants are SNPs. Most of these variants are located at the extreme ends of the *Mcs1b* locus. The list of potential candidate rat variants was further narrowed using WF.COP congenic lines for the *Mcs1b* locus. Genotyping of susceptible congenic *Mcs1b* lines indicated that

Table 11. Sequences for WF and COP alleles of *A102-INDEL-2* and surrounding sequence. Red script indicates sequence variation between the two rat strains.

Strain	Sequence
WF	ATCCAGGGCAGATAGATGATAGATAGATAGATAGATAGATAGATAGATAG ATAGATAGATAGATAGACAGACAGACAGACAGACAGACAGACAGACATAG GAAAGAAGAGTGAGG
COP	ATCCAGGGCAGATAGATGATAGATAGATAGATAGATAGATAGATAGATAG ATAGATAGATAGA--- CAGACAGACAGACAGACAGACAGACAGACATAGGAAAGAAGAGTGAGG

all of the sequence variants at the extreme ends of the *Mcs1b* locus are ruled out as the causative variants. That leaves four candidate *Mcs1b* sequence variants.

The first candidate *Mcs1b* variant is an INDEL called *A102-INDEL-2*. However, analysis of the human orthologous region surrounding *A102-INDEL-2* has not resulted in the identification of a human ortholog. This could be due to several reasons: 1) *A102-INDEL-2* might not have a human ortholog, since it has no functional activity and there is no evolutionary pressure to retain this region, 2) the human ortholog to *A102-INDEL-2* may be located outside of the region tested. When looking for the human orthologs, we restricted the region of interest to 100kb surrounding the human orthologous region to *A102-INDEL-2*. Extending the search to a larger region may result in the identification of an ortholog. Because there is no known human ortholog for *A102-INDEL-2,* the functional analysis of the *Mcs1b* variants will be focused on the candidate SNPs that were identified. However, WF.COP congenic lines that will test the involvement of *A102-INDEL-2* in the *Mcs1b* phenotype are currently being generated.

Three candidate rat SNPs were identified in the *Mcs1b* region. These are *A074-SNP-17*, *A074-SNP-18* and *A046-SNP-A*. All three of these SNPs are located within the rat orthologous region to a human haplotype block that associates with breast cancer risk. The human haplotype block is marked by the SNP *rs889312*, which is in linkage disequilibrium with at least six other SNPs. We hypothesized that since there are multiple breast cancer associated SNPs in the human *MCS1B* region, there are multiple candidate SNPs in the rat *Mcs1b* region. This hypothesis is confirmed with the identification of the three *Mcs1b* candidate SNPs. Functional analysis of *Mcs1b* variants will focus on these three SNPs, since they have a human ortholog in the *rs889312* correlated SNPs. An

analysis of the frequency of the SNP alleles in different rat strains revealed that the variant allele of *A046-SNP-A* is found in both DMBA carcinogenesis resistant and susceptible animals. This suggests that this SNP may not be independently acting on the *Mcs1b* phenotype. The variant allele for *A074-SNP-18* is found in two different susceptible rat strains. The *Mcs1b* causative variant is likely found to be present in the COP rat strain and possibly other DMBA carcinogenesis resistant rat strains. This is due to the fact that the COP allele in the *Mcs1b* region modifies the *Mcs1b* DMBA carcinogenesis phenotype. This would suggest that the variant allele for *A074-SNP-18* is not an independently acting functional genetic element. However, the variant allele for *A074-SNP-17* is only found in the COP rat strain, making this an ideal candidate for the *Mcs1b* mammary cancer resistance. It is possible that all three or a combination of the *Mcs1b* candidate SNPs are involved in modulating mammary cancer susceptibility. Therefore, all three candidate SNPs will be included in a functional analysis of the *Mcs1b* candidate SNPs.

A074-SNP-17 and *A074-SNP-18* are located within two different introns of the gene *Map3k1*. However, analysis of the predicted splice sites for the *Map3k1* gene revealed that the SNPs are not predicted to alter splicing sites. *A046-SNP-A* is located within an intergenic region downstream of the *Map3k1* gene. It is possible that any one of these SNPs or a combination of them are involved in gene regulation of *Mcs1b* genes by being located in a regulatory element.

There are gaps in the *Mcs1b* sequencing data. These gaps result from gaps in the rat reference genome used to make the sequence capture microarrays, from highly repetitive regions, which were not included on the microarray and from low coverage of

the sequencing data. To fill in some of these gaps, we focused on the rat orthologous region to the *rs889312* haplotype block. Out of 75 gaps in this region, we were unable to sequence 27 gaps. This means that we failed to sequence over 10,000 bases that we attempted to sequence in this region. Reasons why sequencing failed were problems with primer design in the regions and problems with getting PCR product in highly repetitive regions. Since there are over 10,000 bases unsequenced in this region, it is possible that there are more genetic variants. It is not possible to sequence these bases with the current techniques available; however, future sequencing techniques may be able to sequence across repetitive regions and can be used to sequence the last remaining bases in this region.

CHAPTER IV

FUNCTIONAL ANALYSIS OF MCS1B GENETIC VARIANTS

Introduction

Sequencing of the *Mcs1b* region resulted in the identification of four candidate *Mcs1b* sequence variants. These are *A102-INDEL-2*, *A046-SNP-46*, *A074-SNP-17* and *A074-SNP-18*. The *Mcs1b* SNPs *A046-SNP-A*, *A074-SNP-17* and *A074-SNP-18* are located within the rat orthologous region to a human haplotype block that is associated with breast cancer risk. These three rat SNPs are our candidate *Mcs1b* rat SNPs. The human haplotype block that is associated with breast cancer risk is shown in Figure 2 and is marked by the SNP *rs889312* [61]. *rs889312* is in linkage disequilibrium with at least six other SNPs as shown in Figure 2. This means that any one or a combination of these SNPs could be the causative one. However, since there are ongoing sequencing experiments such as the 1000 Genomes Project that are identifying new SNPs in the human, it is possible that the causative SNP is not in public databases yet. For this reason, the functional analysis of candidate sequence variants for this aim will mainly focus on the three identified rat SNPs, since all sequence variation between the two rat strains in the *Mcs1b* region is known. Note, the three *Mcs1b* rat SNPs have a human ortholog in the *rs889312* correlated SNPs. There is no known human ortholog for *A102-INDEL-2*. A manual search for regions of high repetitive sequences in the human orthologous region to the *A012-INDEL-2 r*egion did not yield any orthologs.

Therefore, the functional analysis of *Mcs1b* sequence variants will focus on the three candidate rat SNPs identified.

Genetic variants can be located in exons or introns of genes or they can be located in intergenic regions. SNPs can influence a given phenotype through several mechanisms. A SNP that is located within the coding region of a gene could have no effect on t he transcribed protein, result in amino acid substitutions or in truncated proteins. However, none of the *Mcs1b* candidate SNPs are located within the protein-coding region of a *Mcs1b* gene and are therefore not involved in changing the amino acid sequence of a protein.

Some SNPs are located within gene introns. These SNPs can change the splicing pattern of a gene, potentially resulting in a new splice product of a gene. Both *A074-SNP-17* and *A074-SNP-18* are located within introns of *Map3k1*. However, an extensive analysis of potential splicing site changes revealed that neither the COP nor WF allele for both SNPs result in splicing site changes for *Map3k1*. Also there is no additional splice variants annotated in the USCS Genome Browser for *Map3k1*. Furthermore, no splice variants were identified using a QPCR based approcach. Therefore, it is unlikely that these two SNPs are involved in changing the *Map3k1* splicing pattern.

SNPs found in intergenic and intronic regions can be located in enhancer/ repressor and promoter regions. These genetic elements are all involved in regulating gene expression. Proximal promoters are typically located within 1kb of the transcription start site (TSS) and are involved in the basic transcriptional regulation of a particular gene [135]. None of the *Mcs1b* SNPs are located in close proximity to the TSS of any of the *Mcs1b* genes and are therefore not likely to be located in gene promoters. Enhancers

and repressors are genomic regions that are not in close proximity to transcription start sites but can either activate or repress gene transcription through long- range interactions. Enhancers/ repressors bind transcription factors that act to regulate gene expression. Enhancers/ repressors act independently of their location and orientation to a gene promoter. These gene regulatory elements can be over 1Mb away from the targeted gene. The formation of chromatin loops is thought to bring enhancers/ repressors in proximity of gene promoters. Enhancers/ repressors are often found within introns of genes they regulate or within introns of neighboring genes. However, they can also be found within intergenic regions [136, 137]. All three *Mcs1b* SNPs are located either in introns or intergenic regions and could be located within enhancers/ repressors.

Several genes within the *Mcs1b* region are expressed differentially between the WF and COP rat strain. Mammary gland transcript levels of *Mcs1b* genes were compared between 12- week old WF and WF.COP line N3 congenic females. There was a significant expression difference for genes *Gpbp1*, *Map3k1*, *Mier3* and *Il6st* between the two rat strains. When animals were treated with DMBA, a significant expression difference for *Mier3* between the two rat strains was observed [87]. Since there is a difference in the regulation of *Mcs1b* genes between the two rat strains, it is possible that the *Mcs1b* SNPs are involved in regulating the *Mcs1b* genes. Because of their location in respect to the *Mcs1b* genes, it is likely that the *Mcs1b* SNPs are located within enhancer/ repressor regions. Because of this, any functional analysis of the three *Mcs1b* candidate SNPs will focus on potential roles in gene regulation.

The goal of this aim is to perform functional analysis of the *Mcs1b* candidate SNPs in regards to their function in regulation of *Mcs1b* genes. The hypothesis is that the

Mcs1b candidate SNPs are located in enhancer/ repressor regions and regulate *Mcs1b* gene expression levels. The main functional analysis will focus on the rat *Mcs1b* candidate SNPs, since all candidate rat variants in this region are known. However, any positive results for the rat *Mcs1b* SNPs will be tested using the human *rs889312* correlated SNPs as well.

Methods

A. Cloning of candidate *Mcs1b* SNPs and human *rs889312* correlated SNPs into pGL3- Promoter vector

Luciferase assay constructs were designed for the WF and COP *A074-SNP-17*, *A074-SNP-18* and *A046-SNP-A* alleles and the human major and minor *rs889312*, *rs1862625*, *rs1862626*, *rs12697152*, *rs1910020*, *rs4700485* and *rs961847* alleles. The rat SNP alleles were inverted, since the rat *Mcs1b* region is inverted with respect to the human *MCS1B* region. The constructs included the SNP nucleotide and 12bps flanking on either side for a total of 25bps. These 25bps were repeated in tandem five times to increase the signal strength of the luciferase assay. The constructs also included sticky *Kpn1* and *Xho1* restriction sites for cloning purposes. The constructs were generated by Integrated DNA Technologies (IDT) and included *Kpn1* and *XhoI* compatible overhangs. The pGL3- Promoter firefly luciferase reporter vector was digested with *Kpn1* and *Xho1* to generate sticky ends. The constructs were cloned into the multiple- cloning site upstream of the SV40 promoter. The pGL3- Promoter vector was selected because it is designed to test potential enhancer elements. Constructs were annealed at a concentration of 1pmol/µl using a 10mM Tris, 1mM EDTA and 50mM NaCl (pH 8.0) buffer. A Veriti

96-well thermocycler (Life Technologies) was used to anneal the constructs by heating them to 95°C and allowing them to cool down at a rate of 1°C/min for 70 minutes. The annealed constructs were then ligated to the linearized pGL3-Promoter vector using a 3:1 insert: vector ratio and T4 DNA ligase (Promega). Top10 Chemically Competent Cells (Life Technologies) were used in the transduction reaction according to manufacture's protocol. A screen for positive clones was performed using RV3 primers surrounding the multiple cloning site and a FastPCR reaction was performed as previously described. The PCR products were run on a 1% agarose gel and stained using SybrGold (Life Technologies). As a control the pGL3-Promoter vector was used. Clones were extracted using a Spin Miniprep kit (Qiagen) and sequenced to ensure the correct insert has been ligated. Finally, the plasmids were extracted using a PureYield Plasmid Midiprep Kit (Promega) and the entire plasmid was sequenced to ensure no sequence errors. The sequences for the constructs and RV3 primers can be found in Table 12.

B. Cloning of multiple *Mcs1b* rat SNP alleles into the same pGL3-Promoter vector

The rat alleles for *A074-SNP-17* and *A074-SNP-18* were cloned into the multiple cloning site upstream of the SV40 promoter in the pGL3-Promoter vector. The rat alleles for *A046-SNP-A* were cloned into the multiple cloning site downstream of the luciferase gene. Note, *A074-SNP-18* was placed upstream of *A074-SNP17*. Constructing the *A074-SNP-18/A074-SNP-17* constructs was complex. IDT cannot manufacture constructs that are longer than 200bps. The *A074-SNP-18/A074-SNP-17* constructs were each at least 250bps in length. Therefore, the *A074-SNP-18* and *A074-SNP-17* constructs were designed separately with an internal *BglII* site. The *A074-SNP-18* construct contained a

Table 12. Sequences for constructs and RV3 primers for cloning WF and COP *A074-SNP-17*, *A074-SNP-18* and *A046-SNP-A* alleles and human major and minor *rs889312*, *rs1862625*, *rs1862626*, *rs12697152*, *rs1910020*, *rs4700485* and *rs961847* alleles. Nucleotides in red are SNP

Construct/ Primer ID	Sequence
A63-*A046-SNP-A* WF (+)	5'CGAACTTCTACAGGACCTGTGCATTCGAACTTCTACAGGACCTGTGCATTCGAACTTCTACAGG ACCTGTGCATTCGAACTTCTACAGGACCTGTGCATTCGAACTTCTACAGGACCTGTGCATTCC'3
A63-*A046-SNP-A* WF (-)	5'TCGAGGAATGCACAGGTCCTGTAGAAGTTCGAATGCACAGGTCCTGTAGAAGTTCGAATGCACAGGTCCT GTAGAAGTTCGAATGCACAGGTCCTGTAGAAGTTCGAATGCACAGGTCCTGTAGAAGTTCGGTAC'3
A63-*A046-SNP-A* COP (+)	5'CGAACTTCTACAGAACCTGTGCATTCGAACTTCTACAGAACCTGTGCATTCGAACTTCTACAGA ACCTGTGCATTCGAACTTCTACAGAACCTGTGCATTCGAACTTCTACAGAACCTGTGCATTCC'3
A63-*A046-SNP-A* COP (-)	5'TCGAGGAATGCACAGGTTCTGTAGAAGTTCGAATGCACAGGTTCTGTAGAAGTTCGAATGCACAGGTT CTGTAGAAGTTCGAATGCACAGGTTCTGTAGAAGTTCGAATGCACAGGTTCTGTAGAAGTTCGGTAC'3
A91-*A074-SNP-17* WF (+)	5'CATTCCATGACATGTGCAGCTCATCAATTCCATGACATGTGCAGCTCATCAATTCCATGACATG TGCAGCTCATCAATTCCATGACATGTGCAGCTCATCAATTCCATGACATGTGCAGCTCATCAC'3
A91-*A074-SNP-17* WF (-)	5'TCGAGTGATGAGCTGCACATGTCATGGAATTGATGAGCTGCACATGTCATGGAATTGATGAGCTGCAC ATGTCATGGAATTGATGAGCTGCACATGTCATGGAATTGATGAGCTGCACATGTCATGGAATGGTAC'3
A91-*A074-SNP-17* COP (+)	5'CATTCCATGACATTTGCAGCTCATCAATTCCATGACATTTGCAGCTCATCAATTCCATGACATT TGCAGCTCATCA ATTCCATGACATTTGCAGCTCATCAATTCCATGACATTTGCAGCTCATCAC'3
A91-*A074-SNP-17* COP (-)	5'TCGAGGTTTTGGAGATACGTGCACATGGGCGTTTTGGAGATACGTGCACATGGGCGTTTTGGAGATAC GTGCACATGGGCGTTTTGGAGATACGTGCACATGGGCGTTTTGGAGATACGTGCACATGGGCGGTAC'3
A91-*A074-SNP-18* COP (+)	5' CGCCCATGTGCACGTATCTCCAAAACGCCCATGTGCACGTATCTCCAAAACGCCCATGTGCACG TATCTCCAAAACGCCCATGTGCACGTATCTCCAAAACGCCCATGTGCACGTATCTCCAAAACC'3
A91-*A074-SNP-18* COP (-)	5'TCGAGGTTTTGGAGATACGTGCACATGGGCGTTTTGGAGATACGTGCACATGGGCGTTTTGGAGATAC GTGCACATGGGCGTTTTGGAGATACGTGCACATGGGCGTTTTGGAGATACGTGCACATGGGCGGTAC'3
A91-*A074-SNP-18* WF (+)	5'CGCCCATGTGCACATATCTCCAAAACGCCCATGTGCACATATCTCCAAAACGCCCATGTGCACA TATCTCCAAAACGCCCATGTGCACATATCTCCAAAACGCCCATGTGCACATATCTCCAAAACC'3
A91-*A074-SNP-18* WF (-)	5'TCGAGGTTTTGGAGATATGTGCACATGGGCGTTTTGGAGATATGTGCACATGGGCGTTTTGGAGATAT GTGCACATGGGCGTTTTGGAGATATGTGCACATGGGCGTTTTGGAGATATGTGCACATGGGCGGTAC'3
A63-*rs889312* major (+)	5'-CGCTGGAGAAAGGAATGTGCAAATTAGCTGGAGAAAGGAATGTGCAAATTAGCTGGAGAAAGGA ATGTGCAAATTAGCTGGAGAAAGGAATGTGCAAATTAGCTGGAGAAAGGAATGTGCAAATTAC-'3
A63-*rs889312* major (-)	5'- TCGAGTAATTTGCACATTCCTTTCTCCAGCTAATTTGCACATTCCTTTCTCCAGCTAATTTGCACATT CCTTTCTCCAGCTAATTTGCACATTCCTTTCTCCAGCTAATTTGCACATTCCTTTCTCCAGCGGTAC-'3

Table 12 continued.

Construct/ Primer ID	Sequence
A63-*rs889312* minor (+)	5'-CGCTGGAGAAAGGCATGTGCAAATTAGCTGGAGAAAGGCATGTGCAAATTAGCTGGAGAAAGGC ATGTGCAAATTAGCTGGAGAAAGGCATGTGCAAATTAGCTGGAGAAAGGCATGTGCAAATTAC-‘3
A63-*rs889312* minor (-)	5'-TCGAGTAATTTGCACATGCCTTTCTCCAGCTAATTTGCACATGCCTTTCTCCAGCTAATTTGCACATG CCTTTCTCCAGCTAATTTGCACATGCCTTTCTCCAGCTAATTTGCACATGCCTTTCTCCAGCGGTAC-‘3
A75-*rs961847* major (+)	5’-CCTATGCAAATGCAGAGGTACAAGATCTATGCAAATGCAGAGGTACAAGATCTATGCAAATGCA GAGGTACAAGATCTATGCAAATGCAGAGGTACAAGATCTATGCAAATGCAGAGGTACAAGATC-‘3
A75-*rs961847* major (-)	5’-TCGAGATCTTGTACCTCTGCATTTGCATAGATCTTGTACCTCTGCATTTGCATAGATCTTGTACCTCT GCATTTGCATAGATCTTGTACCTCTGCATTTGCATAGATCTTGTACCTCTGCATTTGCATAGGGTAC-‘3
A75-*rs961847* minor (+)	5’-CCTATGCAAATGCGGAGGTACAAGATCTATGCAAATGCGGAGGTACAAGATCTATGCAAATGCG GAGGTACAAGATCTATGCAAATGCGGAGGTACAAGATCTATGCAAATGCGGAGGTACAAGATC -‘3
A75-*rs961847* minor (-)	5-TCGAGATCTTGTACCTCCGCATTTGCATAGATCTTGTACCTCCGCATTTGCATAGATCTTGTACCTCC GCATTTGCATAGATCTTGTACCTCCGCATTTGCATAGATCTTGTACCTCCGCATTTGCATAGGGTAC-‘3
A75-*rs1862626* major (+)	5’-CGCATGGCTAGATTTCAGCCTGTGCTGCATGGCTAGATTTCAGCCTGTGCTGCATGGCTAGATT TCAGCCTGTGCTGCATGGCTAGATTTCAGCCTGTGCTGCATGGCTAGATTTCAGCCTGTGCTC-‘3
A75-*rs1862626* major (-)	5’TCGAGAGCACAGGCTGAAATCTAGCCATGCAGCACAGGCTGAAATCTAGCCATGCAGCACAGGCTGA AATCTAGCCATGCAGCACAGGCTGAAATCTAGCCATGCAGCACAGGCTGAAATCTAGCCATGCGGTAC-‘3
A75-*rs1862626* minor (+)	5’-CGCATGGCTAGATGTCAGCCTGTGCTGCATGGCTAGATGTCAGCCTGTGCTGCATGGCTAGATG TCAGCCTGTGCTGCATGGCTAGATGTCAGCCTGTGCTGCATGGCTAGATGTCAGCCTGTGCTC-‘3
A75-*rs1862626* minor (-)	5’TCGAGAGCACAGGCTGACATCTAGCCATGCAGCACAGGCTGACATCTAGCCATGCAGCACAGGCTGAC ATCTAGCCATGCAGCACAGGCTGACATCTAGCCATGCAGCACAGGCTGACATCTAGCCATGCGGTAC-‘3

Table 12 continued.

Construct/ Primer ID	Sequence
A75-*rs1862625* major (+)	5'CGAGGGGCAGAGGACAAGAAGGGTTAGAGGGGCAGAGGACAAGAAGGGTTAGAGGGGCAGAGGA CAAGAAGGGTTAGAGGGGCAGAGGACAAGAAGGGTTAGAGGGGCAGAGGACAAGAAGGGTTAC-'3
A75-*rs1862625* major (-)	5'-TCGAGTAACCCTTCTTGTCCTCTGCCCCTCTAACCCTTCTTGTCCTCTGCCCCTCTAACCCTTCTTGT CCTCTGCCCCTCTAACCCTTCTTGTCCTCTGCCCCTCTAACCCTTCTTGTCCTCTGCCCCTCGGTAC-'3
A75-*rs1862625* minor (+)	5'-CGAGGGGCAGAGGGCAAGAAGGGTTAGAGGGGCAGAGGGCAAGAAGGGTTAGAGGGGCAGAGGG CAAGAAGGGTTAGAGGGGCAGAGGGCAAGAAGGGTTAGAGGGGCAGAGGGCAAGAAGGGTTAC-'3
A75-*rs1862625* minor (-)	5'-TCGAGTAACCCTTCTTGCCCTCTGCCCCTCTAACCCTTCTTGCCCTCTGCCCCTCTAACCCTTCTTGC CCTCTGCCCCTCTAACCCTTCTTGCCCTCTGCCCCTCTAACCCTTCTTGCCCTCTGCCCCTCGGTAC-'3
A75-*rs1910020* major (+)	5'-CGATATTGTGTCTCTAGATATAAAGGGATATTGTGTCTCTAGATATAAAGGGATATTGTGTCTC TAGATATAAAGGGATATTGTGTCTCTAGATATAAAGGGATATTGTGTCTCTAGATATAAAGGC-'3
A75-*rs1910020* major (-)	5'-TCGAGCCTTTATATCTAGAGACACAATATCCCTTTATATCTAGAGACACAATATCCCTTTATATCTAG AGACACAATATCCCTTTATATCTAGAGACACAATATCCCTTTATATCTAGAGACACAATATCGGTAC-'3
A75-*rs1910020* minor (+)	5'-CGATATTGTGTCTTTAGATATAAAGGGATATTGTGTCTTTAGATATAAAGGGATATTGTGTCTT TAGATATAAAGGGATATTGTGTCTTTAGATATAAAGGGATATTGTGTCTTTAGATATAAAGGC-'3
A75-*rs1910020* minor (-)	5'-TCGAGCCTTTATATCTAAAGACACAATATCCCTTTATATCTAAAGACACAATATCCCTTTATATCTAA AGACACAATATCCCTTTATATCTAAAGACACAATATCCCTTTATATCTAAAGACACAATATCGGTAC-'3
A75-*rs12697152* major (+)	5'-CAGCCAAGGTGGGTGCATAAGAAGGCAGCCAAGGTGGGTGCATAAGAAGGCAGCCAAGGTGGGT GCATAAGAAGGCAGCCAAGGTGGGTGCATAAGAAGGCAGCCAAGGTGGGTGCATAAGAAGGCC-'3
A75-*rs12697152* major (-)	5'-TCGAGGCCTTCTTATGCACCCACCTTGGCTGCCTTCTTATGCACCCACCTTGGCTGCCTTCTTATGCA CCCACCTTGGCTGCCTTCTTATGCACCCACCTTGGCTGCCTTCTTATGCACCCACCTTGGCTGGTAC-'3
A75-*rs12697152* minor (+)	5'-CAGCCAAGGTGGGCGCATAAGAAGGCAGCCAAGGTGGGCGCATAAGAAGGCAGCCAAGGTGGGC GCATAAGAAGGCAGCCAAGGTGGGCGCATAAGAAGGCAGCCAAGGTGGGCGCATAAGAAGGCC-'3
A75-*rs12697152* minor (-)	5'-TCGAGGCCTTCTTATGCGCCCACCTTGGCTGCCTTCTTATGCGCCCACCTTGGCTGCCTTCTTATGCG CCCACCTTGGCTGCCTTCTTATGCGCCCACCTTGGCTGCCTTCTTATGCGCCCACCTTGGCTGGTAC-'3
A75-*rs4700485* major (+)	5'-CCTGGTTTAAACTGAGACGTCTACTCCTGGTTTAAACTGAGACGTCTACTCCTGGTTTAAACTG AGACGTCTACTCCTGGTTTAAACTGAGACGTCTACTCCTGGTTTAAACTGAGACGTCTACTCC-'3
A75-*rs4700485* major (+)	5'-TCGAGGAGTAGACGTCTCAGTTTAAACCAGGAGTAGACGTCTCAGTTTAAACCAGGAGTAGACGTCTC AGTTTAAACCAGGAGTAGACGTCTCAGTTTAAACCAGGAGTAGACGTCTCAGTTTAAACCAGGGTAC-

Table 12 continued.

Constructs/ Primer ID	Sequence
A75-*rs4700485* minor (+)	5'-CCTGGTTTAAACTAAGACGTCTACTCCTGGTTTAAACTAAGACGTCTACTCCTGGTTTAAACTA AGACGTCTACTCCTGGTTTAAACTAAGACGTCTACTCCTGGTTTAAACTAAGACGTCTACTCC-'3
A75-*rs4700485* minor (-)	5'TCGAGGAGTAGACGTCTTAGTTTAAACCAGGAGTAGACGTCTTAGTTTAAACCAGGAGTAGACGTCTT AGTTTAAACCAGGAGTAGACGTCTTAGTTTAAACCAGGAGTAGACGTCTTAGTTTAAACCAGGGTAC-'3
RV3 +	5'-CTA GCA AAA TAG GCT GTC CC-'3
RV3 -	5'-AAC AGT ACC GGA ATG CCA AG-'3

sticky *Kpn1* site at the 5' end and a sticky *BglII* site at the 3' end. The *A074-SNP-17* construct contained a complementary sticky *BglII* site at the 5' end and a sticky *Xho1* site at the 3' end. Both constructs had to ligate to each other and ligate into the plasmid to get a successful ligation reaction. All components were added to the same ligation reaction. The constructs for *A046-SNP-A* were designed as described previously, with the exception of the restriction enzymes used. The *Sal1* and *BamH1* enzymes were used. The sequences for the constructs can be found in Table 13. The pGL3- Promoter vector was digested with the *Kpn1* and *Xho1* enzymes first and the *A074-SNP-18/A074-SNP-17* constructs were cloned first. Cloning reactions and plasmid extractions were performed as previously described. Plasmids were extracted using the Spin Miniprep kit (Qiagen) and sequenced. The plasmids containing the *A074-SNP-18/A074-SNP-17* constructs were digested with the *Sal1* and *BamH1* enzymes and the *A046-SNP-A* constructs were cloned into the vector. Cloning reactions and plasmid extractions were performed as previously described, with the exception of using the RV4 primers for analysis of positive clones. A list of constructs and primers used for cloning is shown in Table 13.

C. Transfection of T47D and MDA-MB-231 cells and luciferase assays

T47D and MDA-MB-231 were ordered from ATCC. T47D cells were grown in RPMI-1640 with the addition of 10% fetal bovine serum (FBS), antibiotics-antimycrobials (Anti-anti, Life Technologies) and 0.2 U nits/ml bovine insulin. MDA-MB-231 cells were grown in DMEM medium with 10% FBS and anti-anti (Life Technologies). Plasmids were transfected into T47D and MDA-MB-231 cells using

Table 13. Sequences of constructs and primers for cloning all three *Mcs1b* candidate SNPs into same pGL3- Promoter. Nucleotides in red are SNP nucleotides.

Construct/ Primer ID	Sequence
A117-*A074-SNP-18* WF (+)	5'CGCCCATGTGCACATATCTCCAAAACGCCCATGTGCACATATCTCCAAAACGCCCATGTGCACA TATCTCCAAAACGCCCATGTGCACATATCTCCAAAACGCCCATGTGCACATATCTCCAAAACA'3
A117-*A074-SNP-18* WF (-)	5'GATCTGTTTTGGAGATATGTGCACATGGGCGTTTTGGAGATATGTGCACATGGGCGTTTTGGAGAT ATGTGCACATGGGCGTTTTGGAGATATGTGCACATGGGCGGTAC'3
A117-*A074-SNP-18* COP (+)	5'CGCCCATGTGCACGTATCTCCAAAACGCCCATGTGCACGTATCTCCAAAACGCCCATGTGCACG TATCTCCAAAACGCCCATGTGCACGTATCTCCAAAACGCCCATGTGCACGTATCTCCAAAACA'3
A117-*A074-SNP-18* COP (-)	5'GATCTGTTTTGGAGATACGTGCACATGGGCGTTTTGGAGATACGTGCACATGGGCGTTTTGGAGATAC GTGCACATGGGCGTTTTGGAGATACGTGCACATGGGCGTTTTGGAGATACGTGCACATGGGCGGTAC'3
A117-*A074-SNP-17* WF (+)	5'GATCTATTCCATGACATGTGCAGCTCATCAATTCCATGACATGTGCAGCTCATCAATTCCATGACATG TGCAGCTCATCAATTCCATGACATGTGCAGCTCATCAATTCCATGACATGTGCAGCTCATCAC'3
A117-*A074-SNP-17* WF (-)	5'TCGAGTGATGAGCTGCACATGTCATGGAATTGATGAGCTGCACATGTCATGGAATTGATGAGCTGCAC ATGTCATGGAATTGATGAGCTGCACATGTCATGGAATTGATGAGCTGCACATGTCATGGAATA'3
A117-*A074-SNP-17* COP (+)	5'GATCTATTCCATGACATTTGCAGCTCATCAATTCCATGACATTTGCAGCTCATCAATTCCATGACATT TGCAGCTCATCAATTCCATGACATTTGCAGCTCATCAATTCCATGACATTTGCAGCTCATCAC'3
A117-*A074-SNP-17* COP (-)	5'TCGAGTGATGAGCTGCAAATGTCATGGAATTGATGAGCTGCAAATGTCATGGAATTGATGAGCTGCAA ATGTCATGGAATTGATGAGCTGCAAATGTCATGGAATTGATGAGCTGCAAATGTCATGGAATA3'
A117-*A046-SNP-A* WF (+)	5' GATCCGAACTTCTACAGGACCTGTGCATTCGAACTTCTACAGGACCTGTGCATTCGAACTTCTACAGG ACCTGTGCATTCGAACTTCTACAGGACCTGTGCATTCGAACTTCTACAGGACCTGTGCATTCG'3
A117-*A046-SNP-A* WF (-)	5'TCGACGAATGCACAGGTCCTGTAGAAGTTCGAATGCACAGGTCCTGTAGAAGTTCGAATGCACAGGTC CTGTAGAAGTTCGAATGCACAGGTCCTGTAGAAGTTCGAATGCACAGGTCCTGTAGAAGTTCG'3
A117-*A046-SNP-A* COP (+)	5' GATCCGAACTTCTACAGAACCTGTGCATTCGAACTTCTACAGAACCTGTGCATTCGAACTTCTACAGA ACCTGTGCATTCGAACTTCTACAGAACCTGTGCATTCGAACTTCTACAGAACCTGTGCATTCG'3
A117-*A046-SNP-A* COP (-)	5'TCGACGAATGCACAGGTTCTGTAGAAGTTCGAATGCACAGGTTCTGTAGAAGTTCGAATGCACAGGTT CTGTAGAAGTTCGAATGCACAGGTTCTGTAGAAGTTCGAATGCACAGGTTCTGTAGAAGTTCG'3
RV4 +	5'GAC GAT AGT CAT GCC CCG CG'3
RV4 -	5'CGC TTC GAG CAG ACA TGA TA'3

Lipofectamine 2000 (Life Technologies) according to manufacture's instructions. 24hrs prior to transfection, 20,000 cells were plated in a 96-well dish using media without Antibiotics- antimycotics and allowed to grow to 95% confluency. A total of 200ng/ well were transfected of the constructs containing plasmids and 10ng/ well of pRL-TK (Renilla expressing vector) were used for T47D cells. A total of 100ng/ well were transfected of the constructs containing plasmids and 5ng/ well of pRL-TK (renilla expressing vector) were used for MDA-MB-231 cells. Transfections were done in triplicate. The luciferase activity was determined 18-24hrs post transfection using the Dual- Luciferase Reporter Assay System (Promega) according to manufacturer's instructions. The reaction products of the luciferase and renilla genes were read on a Biotek Synergy 2 Plate Reader. As a control, the luciferase and renilla activity of the pGL3-Promoter and pGL3-Basic vector were determined.

D. Statistical analysis of luciferase activity

Luciferase and Renilla values were downloaded into Excel. The luciferase activity was divided by the Renilla activity to account for differences in transfection efficiency. An average of the triplicate value was then determined. And the values were divided by the average of the pGL3-Promoter vector to adjust for differences between experiments. Adjusted values were then pooled. Each plasmid transfection had at least nine values from three independent experiments. A Kruskal-Wallis analysis was performed, followed by a Conover- Inman post hoc test to determine the p-values using SYSTAT 13.

E. EMSAs and EMSA supershifts

Oligos used for electrophoretic mobility shift assays (EMSAs) were biotin labeled using the Biotin 3' End DNA Labeling Kit (Thermo Scientific) according to the manufacturer's protocol. In short, oligos were ordered from IDT and resuspended to a concentration of 100μM in water. The oligoes were diluted to a working concentration of 1μM and incubated with TdT reaction buffer, Biotin-N4-CTP and TDT enzyme for 30min at 37°C. The biotin labeled oligos were then extracted using a chloroform: isoamyl alcohol extraction. A list of the oligos used in the EMSA reactions are shown in Table 14. Oligos were annealed by adding equal amounts of forward and reverse oligo and incubating them in a thermocycler. The oligos were allowed to heat to 95°C and were cooled down at a rate of 1°C/ min for 70min.

Nuclear extracts of T47D and MDA-MB-231 cells were performed using NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Scientific) according to manufacturer's protocol. Nuclear extracts were quantified using a Biorad protein assay kit. Nuclear extract aliquots were stored at -80°C.

EMSAs were performed using the LightShift Chemiluminescent EMSA kit (Thermo Scientific) according to manufacturer's protocols. In short, 1μL of biotin labled oligos were incubated with binding buffer, 2.5% glycerol, 100nM $MgCl_2$, 50ng/μL Poly(dI•dC), 0.05% NP-40 and 10μg nuclear extract at room temperature for 20min. 200x unlabeled cold probe was added as competitor. Note, for Figure 14B, 13x cold competitor probe was used. For Figure 18 an increase in cold competitor probe from 10x to 200x was used. A 4 or 5% polyacrylamide gel was pre- run in 0.5% TBE at 100V for 30min. 5μl loading dye was added to the binding reactions and the entire sample was

Table 14. List of oligos used in EMSA and EMSA supershift experiments.

ID	Sequence
A90-*A046-SNP-A* WF (+)	GAATGCACAGGTCCTGTAGAAGTTC
A90-*A046-SNP-A* WF (+)	GAACTTCTACAGGACCTGTGCATTC
A90-*A046-SNP-A* COP (+)	GAATGCACAGGTTCTGTAGAAGTTC
A90-*A046-SNP-A* COP (+)	GAACTTCTACAGAACCTGTGCATTC
A118-*A074-SNP17* WF (+)	TGATGAGCTGCACATGTCATGGAAT
A118-*A074-SNP17* WF (-)	ATTCCATGACATGTGCAGCTCATCA
A118-*A074-SNP17* COP (+)	TGATGAGCTGCAAATGTCATGGAAT
A118-*A074-SNP17* COP (-)	ATTCCATGACATTTGCAGCTCATCA
A118-*A074-SNP18 WF* (+)	GTTTTGGAGATATGTGCACATGGGC
A118-*A074-SNP18* WF (-)	GCCCATGTGCACATATCTCCAAAAC
A118-*A074-SNP18* COP (+)	GTTTTGGAGATACGTGCACATGGGC
A118-*A074-SNP18* COP (-)	GCCCATGTGCACGTATCTCCAAAAC
A90-*rs1862626* major (+)	GCATGGCTAGATTTCAGCCTGTGCT
A90-*rs1862626* major (-)	AGCACAGGCTGAAATCTAGCCATGC
A90-*rs1862626* minor (+)	GCATGGCTAGATGTCAGCCTGTGCT
A90-*rs1862626* minor (-)	AGCACAGGCTGACATCTAGCCATGC
A90-*rs889312* major (+)	GCTGGAGAAAGGAATGTGCAAATTA
A90-*rs889312* major (-)	TAATTTGCACATTCCTTTCTCCAGC
A90-*rs889312* minor (+)	GCTGGAGAAAGGCATGTGCAAATTA
A90-*rs889312* minor (-)	TAATTTGCACATGCCTTTCTCCAGC
A107-*A046-SNP-A* 1bp Deletion (+)	GAATGCACAGGTCTGTAGAAGTTC
A107- *A046-SNP-A* 1bp Deletion (-)	GAACTTCTACAGACCTGTGCATTC
A107-*A046-SNP-A* 3bp deletion (+)	GAATGCACAGGTGTAGAAGTTC
A107-*A046-SNP-A* 3bp deletion (-)	GAACTTCTACACCTGTGCATTC
A107-*A046-SNP-A* 5bp deletion (+)	GAATGCACAGGTAGAAGTTC
A107-*A046-SNP-A* 5bp deletion (-)	GAACTTCTACCTGTGCATTC
A107- *A046-SNP-A* random insert 9bps (+)	GAATGCACCGTCTCTGGAGAAGTTC
A107- *A046-SNP-A* random insert 9bps (-)	CAACTTCTCCAGAGACGGTGCATTC
A107- *A046-SNP-A* random insert 13bps (+)	GAATGCTTGTCGTCCATTTAAGTTC
A107- *A046-SNP-A* random insert 13bps (-)	GAACTTAAATGGACGACAAGCATTC
A107- *A046-SNP-A* random insert 17bps (+)	GAATGCTTTTACTGCCCGTACGTTC
A107- *A046-SNP-A* random insert 17bps (-)	GAACGTACGGGCAGTAAAAGCATTC
A118-*A074-SNP-17* 3bp deletion (+)	TGATGAGCTGCTGTCATGGAAT
A118-*A074-SNP-17* 3bp deletion (-)	ATTCCATGACAGCAGCTCATCA
A118-*A074-SNP-18* 3bp deletion (+)	GTTTTGGAGATTGCACATGGGC
A118-*A074-SNP-18* 3bp deletion (-)	GCCCATGTGCAATCTCCAAAAC
A124- E-box (+)	GCGCTCCCCACGTGGCGGAGGG
A124- E-box (-)	CCCTCCGCCACGTGGGGAGCGC
A124-ARE (+)	CAGTCACAGTGACTCAGCAGAATCT
A124-ARE (-)	AGATTCTGCTGAGTCACTGTGACTG
A124- random oligo (+)	CGCCGGGCGATAGTGGAGTTAGTAG
A124- random (-)	CTACTAACTCCACTATCGCCCGGCG
A128-PRE (+)	GATCCTGTACAGGATGTTCTAGCTACA
A128-PRE (-)	TGTAGCTAGAACATCCTGTACAGGATC
A128- NF1C (+)	AGGTTGGCAAAAAGCCAAGG
A128- NF1C (-)	CCTTGGCTTTTTGCCAACCT
A128- ARRE2 (+)	gatcGGAGGAAAAACTGTTTCATACAGAAG GCGT
A128- ARRE2 (-)	ACGCCTTCTGTATGAAACAGTTTTTCCTCC gatc

loaded onto the gel. The gel was run at 100V for until the blue dye had migrated ¾ down the length of the gel (approximately 45min). The protein-DNA complexes were transferred to a Biodyne B Pre-cut Modified Nylon Membrane (Thermo Scientific) using a TransBlot SD Semi-dry Transfer Cell (Bio-Rad) at 25V for 10min. DNA and proteins were crosslinked using UV crosslinking instruments for 15min. Membranes were washed and chemiluminescence detected using Luminol/ Enhancer Solution, Stable Peroxide Solution and x-ray film.

EMSA supershifts were performed as previously described. However, during the binding reaction 2µg of antibody was added to the reaction, allowed to incubate for 30min before adding biotin labeled probes and incubating for 20 more minutes at room temperature. As a control 2µg of Negative Control for Rabbit IgG Ab-1 (Thermo Scientific) was used. The antibodies used are c-MYC, NRF2, PR, NFIC and ILF2 (Santa Cruz Biotechnology: sc-764x, sc-722x, sc-538x and Abcam: ab86570, ab28772).

F. Luciferase assays of *A074-SNP-17* SNP alleles co-expressed with protein expression vectors

Luciferase assays in T47D cells were performed as described above. An additional 200ng/ well of the protein expression vector was added and luciferase activity was measured 18-24hrs post transfection. The c-MYC protein expressed is the human isoform and the plasmid backbone is the pWZL plasmid. The NRF2 expressed is human isoform and the plasmid backbone is the pCI- neo vector. As a control, the pc3.1 vector was used. The ARE control plasmids used for the NRF2 co-expression luciferase experiments contains the minimum promoter (164 bp) of the rat Glutatione S-transferase

A2 gene in the pGL3- Basic vector. Each experiment was performed in triplicate and the results from three independent experiments were pooled. Results were analyzed as previously described.

G. Western Blot for c-Myc

20μg of each protein extract were used in the Western Blot. Equal volume of Leammli buffer (Biorad) were added to the protein extracts and heated at 95°C for 5min. A prestained protein marker was used (New England Biolabs). Samples were run on a stacking polyacrylamide gel at 60V until samples reached stacking gels' interface. Samples were then run at 100V until pink dye ran off the gel in running buffer (Tris-glycine/ SDS). A PVDF membrane (GE Healthcare) was prepared by soaking in methanol for 20s, then water for 20s and then 10min in Towbin buffer (25 mM Tris, 192 mM glycine, 20% (v/v) methanol (pH 8.3)). Samples were transferred in semi-dry apparatus for 40min at 15V. Membrane was blocked in PBS-tween + milk (0.1% Tween-20, 5% milk) for 1hr. Membranes were incubated with 1/2000 c-Myc antibody (sc-764, Santa Cruz) and 1/4000 GAPDH antibody (A300-641-A, Bethyl) in PBS- tween + milk overnight at 4°C. Membranes were washed twice with PBS-tween and then four times for 5 min. Goat anti-rabbit secondary antibody was added at 1/4000) in PBS-tween for 1hr while shaking. Membranes were washed twice with PBS- tween then three times for 5min and once for 5min in PBS. Equal amounts of Pico reagents (Thermo Scientific) were added for 5min and membrane was exposed to X-ray film.

H. Identification of candidate proteins binding to *A074-SNP-17* using TF search and Mass spectrometry

To identify candidate proteins binding to the WF and COP alleles for *A074-SNP-17,* we used the online program TF Search (http://www.cbrc.jp/research/db/TFSEARCH.html). The 25bp oligos used for the *A074-SNP-17* EMSAs were used for the analysis. The parameters were set to "use all matrices" and the threshold score was set to 75. We noted all proteins that were predicted to bind the two alleles differentially.

For DNA pulldowns, 30µl of Streptavidin Magnetic Particles (Roche) were washed three times with 100µl of TEN100 (10mM Tris-HCl), 1mM EDTA, 100mM NaCl, pH 7.5) on a rotator for 1min at room temperature. Particles were placed into a magnet and the liquid was removed. Particles were resuspended with 100µl TEN100 and 3µg of the *A074-SNP-17* COP, WF or 3bp deletion oligo was added. Oligos used have the same sequence as shown in Table 14, however, they were ordered from IDT with a 3' biotin molecule. The samples were rotated for 10min at room temperature. Samples were washed twice with TEN1000 (10mM Tris-HCl, 1mM EDTA, 1M NaCl, pH 7.5). Samples were blocked with 0.5% milk in TEN100 for 15min on ice and washed once with TEN100. 750µg of T47D nuclear extracts and 700µl of TEN100 were added to the samples and incubated for 30min at room temperature with rotation. The liquid was removed and DNA was eluted using 40µl of buffer C (20mM HEPES, 1.5mM $MgCl_2$, 0.2mM EDTA, 25% v/v glycerol, pH 7.9). A control EMSA using the *A074-SNP-17* COP, WF and MUT was performed using the DNA pulldown and T47D nuclear extract as described previously. 5µl of the DNA pulldown and 3µl of the T47D nuclear extracts

were used. The samples were submitted to the University of Louisville Mass Spectrometry Core and trypsin digestion followed by LC-MS/MS was performed. Proteins were identified through Proteome Discoverer and visualized using Scaffold 3.

I. 5' RACE of *Mier3* and cloning of *Mier3* promoter into pGL3 promoter vector

We used the FirstChoice RLM-RACE Kit (Ambion) to perform 5' Rapid Amplification of cDNA Ends (RACE) on the *Mier3* gene according to the manufacturer's instructions. The mammary gland total RNA from six 12-week old WF females was pooled for the analysis. 10µg total RNA was treated with Tobacco Acid Pyrophosphatase (TAP) for 1hr at 37°C. 5'RACE adapters were ligated and the RNA was reverse-transcribed using M-MVL Reverse Transcriptase. An outer 5'RML-RACE PCR was performed using *Mier3* specific primers, followed by an inner PCR. *Mier3* specific primers for 5'RACE are shown in Table 15. 5µl of the PCR reaction were run on a 2% high resolution agarose gel (GeneMate) and stained with SybrGold (Life Technologies). PCR products were cloned into a pCR2.1 TA cloning vector using the Original TA Cloning Kit (Life Technologies) and 17 clones were sequenced using the University of Louisville Sequencing Core.

The splenic DNA of one WF animal was used for cloning of the *Mier3* promoter. The PCR reaction was performed using the A buffer of the FailSafe PCR Premix Selection Kit (Epicentre), Accuprime PCR enzyme (Life Technologies) and the primers shown in Table 15. The PCR reaction was then cloned into the pCR2.1- TA vector using the Original TA Cloning Kit (Life Technologies). A clone was amplified and extracted

Table 15. Primers for 5'RACE and cloning of *Mier3* promoter.

ID	Sequence
A132- 5'RACE outer primer	CATGTTGTGTTTGATCGTAAAGG
A132- 5'RACE inner primer	CATCTGAGAAGACTGGGTTTC
A132- *Mier3* specific 5' primer	GAAGGAAATATGCCTCTAGAAGAT
A132- Cloning 3F	GGAAGATCTGAACCTGTGGCAACTTGGAT
A132- Cloning 3R	CCCAAGCTTATGACTGGAGGGTGAAGACG

using the Qiagen Spin Miniprep Kit and sequenced. The PCR primers used contain *BglII* and HindIII restriction sites and the correct insert were excised using these two enzymes. The pGL3 promoter vector was digested with the *BglII* and *HindIII* enzymes to remove the SV40 promoter. The linearized vector was gel extracted using the QiaQuick Gel Extraction Kit (Qiagen). Excised insert and linearized pGL3 Promoter vector were ligated using T4 DNA Ligase (Promega) and cloned. Vectors were sequenced for correct insert and extracted using the Qiagen Midiprep Kit. T47D cells were transfected and luciferase activity was determined as described above.

To identify transcription factors binding to the cloned *Mier3* promoter, we used the online program TF SEARCH (http://www.cbrc.jp/research/db/TFSEARCH.html). We set the parameters to a threshold score of 85 and used the vertebrate matrix only. The function of individual transcription factors was looked up using the online database UniProt (http://www.uniprot.org/uniprot/).

J. Preparation of mammary epithelial cell enriched cell preps for 3C and bisulfite sequencing

We extracted mammary epithelial cell enriched preps (MEC preps) from 18 rats in total. Six WF.COP Line N3 congenic and six WF/NHsd WF females received DMBA as previously described. Three WF.COP Line N3 and three WF/NHsd females did not receive DMBA. At twelve-weeks of age, the animals were euthanized. The D and E glands without lymph nodes were collected and pooled for each animal and placed into DMEM/F12 media (Life Technologies). Pooled glands were minced and placed into cell culture flask containing 10ml DMEM/F12 with 0.01g/ml collagenase type III

(Worthington) and incubated for 2.5hr at 37°C with gentle horizontal shaking. 0.2µg/ml deoxyribonuclease I (Wothington) was added and the flasks were incubated for 10min at 37°C with vigorous shaking. Cells were transferred to 15ml tubes and centrifuged for 10min at 1000rpm. The layer of fat was removed and cells were washed once with DMEM/F12. The cells were suspended in 2ml HBSS (Life Technologies) with 0.025% trypsin (Worthington) and 6.8mM EDTA and incubated horizontally for 5min at 37°C. 4ml of DMEM/F12 with 10% FBS was added to stop the reaction. Cells were centrifuged and resuspended in DMEM/F12 with 10% FBS. Cells were allowed to pass through a BD-40 filter and filters were rinsed with 2ml DMEM/F12 with 10% FBS. Cells were centrifuged and resuspended in 200µl PBS. Cells were counted using a hemocytometer using tryptan blue.

K. Chromosome conformation capture

MEC prep cells were diluted in 40ml of PBS and 1.7ml of 37% formaldehyde was added. Cells were left at room temperature for 10min. 2.7ml of 2M glycine was added and the cells were incubated for 5min at room temperature and then on ice for 15min. Cells were centrifuged for 10min at 800 x g and resuspended in 0.5ml of ice-cold lysis buffer (10mM Tris-HCl, 10mM NaCl, 0.2% NP-40 and Halt Protease Inhibitor Cocktail (Thermo Scientific)) for 15min. Cells were Dounce homogenized on ice for 15 strokes, incubated on ice for 1min and then homogenized with an additional 15 strokes. Cell nuclei were centrifuged for 5min at 2,500 x g. Nuclei were washed with 0.5ml of *BglII* buffer (New England Biolabs) and resuspended in 362µl of *BglII* buffer. 38µl of 1% SDS was added and the solution was incubated at 65°C for 10min. 44µl of 10% Triton-X and

400U of *BglII* were added and the solution was incubated overnight at 37°C. 86µl of 10% SDS was added and incubated for 30min at 65°C. Solution was transferred to 15ml conical tubes and 745µl of 10% Triton-X, 745µl of 10x ligation buffer (500mM Tris-HCl, 100mM $MgCl_2$ and 100mM DTT), 80µl of 10mg/ml BSA, 80µl of 100mM ATP, 5960µl of molecular grade water and 4000U T4 Ligase (New England Biolabs) was added and incubated for 2hrs at 16°C. 33µl of 20mg/ml proteinase K (New England Biolabs) was added and solution was incubated overnight at 16°C. 33µl of proteinase K was added and incubated for 2hrs at 42°C. Solutions were transferred to 50ml conical tubes and an equal amount of phenol: chloroform: isoamylalcohol (P:C:I, 25:24:1) was added and vortexed for 30sec. The solution was centrifuged for 5min at 2,460 x g and extracted again using P:C:I, then with chloroform and precipitated with Ethanol/ Sodium acetate. Samples were placed at -20°C for 30min and centrifuged at 10,000 rpm for 30min. The pellet was washed with 70% ethanol and dissolved in 1ml of water. RNAse A (Promega) was added and incubated for 30min at room temperature. Samples were split into two tubes and an additional P:C:I and chloroform extraction was performed. Ethanol/ sodium acetate was added and pellets were washed five times with 70% ethanol. Pellets were dissolved in 300µl TE buffer, incubated for 15min at 37°C and stored at -20°C. A BAC clone containing the entire *Mcs1b* region of interest was used as a positive control. The BAC used was *Rattus norvegicus* CH230-117D7 (Children's Hospital Oakland Research Institute (CHORI)). 20µg of the BAC was digested overnight at 37°C using *BglII* in 1x restriction enzyme buffer, 0.5µl BSA in a total volume of 50µl. The BAC was purified using P:C:I and chloroform extractions, followed by ethanol precipitation. The BAC was ligated using 5µl of 10x T4 DNA ligase buffer and 2U of T4 DNA ligase (Promega) and

was incubated at 16°C overnight. The BAC was then extracted using P:C:I and chloroform extraction followed by an ethanol precipitation. The BAC and DNA concentration was determined using band intensity quantification.

The amount to use in a PCR reaction was determined empirically to be 50ng. The BAC concentration to use in the PCR reaction was determined empirically to be 0.4ng. The PCR reactions were prepared on ice by adding 50ng of DNA, 0.2mM of each dNTP, 0.4µM of each primer, 1x Herculase reaction buffer and 0.3µl Herculase Enhanced polymerase (Agilent) in a total volume of 25µl. The reactions were run on a Veriti 96 well fact thermal cycler (Applied Biosystems) with the following conditions: 95°C for 1min, 34 cycles of 95°C 1min, 60°C for 45sec, 72°C for 2min and one cycle of 95°C for 1min, 60°C for 45sec, 72°C for 8min. 10µl of Ficoll dye (15% Ficoll 400 in water) was added to the PCR reactions and 20µl was added to the wells. For the BAC 10µ of Ficoll dye was added and 10µl was loaded onto the gel. PCR reactions and 500ng of a 100bp DNA ladder were run on a 1% agarose gel. The gel was stained using 0.1ug/ml ethidium bromide for 25min and destained in water for 15min. A picture was taken using a GE Typhoon 9400. PCR products were analyzed using band intensity quantification with ImageJ and divided by the intensity of the PCR band using the BAC. A list of 3C primers used can be found in Table 16.

L. Bisulfate sequencing

We performed bisulfite sequencing on three different groups of rats. Each group contained three females. One group was treated with DMBA, the other two were not. All

Table 16. Primers used for 3C analysis.

ID	Sequence
A136- Fixed 7	ATATAGCTTTCTGCCTGCGTGTAT
A136-Moving B	AGAGCTGACCGTTTGGTATTGTATAAC
A136-Moving E	ACTCCTGACTGTCTGTCTCCCTTTTA
A136-Moving F	GACACATCTATTTTATAGCCACAAACAC
A136-Moving G	AGTATCAGTGAGTGAGGTAGTTCAGAAA
A136-Moving H	AGATTATGGCATTACTGAGTCTGTCTAC
A136-Moving I	CTTGCTACTGGAGAATGTGATGATAAG
A136-Moving J	TAGAATTAAAGGCATATACCACCACAC
A136-Moving K	AGCATGGTCCTACAACTTTTAATACAG
A136-Moving L	ATTGAGAGTTCTTTCTTAGACCTGTAGC
A136-Moving M	CTAAAGTAGAAAATGCACATGGAGGAT

animals were 12- weeks old. MEC preps were prepared as described previously. DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen). DNA was bisulfate converted using the Cell-to-CpG Bisulfite Conversion Kit (Life Technologies) using manufacturer's instructions. 2µg was used in the bisulfite conversion and samples were pooled using equal amounts. The bisulfite conversion was performed under the following conditions: 95°C for 3min, 65°C for 60min, 95°C for 3min, 65°C for 30min. PCR reactions contained 150ng bisulfite converted DNA, 0.2mM of each dNTP, 1x buffer, 0.4µM of forward and reverse primers (list of primers can be found in Table 17), 0.5µl Easy A High- Fidelity PCR Cloning Enzyme (Agilent) in a total volume of 50µl. The PCR reaction were run using the following conditions: one cycle of 95°C for 15min, 18 cycles of 94°C for 45sec, 65°C for 45s (-0.5°C/ cycle)and 72°C for 1.5min, 22 cycles of 94°C for 45sec, 56°C for 45sec and 72°C for 1.5min, one cycle of 72°C for 8min. PCR reactions were run and visualized on a 1% agarose gel. PCR reactions were gel extracted using the Promega Wizard SV Gel and PCR Cleanup kit and cloned into a pCR-2.1 vector using the the Original TA Cloning kit (Life Technologies). Ten clones for each PCR reaction were extracted using the Qiagen Miniprep or the DirectPrep 96 Miniprep kit and sequenced.

M. CRISPR knockout of *A074-SNP-17*

Guide sequences for the CRISPR knockout were designed to be in close proximity to the *A074-SNP-17* targeted site and contained *BbsI* restriction sites. The guide sequences for target constructs are shown in Table 18. The pX335-U6-Chimeric_BB-CBh-HSpCas9n (D10A) vector (Addgene) was digested with the restriction enzyme *BbsI*. Constructs were annealed and cloned as previously described. RBA cells were grown in RPMI 1640

Table 17. Primers for bisulfate sequencing of *Mcs1b* candidate SNPs

ID	Sequence
A139- A074-SNP-17- a1	TTTTTTGAGAATTATTAGGTTAGGAAA
A139- A074-SNP-17- a2	TTCTACAACTACCCTATCAAAACCTTC
A139- A074-SNP-17- b1.1	TGGGTTAGATGGTATGTATTATGTAGTTT
A139- A074-SNP-17- b1.2	ATTACTTCAACAACAATCTTCTAAAAA
A139- A074-SNP-17- b2.1	TTTTTGTAGTTGTTTTGTTAGAGTTTTTT
A139- A074-SNP-17- b2.2	AAACTACATAATACATACCATCTAACCCAT
A139- A074-SNP-18- a1	AGGTTTTTTGGATTAAATTTGGGTA
A139- A074-SNP-18- a2	TCCTTACAAAAAAACTAAATAATTCCT
A139- A074-SNP-18- b1	ATTGTTTTTGTAAAGGGATTGAGTG
A139- A074-SNP-18- b2	AACCTCTTAAACCAAATCTAAACACC
A139- A046-SNP-A- a1	GGGTAGAGTTATAATTTTTGGTGTG
A139- A046-SNP-A- a2	CATCCATTCTAATAAAAATACAAAATACCA
A139- A046-SNP-A- b1	AAAAAGGTGTTTAGGATTATTGGTT
A139- A046-SNP-A- b2	CACAACTCACTCATATAACATACCACA

Table 18. Constructs for CRISPR knockout of *A074-SNP-17*.

ID	Sequence
A141-gSEQ Target 1 ON1	CACCGTTGTTGACCCAGTCGGAATTAAGG
A141-gSEQ Target 1 ON2	AAACCCTTAATTCCGACTGGGTCAACAAC
A141-gSEQ Target 2 ON1	CACCGCTGATGAGCTGCACATGTCATGG
A141-gSEQ Target 2 ON2	AAACCCATGACATGTGCAGCTCATCAGC

(Life Technologies) with the addition of 10% FBS and anti-anti (Life Technologies). LA7 cells were grown in DMEM with the addition of 10% FBS, anti-anti, 4.5g/L glucose, 0.005 mg/ml insulin, 50 nM hydrocortisone and 20 mM HEPES. $5x10^5$ cells were plated into 6- well dishes 24hrs prior of transfection using media without anti-anti. Cells were transfected using Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions. Cells were not transfected, transfected with 4µg of pEGFP-C1, 2µg of target 1 and 2µg of target 2, or mock transfected with Lipofectamine 2000 only. Media was changed every 24hrs. 72hrs post transfection, the cells were harvested with Tripsin- EDTA (Life Technologies), centrifuged for 5min at 500 x g and DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen). The PCR reactions were assembled on ice and contained: 100ng of DNA, 1x Optimase buffer, 0.2mM of each dNTP, 0.4µM of each primer and 1µ Optimase Polymerase (Transgenomics) in a total volume of 50µl. A positive control included with the SURVEYOR Mutation Detection Kit (Transgenomics) was used in the PCR reactions. The PCR reaction was run on a Veriti 96 well fact thermal cycler (Applied Biosystems) under the following conditions: 94°C for 2min, 35 cycles of 94°C for 30sec, 60°C for 30sec, 72°C for 1.5min and 1 cycle of 72°C for 5 min. The PCR reactions were run on a 2% high resolution agarose gel including 500ng of a 100bp ladder and 5µl 1kb ladder (Promega) and the band intensities were quantified. 300ng of each PCR product was used for the heteroduplex formation under the following conditions: 95°C for 10min, 95°C- 85°C (-2.0°C/s), 85°C for 1min, 85°C- 75°C (-0.3°C/s), 75°C for 1min, 75°C- 65°C (-0.3°C/s), 65°C for 1min, 65°C-55°C (-0.3°C/s), 55°C for 1min, 55°C-45°C (-0.3°C/s), 45°C for 1min, 45°C-35°C (-0.3°C/s), 35°C for 1min, 35°C- 25°C (-0.3°C/s), 25°C for 1min. Heteroduplexed DNA

was then digested with SURVEYOR nuclease and run on a 2% high resolution agarose gel. Cells were visualized using fluorescent microscopy using a Olympus IX51 inverted research microscope at 10x.

Results

The goal of this aim is to perform functional analyses on the rat *Mcs1b* candidate SNPs and the human *rs889312* correlated polymorphisms. The hypothesis is that these SNPs are located in regulatory regions that affect expression levels of *Mcs1b/MCS1B* genes. It is known that several *Mcs1b* genes are expressed differently in mammary glands of WF and COP females [87]. The SNPs are most likely located in enhancer/ repressor regions, due to their genomic location and proximity to the nearest gene. A reporter gene assay can be used to test if genomic regions are involved in gene regulation.

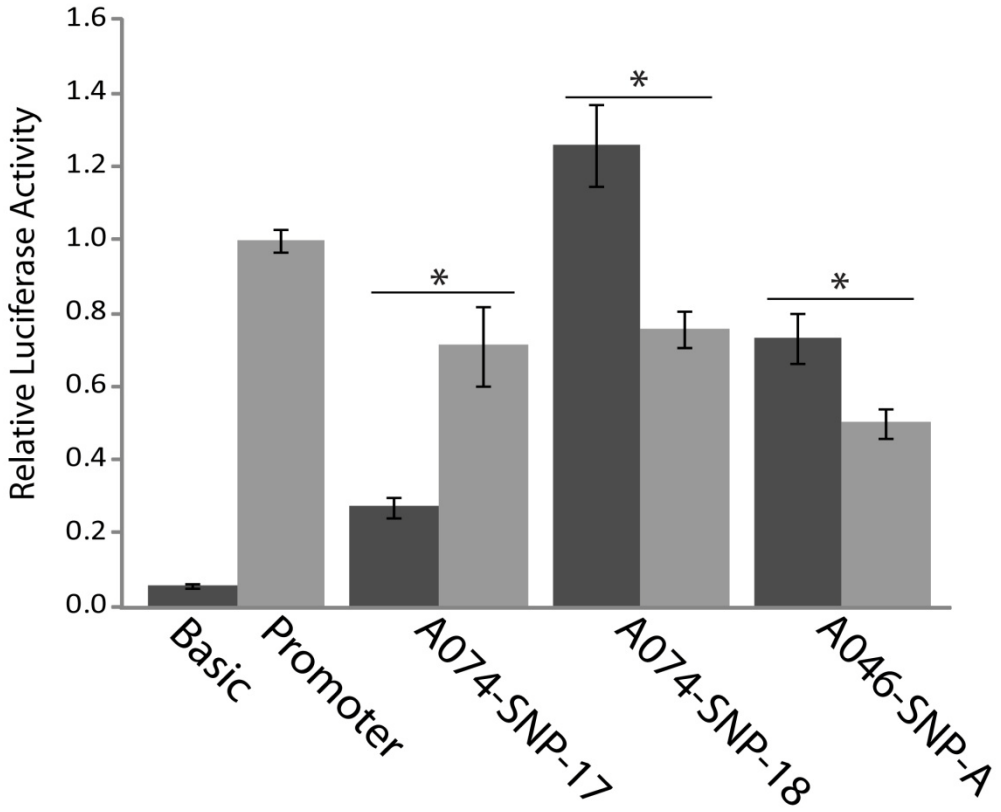
A. Luciferase assays for rat *Mcs1b* candidate SNPs and *rs889312* correlated SNPs

To determine if the rat *Mcs1b* candidate SNPs and the *rs889312* correlated SNPs are involved in gene regulation, a luciferase assay was used. Constructs containing the SNP nucleotide and 12bps flanking on either side were cloned into a pGL3-Promoter vector and the luciferase and Renilla activity was determined. Note, the 25bps of the SNP regions were repeated five times in tandem to increase signal strength. The pGL3-Promoter vector was used, since genomic regions suspected of being enhancers/ repressors can be tested using this vector. The luciferase activity was tested in two different breast cancer cells lines: T47D and MDA-MB-231 cells. T47D cells are luminal A, estrogen receptor (ER) and progesterone receptor (PR) positive, while MDA-MB-231

cells are triple negative, ER, PR and human epidermal growth factor receptor 2 (HER2) negative [138]. *rs889312* appears not to be subtype specific. It is associated with breast cancer risk in both ER positive and triple negative breast cancer subtypes [85, 86]. These two breast cancer cell lines were chosen, because they are examples of two opposing subtypes of breast cancer. A genotyping screen of 40 human breast cancer cell lines revealed that both T47D and MDA-MB-231 cells are heterozygous at SNP *rs889312*, which indicates they are also heterozygous at the six SNPs that *rs889312* tags. The same study also revealed that the expression level of *MAP3K1* is among the highest in T47D cells, while it is among the lowest in MDA-MB-231 cells [139]. Therefore, we chose these two cell lines, since they represent two opposing states of gene expression of a *MCS1B* candidate gene, yet they have the same genotype, indicating that differences in the expression levels are likely due to differences in the cell environment. The luciferase activity for the WF and COP alleles for all three *Mcs1b* candidate SNPs and the major and minor alleles for all seven *rs889312* correlated SNPs were determined in these two cell lines to test for any affects on gene regulation.

The results for the rat *Mcs1b* candidate SNPs in T47D cells are shown in Figure 12A. The luciferase activity is significant lower than the pGL3- Promoter luciferase activity for all SNP alleles except the COP allele of *A074-SNP-18*. This indicates that the three *Mcs1b* candidate SNPs may act as repressors, rather than enhancers, to reduce promoter activity. The luciferase activity is different between the COP and WF alleles for all three *Mcs1b* candidate SNPs. The luciferase activity for *A074-SNP-18* and *A046-SNP-A* is lower for the WF allele than for the COP allele. The luciferase activity for *A074-*

A


Relative Luciferase Activity
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
*
*
*
Basic
Promoter
A074-SNP-17
A074-SNP-18
A046-SNP-A

B

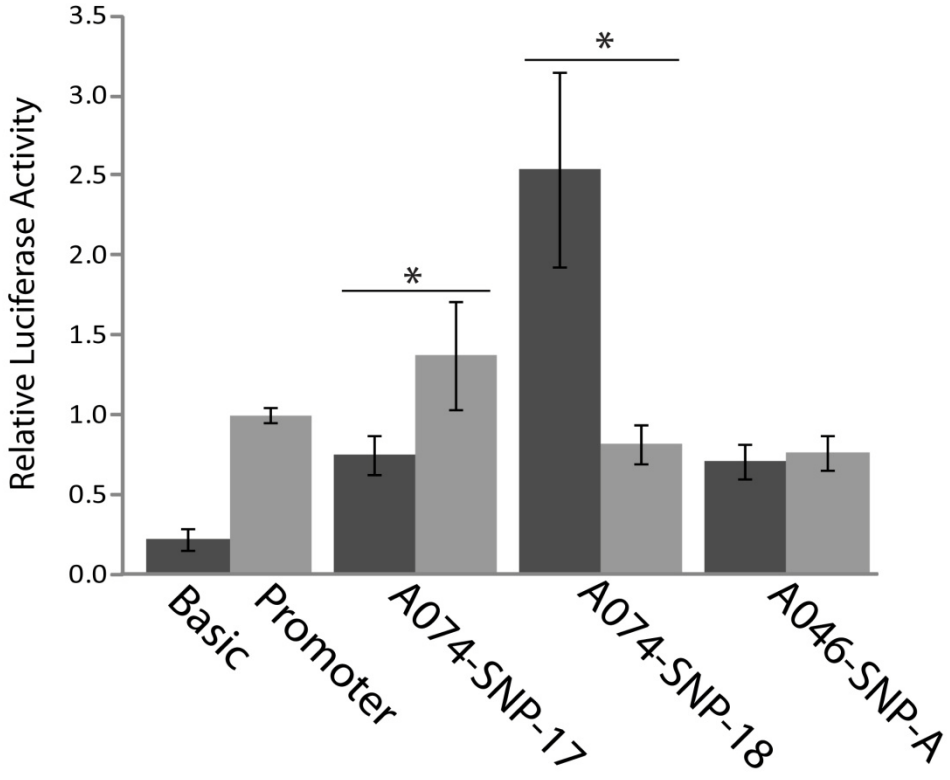

Relative Luciferase Activity
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
*
*
Basic
Promoter
A074-SNP-17
A074-SNP-18
A046-SNP-A

Figure 12. Luciferase assays for rat *Mcs1b* candidate SNPs. (■) Dark grey bars indicate adjusted relative luciferase activity for the COP (resistant) allele. (■) Light grey bars indicate adjusted relative luciferase activity for the WF (resistant) allele. Asterisks indicate luciferase activity that is significantly different between the WF and the COP alleles. All adjusted relative luciferase activities are at least nine values from three independent experiments. Errors bars indicate standard error from pooled experiments. As controls, the luciferase activity for the pGL3-Basic and pGL3-Promoter were determined. The pGL3-Promoter activity was used to adjust the relative luciferase activities to pool values from independent experiments. **A)** Luciferase activities in T47D cells. Luciferase activity is different between the two rat alleles for all three *Mcs1b* candidate SNPs. **B)** Luciferase activities in MDA-MB-231 cells. Luciferase activity was different between the two rat alleles for *A074-SNP-17* and *A074-SNP-18,* but not for *A046-SNP-A*.

SNP-17 shows the opposite pattern, with the COP allele activity being lower than the WF allele activity. This pattern of gene regulation is the same as we see when looking at the expression levels of *Mcs1b* candidate SNPs between the two rat strains. *Map3k1, Gpbp1* and *Mier3* are all expressed higher in the WF mammary gland compared to the COP mammary gland in 12-week old females [87]. This makes *A074-SNP-17* a strong candidate for conferring the difference in expression levels of *Mcs1b* candidate genes between the two rat strains.

The luciferase activities for the *Mcs1b* candidate SNPs in MDA-MB-231 cells are shown in Figure 12B. The luciferase activity for *A074-SNP-17* shows a similar pattern as in T47D cells. The luciferase activity is lower in the COP allele than in the WF allele. For *A074-SNP-18* the pattern is the same as in T47D cells. The luciferase activity for the COP allele is higher than for the WF allele. However, there is no statistical difference between the two rat alleles for *A046-SNP-A*. This SNP showed a luciferase activity that was different between the two alleles in T47D cells. This is likely due to a difference in the cellular environments of the two breast cancer cell lines.

Next, we cloned the major and minor alleles for all seven *rs889312* correlated SNPs into the pGL3-Promoter vector and measured the luciferase activity in both T47D and MDA-MB-231 cells. The luciferase activities for the *rs889312* correlated SNPs in T47D cells are shown in Figure 13A. The luciferase activities between the major and minor alleles of four of the *rs889312* correlated SNPs are statistically different in T47D cells. *rs889312* and *rs12697152* have luciferase activities that show higher major allele activities than the minor allele activities. The major and minor allele luciferase activities

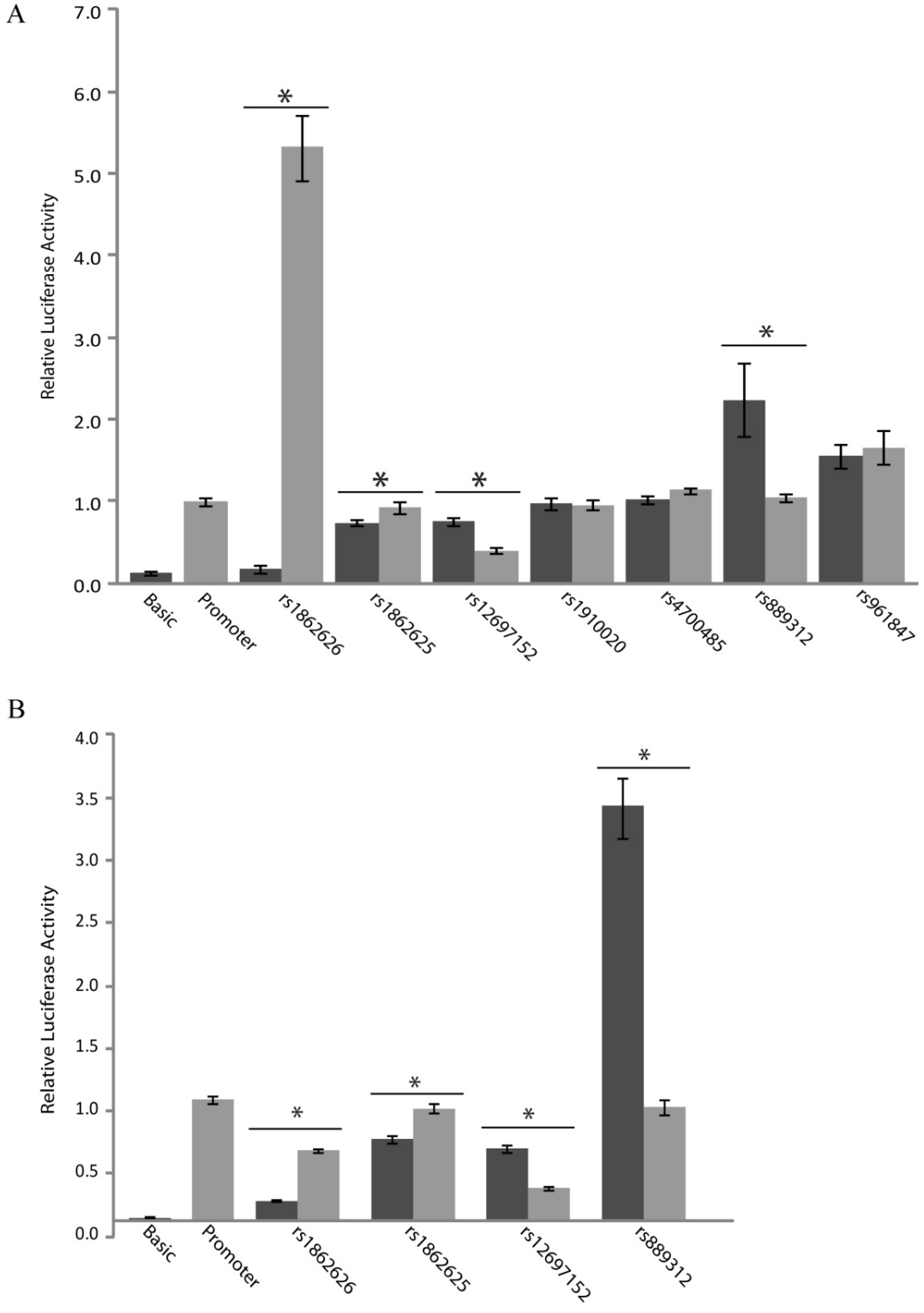

A
Relative Luciferase Activity
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
*
*
*
*
Basic
Promoter
rs1862626
rs1862625
rs12697152
rs1910020
rs4700485
rs889312
rs961847
B
Relative Luciferase Activity
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
*
*
*
*
Basic
Promoter
rs1862626
rs1862625
rs12697152
rs889312

Figure 13. Luciferase assays for human *rs889312* correlated SNPs. ( ) Dark grey bars indicate adjusted relative luciferase activity for the major (resistant) allele. ( ) Light grey bars indicate adjusted relative luciferase activity for the minor (susceptible) allele. Asterisks indicate luciferase activity that is significantly different between the major and the minor alleles. All adjusted relative luciferase activities are at least nine values from three independent experiments. Errors bars indicate standard error from pooled experiments. As controls, the luciferase activity for the pGL3-Basic and pGL3-Promoter were determined. The pGL3-Promoter activity was used to adjust the relative luciferase activities to pool values from independent experiments. **A)** Luciferase activity in T47D cells. Luciferase activity is different between the two major and minor alleles for four *rs889312* correlated SNPs. **B)** Luciferase activity in MDA-MB-231 cells. Luciferase activity is different between the major and minor alleles for the same four *rs889312* correlated SNPs as seen in T47D cells.

for *rs12697152* both are lower than the pGL3-Promoter luciferase activity, suggesting that this SNP may be located in a repressor element. *rs889312* is the SNP used in several GWA studies that identified the human *MCS1B* region. Its luciferase activity is different between the major and minor allele with the major allele having a much higher luciferase activity than the minor allele. The luciferase activity for *rs1862626* and *rs1862625* are also different between the major and minor alleles. Both SNPs have a higher minor allele activity compared to the major allele activity. The major allele activity for *rs1862626* is extremely low compared to the minor allele activity. However, the major allele activity is statistically different than the luciferase activity for pGL3-Basic, indicating that not all promoter activity is being repressed by the *rs1862626* major allele. The pattern of gene regulation, with the major or resistant allele having a lower luciferase activity than the minor or susceptible allele is the same as with the expression level of candidate *Mcs1b* genes and with the rat *Mcs1b* candidate SNPs, *A074-SNP-17*. This makes *rs1862626* a candidate functional ortholog for *A074-SNP-17.* Also, *rs1862625* has the same pattern of gene regulation in the luciferase assay, but its activity is not as exaggerated as with *rs1862626*, therefore *rs1862626* is considered the best candidate for a functional ortholog to *A074-SNP-17.*

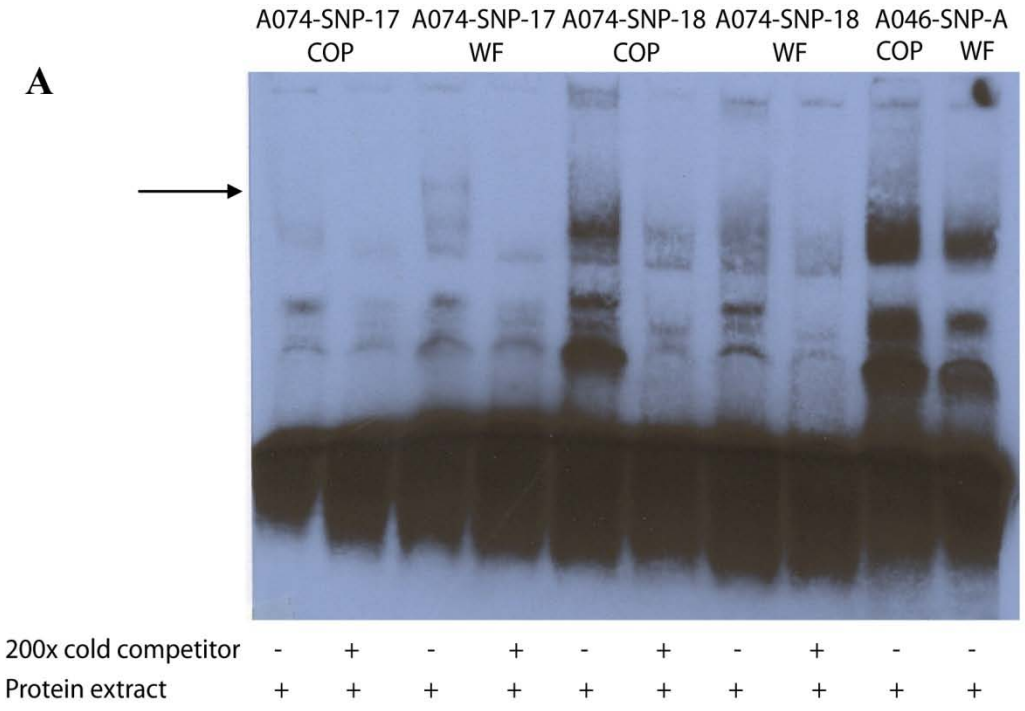
The luciferase activities for *rs889312* correlated SNPs in MDA-MB-231 cells are shown in Figure 13B. Only SNPs that showed a statistically significant difference between the major and minor alleles are shown. All four human SNPs that showed a luciferase activity that is different between the major and minor alleles in T47D cells also do so in MDA-MB-231 cells. There are no changes in pattern or intensity of luciferase activities for SNPs *rs1862625* and *rs12697152* compared to T47D cells. However, while

the pattern of luciferase activity is the same for *rs1862626* and *rs889312* compared to T47D cells, the intensity of the major allele for *rs889312* is much higher and the intensity for the minor allele for *rs1862626* is much lower compared to T47D cells. This is likely due to differences in the cell environments between the two different cell lines. It is possible that different transcription factors bind to the SNPs in the two different cells or that transcription factors are expressed at different levels between the two cell lines.

The differences in luciferase activities between the resistant and susceptible alleles for the rat *Mcs1b* candidate SNPs and *rs889312* correlated SNPs are likely due to differences in transcription factors binding to the alleles. Therefore, we looked at the patterns of DNA binding proteins binding to the rat and human SNP alleles.

B. EMSAs of rat *Mcs1b* candidate SNPs and *rs889312* correlated SNPs

We used electrophoretic mobility shift assays (EMSAs) to determine if there are any DNA binding proteins binding to the SNP regions and if any bind differentially between the WF and COP alleles. We initially focused on using T47D nuclear extracts since, the luciferase assays for all three *Mcs1b* candidate SNPs are different between the two alleles. The EMSAs for the three *Mcs1b* candidate SNPs using T47D nuclear extracts are shown in Figure 14 A and B. We used oligos that contained the SNP nucleotide and 12bps flanking on e ither side, for a total of 25 bps. There are several DNA-protein complexes present for *A046-SNP-A*. One predominant DNA-protein complex (marked with arrow in Figure 14B) can be competed out with cold probe, indicating that this is a specific DNA –protein interaction. The EMSA results indicate that there might be a difference in the intensity of this band between the WF and COP allele, suggesting that


A
A074-SNP-17 COP
A074-SNP-17 WF
A074-SNP-18 COP
A074-SNP-18 WF
A046-SNP-A COP
A046-SNP-A WF
200x cold competitor - + - + - + - + - -
Protein extract + + + + + + + + + +

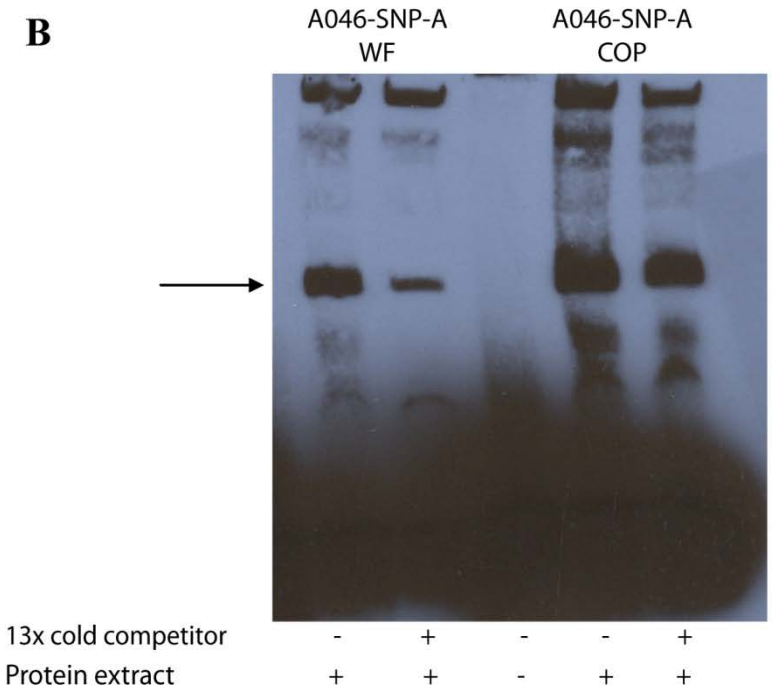

B
A046-SNP-A WF
A046-SNP-A COP
13x cold competitor - + - - +
Protein extract + + - + +

Figure 14. E MSAs for *A074-SNP-17*, *A074-SNP-18* and *A046-SNP-A* using T47D nuclear extracts. DNA- protein complexes were run on a 5% polyacrylamide gel. Protein extracts were T47D nuclear extracts. **A)** EMSAs and competition EMSAs for all three *Mcs1b* SNPs. No competition EMSA for *A046-SNP-A*. Arrow indicates difference between WF and COP alleles for *A074-SNP-17* **B)** Competition EMSA for *A046-SNP-A*. Arrow indicates band of interest for *A046-SNP-A*.

the protein might not bind as tightly to the WF allele. The same is true for *A074-SNP-18*. There are several specific DNA- protein interactions. There may be a difference in the intensity of the band between the WF and COP alleles. There are several specific DNA-protein interactions for *A074-SNP-17*, however, one band is only found in the WF allele and not the COP allele (marked with arrow in Figure 14A). This suggests that this DNA-protein interaction is unique to the WF allele and may represent a transcription factor that binds only to the WF and not the COP allele. This makes *A074-SNP-17* a great candidate for the *Mcs1b* region, since it not only appears to be involved in gene regulation but it also shows a differential pattern of DNA binding proteins for the two rat alleles.

We next performed the EMSA experiments using mutant oligos to determine if any of the DNA-protein binding complexes seen in Figure 14 are due to proteins binding directly to the SNP site and not to the rest of the oligo. We initially tested several different mutant oligos for *A046-SNP-A* to determine which would be the best in reducing specific DNA-protein interactions. We designed the oligos so that the SNP nucleotide was either deleted or replaced by random sequence. The results are shown in Figure 15. Overall, the 3bp deletion oligo worked the best at reducing the intensity of the band of interest for *A046-SNP-A*. A 3bp deletion oligo will be used in subsequent experiments which utilize mutant oligos for the *Mcs1b* candidate SNPs. Several oligos, including the 5bp deletion and 17bp random insertion oligos actually increased the intensity of the band of interest. It is not known, however, if the bands are due to the same DNA- protein complexes.

To determine which DNA- protein complexes are due to proteins binding to the SNP nucleotide, we performed an EMSA experiment using the 3bp deletion oligos

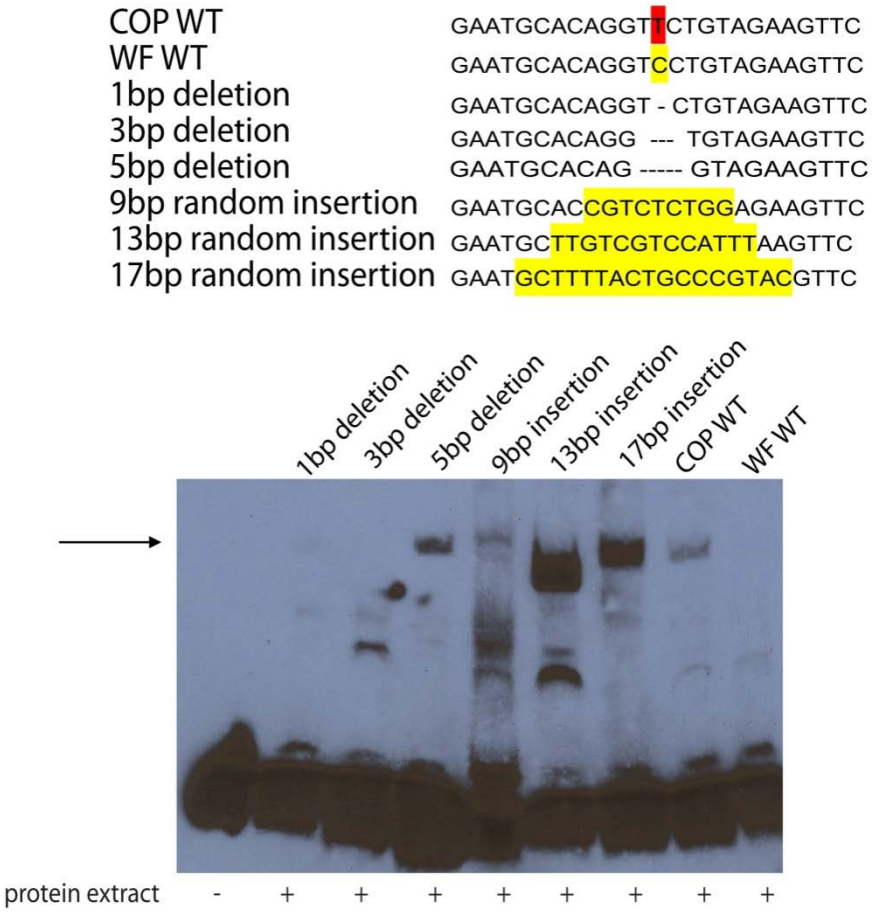
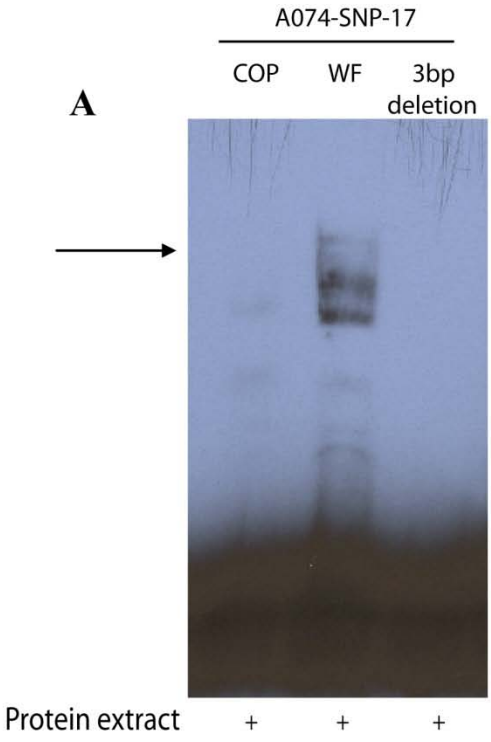


Figure 15. EMSAs with mutant oligos for *A046-SNP-A*. EMSA were run on a 5% polyacrylamide gels. T47D nuclear extracts were used for the EMSA. Mutant oligos are shown above EMSA gel. COP wild type (WT) and WF wild type were added as controls. Arrow indicates band of interest. The 3bp deletion oligo worked best at reducing the intensity of DNA-protein interaction of interest (marked by arrow) and will be used in subsequence experiments.

described above. The results for the EMSA experiments using mutant oligos for all three *Mcs1b* candidate SNPs and T47D nuclear extracts are shown in Figure 16. Figure 16A shows the results for *A074-SNP-17*. The 3bp deletion oligo results in the loss of all DNA-protein complexes, including the complex of interested (indicated by arrow in Figure 16A). The results for *A074-SNP-18* and *A046-SNP-A* indicate loss of the bands of interest (marked by arrow in Figure 16B) when using the 3bp deletion oligo. This indicates that all DNA-protein complexes for *A074-SNP-17* are due to proteins binding to the SNP nucleotide, while only complexes of interest for *A074-SNP-18* and *A046-SNP-A* are due to proteins binding to the SNP nucleotide. It is possible that DNA-protein complexes that are lost with usage of the 3bp deletion oligos are made up of transcription factors.

We next performed the EMSA experiment using MDA-MB-231 nuclear extract as a comparison to the results using T47D nuclear extracts. The results are shown in Figure 17. Note, instead of performing a competition EMSA to identify specific DNA-protein complexes, we used 3bp deletion oligos to determine complexes that bind directly to the SNP nucleotide. Overall, there are fewer DNA-protein complexes when compared to the amount of DNA-protein complexes when using T47D nuclear extracts. *A074-SNP-17* and *A074-SNP-18* show one predominant band that is lost when using the 3bp deletion oligo (arrow #2 in Figure 17). *A046-SNP-A* shows several DNA-protein complexes. The DNA-protein complex marked with arrow #2 in Figure 17 has the same intensity in both the WF and COP allele, but is lost when the 3bp deletion oligo is used. This predominant band migrates at the same size as the predominant band for *A046-SNP-A* using T47D nuclear extracts (data not shown). There is a difference in the intensity of this band


A074-SNP-17
COP
WF
3bp deletion
A
Protein extract
+
+
+

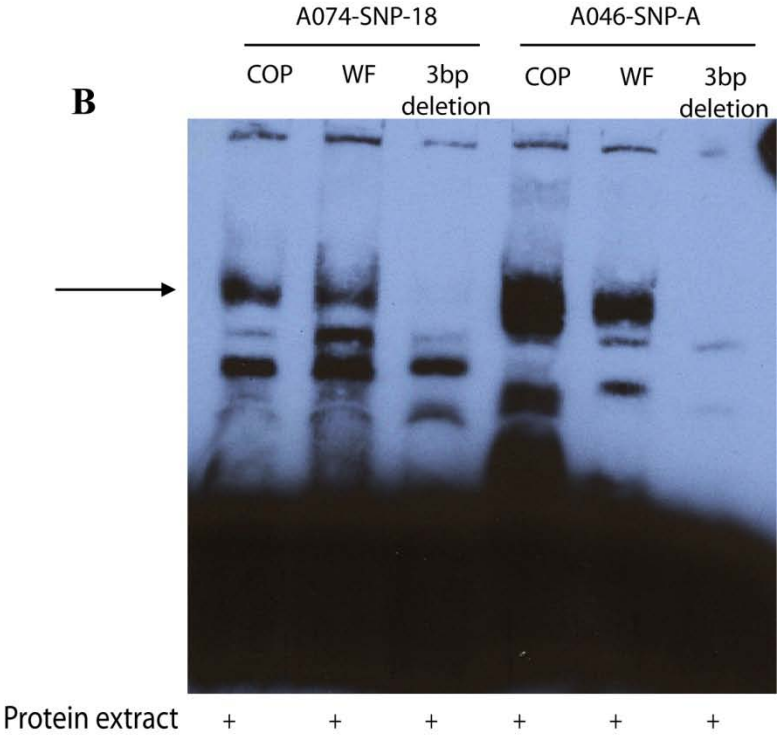

A074-SNP-18
A046-SNP-A
COP
WF
3bp deletion
COP
WF
3bp deletion
B
Protein extract
+
+
+
+
+
+

Figure 16. EMSA of *Mcs1b* candidate SNPs using 3bp deletion oligos and T47D nuclear extracts. EMSA was run on a 5% polyacrylamide gel using T47D nuclear extract. Wild type COP and WF and 3bp deletion probes were used to identify DNA protein complexes that are due to proteins binding to the SNP nucleotide. Arrows indicate bands of interest. **A)** EMSA for *A074-SNP-17* using 3bp deletion oligo. Band of interest is marked with an arrow. All DNA- protein complexes are lost when the SNP nucleotide and 2bp surrounding nucleotides are deleted. **B)** EMSA for *A074-SNP-18* and *A046-SNP-A* using 3bp deletion oligo. Bands of interest are indicated by arrow. Deleting SNP nucleotide results in the loss of the band of interest.

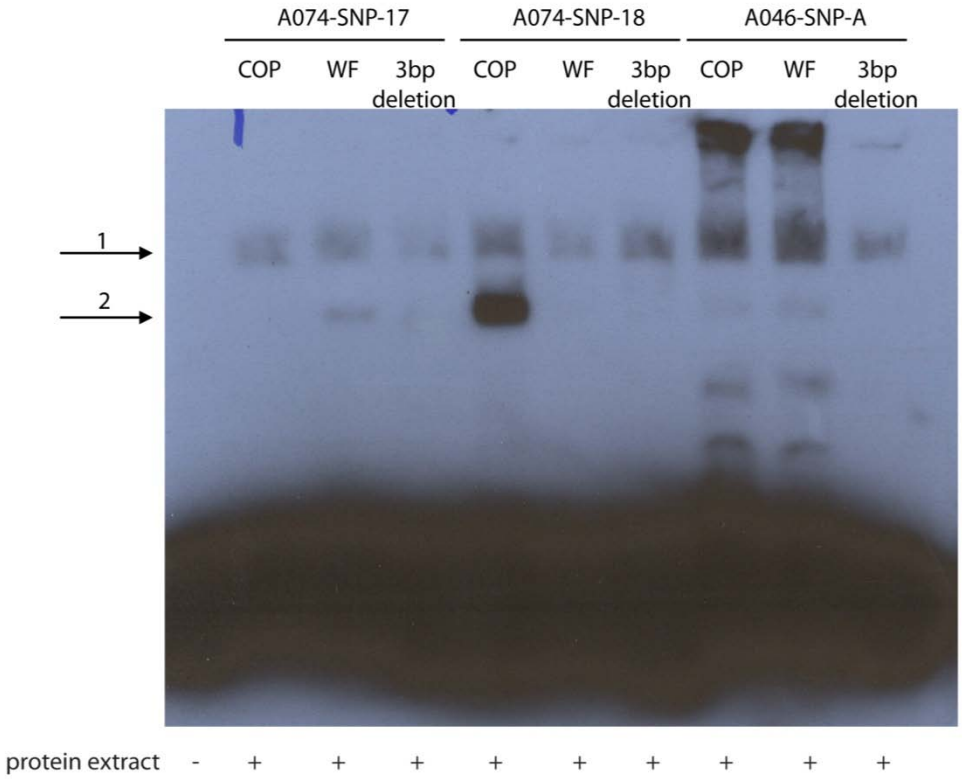


Figure 17. EMSA of *Mcs1b* candidate SNPs using 3bp deletion oligos and MDA-MB-231 nuclear extracts. EMSA was run on a 5% polyacrylamide gel using MDA-MB-231 nuclear extract. Wild- type COP and WF and 3bp deletion probes were used to identify DNA protein complexes that are due to proteins binding to the SNP nucleotide. Bands of interest are marked by arrows. One predominant band (marked by arrow 1) is found in all samples and cannot be competed out, indicating a non-specific interaction. A predominant band (marked by arrow 2) is found in all samples and can be competed out, indicating a specific interaction.

between the WF and COP allele when using T47D nuclear extracts, but the difference in intensity is lost when using MDA-MB-231 nuclear extracts. The luciferase activities for the COP and WF alleles of *A046-SNP-A* were different in T47D cells but the same in MDA-MB-231 cells. It is possible that a transcription factor is not present in MDA-MB-231 cells that is involved in binding to the *A046-SNP-A* SNP and functions as to regulate gene expression. Interestingly, an additional band is found in EMSAs using MDA-MB-231 extracts (marked by arrow #1 in Figure 17). This protein complex binds to all three *Mcs1b* candidate SNPs and is not removed when using the 3bp deletion oligos, suggesting that this is a non-specific interaction.

We also determine the protein binding pattern for two of the four candidate *rs889312* correlated human SNPs using EMSAs. We focused on *rs1862626* because it has a similar pattern of luciferase activity as our candidate *Mcs1b* SNP *A074-SNP-17*. We also performed an EMSA experiment for *rs889312* since it is the SNP used in the initial breast cancer GWA study and the major allele luciferase activity is affected greatly by the cellular environment. The results are shown in Figure 18. There is one predominant band in the *rs1862626* EMSA for the major allele (Figure 18A, marked with arrow). This band can be competed out with increasing amounts of cold probe, indicating that this is a specific DNA- protein interaction. There is one predominant band when using the minor allele probe. This band can also be competed out with increasing amounts of cold competitor probe. There are four bands visible in the *rs889312* EMSA. All of which can be competed out with increasing cold competitor probes (Figure 18B, marked with arrows #1- #4). There appear to be no DNA-protein complexes that are unique to one of

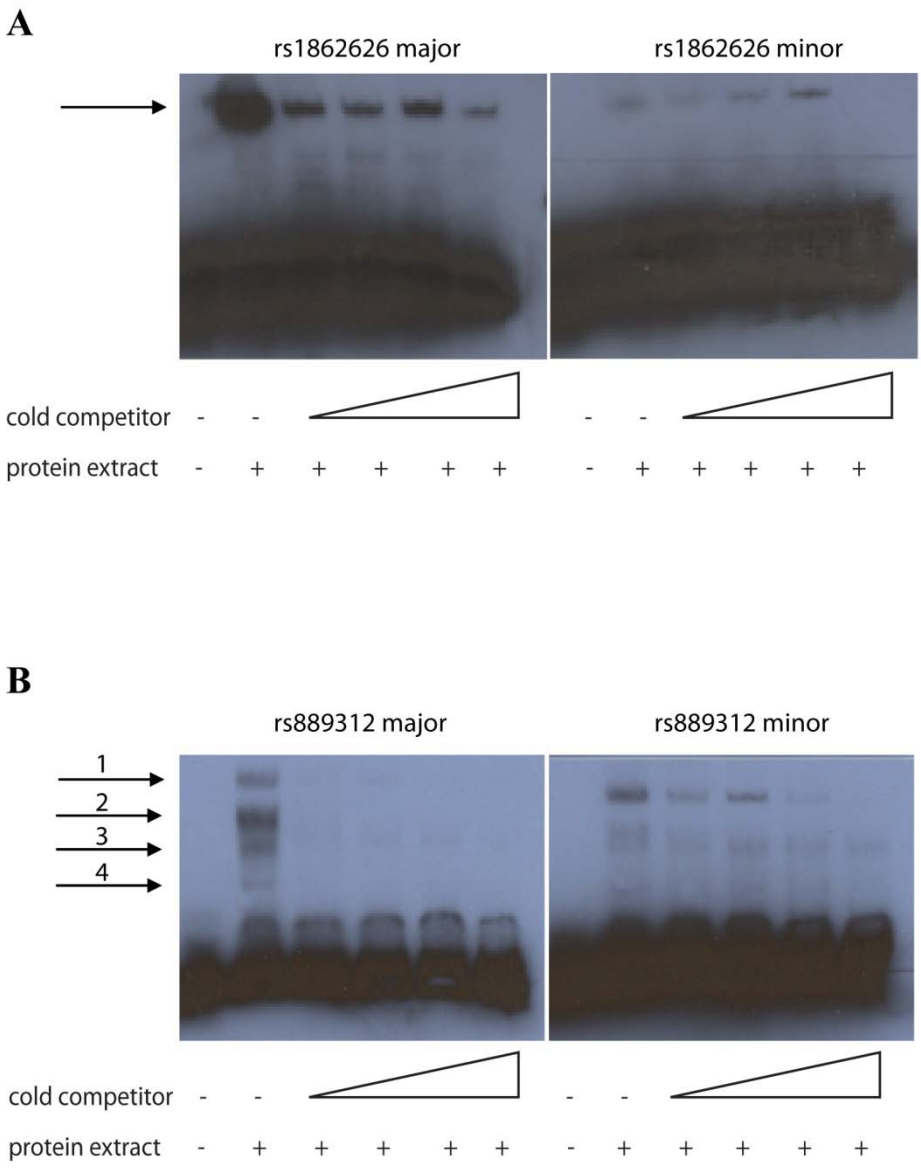


Figure 18. EMSAs for *rs889312* correlated SNPs *rs1862626* and *rs889312*. EMSAs were run on 5% polyscrylamide gel. T47D nuclear extracts were used. Cold competitor concentration was increased from 10x to 200x. **A)** EMSA for *rs1862626* showing one predominant band marked with arrow. **B)** EMSA for *rs889312* with several predominant bands that can be competed out. Bands of interest are marked with arrows.

the alleles in the *rs889312* EMSA. It is possible that there is a difference in the intensity between the two alleles which indicates a difference in the affinity of a DNA binding protein for one of the alleles. There is one DNA-protein binding complex that is found in the *rs1862626* EMSA for both the major and minor allele. It is possible that there is a difference in the intensity of this band between the two alleles, which would indicate a difference in the affinity of the DNA binding protein for the two alleles. It is possible that this DNA-protein complex is responsible for the differential luciferase activity between the major and minor alleles of *rs1862626.*

C. Identification of candidate proteins binding to *A074-SNP-17*

A074-SNP-17 shows a DNA binding protein pattern that is different between the WF and COP alleles. In particular, the WF allele for *A074-SNP-17* shows an extra band in an EMSA using T47D nuclear extract, indicating there is a unique DNA- protein complex binding the WF allele (Figure 14A). Also, the luciferase activity between the WF and COP allele for *A074-SNP-17* is different, suggesting that there might be different transcription factors binding to the two alleles. We wanted to identify which transcription factors bind to the WF and COP alleles for *A074-SNP-17.* We initially used the online program TF SEARCH to identify any DNA binding proteins that are predicted to bind the WF and COP alleles differentially. The results are listed in Table 19. We focused on two transcription factors, c-myc/max and Skn-1, an ortholog of the NRF1, 2 and 3 proteins, because of their role in breast cancer. C- MYC is an important transcription factors that is involved in regulating 15% of all human genes. It is a known proto-oncogene and is involved in cell growth, transformation, angiogenesis and cell- cycle control [140, 141].

Table 19. TF SEARCH results for *A074-SNP-17*. Scores indicate percentage of how close the sequence is to the consensus sequence. Multiple scores indicate multiple predicted binding sites.

Protein	Name	COP score	WF score	Function
SN	Snail	-	86.4	
MyoD	Myblast determining factor	79.8	86.2/79.1	
Skn-1	NRF1,2,3 ortholog	89/77	85.2	Transcription factor
c-myc/max		-	85	Transcriptional activator
USF	Upstream transcription factor 1	-	83.8/82.6/78.3/75.9	Transcriptional activator
Mat alpha	Mating factor alpha 2	-	83.5/77.2	
Cap	Cap signal transcription initiation	83.1/80	80.0/78.7	
N-Myc		-	78.4/76.3	
COUP-TF	COUP-TF/HNF4 heterodimer	81.8	-	
CdxA		78.2	-	
E2F		77	-	Transcription factor
ROR alpha	RAR related orphan receptor alpha	76.9	-	Transcriptional activator
Cre-BP	Cre binding protein 1/ c-jun heterodimer	75.6	-	Transcriptional activator

We performed an EMSA supershift assay to determine if c-MYC binds to the WF and COP alleles of *A074-SNP-17*. The results are shown in Figure 19. There is no supershift of any bands of the COP and WF *A074-SNP-17* alleles, indicating that c-MYC does not bind to *A074-SNP-17* (Figure 19A). However, there is also no supershift for the E-box positive control. E-box is the name of the c-MYC binding site consensus sequence. The consensus sequence is CAC(G/A)TG and is found in the positive control, which has previously been used successfully in a c-MYC supershift assay [142]. We tested our c-MYC antibody (Santa Cruz, sc-764) in a Western blot to determine if it is functional and if c-MYC is expressed in T47D and MDA-MB-231 cells (Figure 19B). The antibody can be used in a western blot and the two breast cancer cell lines express nuclear c-MYC. It is possible that the antibody is not efficient for performing supershift assay and therefore we did not see any supershifts in the positive control. We co-transfected an expression vector containing the human c-MYC gene with the luciferase constructs for *A074-SNP-17*. As shown in Figure 19C, there is no effect of c-MYC on the luciferase activity of the *A074-SNP-17* alleles, further indicating that c-MYC probably does not bind to *A074-SNP-17*. However, it may also be possible that c-MYC is expressed at high levels in T47D cells and further overexpression does not have an effect on *A074-SNP-17* luciferase activity.

We also tested another protein predicted to bind the WF and COP alleles of *A074-SNP-17*. Both the COP and WF allele are predicted to bind the protein Skn-1, which is the *C.elegans* ortholog of the NRF1, 2, 3 proteins. However, the *A074-SNP*-17 COP allele is predicted to bind this protein more often and with a higher affinity (Table 19). Nuclear factor (erythroid derived)- like 2 (NRF2) is a known transcription factor

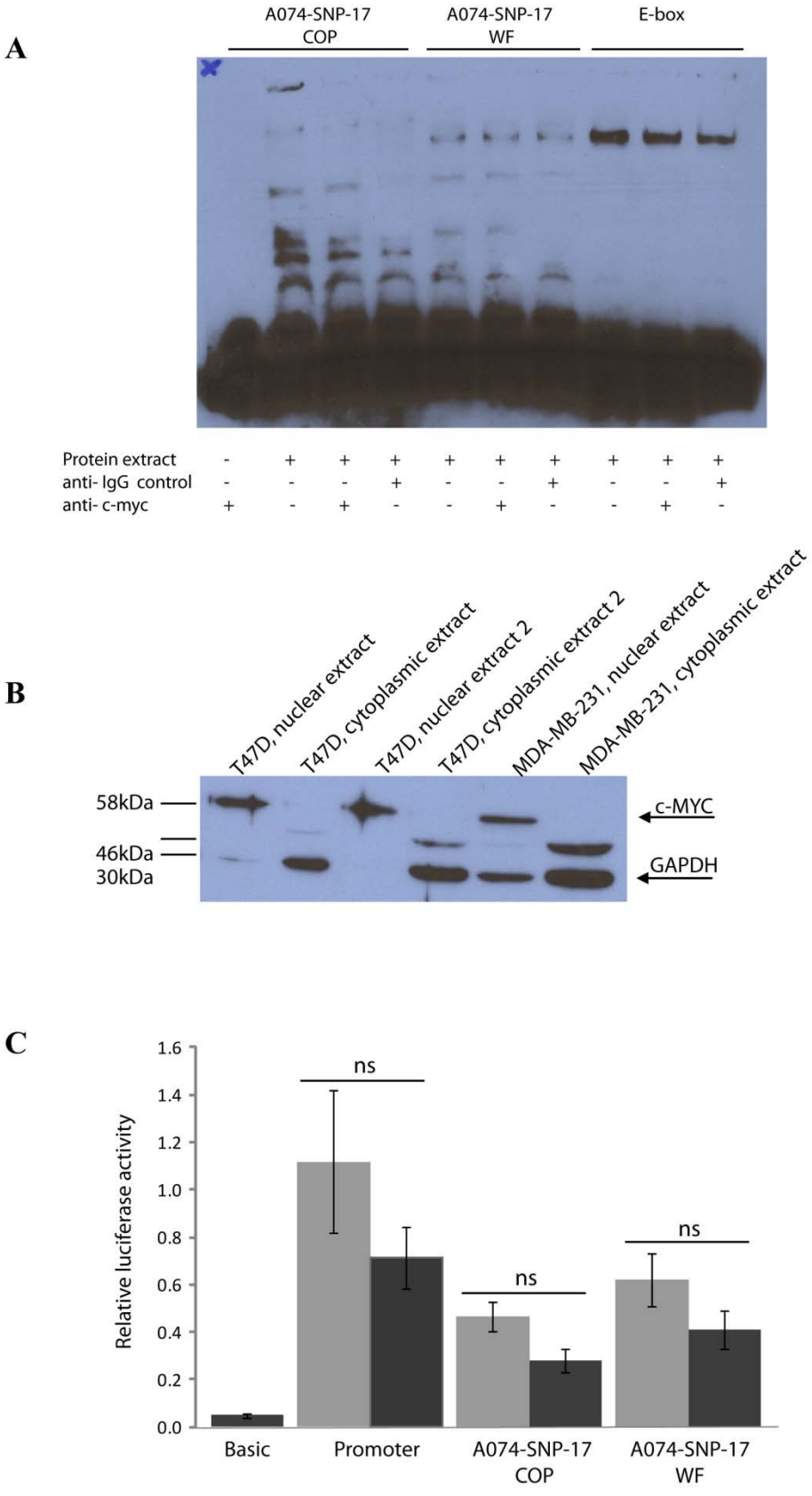

A
A074-SNP-17
COP
A074-SNP-17
WF
E-box
Protein extract - + + + + + + + + +
anti- IgG control - - - + - - + - - +
anti- c-myc + - + - - + - - + -
B
T47D, nuclear extract
T47D, cytoplasmic extract
T47D, nuclear extract 2
T47D, cytoplasmic extract 2
MDA-MB-231, nuclear extract
MDA-MB-231, cytoplasmic extract
58kDa
46kDa
30kDa
c-MYC
GAPDH
C
Relative luciferase activity
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
ns
ns
ns
Basic
Promoter
A074-SNP-17
COP
A074-SNP-17
WF

Figure 19. Analysis of c-MYC binding to *A074-SNP-17*. A) Supershift EMSA for c-MYC binding to *A074-SNP-17*. Supershift EMSA was run on a 5% polyscrylamide gel. E-box oligo contains consensus site for c-MYC. No supershift for *A074-SNP-17* rat alleles or for the E-box consensus sequence was observed. **B)** Western blot for c-MYC using T47D and MDA-MB-231 nuclear and cytoplasmic extracts. T47D and MDA-MB-231 cells express c-MYC and the antibody is specific in a Western blot for c-MYC. **C)** Luciferase assays for co-expression of *A074-SNP-17* luciferase constructs and c-MYC expression vector in T47D cells. ( ) Light grey bars indicate luciferase activity of *A074-SNP-17* allele when co-expressed with empty expression vector. ( ) Dark grey bars indicate luciferase activity of *A074-SNP-17* allele when co-expressed with c-MYC expression vector. Error bars indicate standard error. Ns indicates non- significance. Luciferase activity of the *A074-SNP*-17 allele co-expressed with the empty vector was compared to the luciferase activity of *A074-SNP*-17 co-expressed with the c-MYC expression vector. C-MYC does not have an effect on the luciferase activity of *A074-SNP-17*.

involved in cancer. Under normal conditions, NRF2 is kept in the cytoplasm by the protein KEAP1 and targeted for degradation. Under cellular stress, such as oxidative stress, KEAP1 releases NRF2 and NRF2 locates to the nucleus, where it acts as a transcription factor. There are basal levels of NRF2 found in the nucleus, even under normal conditions [143, 144]. NRF2 can act as a tumor suppressor and oncogene depending on the context. NRF2 protects cells from cytotoxic stress but NRF2 also regulates genes that are involved in cell survival under stress conditions [144]. We performed a supershift EMSA to determine if NRF2 binds to the *A074-SNP-17* alleles using T47D nuclear extract. The results are shown in Figure 20A. NRF2 binds to DNA sequences called Antioxidant Response Elements (AREs). We used a positive control for the NRF2 supershift that contains an ARE [145]. There is a supershift in the presence of NRF2 antibody (Santa Cruz, sc-722x) for all three oligos, marked by an arrow in Figure 20A. This suggests that NRF2 binds both the COP and WF allele of *A074-SNP-17*. The band that is shifted upon addition of NRF2 antibody is of a high molecular weight. We ran the EMSA supershift on a 4% polyacrylamide gel to better visualize the band. However, in competition EMSAs for *A074-SNP-17*, this band could not be competed out, suggesting that this interaction is not specific (data not shown). We therefore designed a random 25bp oligo and performed a supershift EMSA for NRF2. As seen in Figure 20B, using the random oligo also results in a supershift for NRF2. This indicates that NRF2 binds these oligos non-specifically or the antibody cross-reacts with another protein that binds the oligos in a non-specific manner. To ensure that NRF2 does not affect the gene regulation activity of *A074-SNP*-17, we performed a co-expression experiment using a NRF2 expression vector and pGL3 vectors containing an ARE and the COP and WF

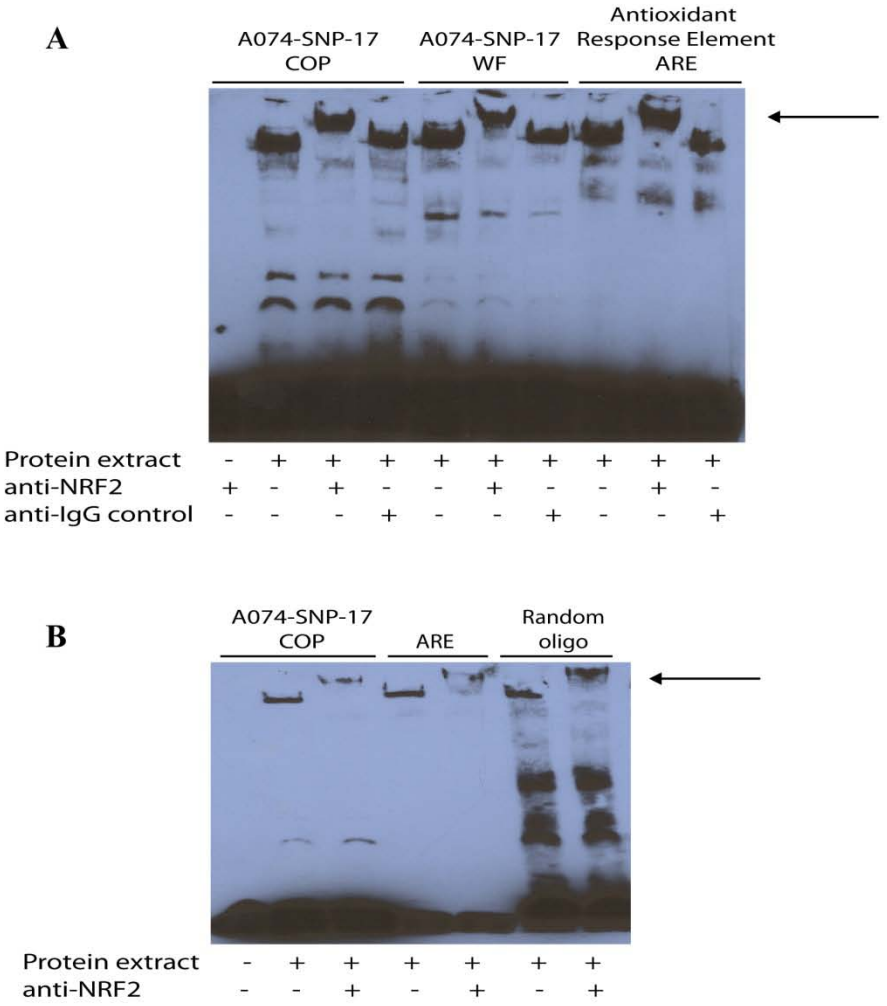

A
A074-SNP-17 COP
A074-SNP-17 WF
Antioxidant Response Element ARE
Protein extract - + + + + + + + + +
anti-NRF2 + - + - - + - - + -
anti-IgG control - - - + - - + - - +
B
A074-SNP-17 COP
ARE
Random oligo
Protein extract - + + + + + +
anti-NRF2 - - + - + - +
C
Relative luciferase activity
20
18
16
14
12
10
8
6
4
2
0
*
ns
ns
*
Basic
Promoter
A074-SNP-17 COP
A074-SNP-17 WF
Antioxidant ResponseElement ARE

Figure 20. Analysis of NRF2 binding to *A074-SNP-17.* EMSAs were run on a 4% polyacrylamide gel. **A)** Supershift EMSA for NRF2 binding to *A074-SNP-17.* ARE oligo contains consensus site for NRF2. There is a supershift present for both rat alleles for *A074-SNP-17* and the ARE consensus site (indicated by arrow) **B)** Supershift EMSA for a random oligo and NRF2. The NRF2 antibody supershifts a random oligo, indicating that the interaction between the oligos and NRF2 is not specific. **C)** Luciferase assays for co-expression of *A074-SNP-17* luciferase constructs and NRF2 expression vector in T47D cells. Co-expression of pGL3 vectors and an empty expression vector is shown using light grey bars ( ). Co-expression of pGL3 vectors with a NRF2 expression vector is shown with dark grey bars ( ). Luciferase activities of co-expression with empty vector was compared to luciferase activity of NRF2 co-expressed with pGL3 vectors. Error bars indicate standard error. Asterisk indicates significance, while ns indicates non-significance. There is no effect of the NRF2 over-expression on the *A074-SNP-17* pGL3 luciferase activity.

A074-SNP-17 alleles. There results are shown in Figure 20C. The luciferase activity of the ARE containing luciferase vector is increased when NRF2 is overexpressed in T47D cells. However, the luciferase activities for *A074-SNP-17* are not changed in the presence of NRF2, suggesting that this protein has no effect on the gene regulation activity of *A074-SNP-17*. Note, NRF2 reduces the luciferase activity of the pGL3-promoter vector, possibly through regulation of another protein that binds the SV40 promoter.

Since we were not able to identify any proteins binding to the *A074-SNP-17* alleles using TF SEARCH, we performed a DNA pulldown and mass spectronomy analysis. The DNA pulldown was performed using T47D nuclear extract and the *A074-SNP-17* COP, WF and 3bp deletion oligos that were used for EMSAs. We lost the DNA-protein complex that is unique to the WF allele in the process of the DNA pulldown (data not shown). This is likely due to the protein being of low abundance and getting removed during the stringent wash steps. We therefore concentrated on proteins that were found in both the COP and WF samples but not in the 3bp deletion sample. Results are shown in Table 20. This procedure resulted in the identification of many proteins binding to both the WF and COP alleles of *A074-SNP*-17. Therefore, results were filtered for proteins that are expressed in the nucleus, are expressed in T47D cells, have known transcriptional regulation function, have known DNA binding domain and are known to be involved in cancer. We chose to test three of these proteins in supershift EMSA experiments. These are the progesterone receptor (PR), nuclear factor 1 C -type (NFIC) and interleukin enhancer- binding factor (ILF2).

The COP allele of *A074-SNP-17* pulled down the progesterone receptor A isoform, while the WF allele pulled down the delta 3+ 6/2 isoform of the progesterone

Table 20. Mass spectrometry results for *A074-SNP-17* using T47D nuclear extracts.

ID	Abbreviation	DNA binding domain	Function
Isoform A of Progesterone receptor	PR-A	C4- type zinc fingers	Transcriptional regulation
Delta 3+6/2 progesterone receptor	PR	C4- type zinc fingers	Transcriptional regulation
Isoform 3 of Nuclear factor 1 C-type	NFIC	CTF/ NF-1	Transcriptional regulation
Interleukin enhancer-binding factor 2	ILF2	DZF	Transcriptional regulation
Cell division cycle and apoptosis regulator protein 1	CCAR1	SAP	Transcriptional regulation
RuvB-like 2	RUVBL2	-	Transcriptional regulation
Isoform 2 of Casein kinase I isoform delta	CSNK1D	-	serine/threonine-protein kinase
Superkiller viralicidic activity 2-like 2	SKIV2L2	-	mRNA splicing

receptor. The progesterone receptor is activated by progesterone and is involved in the regulation of various genes. Progesterone has proliferative and carcinogenic effects on breast tissue [146]. The progesterone receptor B differs from the A isoform through an additional 164 amino acids at the N-terminal. These 164 amino acids contain an additional transcription activation function. The two isoforms of the progesterone receptor have distinct functions and the isoform A has been shown to be a transdominant inhibitor of the B isoform and shows anti-estogenic effects [147]. The delta 3+ 6/2 isoform of the progesterone receptor lacks exon 3 and 52bps in exon 6. It is unknown how this affects the function of the protein [148]. The progesterone receptor binds to progesterone response elements (PRE), which are characterized by the sequence 5'GNAGANNNTGTNC'3 [149, 150]. We used an antibody that can recognize both A and B isoforms of the progesterone receptor for the supershift EMSA experiment (Santa Cruz, sc-538x). The results are shown in Figure 21A. There was a supershift detected for the PRE positive control (marked by arrow in Figure 21A), suggesting that the supershift worked correctly. However, there was no supershift detected for the *A074-SNP-17* alleles, suggesting that the progesterone receptor does not bind to *A074-SNP-17*.

We also performed a supershift EMSA for the nuclear factor 1 C (NFIC) protein. NFIC is part of the nuclear factor 1 family of transcription factors. Other members include NFIA, NFIB and NFIX. These transcription factors bind to the sequence 5'-TTGGCNNNNNGCCAA-3 [151]. A positive control has been generated from this sequence and is named NFIC consensus binding site in Figure 21B. No supershift was detected for the *A074-SNP-17* and positive control oligos. There is a large band for the

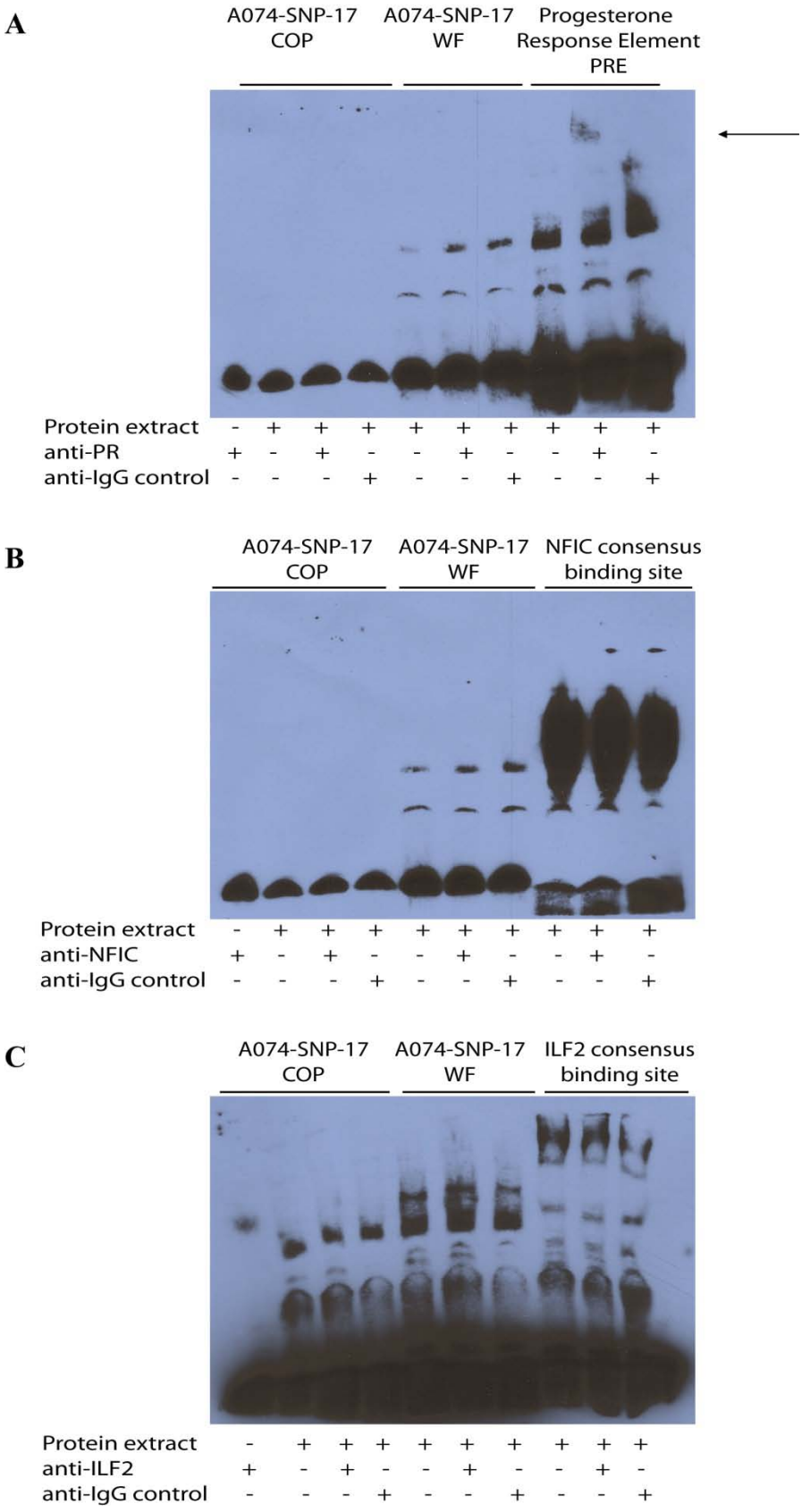

A
A074-SNP-17 COP
A074-SNP-17 WF
Progesterone Response Element PRE
Protein extract - + + + + + + + + +
anti-PR + - + - - + - - + -
anti-IgG control - - - + - - + - - +
B
A074-SNP-17 COP
A074-SNP-17 WF
NFIC consensus binding site
Protein extract - + + + + + + + + +
anti-NFIC + - + - - + - - + -
anti-IgG control - - - + - - + - - +
C
A074-SNP-17 COP
A074-SNP-17 WF
ILF2 consensus binding site
Protein extract - + + + + + + + + +
anti-ILF2 + - + - - + - - + -
anti-IgG control - - - + - - + - - +

Figure 21. Supershift EMSAs of PR, NFIC and ILF2 for *A074-SNP-17.* EMSAs were run on a 5% polyacrylamide gel. T47D nuclear extract was used for all EMSAs. **A)** supershift EMSA for the progesterone receptor (PR). There is a supershift for the positive PRE control oligo but not for the *A074-SNP-17* alleles. Arrow indicates supershift **B)** Supershift EMSA for nuclear factor I C (NFIC). There is no supershift detected for any of the oligos used. **C)** Supershift EMSA for interleukin enhancer binding factor (ILF2). There is no supershift detected for any of the oligos used.

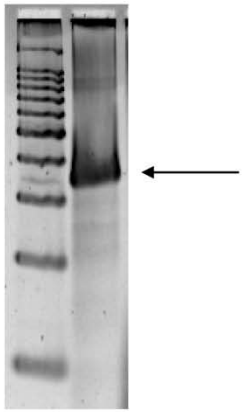
NFIC positive control oligo. All four members of the nuclear factor 1 family bind to the same consensus sequence and it is possible that a supershift overlaps this large band. The antibody (Abcam, ab86570) used has not been tested for gel supershifts and may not be appropriate to use for this experiment.

The *A074-SNP-17* alleles pulled down the interleukin enhancer binding factor 2 (ILF2) protein. ILF2 is the 45kda subunit of the nuclear factor of activated T-cells (NFAT) and forms a complex with interleukin enhancer binding factor 3 (ILF3). Together these proteins work as transcription factors for a variety of genes [152]. A positive control that contains a ILF2 consensus binding site was used for the supershift EMSA [153]. There were no supershifts detected for any of the oligos, including the positive control oligo (Figure 21C). This may be due to the antibody used having not been tested in supershift EMSAs and may be inappropriate for this method (Abcam, ab28772). It is possible that ILF2 does not bind to the *A074-SNP-17* alleles, however, this supershift EMSA is inconclusive. Note, there appears to be an extra band in the WF allele when ILF2 antibody is added. However, the ILF2 antibody gives off a background chemiluminescence (lane 1 of the supershift EMSA) that runs at the same size as the extra band for the WF allele. This means the extra band is likely to be background chemiluminescence from the ILF2 antibody. We did not have access to expression vectors for the three genes tested and therefore could not perform any co-expression studies with protein expression vectors and the *A074-SNP-17* luciferase assay vectors.

D. Luciferase assays for *Mier3* promoter

We previously performed luciferase assays for the three candidate *Mcs1b* SNPs in a pGL3- Promoter vector, which contains a strong SV40 promoter. We wanted to test how the *Mcs1b* candidate SNPs influence an endogenous promoter. We chose to use the promoter of the *Mesoderm induction early response protein 3* (*Mier3)*, since it is the candidate gene for the *Mcs1b* locus. *Mier3* is expressed higher in the mammary glands of 12-week old WF compared to COP females both with and without DMBA [87]. There are multiple annotated *Mier3* mRNAs in public databases, so we wanted to determine the transcription start site (TSS) for the *Mier3* gene in WF mammary glands. There is no sequence difference in the *Mier3* promoter between the WF and COP rat strains, therefore we focused on the WF rat strain only. We used 5'RACE to determine the TSS for the *Mier3* gene. The only known annotated *Mier3* gene products differ in their exon 1. Therefore, we considered exon 2 and beyond common to all *Mier3* isoforms and designed primers against common regions. We used pooled RNA from the mammary glands of six WF females. Results are shown in Figure 22A. There was one predominant PCR product (marked by arrow) after 5'RACE PCR. Several other faint bands can be seen on the gel, indicating there may be multiple different splice forms of the *Mier3* gene. However, after cloning the PCR reaction, only one splice form could be found, which corresponds in size to the predominant PCR product shown in Figure 22A. The alignment of the results from the 5'RACE and the annotated *Mier3* isoforms are shown in Figure 22B. The predicted exons 1 of the isoforms are highlighted in gray. The results from the 5' RACE overlap the first annotated *Mier3* transcript, except the 5' untranslated region (UTR) is much shorter. The coding regions highlighted in the figure are predicted, it is possible that the coding regions starts downstream of the region highlighted here. There are several more

A



B

1st Annotated rat Mier3 ttgggtcttt gtcttctgag gatcatgatt ttgaccccac tgctgagatg ctggtccatg . . .

5′ Race results for Mier3 ac tgctgagatg ctggtccatg . . .

2nd annotated rat Mier 3 . . . at gactctgtgt ttcattccag ttgggtcttt gtcttctgag gatcatgatt ttgaccccac tgctgagatg ctggtccatg . . .

Figure 22. 5'RACE results for the *Mier3* gene. A) Results from outer and inner PCR specific to the *Mier3* gene. PCR reaction results in one predominant PCR product marked with an arrow. Several other PCR products formed, however, none of them were found in the sequencing results. **B)** Results for the 5'RACE of *Mier3*. The two annotated rat versions of the *Mier3* gene are shown. These versions were found in the Nucleotide NCBI database. The gray highlighted areas indicate predicted coding regions of the *Mier3* gene.

"ATG" start codons that are in- frame with the one highlighted in the figure. It is likely that this initial "ATG" is not where translation starts, but that a downstream, in-frame "ATG" is used. The second annotated *Mier3* transcript contains a much larger exon 1 and 5' UTR. Only regions that overlap with the first annotated *Mier3* transcript are shown in Figure 22B. We used the results from the 5'RACE to determine which region to clone as the *Mier3* promoter. We picked a region up to 1.5kb upstream of the TSS determined from the 5'RACE results and 0.5kb downstream of the TSS. We excised the SV40 promoter out of the pGL3 Promoter vector and replaced it with the 2kb region of the *Mier3* promoter. Interestingly, there are two separate regions containing a string of thymidines in this region. We were not able to find a clone that contained the correct sequence for both regions. We picked two independent clones. Each clone contained the correct sequence for one of the regions and performed the luciferase activity with both clones to determine if there is a difference. The results are shown in Figure 23A. There is no significant difference in the luciferase activity of the two independent *Mier3* promoter clones. This suggests that either one of the clones can be used for the analysis. However, there is a stark difference in the luciferase activity of the cloned *Mier3* promoter and the original SV40 promoter. The luciferase activity of the cloned *Mier3* promoter is much lower than the SV40 promoter activity. More importantly, there is no significant difference in luciferase activities between the cloned *Mier3* promoter and the pGL3-Basic vector. This suggests that the region of the *Mier3* gene that we cloned has no promoter activity. The cloned region is located on chr*2*: 43,002,349-43,004,390. We only cloned about 1.5kb upstream of the *Mier3* TSS. This is because there is a *BglII* site located at the 5' end of this region, making it difficult to clone more of this region using

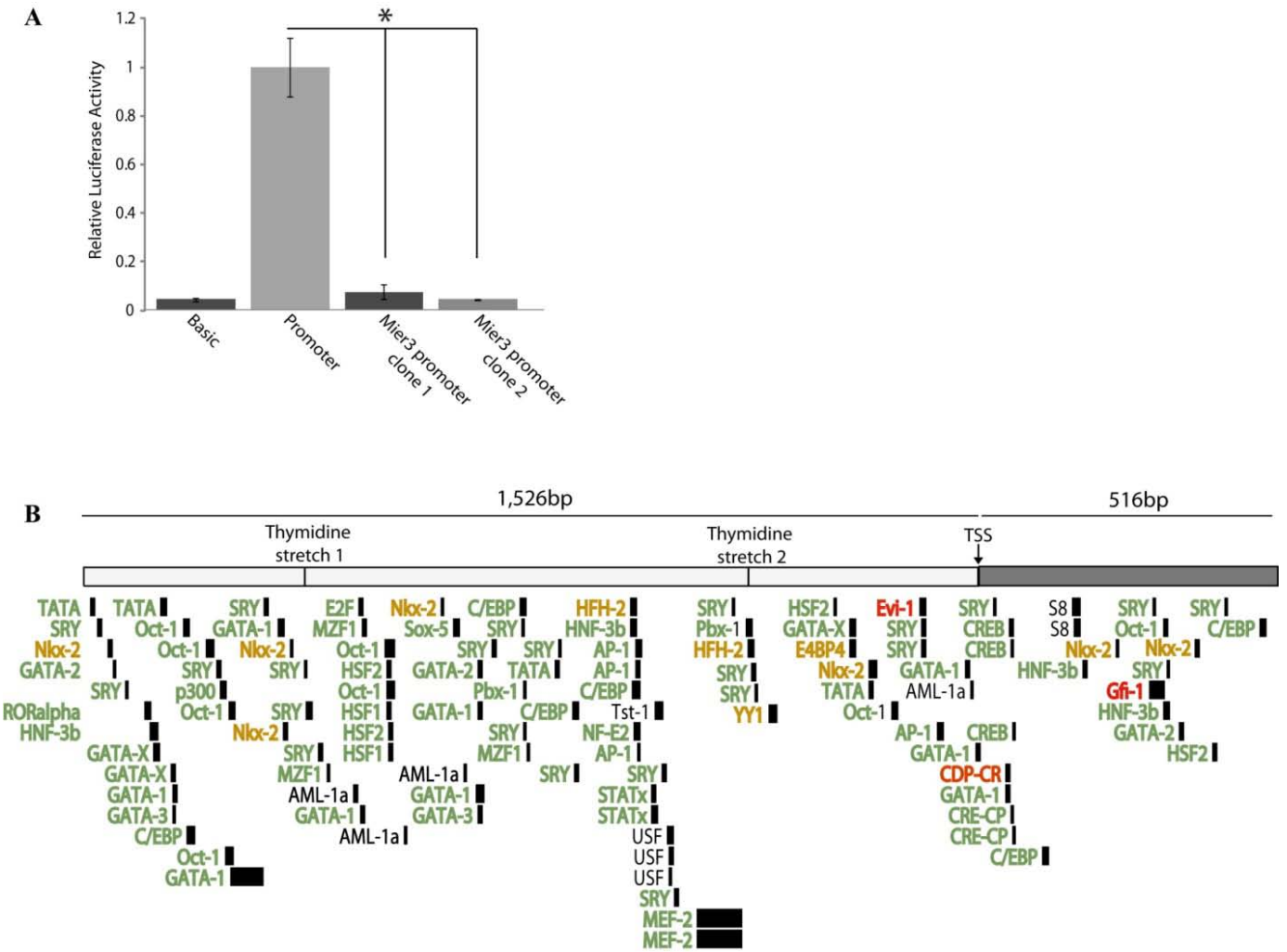

A
Relative Luciferase Activity
1.2
1
0.8
0.6
0.4
0.2
0
*
Basic
Promoter
Mier3 promoter clone 1
Mier3 promoter clone 2
B
1,526bp
516bp
Thymidine stretch 1
Thymidine stretch 2
TSS
TATA
SRY
Nkx-2
GATA-2
RORalpha
HNF-3b
GATA-X
GATA-1
GATA-3
C/EBP
Oct-1
p300
GATA-A
E2F
MZF1
HSF2
HSF1
AML-1a
Sox-5
GATA-3
HFH-2
HNF-3b
AP-1
TATA
Pbx-1
NF-E2
STATx
USF
MEF-2
Pbx-1
YY1
HSF2
E4BP4
Evi-1
CREB
Gfi-1
CDP-CR
CRE-CP
S8
Tst-1

Figure 23. Luciferase assay results for the *Mier3* promoter. A) Luciferase assay results for the Mier3 promoter. Results are pooled from three independent experiments. There is a significant difference between the SV40 and Mier3 promoter. There is no significant difference between the pGL3- Basic vector and the *Mier3* promoter. **B)** Transcription factors that bind to the cloned *Mier3* promoter using the program TF SEARCH. Transcription factors shown in green are known activators. Transcription factors shown in red are known repressors. Transcription factors shown in gold can act as both activators and repressors. The type of gene regulation is unknown for transcription factors in black. Marked on the figure are also the two thymidine stretches that could not both be sequenced correctly in a clone and the transcription start site (TSS).

this enzyme. Shown in Figure 23B are transcription factors predicted to bind to the cloned region using the program TF SEARCH. There are several well known activators predicted to bind this region, including AP-1 and E2F. However, there are also several known repressors predicted to bind this region that could overpower the effects of any activators binding to this region. However, it is more likely that this region is not sufficient in activating transcription and that important transcriptional elements are found either upstream or downstream of the cloned region. Because the cloned *Mier3* promoter region did not have an effect on luciferase activity, we did not continue with cloning this region into the luciferase reporter vectors that contain the *Mcs1b* candidate SNP alleles. The *Mcs1b* candidate SNPs lower the luciferase activity of the SV40 promoter and we would therefore not be likely to detect any effects of the *Mcs1b* candidate SNPs on the cloned *Mier3* promoter region, since the luciferase activity of this region is already so low.

E. Luciferase assays for all three SNPs in the same pGL3-promoter vector

We initially cloned the three *Mcs1b* candidate SNPs into the pGL3- Promoter vector individually (Figures 12). We wanted to know what happens if we clone all three *Mcs1b* candidate SNPs into the same pGL3- Promoter vector. The idea is that we wanted to test if the luciferase activity of any one of the SNPs will dominate the luciferase activity when all three SNP alleles are present. We cloned *A074-SNP-17* and *A074-SNP-18* into the first multiple cloning site of the pGL3- Promoter vector, upstream of the SV40 promoter. We cloned *A046-SNP-A* into the second multiple cloning site downstream of the luciferase gene. We selected this configuration because it mimics the

physical location of these SNPs along rat chromosome *2*. The configuration of this vector can be seen in Figure 24. The results for the luciferase assays can be seen in Figure 25. As a reference the luciferase assay results from each *Mcs1b* candidate SNP individually have been added to the figure. These are the same results as shown in Figure 12. The results in Figure 25A show that in T47D cells there is no difference in the luciferase activity between the WF and COP alleles when all three SNP alleles are cloned into the same vector. The COP allele activity for all three SNPs combined mimics the COP allele activity for *A074-SNP-17* individually. The WF allele activity for all three SNPs combined is much lower than for any of the three SNPs individually. However, in all three SNPs individually the WF allele lowers the luciferase activity compared to the pGL3-Promoter vector by itself. It is possible that all three SNPs have an additive effect and the low luciferase activity we see for the WF allele when all three SNP alleles are present is due to a limit as to how low the SV40 activity can be reduced. This would indicate that the SV40 promoter activity could not be reduced further than it already is by the presence of all three SNPs in the same vector. The results are similar when repeating the experiment in MDA-MB-231 cells (Figure 25B).

Interestingly, there is no difference in the COP and WF luciferase activity when all three SNPs are present in the same vector, while there is a difference when all three SNPs are tested individually. This system is very artificial and all information on the natural chromatin state within the cell is lost when using a plasmid- based system. It is possible that there is chromatin looping in this region, which could put one or more of the *Mcs1b* candidate SNPs in an insulator loop. It is also possible that one or more of these SNPs is located within a region that is methylated in the cell. Therefore, experiments will

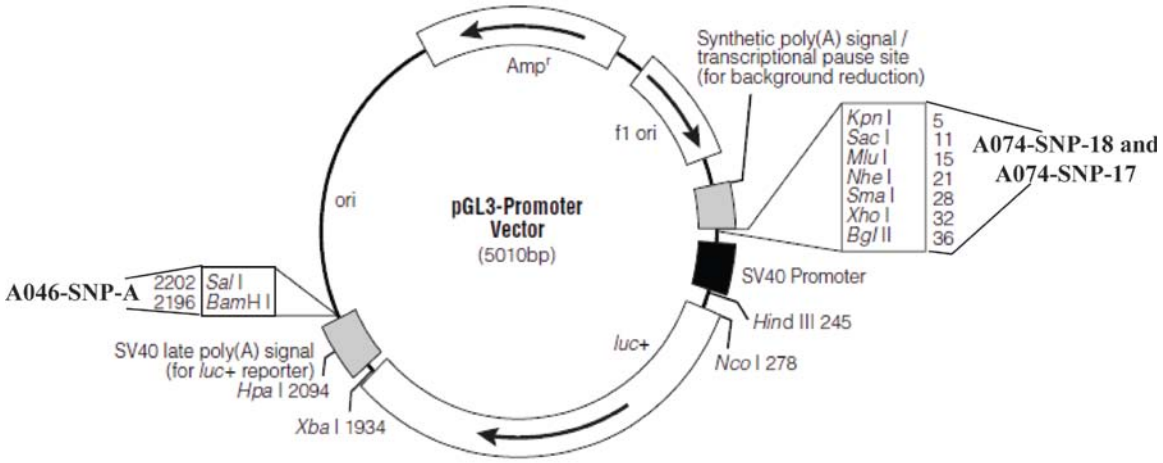
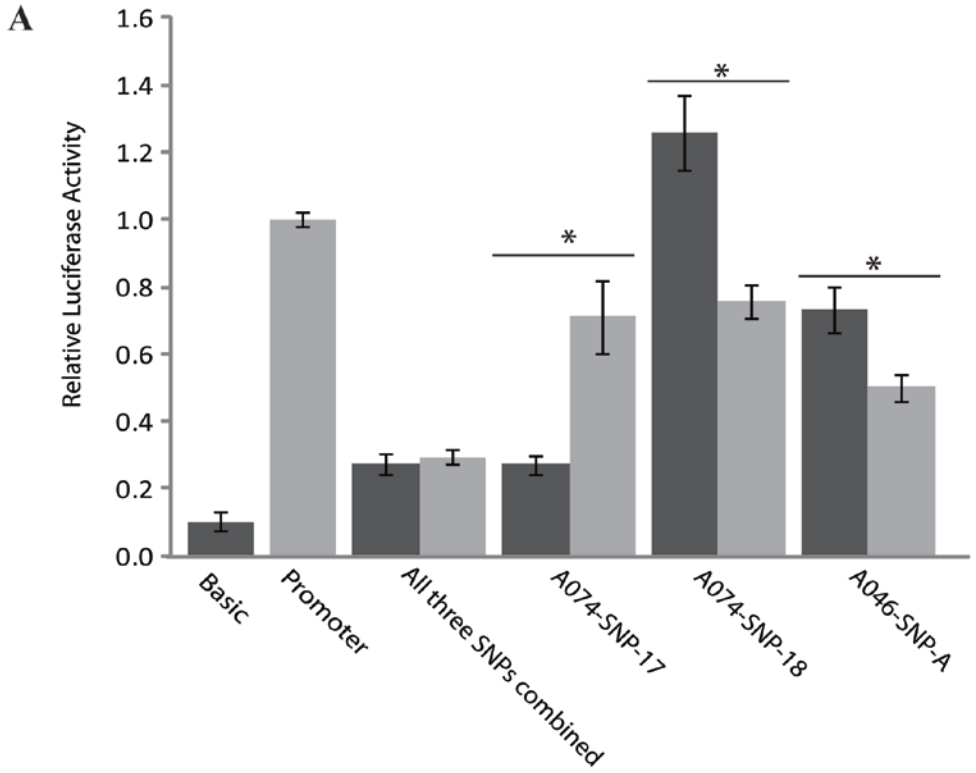


Figure 24. p GL3- Promoter vector configuration for cloning all three *Mcs1b* candidate SNPs into same vector. *A074-SNP-18* and *A074-SNP-17* were cloned into the first multiple cloning site upstream of SV40 promoter. *A046-SNP-A* was cloned into second multiple cloning site downstream of luciferase gene.


A
Relative Luciferase Activity
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
*
*
*
Basic
Promoter
All three SNPs combined
A074-SNP-17
A074-SNP-18
A046-SNP-A

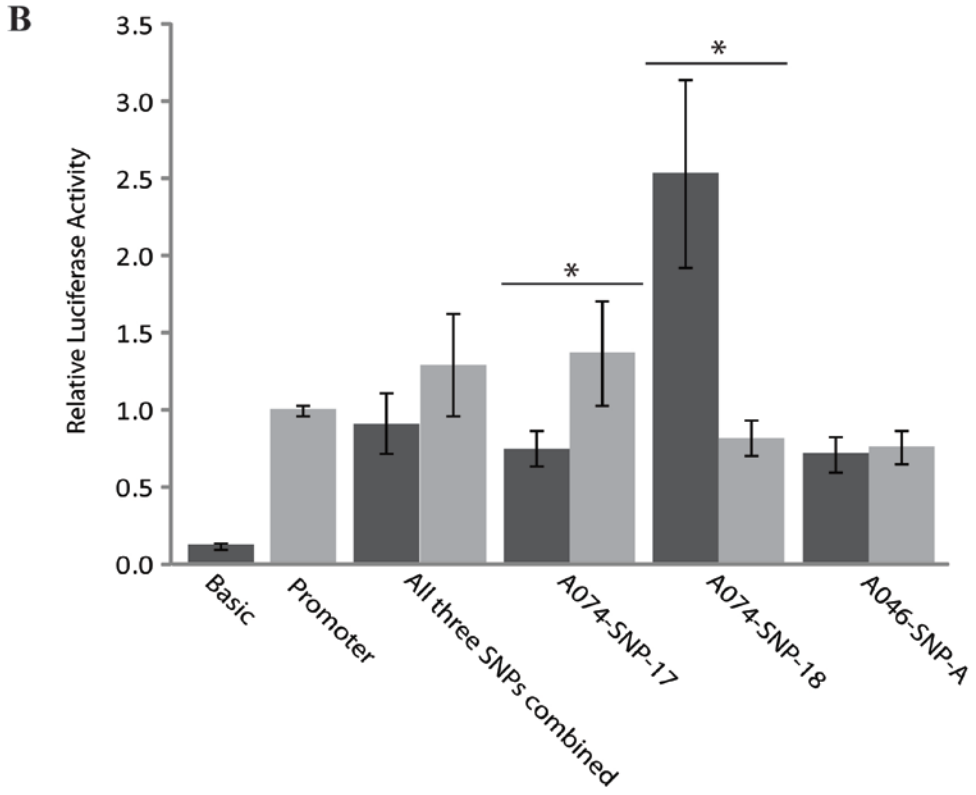

B
Relative Luciferase Activity
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
*
*
Basic
Promoter
All three SNPs combined
A074-SNP-17
A074-SNP-18
A046-SNP-A

Figure 25. Luciferase assay results for all three *Mcs1b* candidate SNPs in the same pGL3- Promoter vector. (Dark grey bars indicate adjusted relative luciferase activity for the COP (resistant) allele. (Light grey bars indicate adjusted relative luciferase activity for the WF (susceptible) allele. Asterisks indicate luciferase activity that is significantly different between the major and the minor alleles. All adjusted relative luciferase activities are the mean of nine values from three independent experiments. Errors bars indicate standard error from pooled experiments. As controls, the luciferase activity for the pGL3-Basic and pGL3-Promoter were determined. The pGL3-Promoter activity was used to adjust the relative luciferase activities to pool values from independent experiments. As a reference the luciferase activity of the COP and WF alleles for each of the *Mcs1b* SNPs individually is shown. These results are equivalent to Figures 12 and 13. **A)** Luciferase activity in T47D cells for all three *Mcs1b* candidate SNPs in the same pGL3- Promoter vector. There is no significant difference in the activity between the WF and COP alleles for all three *Mcs1b* candidate SNPs. **B)** Luciferase activity in MDA-MB-231 cells for all three *Mcs1b* candidate SNPs in the same pGL3- Promoter vector. There is no significant difference in the activity between the WF and COP alleles for all three *Mcs1b* candidate SNPs.

have to be done that do not use a plasmid-based system to determine the nature of these SNPs in a less- artificial system.

F. Chromosome conformation capture (3C) of the *Mcs1b* locus

Chromatin looping is a potential mechanism by which intronic or intergenic SNPs can affect the regulations of genes. A study looking at the expression levels of *MCS1B* genes in different tissues of the cardiovascular system revealed that there is a high correlation in the expression levels of *MIER3*, *MAP3K1* and *GPBP1*. The study also revealed that the expression level of *SETD9* is inversely correlated with the other *MCS1B* genes in the tissues tested. This suggests that *MIER3*, *MAP3K1* and *GPBP1* are regulated by a similar mechanism and that there is the potential for a chromatin loop containing *SETD9*, which shields it from a common regulatory mechanism. Because of the position of *SETD9* it is possible that this chromatin loop is formed in-between *MAP3K1* and *MIER3*. The results of this study can be seen in Figure 26 [154]. The CCCTC- binding factor (CTCF) is a protein that is essential for the formation of long-range chromatin loops [155]. We therefore scanned the ENCODE project for CTCF sites within the *MCS1B* region, to provide evidence for a chromatin loop within the vicinity of the *MIER3* and *MAP3K1* genes. The results are shown in Figure 27. There are three CTCF binding sites within the region of interest that were identified using the ENCODE project data. It is possible that the *MAP3K1* gene loops towards the *MIER3* gene placing the *MIER3* and *SETD9* gene into a chromatin loop. We used the human ENCODE data as a guide to design a chromosome confirmation capture (3C) experiment to determine any chromatin looping within the rat *Mcs1b* region. The hypothesis is that there is a

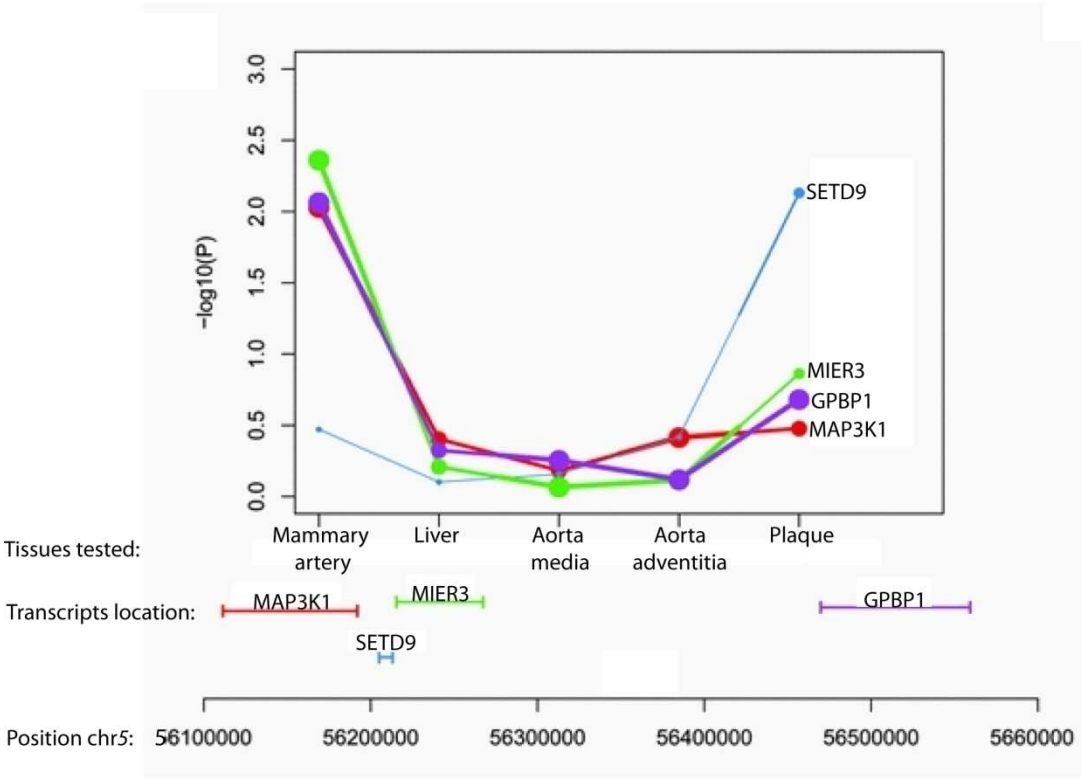


Figure 26. Correlation of *MCS1B* transcript levels in different cardiovascular tissues. Adapted from Folkersen et al. (2010) [154]. Expression levels of *MCS1B* genes in different tissues of the cardiovascular system. The expression levels of MAP3K1, MIER3 and GPBP1 are correlated in different tissue types. However, in some tissues there appears to be an inverse correlation between the expression levels of SETD9 and other *MCS1B* genes. It is possible that SETD9 is located in a chromatin loop in some of these tissues and its expression is regulated through other mechanisms.

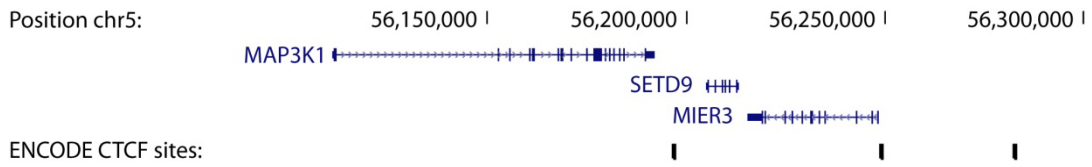


Figure 27. Location of human ENCODE identified CTCF sites. ENCODE predicted CTCF sites were identified using the UCSC Genome Browser. Location along human chromosome *5* is shown. Also, marked in blue are locations of *MCS1B* transcripts. ENCODE predicted CTCF sites were identified in a wide- array of human cell lines. The CTCF site inbetween *MAP3K1* and *MIER3* was identified in T47D cells.

chromatin loop forming between the *Map3k1* gene and the *Mier3* gene within the rat mammary gland epithelial cells, which would bring the *Mcs1b* candidate SNPs in close proximity to the *Mier3* promoter. A diagram of the 3C hypothesis can be seen in Figure 28A. We designed the 3C experiment in such as way that a chromatin loop between the *Mier3* and *Map3k1* gene can be tested. For this we selected a fixed primer downstream of one of the CTCF sites (Figure 28B). Ten moving primers were designed that move way from the fixed primer and span the distance between the *Mier3* and *Map3k1* gene. The setup for the 3C experiment can be seen in Figure 28B.

We performed the 3C experiment using rat mammary epithelial cell enriched preps (MEC prep). It is known that the *Mcs1b* locus acts within the mammary gland [87]. We believe that the mammary epithelial cells are the causative *Mcs1b* cell type. We initially extracted the MEC preps from 18 different animals and divided them into six groups of three rats each. However, after performing a PCR titration with primers in close proximity to each other, we discovered that only three groups resulted in PCR product. We used two DMBA treated groups, one contained COP females, the other WF females. We also used one COP female group that did not receive DMBA. However, this group consisted of only two animals. We started our analysis by performing the PCR reaction using pooled DNA for each group. The results are shown in Figure 29A. There was no difference in the pattern of the relative interaction frequencies between the different groups. We therefore pooled the data from all three groups as shown in Figure 29B.

There are two increases in the relative interaction frequency above background. The first one is seen with moving primer E. In a typical 3C experiment, there is an

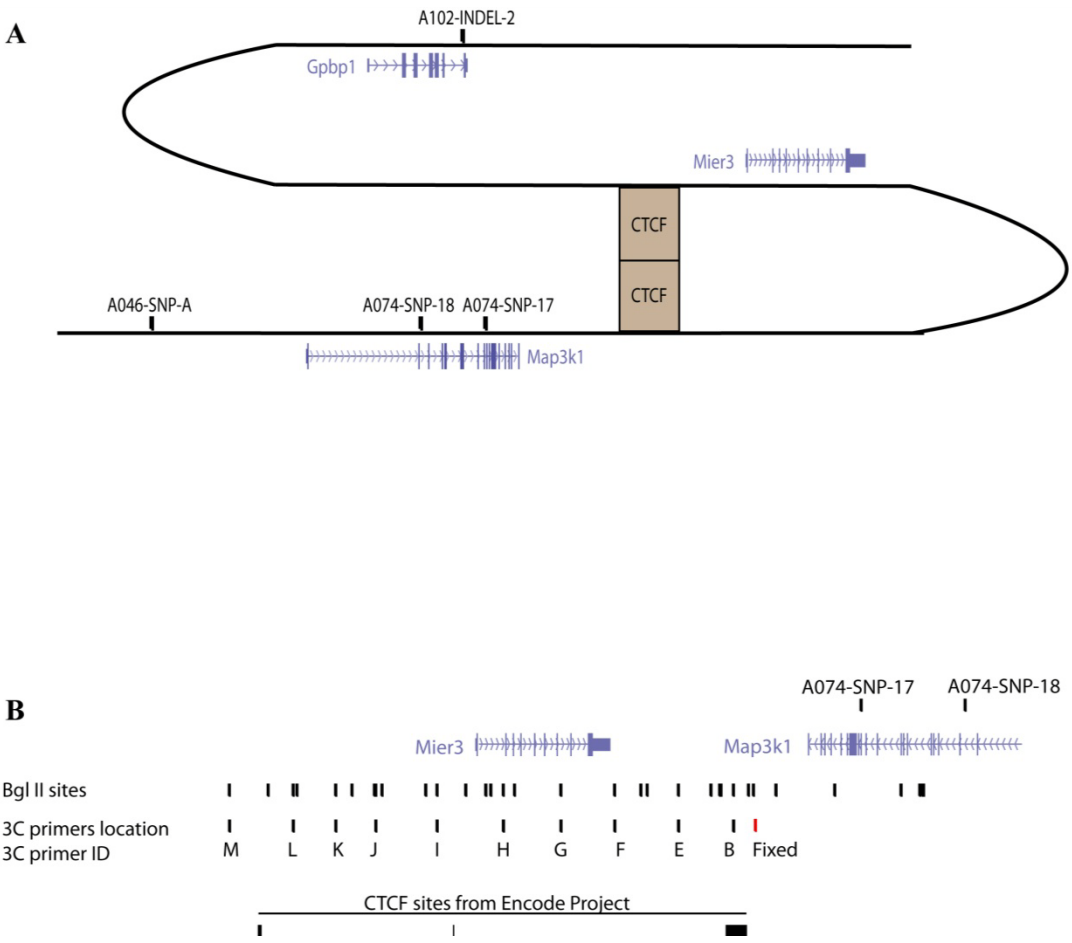


Figure 28. Hypothesis and experimental design for the 3C experiment. A) Hypothesis for the 3C experiment. We used the human ENCODE data as a guide for the location of a potential chromatin loop in the rat *Mcs1b* region. The loop would place rat SNPs in close proximity to the *Mier3* promoter. **B)** Experimental design for 3C in using rat mammary epithelial cell enriched preps. Shown are the location of the *Mcs1b* candidate SNPs, and shown in blue are the transcript locations for Mier3 and Map3k1. Black marks indicate the location of *BglII* sites within this region and the location of the 3C primers used. Marked in red is the fixed primer used. Note, the ID of the 3C primer is marked right underneath its location. As a reference, the rat orthologous regions to the ENCODE identified CTCF sites are shown.

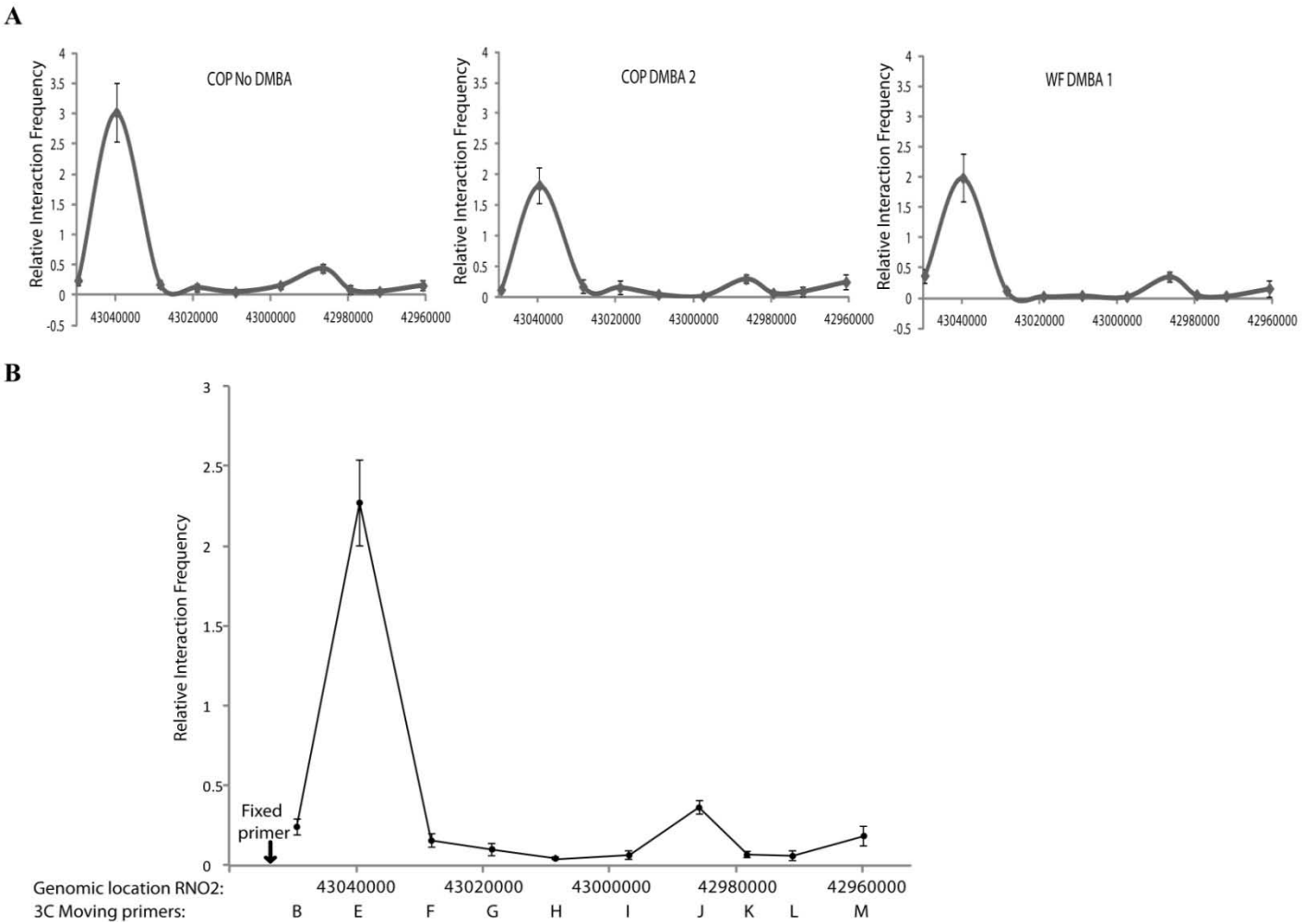

A
COP No DMBA
COP DMBA 2
WF DMBA 1
Relative Interaction Frequency
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
43040000
43020000
43000000
42980000
42960000
B
Relative Interaction Frequency
3
2.5
2
1.5
1
0.5
0
Fixed primer
Genomic location RNO2:
43040000
43020000
43000000
42980000
42960000
3C Moving primers:
B
E
F
G
H
I
J
K
L
M

Figure 29. 3C results for the *Mcs1b* locus. A) 3C results for each group separately. Results are pooled from three independent PCR reactions. COP no DMBA contains 2 females, while the other groups are made up of three females each. Relative interaction frequency is the band intensity of the sample divided by the band intensity of the BAC control for each primer pair. Error bars indicate standard error. **B)** Since there is no difference in the pattern of relative interaction frequencies among the groups, the results were pooled. Each value is the mean of nine separate data points. Error bars indicate standard error. Separation between each value indicates relative distances. The location of the fixed primer is indicated with an arrow and the location of moving primers is marked under x-axis.

increase in the interaction frequency when primers in close proximity to the fixed primer are used. Usually, primer pairs within 10kb of each other have a higher relative interaction frequency than primers further apart, because the smaller the distance is, the more likely it is that these fragments come together without the need of looping structures. Moving primer E is 13kb away from the fixed primer. It is possible that this interaction indicates a chromatin loop but it is also possible that this is a remnant of the higher interaction frequency due to close proximity of the primers. An additional experiment typically done after 3C is to perform chromatin immunoprecipitation (ChIP) for CTCF. This experiment would indicate if there is indeed a chromatin loop forming between the fixed primer and primer E. However, the second increase in relative interaction frequency at moving primer J indicates a chromatin looping structure. This chromatin loop is shown in Figure 28A. This chromatin loop would bring the *Mcs1b* candidate SNPs in close proximity to the *Mier3* promoter and may indicate a possible mechanism as to how these SNPs act on the *Mcs1b* candidate genes. Interestingly, moving primer I is closest to the rat orthologous region to a CTCF binding site in the human, however, no looping is detected using moving primer I (Figure 28B). It is possible that an alternative CTCF site exists in the rat that is in close proximity to moving primer J. While the 3C makes a compelling case for a chromatin loop forming between the fixed primer and primer J, a follow-up ChIP experiment for CTCF is needed to verify that CTCF binds in this region and facilitates the formation of a chromatin loop.

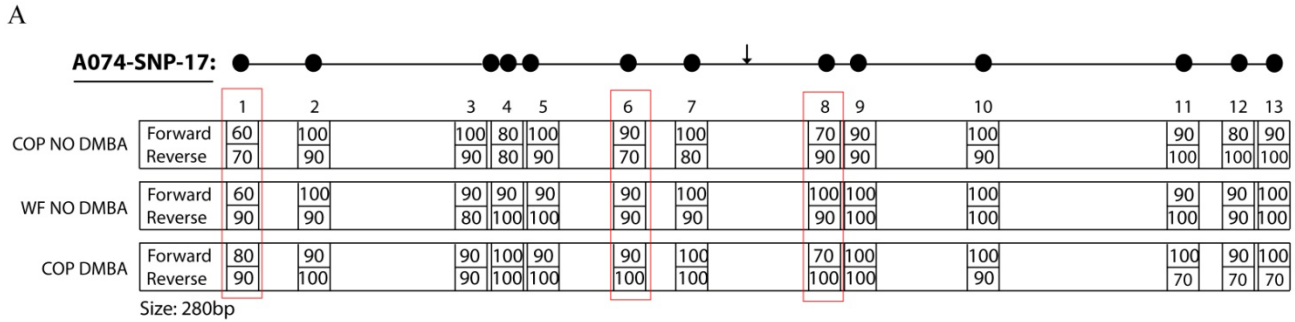
It is interesting that this chromatin looping would place *Mier3* into the chromatin loop. Chromatin loops can both increase or decrease the expression levels of transcripts within them, but shield transcripts from the effects of elements outside of the loop. It is

unknown how the Mcs1b candidate SNPs would interact with a gene located within a chromatin loop. Overall, the 3C results indicate a chromatin loop between the *Map3k1* and *Mier3* gene, which would bring the *Mcs1b* candidate SNPs in proximity to the *Mier3* promoter.

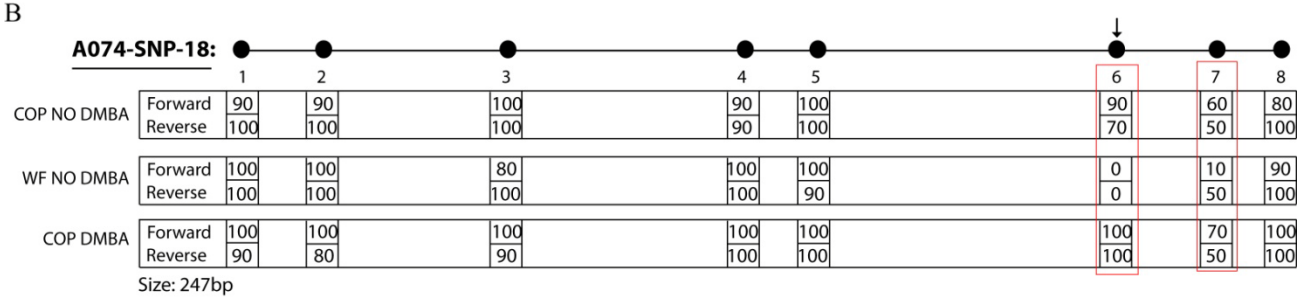
G. Bisulfite sequencing of *Mcs1b* candidate SNPs

A074-SNP-17 is located within a predicted CpG island. CpG islands are regions of high cytosine and guanine content. Cytosines in CpG pairs can be methylated. CpG islands are often found in gene promoter and promoter methylation is often associated with gene silencing [156]. We therefore wanted to test the methylation status of the *Mcs1b* candidate SNPs. As previously done with the 3C experiment, we focused on MEC preps, since we believe that the causative cell type is the mammary epithelial cell. We tested three different rat groups. One group contained females treated with DMBA. Results are shown in Figure 30. Overall, there is a high degree of methylation in these regions. There are differences in the methylation status among different groups for some of the CpG moieties tested. For *A074-SNP-17* some CpG moieties are differentially methylated (1,6 and 8), with the COP females that were not treated with DMBA having a lower methylation status. One of the CpG moieties for *A074-SNP-18* contains the SNP nucleotide, therefore the WF allele cannot be methylated. However, the COP allele is highly methylated. For *A046-SNP-A* there appears to be a lower degree of methylation in WF females compared to both of the COP groups. Overall, there is a high degree of methylation in this regions, however, some individual CPG moieties are differentially

A


A074-SNP-17:
1 2 3 4 5 6 7 8 9 10 11 12 13
COP NO DMBA
Forward 60 100 100 80 100 90 100 70 90 100 90 80 90
Reverse 70 90 90 80 90 70 80 90 90 90 100 100 100
WF NO DMBA
Forward 60 100 90 90 90 90 100 100 100 100 90 90 100
Reverse 90 90 80 100 100 90 90 90 100 100 100 90 100
COP DMBA
Forward 80 90 90 100 90 90 100 70 100 100 100 90 100
Reverse 90 100 90 100 100 100 100 100 100 90 70 70 70
Size: 280bp

B


A074-SNP-18:
1 2 3 4 5 6 7 8
COP NO DMBA
Forward 90 90 100 90 100 90 60 80
Reverse 100 100 100 90 100 70 50 100
WF NO DMBA
Forward 100 100 80 100 100 0 10 90
Reverse 100 100 100 100 90 0 50 100
COP DMBA
Forward 100 100 100 100 100 100 70 100
Reverse 90 80 90 100 100 100 50 100
Size: 247bp

C

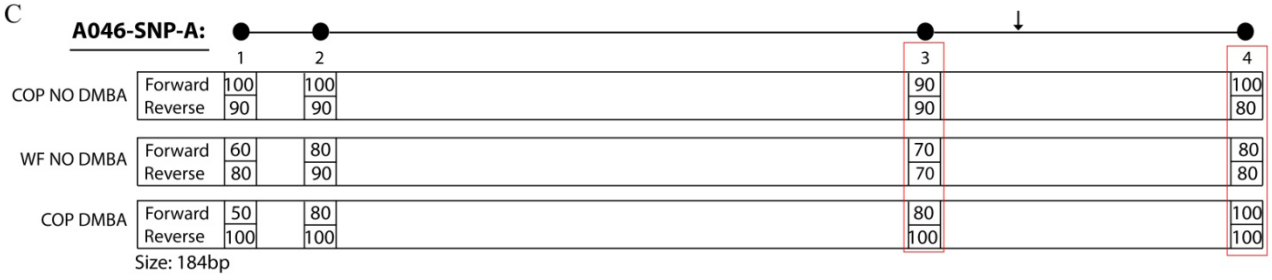

A046-SNP-A:
1 2 3 4
COP NO DMBA
Forward 100 100 90 100
Reverse 90 90 90 80
WF NO DMBA
Forward 60 80 70 80
Reverse 80 90 70 80
COP DMBA
Forward 50 80 80 100
Reverse 100 100 100 100
Size: 184bp

Figure 30. Results for bisulfite sequencing of *Mcs1b* candidate SNPs. Black circles indicate individual CpG moieties within region tested. For each CpG moiety we sequenced 10 individual clones. Values indicate the percentages of clones that were methylated. Both the forward and reverse strands were sequenced. Arrows indicate the locations of individual SNPs. Red rectangles indicate CpG moieties where there is a difference in the methylation status among the groups tested. **A)** Methylation status of *A074-SNP-17*. Overall, there is a high degree of methylation in this region. A lower methylation status is observed in the COP groups for some of the CpG moieties. **B)** Methylation status of *A074-SNP-18*. The WF allele of *A074-SNP-18* cannot be methylated, however, the COP allele is highly methylated. **C)** Methylation status for *A064-SNP-A*. Overall, there is a high degree of methylation. There is a reduction in the methylation status in the WF allele compared to the COP allele.

methylated among the groups. Some of these CpG moieties are located in the vicinity of SNP sites.

H. CRISPR knockout of *A074-SNP-17* in rat mammary gland cell lines

A074-SNP-17 is our most promising *Mcs1b* candidate SNP because of its apparent role in gene regulation through differential binding of DNA binding proteins. We wanted to know if gene editing of *A074-SNP-17* can be achieved in rat mammary gland cancer cells, with the ultimate goal being genetic manipulation of *A074-SNP-17* in WF and COP rats. We used the clustered regularly interspaced short palindromic repeats (CRISPR) system to try to knockout *A074-SNP-17* in rat mammary tumor cell lines. This system uses single guide RNAs that guide the CAS9 nuclease to regions of interest resulting in the formation of double- stranded breaks [157]. We transfected two different cell lines with the *A074-SNP-17* targeted CRISPR system. These were RBA and LA7 rat mammary tumor cell lines. Both cell lines were derived from the Sprague- Dawley rat strain and share the same genotype with WF rats at all three *Mcs1b* candidate SNPs. The cell lines were derived from DMBA induced mammary tumors.

In order to determine the effectiveness of the *A074-SNP-17* CRISPR system, we used the SURVEYOR mismatch- specific DNA nuclease, which can identity mismatched DNA regions due to INDELs and will cleave regions containing mismatches. The results are shown in Figure 31 A and B. If there is a successful induction of an INDEL in the *A074-SNP-17* region, the *A074-SNP-17* PCR product should be digested into two fragments of about 107 and 226bp. There was no formation of the two digested fragments visible for the two cell lines. A positive control, which contains plasmid DNA that have a

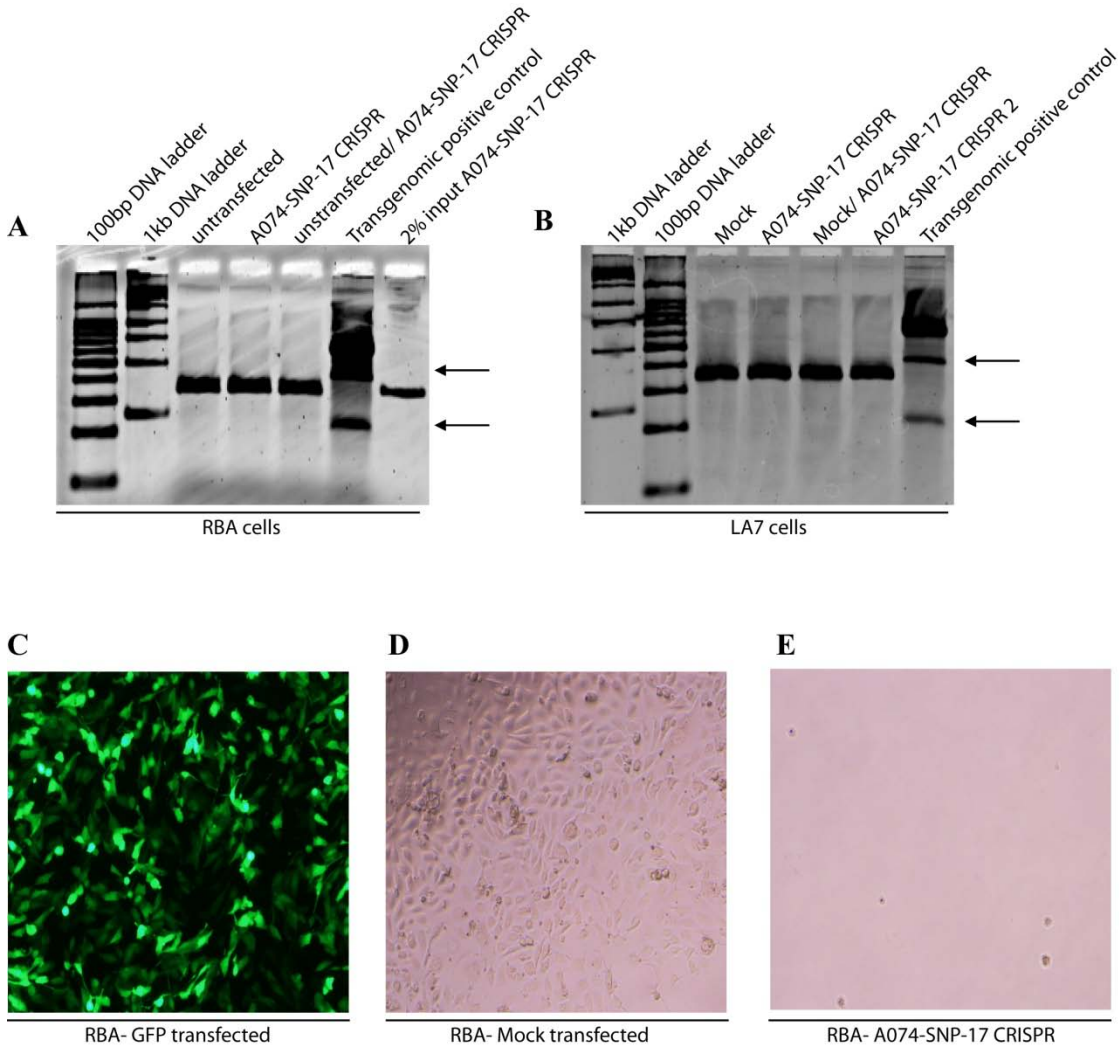


Figure 31. Results for CRISPR knockout of *A074-SNP-17* in rat mammary cancer cell lines. A) and **B)** SURVEYOR nuclease results in RBA and LA7cells respectively. Fragmentation resulted in the positive control only (marked with arrows). Agarose gel was 2% high resolution stained with ethidium bromide. **C)** RBA cells transfected with GFP show high transfection efficiency. **D)** and **E)** RBA cells mock transfected or transfected with *A074-SNP-17* CRISPR system. Few cells survive transfection with *A074-SNP-17* CRISPR system.

single nucleotide mismatch should result in the formation of two fragments, which are 217and 416bp in size. Figure 31 A and B shows that the positive control resulted in the formation of the appropriate fragments, indicating that the SURVEYOR enzyme was functioning properly. It is possible that we didn't detect any *A074-SNP-17* PCR fragments, because of low efficacy of the CRISPR system and not enough mismatches are formed to be detected with this system. Low efficacy can be due to the system not being specific enough or low transfection efficiency of the cells. Transfection efficiency was not a problem. We were able to transfect both cell lines using a GFP vector as can be seen in Figure 31C. Transfection of RBA cells with GFP resulted in high GFP expression. However, we did notice that within the 72 hours post transfection, a lot of the cells transfected with the *A074-SNP-17* CRISPR vectors died when compared to cells that were mock transfected with Lipofectamine 2000 (Figure 31 D and E). Therefore, it is plausible that the extracted DNA used in SURVEYOR assays came from untransfected cells and these cells did not undergo genome editing. It is also possible that cells that were transfected with the *A074-SNP-17* system died because of too much DNA damage that was induced with the CRISPR system and could not be repaired on time. There is no evidence that either one of these cells are deficient in DNA repair pathways, however, it is possible that the CRISPR system resulted in a lot of off- target DNA damage that could not be repaired before DNA extraction.

Discussion

Sequencing of the *Mcs1b* region between the WF and COP rat strain resulted in the identification of three *Mcs1b* candidate SNPs. Since none of the SNPs are found

within coding regions of genes, we hypothesized that the three SNPs are located within gene regulatory regions and affect the expression levels of neighboring genes. This idea is supported by the fact that *Mcs1b* candidate genes are expressed at different levels within the mammary glands of the WF and COP rat strains [87]. We tested the ability of the SNPs to regulate a luciferase gene in a reporter assay. All three candidate SNPs showed a l uciferase activity that was different between the COP and WF alleles. However, *A074-SNP-17* is the only candidate SNP that had a pattern of gene regulation that mimics the gene expression data of *Mcs1b* candidate genes for the two rat strains. The *Mcs1b* locus has a human ortholog. Performing the same luciferase assay for all seven *rs889312* correlated SNPs indicated that *rs1862626* has a p attern of gene regulation that is the same as for *A074-SNP-17*. This indicates that *rs1862626* may be the functional ortholog to *A074-SNP-17*. The luciferase activity for *A074-SNP-17* is similar in T47D and MDA-MB-231 breast cancer cells. These two cell lines differ drastically in their cellular environments. This may indicate that *A074-SNP-17* has a u niversal function. However, several other SNPs showed vastly different luciferase activities in the two cellular environments, which indicates that these SNPs may bind transcription factors that may be expressed differentially between the two cell lines.

A search for transcription factors binding to candidate SNPs revealed that the WF allele of *A074-SNP-17* binds a DNA binding protein complex that does not bind to the COP allele. This DNA binding protein may be involved in the regulatory function of this SNP region. Our EMSAs do not support *rs1862626* being ortholog to *A074-SNP-17*, because both major and minor alleles of *rs1862626* result in the same DNA binding protein shift pattern. However, *rs1862626* could still be the functional ortholog, since the

two alleles appeared to show a differential affinity for binding a DNA binding protein. A hunt for DNA binding protein for *A074-SNP-17* did not result in a positive identification of a transcription factor binding to this SNP. NRF2 does bind to this SNP; however, the binding appears to be non-specific. We focused on a small amount of transcription factors from our mass spectrometry data; studying more of the transcription factors may be fruitful.

A luciferase assay of all three *Mcs1b* candidate SNP in the same luciferase reporter revealed that there is an additive effect of the suppressive activity of all three SNPs. This made us wonder how all three SNPs affect the expression of the same set of genes. We performed chromosome conformation capture and showed that there is a loop forming between the *Map3k1* and *Mier3* gene, which would bring the three candidate SNPs in close proximity to the *Mier3* promoter. However, a follow-up experiment using ChIP of CTCF is needed to confirm the 3C results. A luciferase assay of 2kb of the region just upstream and downstream of the *Mier3* transcription start site revealed that there is no intrinsic promoter activity in this region. It is possible that DNA looping is required for the three candidate SNPs to get in close proximity to the *Mier3* promoter and this is needed for promoter activation for the *Mier3* gene.

A074-SNP-17 is located with a predicted CpG island. We therefore checked the methylation status of CpG moieties for the *Mcs1b* candidate SNPs. We tested the methylation status using DNA extracted from MECs. We found that overall, there was a high degree of methylation in these regions. There appears to be a difference in the methylation status for some of the CpG moieties in the regions between the WF and COP rats. For *A074-SNP-17* there appears to be less methylation in the COP, while for *A046-*

SNP-A there appears to be less methylation in the WF rat. A high degree of methylation indicates silencing of specific regions. This could indicate that both SNPs are located within important regions that may be involved in gene regulation and are tightly regulated. One of the CpG moieties located in the *A074-SNP-18* region contains the SNP nucleotide. Therefore, the WF allele cannot be methylated, while the COP allele can. There is a high degree of methylation of this CpG moiety in the COP rat. This might indicate that this SNP does not have an effect in the COP rat because it is highly methylated and therefore silenced.

Overall, we believe that *A074-SNP-17* is a great candidate for the *Mcs1b* phenotype. In order to provide further evidence that *A074-SNP-17* is the causative SNP for the *Mcs1b* region, we want to perform genome editing in the rat to knockout or exchange alleles for *A074-SNP-17*. We tested the ability of an *A074-SNP-17* CRISPR construct to knockout the SNP region in rat mammary tumor cell lines. However, we were not able to knockout this region efficiently, likely because of the death of transfected cells. A different screening method or another method for genome editing such as TALENs or zinc fingers may prove to be more fruitful than the CRISPR system.

CHAPTER V

SIGNIFICANT OVERLAP BETWEEN HUMAN GENOME-WIDE ASSOCIATION NOMINATED BREAST CANCER RISK ALLELES AND RAT MAMMARY CANCER SUSCEPTIBILITY LOCI

Adapted from Sanders et al. (2014) [158]. See Appendix.

Introduction

Breast cancer is a complex disease characterized by environmental, genetic, and epigenetic factors. Due to the complexity of developing this disease a woman's individual risk may vary greatly from population risk estimates. The familial relative risk of developing breast cancer increases with the number of affected relatives, suggesting that there is a strong genetic component associated with this disease [54, 159]. High-penetrance breast cancer risk mutations such as those of *BRCA1* and *BRCA2* have been identified [63, 64, 160]. Population frequencies of mutations with high-penetrance toward risk are rare due to their severe effects on individuals; and thus, these mutations account for only a small percentage of population risk. Risk alleles with moderate penetrance and minor allele population frequencies of 0.005-0.01 (e.g. *PALB2*) are estimated to account for approximately 3% of risk. Therefore, a majority of population-based breast cancer risk is likely explained by low penetrance alleles with rare to common population frequencies [60]. Genome-wide association (GWA) studies have been used to identify several low penetrance breast cancer risk alleles [57]. Due to a need to control for

numerous multiple comparisons made in GWA studies, a Bonferroni correction based p-value cut-off of $\leq 1x10^{-7}$ is typically required for an association to be considered genome-wide significant. It has been suggested that this approach is too stringent as it may result in many false negative associations [161]. Furthermore, while GWA studies are unbiased approaches to identify genomic regions associated with breast cancer risk, these epidemiology- based approaches cannot easily determine risk genes or genetically determined mechanisms of susceptibility. Currently, only a small percentage of breast cancer heritability is explained by published studies suggesting that considerable genetic variation associated with breast cancer risk remains to be identified [60, 162].

Comparative genetics between rats and humans has also been used to identify breast cancer risk alleles [111]. In general, the laboratory rat is a good experimental organism to model breast cancer. Compared to induced mammary tumors in mice, rats develop mammary carcinomas of ductal origin, which is similar to a majority of human breast cancers. Also, rat mammary tumors are responsive to estrogen, just as a majority of human breast tumors [97, 163]. Most importantly, the laboratory rat is a versatile organism to study breast cancer susceptibility, as experiments can be controlled at genetic and environmental levels. Inbred rat strains exhibit differential susceptibility to chemically induced carcinogenesis using 7,12-dimethylbenz[a]anthracene (DMBA) [92, 97, 98, 100]. Copenhagen (COP) and Wistar- Kyoto (WKY) rat strains are resistant to DMBA, *N*-Nitroso-*N*-methylurea (NMU), and oncogene induced mammary carcinomas, while the Wistar- Furth (WF) rat strain is susceptible.

Previous genetic studies using rats have identified eight *Mammary carcinoma susceptibility* (*Mcs*) loci, named *Mcs 1-8* [106-108, 123]. A ($WFxCOP)_{F1}$ x WF

backcross design was used to identify *Mcs 1-4*. Copenhagen alleles at *Mcs 1-3* are associated with decreased mammary tumor multiplicity, while the *Mcs 4* COP allele is associated with increased tumor development [106]. Further analysis of the *Mcs1* locus using WF.COP congenic lines, spanning different regions of the quantitative trait locus (QTL), identified three independent loci associated with mammary carcinoma susceptibility, named *Mcs 1a-c* [123]. Another linkage analysis study using WF and WKY rat strains revealed four additional QTLs associated with mammary carcinoma susceptibility, named *Mcs 5-8*. Additionally, a modifier of *Mcs8, Mcsm1*, partially counteracts the resistance phenotype conferred by *Mcs8* [107, 108]. Further analysis of the *Mcs5* locus using WF.WKY congenic rat lines resulted in the identification of four subloci named *Mcs5a1*, *Mcs5a2*, *Mcs5b* and *Mcs5c* [110, 111]. Additional linkage analysis using the SPRD-Cu3 rat strain (DMBA-induced mammary carcinogenesis susceptible) and the resistant WKY rat strain resulted in the identification of three more rat QTLs associated with mammary cancer named *Mcstm1*, *Mcstm2/Mcsta2* and *Mcsta1* [115, 116]. Several rat genomic regions that associate with mammary cancer susceptibility were identified using beta-estradiol instead of DMBA to induce carcinogenesis. These QTLs were identified using the August Copenhagen Irish (ACI) rat strain, which is susceptible to beta-estradiol carcinogenesis and the COP and Brown Norway (BN) rat strains, which are resistant. These loci are named *Estrogen- induced mammary cancer* loci or *Emca 1-2* and *Emca 4-8* [117, 118].

Comparative genomics between human breast and rat mammary cancer risk alleles will continue to be warranted, especially if appreciable overlap in genetic susceptibility exists between these species. In this study, genomic locations of human

breast cancer risk GWA study-identified polymorphisms were compared to the rat genome to determine if positive associations were more often located at orthologs to rat mammary cancer risk loci than at randomly selected regions not known to be associated with rat mammary cancer susceptibility. The hypothesis is that positive associations map to orthologs of rat mammary cancer risk loci more often than to randomly selected rat genomic region.

Methods

A. Converting rat mammary cancer associated loci to human orthologous regions

Previously published information on rat mammary cancer associated loci was used. Human orthologous regions of rat regions that associate with mammary cancer susceptibility listed in Table 21 were determined using the "In other genomes (convert)" function available at the UCSC genome browser [119]. Rat Nov. 2004 (Baylor 3.4/rn4) and human Feb. 2009 (GRCh37/hg19) genome assemblies were used. If a rat mammary cancer locus split into multiple human orthologous regions, we noted all orthologous regions until they reached less than 1% of the bases and spanned less than 1% of the original rat mammary cancer locus using the UCSC genome browser.

B. Identification of random rat regions

To determine if human GWA study identified polymorphisms map to rat mammary cancer loci more frequently than to random regions of the rat genome, we selected rat genome segments that have not shown an association with mammary cancer risk. These rat genomic regions were named "random rat regions" and are listed in Table

Table 21. Location of rat mammary cancer susceptibility loci and human orthologous regions used.

Rat *Mcs* locus (overlap)	Boundary Markers	Rat chr	Region (UCSC rat assembly 2004)	Reference	Human Orthologous Region (UCSC human assembly 2009)
DMBA induced mammary carcinogenesis					
Mcs1a	D2Mit29 to D2Uwm14	RNO2	5,601,528-10,539,344	Haag et al. [123]	*Chr5*: 89,216,702-93,113,337
Mcs1b	ENSRNOSNP2740854 to g2Ul2-27	RNO2	42,364,155-44,195,382	DenDekker et al. [87]	*Chr5*: 54,816,178-57,003,049
Mcs1c	D2M13Mit286 to D2Uia5	RNO2	13,909,383-20,666,092	Haag et al. [123]	*Chr5*: 81,891,633-86,857,442 *Chr5*: 86,171,198-86,251,067
Mcs2 (overlaps *Mcs6, Emca4*)	D7rat39 to D7Uwm12	RNO7	4,936,704-86,028,057	Sanders et al. [108]	*Ch12*: 57,316,160-108,177,690 *Chr8*: 97,242,984-115,650,989 *Chr19*: 281,161-2,497,331 *Chr19*: 15,059,910-15,808,112
Mcs3	D1Rat27 to D1Mit12	RNO1	90,282,174-156,954,117	Shepel et al. [106]	*Chr15*: 80,282,370-102,265,870 *Chr15*: 25,574,935-28,567,541 *Chr11*: 17,403,456-22,898,646 *Chr11*: 74,958,193-89,350,902 *Chr19*: 48,799,986-51,921,957 *Chr19*: 28,701,413-30,656,003
Mcs4	D8Rat164 to D8Rat108	RNO8	28,414,100-72,403,639	Shepel et al. [106]	*Chr11*: 107,453,990-132,383,506 *Chr15*: 62,105,069-76,028,735 *Chr15*: 76,091,658-78,185,872 *Chr15*: 78,380,119-78,998,961 *Chr15*:51,349,646-51,942,505
Mcs5a1 (overlaps *Mcstm1, Emca8*)	SNP-61634906 to SNP-61666918	RNO5	61,634,727-61,666,739	Samuelson et al. [111]	*Chr9*: 37,562,516-37,589,491
Mcs5a2 (overlaps *Mcstm1, Emca8*)	SNP-61667232 to gUwm23-29	RNO5	61,667,053-61,751,614	Samuelson et al. [111]	*Chr9*: 37,590,988-37,654,512

Table 21 continued.

Rat *Mcs* locus (overlap)	Boundary Markers	Rat chr	Region (UCSC rat assembly 2004)	Reference	Human Orthologous Region (UCSC human assembly 2009)
Mcs5b (overlaps *Mcstm1, Emca8*)	gUwm50-20 to D5Got9	RNO5	65,498,190-67,464,050	Samuelson et al. [110]	*Chr9*: 103,492,712-105,220,552
Mcs5c (overlaps *Mcstm1, Emca8*)	gUwm74-1 to gUwm54-8	RNO5	81,118,457-81,295,367	Veillet et al. [164]	*Chr9*: 118,231,525-118,416,951 *Chr12*: 72,033,141-72,033,263
Mcs6 (overlaps *Mcs2*)	D7Rat171 to gUwm64-3	RNO7	22,382,725-55,384,873	Sanders et al. [108]	*Chr12*: 71,270,266-105,502,699
Mcs7 (overlaps *Mcsta1*)	D10Got124 to gUwm58-136	RNO10	89,575,060-100,335,500	Cotroneo et al. [165]	*Chr17*: 40,183,547-67,946,104
Mcs8	D14Mit1 to D14Rat99	RNO14	12,386,493-26,416,791	Lan et al. [107]	*Chr4*: 65,556,457-81,559,483
Mcsm1 (overlaps *Emca7*)	D6Mit9 to D6Rat12	RNO6	34,039,303-114,032,192	Lan et al. [107]	*Chr14*: 25,151,530-80,417,386 *Chr2*: 334,41-18,603,019 *Chr7*: 12,561,599-19,619,365 *Chr7*: 107,770,320-111,916,436
Mcstm1 (overlaps *Mcs5,Emca1, Emca8)*	D5rat124 to *Pla2g2a*	RNO5	19,206,257-157,657,360	Piessevaux et al. [116]	*Chr1:* 20,301,931-59,012,763 *Chr1:* 59,119,520-67,602,141 *Chr9:* 27,325,071-123,488,955 *Chr6:* 87,792,854-100,245,025 *Chr8:* 87,055,841-97,247,307 *Chr8:* 58,994,818-62,700,945
Mcstm2 (overlaps *Emca2*)	D18Wox8 to D18Rat44	RNO18	32,458,819-86,863,412	Piessevaux et al. [116]	*Chr18:* 10,202,644-13,129,349 *Chr18:* 41,356,963-54,158,113

Table 21 continued.

Rat *Mcs* locus (overlap)	Boundary Markers	Rat chr	Region (UCSC rat assembly 2004)	Reference	Human Orthologous Region (UCSC human assembly 2009)
Mcsta1 (overlaps *Mcs7*)	D10Rat91 to D10Rat97	RNO10	9,762,188-108,776,963	Piessevaux et al. [116]	*Chr5:* 130,482,861-173,663,969 *Chr5:* 177,530,539-180,675,650 *Chr17:* 690,639-15,624,409 *Chr17:* 16,916,926-20,222,700 *Chr17:* 25,525,650-78,247,249 *Chr16:* 78,402-6,094,950
***β*-estradiol induced mammary carcinogenesis**					
Emca1 (overlaps *Mcstm1, Emca8*)	D5Rat53 to D5Rat57	RNO5	103,677,474-155,121,024	Gould et al. [117]	*Chr1:* 23,607,020-59,012,763 *Chr1:* 59,119,520-67,602,141 *Chr9:* 17,037,252-27,300,264
Emca2 (overlaps *Mcstm2*)	D18Rat27 to D18Rat43	RNO18	18,562,643-66,652,947	Gould et al. [117]	*Chr5:* 110,259,180-130,363,372 *Chr5:* 137,224,929-147,624,793 *Chr5:* 147,647,196-150,176,352 *Chr18:* 10,202,644-13,129,349 *Chr18:* 35,982,130-41,016,602 *Chr18:* 52,597,120-54,158,113 *Chr18:* 54,267,924-58,201,561 *Chr2:* 127,805,417-128,786,719
Emca4 (overlaps *Mcs2*)	D7Rat44 to D7Rat15	RNO7	66,201,980-107,428,439	Schaffer et al. [118]	*Chr8:* 97,242,984-137,409,536 *Chr12:* 57,316,160-59,093,375
Emca5	D3Rat227 to D3Rat210	RNO3	41,054,012-171,063,335	Schaffer et al. [118]	*Chr20:* 1,746,912-62,907,504 *Chr2:* 110,841,402-113,650,057 *Chr2:* 159,530,076-188,395,371 *Chr11:* 26,296,319-57,753,858 *Chr15:* 32,905,485-34,664,466 *Chr15:* 34,933,152-51,298,144

Table 21 continued.

Rat *Mcs* locus (overlap)	Boundary Markers	Rat chr	Region (UCSC rat assembly 2004)	Reference	Human Orthologous Region (UCSC human assembly 2009)
Emca6	D4Rat14 to D4Rat202	RNO4	41,729,583-159,115,617	Schaffer et al. [118]	*Chr7:* 23,252,368-33,103,107 *Chr7:* 115,026,301-150,558,396 *Chr3:* 88,756-12,883,445 *Chr3:* 13,004,609-15,163,132 *Chr3:* 64,018,604-75,322,612 *Chr3:* 125,977,400-128,219,297 *Chr2:* 68,713,643-89,165,869 *Chr4:* 89,504,626-95,273,083 *Chr4:* 120,978,632-122,320,931 *Chr12:* 156,786-2,821,588 *Chr10:* 43,277,230-46,218,580
Emca7 (overlaps *Mcsm1*)	D6Rat68 to D6Rat81	RNO6	2,802,670-111,967,837	Schaffer et al. [118]	*Chr14:* 25,151,530-78,362,253 *Chr2:* 33,441-35,642,893 *Chr2:* 38,644,737-51,698,454 *Chr7:* 12,561,599-19,619,365 *Chr7:* 105,197,211-111,916,436
Emca8 (overlaps *Mcs5*, *Mcstm1*, *Emca1*)	D5Rat134 to D5Rat37	RNO5	52,434,178-148,460,381	Schaffer et al. [118]	*Chr9:* 6,756,013-27,300,264 *Chr9:* 27,925,947-123,488,955 *Chr1:* 33,159,021-59,012,763 *Chr1:* 59,119,520-67,602,141

22. We initially focused on f ourteen *Mcs/Mcsm* regions with an average size of 22,322,710 bp as these are generally smaller in size than other rat mammary cancer associated loci identified. Fourteen random rat genome regions, each 22,322,710 bps in size were used for comparison. Random rat regions were selected by picking a chromosome using a random number generator function of Microsoft Excel. The range of chromosomes entered into the random number generator function was 1 through 21 (rats have 21 chromosomes, including a sex chromosome). The start position for each random rat region was determined using a random number generator function of Excel. The rat genome is 2.75 Gb in size [166]; or, 130,952,381bp per chromosome if divided equally across chromosomes. Therefore, values for the rat genome start-position were chosen from 1-130,952,381 using a random number generator. Following, 22,322,709 bps were added to each random start position to obtain the desired full length. The 14 random rat genome regions were then entered into the UCSC genome browser, and the human orthologous regions were determined using the "in other genomes (convert)" function, as described above [119]. Randomly-generated rat genome segments were used as controls if the human orthologous segment did not contain sequence that was also orthologous to a known rat mammary cancer associated locus. We also verified, using the UCSC genome browser, that human orthologous regions to random rat regions were not located at human centromeric regions, as genetic variation in these chromosomal regions is underrepresented in GWA studies [167, 168]. Total sizes and percentages of rat genome covered by rat mammary cancer loci and random genome regions used are in Table 23.

C. Determining human GWA study nominated polymorphisms

Table 22. Random rat genomic segments and human orthologous regions used.

Rat *Mcs* locus	Rat *chr*	Region (UCSC rat assembly 2004)	Human orthologous region (UCSC human assembly 2009)
Random rat region 1	RNO9	20,000,000-44,322,711	*Chr2:* 97,158,323-106,711,249 *Chr6:* 56,223,874-73,919,999 *Chr2:* 189,182,486-189,878,065 *Chr13:* 103,235,577-103,556,495 *Chr2:* 128,848,569-129,254,860
Random rat region 2	RNO15	60,000,001-84322711	*Chr13:* 53,226,266-74,878,291 *Chr13:* 42,064,282-42,529,444
Random rat region 3	RNO16	68,621,246-92,943,956	*Chr13:* 103,539,456-115,092,822 *Chr8*: 36,627,241-42,308,840 *Chr8:* 638,582-6,693,649 *Chr8:* 42,690,588-43,056,179 *Chr13:* 52,753,969-53,050,606
Random rat region 4	RNO9	91,398,460-115,721,170	*Chr5:* 98,385,946-110,062,886 *Chr18:* 612,848-9,957,727 *Chr2*: 240,340,012-242,806,427
Random rat region 5	RNO13	55,373,307-79,696,017	*Chr1:* 169,844,936-194,938,667
Random rat region 6	RNO11	39,408,000-63,730,710	*Chr3:* 95,108,010-118,895,417
Random rat region 7	RNO17	68,384,015-92,706,72	*Chr10:* 138,740-22,530,353 *Chr1:* 236,673,870-240,084,642
Random rat region 8	RNO3	12,585,543-36,908,253	*Chr2:* 140,246,548-155,465,845 *Chr9:* 123,526,091-129,443,210
Random rat region 9	RNO19	34,130,390-58,453,100	*Chr16:* 66,968,878-90,107,058 *Chr10:* 33,502,588-35,153,585 *Chr1:* 229,402,942-235,324,796 *Chr4:* 150,548,912-150,855,848
Random rat region 10	RNO12	18,203,110-42,525,820	*Chr12:* 110,503,298-120,870,994 *Chr12:* 121,578,435-132,335,900 *Chr7:* 66,878,689-71,941,664 *Chr7:* 101,137,811-102,184,451 *Chr7:* 99,995,220-100,350,712 *Chr7:* 72,707,443-74,223,683 *Chr7:* 75,027,443-76,145,496
Random rat region 11	RNO20	30,416,373-54,739,083	*Chr6:* 101,086,446- 116,620,662 *Chr6:* 117,266,139-123,147,126 *Chr2:* 109,065,537-109,613,060 *Chr6:* 116,688,407-116,905,609
Random rat region 12	RNO13	955,085-25,277,795	*Chr18:* 58,351,906-63,553,937 *Chr2:* 124,758,685-125,682,595
Random rat region 13	RNO1	1,136,860- 25,459,569	*Chr6:* 128,011,342-150,185,813 *Chr6*: 123,315,387-124,317,854
Random rat region 14	RNO2	182,078,762- 206,401,472	*Chr1:* 107,259,608-154,441,176

Table 23. Total size and percentage of rat genome covered by rat mammary cancer loci and random rat regions

Region	Loci	Total size (bases)	Total overlapping bases	Total unique bases	Rat Genome Portion (based on total unique bases)
***Mcs/Mcsm* only**	14	345,323,605	33,002,148	312,321,457	11.4%
All known rat mammary cancer loci	24	1,230,487,116	325,386,323	905,100,793	32.9%
Random rat regions	14	312,517,940	-	312,517,940	11.4%

Human breast cancer risk GWA studies considered were published through March 2013. In the first round of analysis we picked GWA studies with a clearly defined study population of subjects of European descent. In the second round of analysis, the defined population was broader and included studies that tested populations of non-European descent. Studies that included non-European descent populations were subdivided into respective populations used. The GWA studies evaluated are listed in Tables 24 and 25. Results from GWA studies used consisted of multiple stages (2-4 stages) to evaluate breast cancer risk association. In our analysis, all SNPs that entered the final stage of their respective study were compared to the rat genome. A tested SNP was called either "associated" if it reached genome wide significance in its respective study or "potentially associated" if it failed to meet the respective study statistical criterion following the final stage of analysis. Conventionally, a p-value level for an association to be considered statistically significant in a GWA study is 1×10^{-7}. This stringency is to protect from false-positives due to multiple comparisons on a genome-wide scale. It has been argued that this low p-value requirement results in numerous false negative associations [161]. Therefore, we queried supplemental material of each published GWA study considered and picked polymorphisms that failed the validation stage in the respective study. We also included polymorphisms that did reach genome-wide significance. We considered 533 SNPs from studies that included populations of European descent, and 285 SNPs from studies of non-European descent populations. Human genomic locations of polymorphisms were found using dbSNP (GRCh37 assembly) [169]. These were compared to locations of the human orthologous regions of rat mammary cancer loci and

Table 24. Breast cancer risk genome-wide association studies using populations of European descent.

GWAS	Population	Stages	Cases/ Controls stage 1	Cases/ Controls stage 2	Cases/ Controls stage 3	Cases/ Controls stage 4	Study p-value cut-off for significance
Ahmed et al. [170]	European descent	4	390/364	3,990/3,928	3,878/3,928	33,134/ 36,141	P<E-07
Antoniou et al. [122]	European descent	2	1,193/ 1,190	5,986/2,974			P<E-07
Easton et al. [61]	European descent	3	408/400	3,990/3,916	21,860/22,578		P<E-07
Fletcher et al. [171]	European descent	3	3,981/ 2,365	4,804/3,936	4,237/5,044		-
Garcia-Closas et al. [172]	European descent	2	4,193/35,194	6,514/41,455			P<5E-08
Gaudet et al. [76]	European descent	2	899/804	1,264/1,222			P<E-05
Ghoussaini et al. [173]	European descent	2	56,989/58,098	69,564/68,150			P<E-04
Haiman et al. [174]	European descent / African descent	2	African descent (1,004/2,745), European descent (1,718/3,670)	European descent (2,292/16,901)			-
Hunter et al. [175]	European descent	2	1,145/1,142	1,776/2,072			P<2E-05
Li et al. [176]	European descent	2	617/4,583	1,011/7,604			P<E-05

Table 24 continued.

GWAS	Population	Stages	Cases/ Controls stage 1	Cases/ Controls stage 2	Cases/ Controls stage 3	Cases/ Controls stage 4	Study p-value cut-off for significance
Li et al. [177]	European descent	2	2,702/ 5,726	?			P<E-06
Mavaddat et al. [178]	European descent	2	4,470/4,560	?			P<5E-02/6.25E-03
Michailidou et al. [62]	European descent	2	10,052/ 12,575	45,290/ 41,880			P<5E-08
Murabito et al. [120]	European descent	1	250/1,345				P<5E-08
Sehrawat et al. [179]	European descent	2	348/348	1,153/1,215			P<6.4E-08
Stacey et al. [180]	European descent	2	1,600/11,563	4,554/17,577			P<E-07/ P<6.8E-08
Stacey et al. [181]	European descent	2	6,145/33,016	5,028/32,090			-
Thomas et al. [121]	European descent	3	1,145/1,142	4,547/4,434	4,078/5,223		P<5E-07
Turnbull et al. [182]	European descent	2	3,659/4,897	12,576/12,223			P<E-04

Table 25. Breast cancer risk genome-wide association studies of non-European descent populations.

GWAS	Population	Stages	Cases/ Controls stage 1	Cases/ Controls stage 2	Cases/ Controls stage 3	Cases/ Controls stage 4	Study p-value cut-off for significance
Cai et al. [183]	Asian descent	4	2,062/2,066	4,146/1,823	6,436/6,716	4,509/6,338	-
Chen et al. [184]	African- American descent	2	3,153/ 2,831	3,607/ 11,330			P<5E-08
Gold et al. [77]	Ashkenazi Jewish descent	3	249/299	950/979	243/187		P<E-05
Haiman et al. [174]	African descent/ European descent	2	African descent (1,004/2,745), European descent (1,718/3,670)	European descent (2,292/16,901)			-
Kim et al. [78]	Asian descent	3	2,273/2,052	2,052/2,169	1,997/1,676		P<5E-04
Long et al. [185]	Asian descent/ European descent	3	2,073/2,084	4,425/1,915	Asian descent (6,173/6,340), European descent (2,797/2,662)		-
Long et al. [186]	Asian descent	4	2,918/2,324	3,972/3,852	5,203/5,138	7,489/9,934	P<5E-08
Zheng et al. [187]	Asian descent	3	1,505/1,522	1,554,1,576	3,472/900		P<5E-08
Zheng et al. [124]	Asian descent	2	23,637/25,579				P<1.5E-03

random rat regions. If a polymorphism mapped to a region of interest, the name, location, odds ratio,95% confidence interval, and p-value were noted.

D. Statistics

The number of human polymorphisms that mapped to orthologous regions containing rat mammary cancer loci (observed) was compared to the number of human polymorphisms that mapped to random rat regions (expected) using a chi-square analysis with one degree of freedom. Several rat mammary cancer loci overlap extensively and subsequently several human polymorphisms mapped to multiple rat loci. Currently, it is not known if these overlapping rat mammary cancer loci would fine-map to the same locus or independent loci. For this study, human polymorphisms mapping to overlapping rat mammary cancer susceptibility associated sequences were counted only once. For analysis of associated (passed genome-wide significance level) versus potentially associated (did not pass genome-wide significance level) associations, a logistic regression was performed using SYSTAT 13™ statistical software. A threshold of associated or potentially associated was used as the independent variable and outcome was either the SNP mapped to a rat mammary cancer locus or it mapped to a random rat region.

Results

A. Significantly more breast cancer risk GWA study nominated SNPs are located at orthologs of rat *Mcs/Mcsm* loci compared to random rat genomic regions

We picked 28 GWA studies of breast cancer risk in which well defined populations were analyzed (Table 21). Physical locations of polymorphisms that failed the final validation step and polymorphisms that reached genome-wide significance were determined using dbSNP [169]. We included SNPs that failed the final validation step of the respective study, because it has been suggested that many true associations are ruled out in a GWA study due to stringent statistical analysis methods [161]. We determined if sequences containing these polymorphisms were located at either a human genome region orthologous to a known rat mammary cancer locus or to a randomly selected region of the rat genome. Our goal was to determine if GWA study-nominated potentially-associated (did not pass final validation) and associated (genome-wide significant) risk polymorphisms, together map more often to human orthologous regions of rat mammary cancer susceptibility loci than to randomly selected rat genome segments of similar size. If yes, it would suggest that human GWA information combined with rat genetic susceptibility information is broadly useful to determine true genetic associations. Overall, rat *Mcs/Mcsm* loci are mapped to shorter genomic segments than other rat mammary cancer risk loci; therefore, we first compared overlap between human GWAS nominated breast cancer risk SNPs and rat *Mcs/Mcsm* loci to overlap of human associated SNPs with randomly selected rat genomic regions not known to contain mammary cancer susceptibility loci (Figure 32). Human GWA studies were grouped by population of descent for comparison. There was a significant difference between the number of GWAS nominated SNPs mapping to rat *Mcs/Mcsm* loci compared to random rat regions in studies analyzing populations of European descent (66 SNPs to 51 SNPs respectively, p-value <0.05). This result supports previous studies indicating rat genetic susceptibility

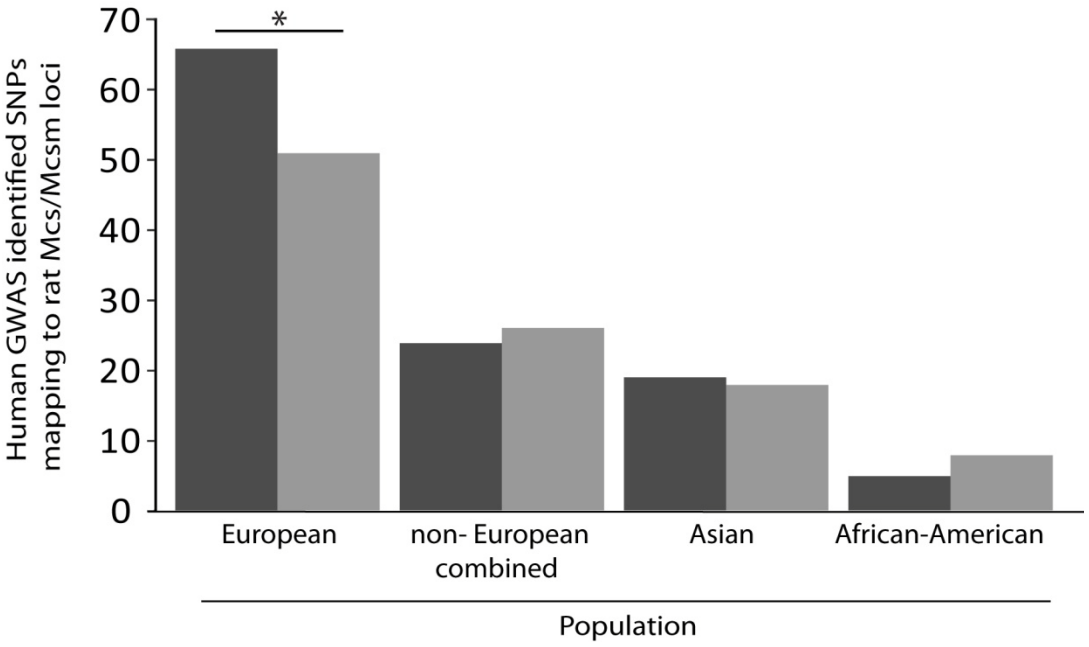


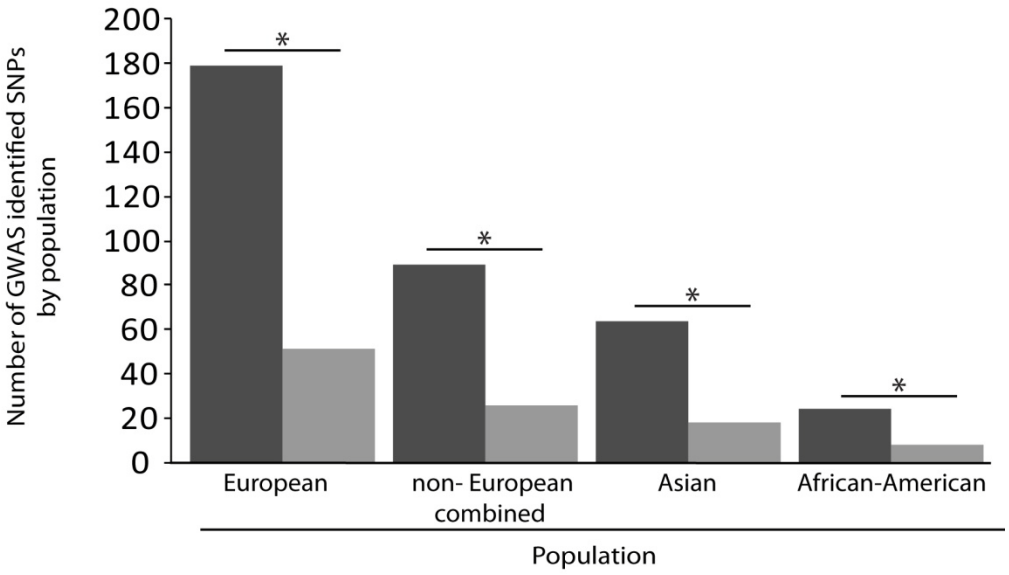
Figure 32. Number of breast cancer risk GWA study nominated SNPs mapping to rat *Mcs/Mcsm* regions. Adapted from Sanders et al. (2014) [158]. Number of GWA study nominated SNPs mapping to orthologs of rat *Mcs/Mcsm* loci and rat random regions. (■) Dark grey columns represent the number of GWA study nominated human SNPs mapping to the human orthologous regions of the *Mcs/Mcsm* loci. (■) Light grey columns represent the number of GWA study nominated human SNPs mapping to the human orthologous regions of the random rat control regions. The difference between risk associated SNPs mapping to rat *Mcs/Mcsm* and random rat regions was statistically significant for European populations. Asterisk indicates P-value <0.05 using chi-square analysis with number of SNPs mapping to *Mcs/Mcsm* set as the observed value and number of SNPs mapping to random rat regions as the expected value.

is useful to predict and study human breast cancer risk loci. There was no difference in Asian or African-American descent populations. This is likely due to a limited number of published population-based breast cancer risk genetic-association studies using these populations.

B. Breast cancer risk GWA study nominated polymorphisms map more often to orthologs of all known rat mammary cancer loci than to randomly selected regions

Next, we included additional rat mammary cancer susceptibility loci that have been identified, but span large genomic segments. Loci added were *Mcstm1, Mcstm2, Mcsta1, Emca1-2* and *Emca4-8* [115-118]. The same random rat genomic regions used previously were used in this analysis to be consistent. Respectively, 179 and 51 GWA study nominated polymorphisms were located in human orthologous regions to rat mammary cancer loci and randomly selected rat regions (Figure 33A) when studies using populations of European descent were considered. This difference was statistically significant ($P < 0.01$). Note, some rat mammary cancer loci identified in independent studies have long regions of overlap. Consequently, several human GWA study identified polymorphisms mapped to human sequence orthologous to overlapping rat susceptibility loci. As it is not known if these rat loci contain unique sub-loci, human risk associated polymorphisms mapping to overlapping rat regions were counted only once. The size of the rat genome covered by all known rat mammary cancer susceptibility loci compared to control loci was disproportionate (Table 23). However, the ratio of breast cancer risk associated human SNPs at orthologs to rat mammary cancer susceptibility loci

A


Number of GWAS identified SNPs by population
200
180
160
140
120
100
80
60
40
20
0
*
*
*
*
European
non- European combined
Asian
African-American
Population

B

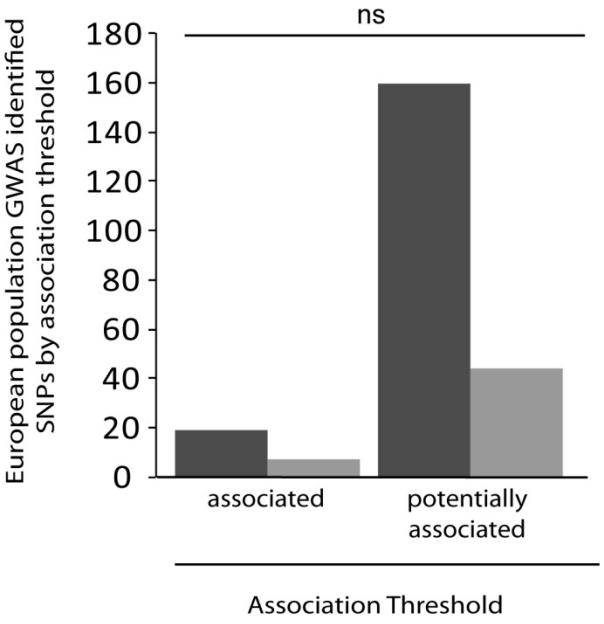

European population GWAS identified SNPs by association threshold
180
160
140
120
100
80
60
40
20
0
ns
associated
potentially associated
Association Threshold

Figure 33. Number of breast cancer risk GWA study nominated SNPs mapping to orthologs of rat mammary cancer loci or randomly selected rat genomic segments. Adapted from Sanders et al. (2014) [158]. ( ) Dark grey columns indicate GWA study nominated SNPs that map to human orthologous regions of rat mammary cancer loci. ( ) Light grey columns indicate GWA study nominated SNPs that mapped to human orthologous regions of randomly selected rat genomic regions. **A.** Studies by population descent. Asterisks indicate statistical significance ($p < 0.01$). The difference between risk associated SNPs mapping to rat mammary cancer loci and random rat regions in studies of European, Asian and African- American descent populations was significant (P-values <0.01 using chi-square analysis with number of SNPs mapping to rat mammary cancer loci set as the observed value and number of SNPs mapping to random rat regions as the expected value). **B.** Associated and potentially associated SNPs identified in populations of European descent that mapped to rat regions of interest were compared using logistic regression. Threshold of association was not a significant predictor of whether a SNP mapped to an ortholog of a rat mammary cancer locus or a random rat region. "ns" indicates a comparison was not statistically significant.

to SNPs at random segments was higher than the ratio of susceptibility loci bases to random bases (3.5 vs. 2.9). This result was relatively proportionate to the previous result when only rat *Mcs/Mcsm* loci were considered (1.29 for *Mcs/Mcsm* and 1.21 for all susceptibility loci), suggesting a potential bias was not introduced by the increase in total genomic coverage.

Not surprisingly, only 179 of 533 or 33.6% of the total human GWA study identified SNPs using populations of European descent were located at orthologs to rat mammary cancer associated loci. It is notable that 57 of the 533 total SNPs evaluated were reported in more than one GWA study; a majority of these were potential associations that failed the final validation step of the respective study. These results further suggest there are several breast cancer risk associated SNPs not reaching genome-wide statistical significance in human population-based genetic studies.

Since more breast cancer risk polymorphisms nominated from GWA studies of populations of European descent mapped to orthologs of rat mammary cancer loci than to randomly selected regions of the rat genome, we determined if this was the case for association studies using non-European descent populations. We queried the nine GWA studies of populations of non-European ancestry that are listed in Table 25. These were GWA studies using populations of African, African-American, Ashkenazi Jewish, and Asian descent; however, only polymorphisms from studies using African-American, Ashkenazi Jewish and Asian descent populations mapped to any of the human orthologous segments to rat genomic regions picked for this study. First, results from all studies of non-European descent populations were combined (Figure 33A). Eighty-nine risk associated SNPs mapped to orthologs of rat mammary cancer loci and 26 SNPs were

located at randomly selected rat regions. Next, studies using populations of Asian, Ashkenazi Jewish and African-American descent were considered separately. This resulted in 64 Asian descent population SNPs mapping to orthologs of rat mammary cancer loci and 18 SNPs to random rat regions. Twenty-four SNPs identified in studies of African-American descent populations were located at orthologs to rat mammary cancer loci and eight SNPs in random rat regions. The difference between rat mammary cancer loci and random regions was statistically significant ($p < 0.01$) for both populations (Figure 33A). Interestingly, one SNP from a study of an Ashkenazi Jewish population mapped to the human orthologous region of rat *Mcsta1,* but no GWA study nominated SNP from that study mapped to a rat random region [77]. The lack of human SNPs mapping to orthologs of rat mammary cancer loci from populations of African and Ashkenazi Jewish decent may be due to a limited number of studies conducted on these populations. On the other hand, it may indicate that susceptibility alleles different from those currently identified in laboratory rats are segregating in these populations. Out of 285 SNPs considered from studies using populations of non-European descent, 89 SNPs or 31% mapped to orthologs of rat mammary cancer loci. Fifteen risk associated SNPs were represented in more than one human GWA study.

Next, GWA-study nominated variants from populations of European descent were separated by associated (reached genome-wide significance) and potentially associated (did not reach genome-wide significance after the final stage) variants (Figure 33B). Nineteen associated SNPs were located at rat mammary cancer loci compared to seven SNPs that mapped to random rat regions. Comparatively, 160 potentially associated SNPs mapped to rat mammary cancer susceptibility loci compared to 44 SNPs that

mapped to random rat regions. A logistic regression was performed using threshold of association (associated or potentially associated) as the independent variable and rat genome location (ortholog of a rat mammary cancer risk locus or a randomly selected locus) as the dependent variable. Threshold of association was not a significant effect (P-value = 0.54). This result, that both associated and potentially associated breast cancer risk variants map more often to orthologs of rat mammary cancer risk loci than rat regions not associated with susceptibility, strongly supports that comparative genomics between humans and rats may be an effective integrative approach to determine which potential associations nominated by human association studies are true positives.

Human populations have been studied more extensively for breast cancer genetic risk than have rat populations; therefore, it is not surprising that human studies have yielded a considerable number of genome-wide significantly associated SNPs in alleles where it is not known if the rat genome contains a concordant allele. Interestingly, seven strongly associated human SNPs were in sequences orthologous to the randomly selected rat genome regions that are not known to associate with rat mammary cancer based on studies evaluating specific rat strains; thus, it is possible that a portion of the rat genome used in this study as rat random-genome control regions may actually associate with unidentified rat mammary cancer susceptibility loci. Thus, more rat genomic regions associated with mammary cancer risk may be identified with additional rat genetic studies. To date, only six inbred rat strains have been used to identify rat genomic regions associated with mammary cancer risk [106, 107, 115-118]. Therefore, it is highly likely that more mammary cancer susceptibility loci may be identified by incorporating additional rat strains. It is also possible that more extensive analysis of previously

studied rat strains may yield additional susceptibility loci by using a higher density of genetic markers for example.

Twenty-one of the 24 known rat mammary cancer associated loci are orthologous to human loci containing SNPs that are either associated or potentially associated to breast cancer risk. F ourteen of the known rat mammary cancer associated loci are orthologous to human risk alleles marked by GWA study nominated SNPs reaching genome-wide significance. Human GWA study designs do not definitively determine causative genes or mechanisms. The laboratory rat is a versatile experimental organism to complement human studies of breast cancer. For example, inbred rat strains provide a model with reduced genetic variation that can be genetically manipulated and environmentally controlled. The overlap between human breast and rat mammary cancer susceptibility associated loci suggests rats can be used extensively to study genetically determined mechanisms and environment interactions that will translate directly to human breast cancer risk and prevention.

C. Human GWAS nominated breast cancer risk SNPs map similarly to rat mammary cancer associated loci identified using 7,12-dimethylbenz[a]anthracene (DMBA) or beta-estradiol

Several rat mammary cancer loci used in this study were identified using DMBA to induce mammary tumors. These are *Mcs1a-c, Mcs2-4, Mcs5a1, Mcs5a2, Mcs5b-c, Mcs6-Mcs8, Mcsm1, Mcstm1-2* and *Mcsta1*. The remaining rat mammary cancer loci considered were identified using beta- estradiol to induce mammary carcinogenesis. Estradiol associated susceptibility loci are *Emca1-2* and *Emca4-8*. While DMBA is

representative of environmental polycyclic aromatic hydrocarbons, this synthesized mammary carcinogen is not found in nature. Conversely, estradiol is an endogenous environmental exposure associated with breast cancer risk. H uman GWA study nominated SNPs mapping to orthologs of rat mammary cancer loci identified using DMBA were compared to those identified using beta-estradiol. We considered SNPs from all GWA studies, irrespective of the population used. We noted that many DMBA and beta-estradiol identified rat mammary cancer loci overlap. In fact, seven of the 14 DMBA associated rat mammary cancer loci overlap at least one beta-estradiol associated rat mammary cancer risk locus, and five of the seven beta-estradiol loci overlap rat mammary cancer loci identified using DMBA. To account for this overlap, human SNPs mapping to overlapping rat mammary cancer loci, one identified using DMBA and the other using beta-estradiol, were included once in the "DMBA" group and once in the "beta-estradiol" group. These results are shown in Figure 34. A relatively similar number of GWA study nominated SNPs mapped to orthologs of rat mammary cancer loci that were identified using DMBA (181 SNPs) and beta estradiol (146 SNPs). This suggests that different mammary carcinoma induction methods can effectively identify rat susceptibility loci relevant to human disease risk, and it also suggests that a plethora of carcinogenesis mechanisms may be genetically determined.

Discussion

It has been suggested that the use of Bonferroni-based correction procedures to protect against multiple comparisons in genome-wide association studies is too stringent and results in an abundance of false negative associations with little recourse to sort these

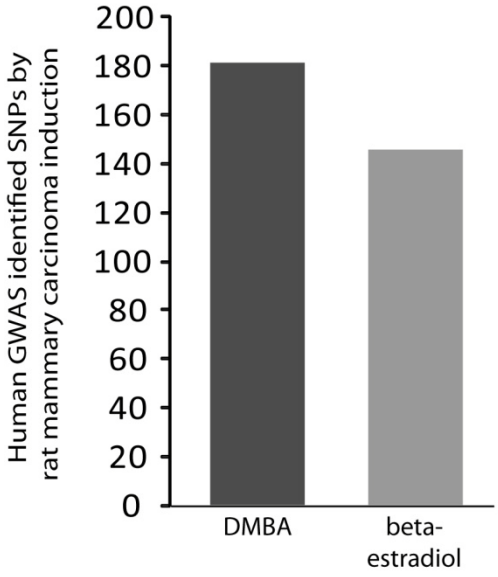


Figure 34. Number of breast cancer risk GWA study nominated SNPs mapping to regions identified using DMBA or beta-estradiol. Adapted from Sanders et al. (2014) [158]. Number of GWA study nominated SNPs mapping to rat mammary cancer loci separated by method of mammary carcinogenesis induction. Slightly more SNPs mapped to orthologs of rat loci that were identified using DMBA than beta-estradiol.

from true-negative associations. Therefore, we considered associated and potentially associated human SNPs from breast cancer risk GWA studies to determine if SNPs that failed validation and SNPs that reached genome- wide significance map to respective regions of the rat genome known to associate with rat mammary cancer risk more often than to regions of the rat genome that are not known to associate with susceptibility. Results presented here indicate that the rat genome is useful to prioritize and rank human alleles potentially associated with risk. The rat genome is useful regardless of the human population studied. Significantly more SNPs from GWA studies of populations of European, Asian, and African-American descent map to human orthologous regions of rat mammary cancer loci than to human orthologous regions of randomly selected rat genomic regions not known to associate with mammary cancer susceptibility. This supports the general idea that there are SNPs associated with breast cancer risk that are missed due to conservative statistical methods used in GWA studies, and that the rat is useful to parse out important genetic variation in susceptibility to mammary carcinogenesis.

Interestingly, we were unable to map GWA study nominated SNPs to three of the 24 known rat mammary cancer loci. These were *Mcs1a, Mcs5a1,* and *Mcs5c*. However, using a genome-targeted population-based genetic association study, a human SNP associated with breast cancer risk has been identified at human *MCS5A1* [111]. The risk associated SNPs at *MCS5A1* are adjacent to a breast cancer risk associated SNP at *MCS5A2*, which was identified in two independent human population based studies [111, 178]. Taken together, there is a high correlation between genetics of breast cancer susceptibility in humans and mammary cancer susceptibility in rats. Interestingly, there

are several human genomic regions that are human GWA study nominated hotspots (*e.g.* 19q13, *FGFR2*) that are not known to have concordant rat orthologs. An explanation is that human breast and rat mammary cancer susceptibility are controlled by overlapping and non-overlapping genetic mechanisms. Another explanation is that there are rat genomic regions associated with mammary cancer risk yet to be discovered by using additional inbred strains, more extensive analysis of strains previously studied, and different methods of carcinogenesis induction.

CHAPTER VI

CONCLUDING REMARKS

Breast cancer is a complex disease, characterized by genetic, epigenetic and environmental factors. The laboratory rat has been used extensively as a model to study the genetic component of breast cancer. Several rat genomic regions have been identified that associate with mammary cancer risk. This dissertation focuses on two of these regions, *Mcs6* and *Mcs1b*. The *Mcs6* locus was initially mapped to a large region of about 33Mb. This makes this locus too large for functional studies. We focused on narrowing this locus to a smaller interval using phenotyping of congenic animals. We were able to conclude that the *Mcs6* locus is located in an 8.5Mb region on rat chromosome *2*: 46,915,037-55,364,398. This significantly reduced the size of this locus. Several GWAS identified SNPs map to this location, making it an important locus to study. Overall, we were able to generate the model needed to study this locus with this project. Studying this locus is highly relevant, because there are no known breast cancer susceptibility genes in this region and studying this locus may lead to the identification of a novel breast cancer susceptibility gene.

The second rat genomic region, this dissertation focuses on, is the *Mcs1b* locus. We used next-generation sequencing to identify 72 variants between the two rat strains in this region. Most of the variants are located at the extreme ends of the *Mcs1b* locus. We used genotyping of existing congenic animals to rule out most of the variants. This

resulted in the identification of three candidate *Mcs1b* SNPs and one INDEL within this locus that have the potential of being responsible for the *Mcs1b* phenotypes. We focused on the *Mcs1b* candidate SNPs because there are known orthologous human SNPs that associate with breast cancer. Overall, we were able to use sequence capture using rat DNA, which to our knowledge had not been done previously and we were able to use sequencing, genetic analysis and bioinformatics to generate a list of candidate *Mcs1b* variants.

Functional analysis of the candidate *Mcs1b* SNPs revealed that *A074-SNP-17* shows a gene regulatory pattern that mimics a gene regulatory pattern that is seen endogenously in the rat strains used for this analysis. Furthermore, a DNA binding protein analysis revealed that some protein complexes appear only to bind to the SNP allele of the mammary cancer susceptible rat strain, indicating that there is a difference in the DNA binding proteins that bind to *A074-SNP-17*. Based on these data, *A074-SNP-17* is considered the strongest candidate for conferring the *Mcs1b* phenotypes. Data from 3C experiments reveal that there appears to be a chromatin loop forming in the *Mcs1b* region, which would bring our candidate *Mcs1b* SNPs in close proximity to the *Mcs1b* candidate gene *Mier3*. There does not appear to be a difference in chromatin loop formation in the *Mcs1b* region between the mammary cancer resistant and susceptible rat strains used in our analysis.

Overall, our model for the *Mcs1b* locus is that the *Mcs1b* candidate SNPs are located in enhancer regions and affect the transcriptional regulation of nearby genes. The luciferase activity for all three Mcs1b candidate SNPs is different between the susceptible WF and resistant COP allele, suggesting that all three SNPs act as transcriptional

regulators. *A074-SNP-17* shows a luciferase activity pattern that is similar to the pattern of differential gene expression of *Mcs1b* candidate SNPs. This indicates that *A074-SNP-17* may be involved in regulating the expression levels of *Mcs1b* candidate genes. Also, analysis of DNA binding proteins revealed that a DNA binding protein binds to the WF allele of *A074-SNP-17* but not the COP allele. This makes *A074-SNP-17* the most likely candidate for conferring the *Mcs1b* phenotype. There is evidence that a chromatin loop forms in the *Mcs1b* locus, which brings the candidate *Mcs1b* SNPs in close proximity to one of the *Mcs1b* candidate genes. This would be a mechanism as to how *A074-SNP-17* can affect the expression level of the *Mcs1b* candidate gene.

GWA studies have a great potential of identifying candidate SNPs associated with breast cancer. However, they provide no information on the causative gene, variant or what the underlying mechanism is. We were able to generate an animal model to study one GWA identified SNP, *rs889312*, and the SNPs it tags. We were able to study the role of these SNPs in gene regulation and provide a mechanism by which they could affect the transcript levels of candidate genes. The only other GWA identified SNPs have been studied for function reside within the *FGFR2* gene. Therefore, studying these SNPs will result in the identification of novel breast cancer susceptibility genes and will elucidate the mechanisms underlying genetic susceptibility to breast cancer.

An analysis of the overlap between human breast and rat mammary cancer susceptibility revealed that there is extensive genomic overlap. The rat genome may provide utility to identify true-positive associations regardless of the human population used for a GWA study. The laboratory rat will continue to be an important model organism for researching genetically determined mechanisms of mammary cancer

susceptibility that may translate directly to human susceptibility. An appreciable number of GWA study nominated SNPs not meeting genome-wide significance levels have genomic overlap with rat mammary cancer susceptibility loci. This supports the general idea that Bonferroni-based multiple-comparison correction procedures are too stringent and complementary approaches that integrate rat genomics would be highly efficacious to prioritize breast cancer risk associated alleles.

REFERENCES

1. DeSantis C, Ma J, Bryan L, Jemal A: **Breast cancer statistics, 2013**. *CA: A Cancer Journal for Clinicians* 2013, **64**(1):52-62.
2. Siegel R, Ma J, Zou Z, Jemal A: **Cancer statistics, 2014**. *CA: A Cancer Journal for Clinicians* 2014, **64**(1):9-29.
3. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ: **Projecting individualized probabilities of developing breast cancer for white females who are being examined annually**. *J Natl Cancer Inst* 1989, **81**:1879-1886.
4. Thomson CA: **Diet and Breast Cancer: Understanding Risks and Benefits**. *Nutrition in Clinical Practice* 2012, **27**(5):636-650.
5. Mahabir S: **Association Between Diet During Preadolescence and Adolescence and Risk for Breast Cancer During Adulthood**. *Journal of Adolescent Health* 2013, **52**(5, Supplement):S30-S35.
6. Jansen-van der Weide M, Greuter MW, Jansen L, Oosterwijk J, Pijnappel R, Bock G: **Exposure to low-dose radiation and the risk of breast cancer among women with a familial or genetic predisposition: a meta-analysis**. *Eur Radiol* 2010, **20**(11):2547-2556.
7. Heyes GJ, Mill AJ, Charles MW: **Mammography—oncogenecity at low doses**. *Journal of Radiological Protection* 2009, **29**(2A):A123.
8. Rozenberg S, Vandromme J, Antoine C: **Postmenopausal hormone therapy: risks and benefits**. *Nat Rev Endocrinol* 2013, **9**(4):216-227.
9. Liang J, Shang Y: **Estrogen and Cancer**. *Annual Review of Physiology* 2013, **75**(1):225-240.
10. Reynolds P: **Smoking and Breast Cancer**. *J Mammary Gland Biol Neoplasia* 2013, **18**(1):15-23.
11. Pelucchi C, Tramacere I, Boffetta P, Negri E, Vecchia CL: **Alcohol Consumption and Cancer Risk**. *Nutrition & Cancer* 2011, **63**(7):983-990.
12. Brody JG, Moysich KB, Humblet O, Attfield KR, Beehler GP, Rudel RA: **Environmental pollutants and breast cancer**. *Cancer* 2007, **109**(S12):2667-2711.
13. Ronckers C, Erdmann C, Land C: **Radiation and breast cancer: a r eview of current evidence**. *Breast Cancer Res* 2005, **7**(1):21 - 32.
14. Carmichael A, Sami AS, Dixon JM: **Breast cancer risk among the survivors of atomic bomb and patients exposed to therapeutic ionising radiation**. *European Journal of Surgical Oncology (EJSO)* 2003, **29**(5):475-479.
15. Williams D: **Radiation carcinogenesis: lessons from Chernobyl**. *Oncogene* 2008, **27**(S2):S9-S18.

16. NTP. 2011. Report on Carcinogens TERTP, NC: U.S. Department of Health and Human Services, Public Health Service, National Toxicology Program. 499 pp. In.
17. Pike MC, Krailo MD, Henderson BE, Casagrande JT, Hoel DG: **Hormonal risk factors, breast tissue age and the age-incidence of breast cancer**. *Nature* 1983, **303**(5920):767-770.
18. Travis R, Key T: **Oestrogen exposure and breast cancer risk**. *Breast Cancer Res* 2003, **5**(5):239 - 247.
19. Narod SA: **Hormone replacement therapy and the risk of breast cancer**. *Nat Rev Clin Oncol* 2011, **8**(11):669-676.
20. Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, Vogel V, Robidoux A, Dimitrov N, Atkins J *et al*: **Tamoxifen for Prevention of Breast Cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study**. *Journal of the National Cancer Institute* 1998, **90**(18):1371-1388.
21. Woolcott C, Koga K, Conroy S, Byrne C, Nagata C, Ursin G, Vachon C, Yaffe M, Pagano I, Maskarinec G: **Mammographic density, parity and age at first birth, and risk of breast cancer: an analysis of four case–control studies**. *Breast Cancer Res Treat* 2012, **132**(3):1163-1171.
22. Gompel A, Somaï S, Chaouat M, Kazem A, Kloosterboer HJ, Beusman I, Forgez P, Mimoun M, Rostène W: **Hormonal regulation of apoptosis in breast cells and tissues**. *Steroids* 2000, **65**(10–11):593-598.
23. Bendrik C, Dabrosin C: **Estradiol Increases IL-8 Secretion of Normal Human Breast Tissue and Breast Cancer In Vivo**. *The Journal of Immunology* 2009, **182**(1):371-378.
24. Jefcoate CR, Liehr JG, Santen RJ, Sutter TR, Yager JD, Yue W, Santner SJ, Tekmal R, Demers L, Pauley R *et al*: **Chapter 5: Tissue-Specific Synthesis and Oxidative Metabolism of Estrogens**. *JNCI Monographs* 2000, **2000**(27):95-112.
25. Lewis JS, Jordan VC: **Selective estrogen receptor modulators (SERMs): Mechanisms of anticarcinogenesis and drug resistance**. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 2005, **591**(1–2):247-263.
26. Wardell SE, Marks JR, McDonnell DP: **The turnover of estrogen receptor α by the selective estrogen receptor degrader (SERD) fulvestrant is a saturable process that is not required for antagonist efficacy**. *Biochemical Pharmacology* 2011, **82**(2):122-130.
27. IARC: **Alcohol Consumption and Ethyl Carbamate**. In: *International Agency for Research on Cancer*. Lyon, France; 2010.
28. Bagnardi V, Rota M, Botteri E, Tramacere I, Islami F, Fedirko V, Scotti L, Jenab M, Turati F, Pasquali E *et al*: **Light alcohol drinking and cancer: a meta-analysis**. *Annals of Oncology* 2013, **24**(2):301-308.
29. Oyesanmi O, Snyder D, Sullivan N, Reston J, Treadwell J, K S: **Alcohol Consumption and Cancer Risk: Understanding Possible Causal Mechanisms for Breast and Colorectal Cancers.** In. Rockville (MD): Agency for Healthcare Research and Quality; 2010.

30. Seitz HK, Becker P: **Alcohol metabolism and cancer risk**. *Alcohol Research & Health* 2007, **30**(1):38-47.
31. Johnson KC, Miller AB, Collishaw NE, Palmer JR, Hammond SK, Salmon AG, Cantor KP, Miller MD, Boyd NF, Millar J *et al*: **Active smoking and secondhand smoke increase breast cancer risk: the report of the Canadian Expert Panel on Tobacco Smoke and Breast Cancer Risk (2009)**. *Tobacco Control* 2011, **20**(1):e2.
32. Hecht SS: **Tobacco smoke carcinogens and breast cancer**. *Environmental and Molecular Mutagenesis* 2002, **39**(2-3):119-126.
33. Cox DG, Dostal L, Hunter DJ, Le Marchand L, Hoover R, Ziegler RG, Thun MJ: **N-Acetyltransferase 2 Polymorphisms, Tobacco Smoking, and Breast Cancer Risk in the Breast and Prostate Cancer Cohort Consortium**. *American Journal of Epidemiology* 2011, **174**(11):1316-1322.
34. Terry PD, Goodman M: **Is the Association between Cigarette Smoking and Breast Cancer Modified by Genotype? A Review of Epidemiologic Studies and Meta-analysis**. *Cancer Epidemiology Biomarkers & Prevention* 2006, **15**(4):602-611.
35. Rudel RA, Attfield KR, Schifano JN, Brody JG: **Chemicals causing mammary gland tumors in animals signal new directions for epidemiology, chemicals testing, and risk assessment for breast cancer prevention**. *Cancer* 2007, **109**(S12):2635-2666.
36. Safe S, Bandiera S, Sawyer T, Robertson L, Safe L, Parkinson A, Thomas PE, Ryan DE, Reik LM, Levin W *et al*: **PCBs: Structure-Function Relationships and Mechanism of Action**. *Environmental Health Perspectives* 1985, **60**(ArticleType: research-article / Full publication date: May, 1985 / Copyright © 1985 The National Institute of Environmental Health Sciences (NIEHS)):47-56.
37. Lauby-Secretan B, Loomis D, Grosse Y, Ghissassi FE, Bouvard V, Benbrahim-Tallaa L, Guha N, Baan R, Mattock H, Straif K: **Carcinogenicity of polychlorinated biphenyls and polybrominated biphenyls**. *The Lancet Oncology* 2013, **14**(4):287-288.
38. Prentice RL, Shaw PA, Bingham SA, Beresford SAA, Caan B, Neuhouser ML, Patterson RE, Stefanick ML, Satterfield S, Thomson CA *et al*: **Biomarker-calibrated Energy and Protein Consumption and Increased Cancer Risk Among Postmenopausal Women**. *American Journal of Epidemiology* 2009, **169**(8):977-989.
39. World Cancer Research Fund, American Institute for Cancer Research: **Food, Nutrition, Physical Activity, and the Prevention of Cancer: A Global Perspective**. In. Washington D.C. : American Institute for Cancer Research; 2007.
40. Turner LB: **A meta-analysis of fat intake, reproduction, and breast cancer risk: An evolutionary perspective**. *American Journal of Human Biology* 2011, **23**(5):601-608.
41. Schulz M, Hoffmann K, Weikert C, Nöthlings U, Schulze MB, Boeing H: **Identification of a dietary pattern characterized by high-fat food choices associated with increased risk of breast cancer: the European Prospective**

Investigation into Cancer and Nutrition (EPIC)-Potsdam Study. *The British Journal of Nutrition* 2008, **100**(5):942-946.

42. Young LR, Kurzer MS, Thomas W, Redmon JB, Raatz SK: **Effect of Dietary Fat and Omega-3 Fatty Acids on Urinary Eicosanoids and Sex Hormone Concentrations in Postmenopausal Women: A Randomized Controlled Feeding Trial**. *Nutrition and Cancer* 2011, **63**(6):930-939.
43. Dong J-Y, He K, Wang P, Qin L-Q: **Dietary fiber intake and risk of breast cancer: a meta-analysis of prospective cohort studies**. *The American Journal of Clinical Nutrition* 2011, **94**(3):900-905.
44. Trock BJ, Hilakivi-Clarke L, Clarke R: **Meta-Analysis of Soy Intake and Breast Cancer Risk**. *Journal of the National Cancer Institute* 2006, **98**(7):459-471.
45. Kushi LH, Doyle C, McCullough M, Rock CL, Demark-Wahnefried W, Bandera EV, Gapstur S, Patel AV, Andrews K, Gansler T *et al*: **American Cancer Society guidelines on nutrition and physical activity for cancer prevention**. *CA: A Cancer Journal for Clinicians* 2012, **62**(1):30-67.
46. Warri A, Saarinen NM, Makela S, Hilakivi-Clarke L: **The role of early life genistein exposures in modifying breast cancer risk**. *Br J Cancer* 2008, **98**(9):1485-1493.
47. Dolinoy DC, Weidman JR, Waterland RA, Jirtle RL: **Maternal Genistein Alters Coat Color and Protects Avy Mouse Offspring from Obesity by Modifying the Fetal Epigenome**. *Environmental Health Perspectives* 2006, **114**(4):567-572.
48. Day JK, Bauer AM, desBordes C, Zhuang Y, Kim B-E, Newton LG, Nehra V, Forsee KM, MacDonald RS, Besch-Williford C *et al*: **Genistein Alters Methylation Patterns in Mice**. *The Journal of Nutrition* 2002, **132**(8):2419S-2423S.
49. Wu H-C, Delgado-Cruzata L, Flom JD, Perrin M, Liao Y, Ferris JS, Santella RM, Terry MB: **Repetitive element DNA methylation levels in white blood cell DNA from sisters discordant for breast cancer from the New York site of the Breast Cancer Family Registry**. *Carcinogenesis* 2012, **33**(10):1946-1952.
50. Steinhoff C, Schulz WA: **Transcriptional regulation of the human LINE-1 retrotransposon L1.2B**. *Mol Genet Genomics* 2004, **270**(5):394-402.
51. Yu F, Zingler N, Schumann G, Strätling WH: **Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription**. *Nucleic Acids Research* 2001, **29**(21):4493-4501.
52. Palmer JR, Wise LA, Hatch EE, Troisi R, Titus-Ernstoff L, Strohsnitter W, Kaufman R, Herbst AL, Noller KL, Hyer M *et al*: **Prenatal Diethylstilbestrol Exposure and Risk of Breast Cancer**. *Cancer Epidemiology Biomarkers & Prevention* 2006, **15**(8):1509-1514.
53. de Assis S, Warri A, Cruz MI, Laja O, Tian Y, Zhang B, Wang Y, Huang TH-M, Hilakivi-Clarke L: **High-fat or ethinyl-oestradiol intake during pregnancy increases mammary cancer risk in several generations of offspring**. *Nat Commun* 2012, **3**:1053.
54. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland**. *New England Journal of Medicine* 2000, **343**(2):78-85.

55. Ahlbom A, Lichtenstein P, Malmström H, Feychting M, Pedersen NL, Hemminki K: **Cancer in Twins: Genetic and Nongenetic Familial Risk Factors**. *Journal of the National Cancer Institute* 1997, **89**(4):287-293.
56. Verkasalo PK, Kaprio J, Koskenvuo M, Pukkala E: **Genetic predisposition, environment and cancer incidence: A nationwide twin study in Finland, 1976–1995**. *International Journal of Cancer* 1999, **83**(6):743-749.
57. Hindorff LA, Gillanders EM, Manolio TA: **Genetic architecture of cancer and other complex diseases: lessons learned and future directions**. *Carcinogenesis* 2011, **32**(7):945-954.
58. Harris TJR, McCormick F: **The molecular pathology of cancer**. *Nat Rev Clin Oncol* 2010, **7**(5):251-265.
59. Ghoussaini M, Pharoah PDP, Easton DF: **Inherited Genetic Susceptibility to Breast Cancer: The Beginning of the End or the End of the Beginning?** *The American Journal of Pathology* 2013, **183**(4):1038-1051.
60. Mavaddat N, Antoniou AC, Easton DF, Garcia-Closas M: **Genetic susceptibility to breast cancer**. *Molecular Oncology* 2010, **4**(3):174-191.
61. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R *et al*: **Genome-wide association study identifies novel breast cancer susceptibility loci**. *Nature* 2007, **447**(7148):1087-1093.
62. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK *et al*: **Large-scale genotyping identifies 41 new loci associated with breast cancer risk**. *Nat Genet* 2013, **45**(4):353-361.
63. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, et al.: **Linkage of Early-Onset Familial Breast Cancer to Chromosome 17q21**. *Science* 1990, **250**(4988):1684-1684.
64. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W *et al*: **A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1**. *Science* 1994, **266**(5182):66-71.
65. Deng C-X, Brodie SG: **Roles of BRCA1 and its interacting proteins**. *BioEssays* 2000, **22**(8):728-737.
66. Caestecker KW, Van de Walle GR: **The role of BRCA1 in DNA double-strand repair: Past and present**. *Experimental Cell Research* 2013, **319**(5):575-587.
67. Rahman N, Stratton MR: **THE GENETICS OF BREAST CANCER SUSCEPTIBILITY**. *Annual Review of Genetics* 1998, **32**(1):95-121.
68. Wooster R, Neuhausen S, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D *et al*: **Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13**. *Science* 1994, **265**(5181):2088-2090.
69. Jensen RB: **BRCA2: One Small Step for DNA Repair, One Giant Protein Purified**. *Yale J Biol Med* 2013, **86**(4):479–489.
70. Karami F, Mehdipour P: **A Comprehensive Focus on Global Spectrum of BRCA1 and BRCA2 Mutations in Breast Cancer**. *BioMed Research International* 2013, **2013**:21.

71. Euhus DM, Robinson L: **Genetic Predisposition Syndromes and Their Management**. *Surgical Clinics of North America* 2013, **93**(2):341-362.
72. Reis LO, Dias FG, Castro MA, Ferreira U: **Male breast cancer**. In: *Aging Male.* vol. 14: Taylor & Francis Ltd; 2011: 99-109.
73. Apostolou P, Fostira F: **Hereditary Breast Cancer: The Era of New Susceptibility Genes**. *BioMed Research International* 2013, **2013**:11.
74. Bradbury A, Olopade O: **Genetic susceptibility to breast cancer**. *Rev Endocr Metab Disord* 2007, **8**(3):255-267.
75. Antoniou AC, Sinilnikova OM, McGuffog L, Healey S, Nevanlinna H, Heikkinen T, Simard J, Spurdle AB, Beesley J, Chen X *et al*: **Common variants in LSP1, 2q35 and 8q24 and breast cancer risk for BRCA1 and BRCA2 mutation carriers**. *Human Molecular Genetics* 2009, **18**(22):4442-4456.
76. Gaudet MM, Kirchhoff T, Green T, Vijai J, Korn JM, Guiducci C, Segrè AV, McGee K, McGuffog L, Kartsonaki C *et al*: **Common Genetic Variants and Modification of Penetrance of BRCA2-Associated Breast Cancer**. *PLoS Genet* 2010, **6**(10):e1001183.
77. Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, Friedman E, Narod S, Olshen AB, Gregersen P *et al*: **Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33**. *Proceedings of the National Academy of Sciences* 2008, **105**(11):4340-4345.
78. Kim H-c, Lee J-Y, Sung H, Choi J-Y, Park S, Lee K-M, Kim Y, Go M, Li L, Cho Y *et al*: **A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study**. *Breast Cancer Research* 2012, **14**(2):R56.
79. Meyer KB, Maia A-T, O'Reilly M, Teschendorff AE, Chin S-F, Caldas C, Ponder BAJ: **Allele-Specific Up-Regulation of FGFR2 Increases Susceptibility to Breast Cancer**. *PLoS Biol* 2008, **6**(5):e108.
80. Meyer Kerstin B, O'Reilly M, Michailidou K, Carlebur S, Edwards Stacey L, French Juliet D, Prathalingham R, Dennis J, Bolla MK, Wang Q *et al*: **Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1**. *The American Journal of Human Genetics* 2013, **93**(6):1046-1060.
81. Jara L, Gonzalez-Hormazabal P, Cerceño K, Di Capua G, Reyes J, Blanco R, Bravo T, Peralta O, Gomez F, Waugh E *et al*: **Genetic variants in FGFR2 and MAP3K1 are associated with the risk of familial and early-onset breast cancer in a South-American population**. *Breast Cancer Res Treat* 2013, **137**(2):559-569.
82. Chan M, Ji SM, Liaw CS, Yap YS, Law HY, Yoon CS, Wong CY, Yong WS, Wong NS, Ng R *et al*: **Association of common genetic variants with breast cancer risk and clinicopathological characteristics in a Chinese population**. *Breast Cancer Res Treat* 2012, **136**(1):209-220.
83. Shan J, Mahfoudh W, Dsouza S, Hassen E, Bouaouina N, Abdelhak S, Benhadjayed A, Memmi H, Mathew R, Aigha I *et al*: **Genome-Wide Association Studies (GWAS) breast cancer susceptibility loci in Arabs: susceptibility and prognostic implications in Tunisians**. *Breast Cancer Res Treat* 2012, **135**(3):715-724.

84. Garcia-Closas M, Chanock S: **Genetic Susceptibility Loci for Breast Cancer by Estrogen Receptor Status**. *Clinical Cancer Research* 2008, **14**(24):8000-8009.
85. Broeks A, Schmidt MK, Sherman ME, Couch FJ, Hopper JL, Dite GS, Apicella C, Smith LD, Hammet F, Southey MC *et al*: **Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium**. *Human Molecular Genetics* 2011, **20**(16):3289-3303.
86. Garcia-Closas M, Hall P, Nevanlinna H, Pooley K, Morrison J, Richesson DA, Bojesen SE, Nordestgaard BG, Axelsson CK, Arias JI *et al*: **Heterogeneity of Breast Cancer Associations with Five Susceptibility Loci by Clinical and Pathological Characteristics**. *PLoS Genet* 2008, **4**(4):e1000054.
87. denDekker AD, Xu X, Vaughn MD, Puckett AH, Gardner LL, Lambring CJ, Deschenes L, Samuelson DJ: **Rat Mcs1b Is Concordant to the Genome-Wide Association-Identified Breast Cancer Risk Locus at Human 5q11.2 and MIER3 is a Candidate Cancer Susceptibility Gene**. *Cancer Research* 2012, **72**(22):6002-6012.
88. Development Organization Workgroup Asn ATCCS: **Cell line misidentification: the beginning of the end**. *Nat Rev Cancer* 2010, **10**(6).
89. Burdall S, Hanby A, Lansdown M, Speirs V: **Breast cancer cell lines: friend or foe?** *Breast Cancer Res* 2003, **5**(2):89 - 95.
90. Pinho SS, Carvalho S, Cabral J, Reis CA, Gärtner F: **Canine tumors: a spontaneous animal model of human carcinogenesis**. *Translational Research* 2012, **159**(3):165-172.
91. Borowsky AD: **Choosing a Mouse Model: Experimental Biology in Context—The Utility and Limitations of Mouse Models of Breast Cancer**. *Cold Spring Harb Perspect Biol* 2011, **3**(9):a009670.
92. Gould MN: **Rodent models for the study of etiology, prevention and treatment of breast cancer**. *Seminars in Cancer Biology* 1995, **6**(3):147-152.
93. Kretschmann KL, Welm AL: **Mouse models of breast cancer metastasis to bone**. *Cancer and Metastasis Reviews* 2012, **31**(3-4):579-583.
94. Brodie SG, Xu X, Qiao W, Li WM, Cao L, Deng CX: **Multiple genetic changes are associated with mammary tumorigenesis in Brca1 conditional knockout mice**. *Oncogene* 2001, **20**(51):7514-7523.
95. Ludwig T, Fisher P, Murty V, Efstratiadis A: **Development of mammary adenocarcinomas by tissue-specific knockout of Brca2 in mice**. *Oncogene* 2001, **20**(30):3937-3948.
96. Blackburn A, Jerry D: **Map Making in the 21st Century: Charting Breast Cancer Susceptibility Pathways in Rodent Models**. *J Mammary Gland Biol Neoplasia* 2011, **16**(1):57-64.
97. Russo IH, Russo J: **Role of Hormones in Mammary Cancer Initiation and Progression**. *J Mammary Gland Biol Neoplasia* 1998, **3**(1):49-61.
98. Russo J, Russo IH: **Atlas and Histologic Classification of Tumors of the Rat Mammary Gland**. *J Mammary Gland Biol Neoplasia* 2000, **5**(2):187-200.
99. Medina D: **Chemical Carcinogenesis of Rat and Mouse Mammary Glands**. *Breast Disease* 2007, **28**(1):63-68.

100. Shull JD: **The Rat Oncogenome: Comparative Genetics and Genomics of Rat Models of Mammary Carcinogenesis**. *Breast Disease* 2007, **28**(1):69-86.
101. Huggins C, Grand LC, Brillantes FP: **Mammary Cancer Induced by a Single Feeding of Polynuclear Hydrocarbons, and its Suppression**. *Nature* 1961, **189**(4760):204-207.
102. Moore CJ, Bachhuber AJ, Gould MN: **Relationship of Mammary Tumor Susceptibility, Mammary Cell-Mediated Mutagenesis, and Metabolism of Polycyclic Aromatic Hydrocarbons in Four Types of Rats**. *Journal of the National Cancer Institute* 1983, **70**:777-784.
103. Isaacs JT: **Genetic Control of Resistance to Chemically Induced Mammary Adenocarcinogenesis in the Rat**. *Cancer Research* 1986, **46**(8):3958-3963.
104. Haag JD, Newton MA, Gould MN: **Mammary carcinoma suppressor and susceptibility genes in the Wistar–Kyoto rat**. *Carcinogenesis* 1992, **13**(10):1933-1935.
105. Hsu L-C, Kennan WS, Shepel LA, Jacob HJ, Szpirer C, Szpirer J, Lander ES, Gould MN: **Genetic Identification of Mcs-1, a Rat Mammary Carcinoma Suppressor Gene**. *Cancer Research* 1994, **54**(10):2765-2770.
106. Shepel LA, Lan H, Haag JD, Brasic GM, Gheen ME, Simon JS, Hoff P, Newton MA, Gould MN: **Genetic Identification of Multiple Loci That Control Breast Cancer Susceptibility in the Rat**. *Genetics* 1998, **149**(1):289-299.
107. Lan H, Kendziorski CM, Haag JD, Shepel LA, Newton MA, Gould MN: **Genetic Loci Controlling Breast Cancer Susceptibility in the Wistar-Kyoto Rat**. *Genetics* 2001, **157**(1):331-339.
108. Sanders J, Haag JD, Samuelson DJ: **Physical Confirmation and Mapping of Overlapping Rat Mammary Carcinoma Susceptibility QTLs, Mcs2 and Mcs6**. *PloS One* 2011, **6**(5):e19891.
109. Samuelson DJ, Haag JD, Lan H, Monson DM, Shultz MA, Kolman BD, Gould MN: **Physical evidence of Mcs5, a Q TL controlling mammary carcinoma susceptibility, in congenic rats**. *Carcinogenesis* 2003, **24**(9):1455-1460.
110. Samuelson DJ, Aperavich BA, Haag JD, Gould MN: **Fine Mapping Reveals Multiple Loci and a Possible Epistatic Interaction within the Mammary Carcinoma Susceptibility Quantitative Trait Locus, Mcs5**. *Cancer Research* 2005, **65**(21):9637-9642.
111. Samuelson DJ, Hesselson SE, Aperavich BA, Zan Y, Haag JD, Trentham-Dietz A, Hampton JM, Mau B, Chen K-S, Baynes C *et al*: **Rat Mcs5a is a compound quantitative trait locus with orthologous human loci that associate with breast cancer risk**. *Proceedings of the National Academy of Sciences* 2007, **104**(15):6299-6304.
112. Smits B, Sharma D, Samuelson D, Woditschka S, Mau B, Haag J, Gould M: **The non-protein coding breast cancer susceptibility locus Mcs5a acts in a non-mammary cell-autonomous fashion through the immune system and modulates T-cell homeostasis and functions**. *Breast Cancer Research* 2011, **13**(4):R81.
113. Smits BMG, Traun BD, Devries TL, Tran A, Samuelson D, Haag JD, Gould M: **An insulator loop resides between the synthetically interacting elements of**

the human/rat conserved breast cancer susceptibility locus MCS5A/Mcs5a. *Nucleic Acids Research* 2012, **40**(1):132-147.

114. Kim HY, Stewart TP, Wyatt BN, Siriwardhana N, Saxton AM, Kim JH: **Gene expression profiles of a mouse congenic strain carrying an obesity susceptibility QTL under obesigenic diets**. *Genes Nutr* 2010, **5**(3):237- 250.
115. Stieber D, Piessevaux G, Rivière M, Laes J-F, Quan X, Szpirer J, Szpirer C: **Isolation of two regions on rat chromosomes 5 and 18 affecting mammary cancer susceptibility**. *International Journal of Cancer* 2007, **120**(8):1678-1683.
116. Piessevaux G, Lella V, Rivière M, Stieber D, Drèze P, Szpirer J, Szpirer C: **Contrasting epistatic interactions between rat quantitative trait loci controlling mammary cancer development**. *Mamm Genome* 2009, **20**(1):43-52.
117. Gould KA, Tochacek M, Schaffer BS, Reindl TM, Murrin CR, Lachel CM, VanderWoude EA, Pennington KL, Flood LA, Bynote KK *et al*: **Genetic Determination of Susceptibility to Estrogen-Induced Mammary Cancer in the ACI Rat: Mapping of Emca1 and Emca2 to Chromosomes 5 and 18**. *Genetics* 2004, **168**(4):2113-2125.
118. Schaffer BS, Lachel CM, Pennington KL, Murrin CR, Strecker TE, Tochacek M, Gould KA, Meza JL, McComb RD, Shull JD: **Genetic Bases of Estrogen-Induced Tumorigenesis in the Rat: Mapping of Loci Controlling Susceptibility to Mammary Cancer in a Brown Norway × ACI Intercross**. *Cancer Research* 2006, **66**(15):7793-7800.
119. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler, David: **The Human Genome Browser at UCSC**. *Genome Research* 2002, **12**(6):996-1006.
120. Murabito J, Rosenberg C, Finger D, Kreger B, Levy D, Splansky G, Antman K, Hwang S-J: **A genome-wide association study of breast and prostate cancer in the NHLBI's Framingham Heart Study**. *BMC Medical Genetics* 2007, **8**(Suppl 1):S6.
121. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K *et al*: **A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1)**. *Nat Genet* 2009, **41**(5):579-584.
122. Antoniou AC, Wang X, Fredericksen ZS, McGuffog L, Tarrell R, Sinilnikova OM, Healey S, Morrison J, Kartsonaki C, Lesnick T *et al*: **A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population**. *Nat Genet* 2010, **42**(10):885-892.
123. Haag JD, Shepel LA, Kolman BD, Monson DM, Benton ME, Watts KT, Waller JL, Lopez-Guajardo CC, Samuelson DJ, Gould MN: **Congenic Rats Reveal Three Independent Copenhagen Alleles within the Mcs1 Quantitative Trait Locus That Confer Resistance to Mammary Cancer**. *Cancer Research* 2003, **63**(18):5808-5812.
124. Zheng W, Zhang B, Cai Q, Sung H, Michailidou K, Shi J, Choi J-Y, Long J, Dennis J, Humphreys MK *et al*: **Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast**

cancer cases and 25 579 c ontrols. *Human Molecular Genetics* 2013, **22**(12):2539-2550.
125. Nigam R, Laulederkind SJF, Hayman GT, Smith JR, Wang S-J, Lowry TF, Petri V, Pons JD, Tutaj M, Liu W *et al*: **Rat Genome Database: a unique resource for rat, human, and mouse quantitative trait locus data**. *Physiological Genomics* 2013, **45**(18):809-816.
126. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics* 2005, **21**(2):263-265.
127. Webb JD, Murányi A, Pugh CW, Ratcliffe PJ, Coleman ML: **MYPT1, the targeting subunit of smooth-muscle myosin phosphatase, is a s ubstrate for the asparaginyl hydroxylase factor inhibiting hypoxia-inducible factor (FIH)**. *Biochemical Journal* 2009, **420**(2):327-333.
128. Cho H-S, Suzuki T, Dohmae N, Hayami S, Unoki M, Yoshimatsu M, Toyokawa G, Takawa M, Chen T, Kurash JK *et al*: **Demethylation of RB Regulator MYPT1 by Histone Demethylase LSD1 Promotes Cell Cycle Progression in Cancer Cells**. *Cancer Research* 2011, **71**(3):655-660.
129. Singaraja RR, Hadano S, Metzler M, Givan S, Wellington CL, Warby S, Yanai A, Gutekunst C-A, Leavitt BR, Yi H *et al*: **HIP14, a n ovel ankyrin domain-containing protein, links huntingtin to intracellular trafficking and endocytosis**. *Human Molecular Genetics* 2002, **11**(23):2815-2828.
130. Carlsson E, Ranki A, Sipila L, Karenko L, Abdel-Rahman WM, Ovaska K, Siggberg L, Aapola U, Assamaki R, Hayry V *et al*: **Potential role of a navigator gene NAV3 in colorectal cancer**. *Br J Cancer* 2012, **106**(3):517-524.
131. Carlsson E, Krohn K, Ovaska K, Lindberg P, Häyry V, Maliniemi P, Lintulahti A, Korja M, Kivisaari R, Hussein S *et al*: **Neuron navigator 3 alterations in nervous system tumors associate with tumor malignancy grade and prognosis**. *Genes, Chromosomes and Cancer* 2013, **52**(2):191-201.
132. Karenko L, Hahtola S, Päivinen S, Karhu R, Syrjä S, Kähkönen M, Nedoszytko B, Kytölä S, Zhou Y, Blazevic V *et al*: **Primary Cutaneous T-Cell Lymphomas Show a Deletion or Translocation Affecting NAV3, the Human UNC-53 Homologue**. *Cancer Research* 2005, **65**(18):8101-8110.
133. Maliniemi P, Carlsson E, Kaukola A, Ovaska K, Niiranen K, Saksela O, Jeskanen L, Hautaniemi S, Ranki A: **NAV3 copy number changes and target genes in basal and squamous cell cancers**. *Experimental Dermatology* 2011, **20**(11):926-931.
134. Scholz C-J, Kurzeder C, Koretz K, Windisch J, Kreienberg R, Sauer G, Deissler H: **Tspan-1 is a tetraspanin preferentially expressed by mucinous and endometrioid subtypes of human ovarian carcinomas**. *Cancer Letters* 2009, **275**(2):198-203.
135. Epstein DJ: **Cis-regulatory mutations in human disease**. *Briefings in Functional Genomics & Proteomics* 2009, **8**(4):310-316.
136. Bulger M, Groudine M: **Functional and Mechanistic Diversity of Distal Transcription Enhancers**. *Cell* 2011, **144**(3):327-339.
137. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A: **Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq**. *Science* 2013, **339**(6123):1074-1077.

138. Holliday D, Speirs V: **Choosing the right cell line for breast cancer research**. *Breast Cancer Research* 2011, **13**(4):215.
139. Riaz M, Elstrodt F, Hollestelle A, Dehghan A, Klijn J, Schutte M: **Low-risk susceptibility alleles in 40 human breast cancer cell lines**. *BMC Cancer* 2009, **9**(1):236.
140. Chen Y, Olopade OI: **MYC in breast tumor progression**. *Expert Review of Anticancer Therapy* 2008, **8**(10):1689-1698.
141. Shiu RP, Watson PH, Dubik D: **c-myc oncogene expression in estrogen-dependent and -independent breast cancer**. *Clinical Chemistry* 1993, **39**(2):353-355.
142. Kyo S, Takakura M, Taira T, Kanaya T, Itoh H, Yutsudo M, Ariga H, Inoue M: **Sp1 cooperates with c-Myc to activate transcription of the human telomerase reverse transcriptase gene (hTERT)**. *Nucleic Acids Research* 2000, **28**(3):669-677.
143. Loignon M, Miao W, Hu L, Bier A, Bismar TA, Scrivens PJ, Mann K, Basik M, Bouchard A, Fiset PO *et al*: **Cul3 overexpression depletes Nrf2 in breast cancer and is associated with sensitivity to carcinogens, to oxidative stress, and to chemotherapy**. *Molecular Cancer Therapeutics* 2009, **8**(8):2432-2440.
144. Sporn MB, Liby KT: **NRF2 and cancer: the good, the bad and the importance of context**. *Nat Rev Cancer* 2012, **12**(8):564-571.
145. Wang W, Kwok AM, Chan JY: **The p65 Isoform of Nrf1 Is a D ominant Negative Inhibitor of ARE-mediated Transcription**. *Journal of Biological Chemistry* 2007, **282**(34):24670-24678.
146. Kim JJ, Kurita T, Bulun SE: **Progesterone Action in Endometrial Cancer, Endometriosis, Uterine Fibroids, and Breast Cancer**. *Endocrine Reviews* 2013, **34**(1):130-162.
147. Graham JD, Clarke C: **Progesterone receptors - animal models and cell signaling in breast cancer: Expression and transcriptional activity of progesterone receptor A and progesterone receptor B in mammalian cells**. *Breast Cancer Res* 2002, **4**(5):187 - 190.
148. Hisatomi H, Kohno N, Wakita K, Nagao K, Hirata H, Hikiji K, Harada S: **Novel alternatively spliced variant with a d eletion of 52 B P in exon 6 of the progesterone receptor gene is observed frequently in breast cancer tissues**. *International Journal of Cancer* 2003, **105**(2):182-185.
149. Brayman MJ, Julian J, Mulac-Jericevic B, Conneely OM, Edwards DP, Carson DD: **Progesterone Receptor Isoforms A and B Differentially Regulate MUC1 Expression in Uterine Epithelial Cells**. *Molecular Endocrinology* 2006, **20**(10):2278-2291.
150. Yin P, Roqueiro D, Huang L, Owen JK, Xie A, Navarro A, Monsivais D, Coon V JS, Kim JJ, Dai Y *et al*: **Genome-Wide Progesterone Receptor Binding: Cell Type-Specific and Shared Mechanisms in T47D Breast Cancer Cells and Primary Leiomyoma Cells**. *PloS One* 2012, **7**(1):e29021.
151. Xu H, Uno JK, Inouye M, Collins JF, Ghishan FK: **NF1 transcriptional factor(s) is required for basal promoter activation of the human intestinal NaPi-IIb cotransporter gene**. *American Journal of Physiology - Gastrointestinal and Liver Physiology* 2005, **288**(2):G175-G181.

152. Shamanna RA, Hoque M, Pe'ery T, Mathews MB: **Induction of p53, p21 and apoptosis by silencing the NF90/NF45 complex in human papilloma virus-transformed cervical carcinoma cells**. *Oncogene* 2013, **32**(43):5176-5185.
153. Sabapathy K, Kallunki T, David J-P, Graef I, Karin M, Wagner EF: **C-Jun Nh2-Terminal Kinase (Jnk)1 and Jnk2 Have Similar and Stage-Dependent Roles in Regulating T Cell Apoptosis and Proliferation**. *The Journal of Experimental Medicine* 2001, **193**(3):317-328.
154. Folkersen L, Hooft Fvt, Chernogubova E, Agardh HE, Hansson GK, Hedin U, Liska J, Syvänen A-C, Paulsson-Berne G, Franco-Cereceda A *et al*: **Association of Genetic Risk Variants With Expression of Proximal Genes Identifies Novel Susceptibility Genes for Cardiovascular Disease**. *Circulation: Cardiovascular Genetics* 2010, **3**(4):365-373.
155. Bell AC, West AG, Felsenfeld G: **The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators**. *Cell* 1999, **98**(3):387-396.
156. Rose NR, Klose RJ: **Understanding the relationship between DNA methylation and histone lysine methylation**. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 2014.
157. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O *et al*: **DNA targeting specificity of RNA-guided Cas9 nucleases**. *Nat Biotech* 2013, **31**(9):827-832.
158. Sanders J, Samuelson D: **Significant overlap between human genome-wide association-study nominated breast cancer risk alleles and rat mammary cancer susceptibility loci**. *Breast Cancer Research* 2014, **16**(1):R14.
159. Hemminki K, Vaittinen P, Kyyrönen P: **Age-specific familial risks in common cancers of the offspring**. *International Journal of Cancer* 1998, **78**(2):172-175.
160. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G *et al*: **Identification of the breast cancer susceptibility gene BRCA2**. *Nature* 1995, **378**(6559):789-792.
161. Thomas DC, Haile RW, Duggan D: **Recent Developments in Genomewide Association Scans: A Workshop Summary and Review**. *The American Journal of Human Genetics* 2005, **77**(3):337-345.
162. Ioannidis JPA, Castaldi P, Evangelou E: **A Compendium of Genome-Wide Associations for Cancer: Critical Synopsis and Reappraisal**. *Journal of the National Cancer Institute* 2010, **102**(12):846-858.
163. Russo J, Tait L, Russo IH: **Susceptibility of the mammary gland to carcinogenesis. III. The cell of origin of rat mammary carcinoma**. *Am J Pathol* 1983, **113**(1):50-66.
164. Veillet AL, Haag JD, Remfert JL, Meilahn AL, Samuelson DJ, Gould MN: **Mcs5c: A Mammary Carcinoma Susceptibility Locus Located in a Gene Desert that Associates with Tenascin C Expression**. *Cancer Prevention Research* 2011, **4**(1):97-106.
165. Cotroneo MS, Merry GM, Haag JD, Lan H, Shepel LA, Gould MN: **The Mcs7 quantitative trait locus is associated with an increased susceptibility to mammary cancer in congenic rats and an allele-specific imbalance**. *Oncogene* 2006, **25**(36):5011-5017.

166. Gibbs R, Weinstock G, Metzker M, Muzny D, Sodergrin E, Scherer S, Graham S, Scheffen D, Worley K, Burch P *et al*: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution**. *Nature* 2004, **428**(6982):493-521.
167. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The Sequence of the Human Genome**. *Science* 2001, **291**(5507):1304-1351.
168. Lander E, Linton L, Birren B, Nusbaum C, Zodi M, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
169. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Research* 2001, **29**(1):308-311.
170. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, Morrison J, Maranian M, Pooley KA, Luben R *et al*: **Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2**. *Nat Genet* 2009, **41**(5):585-590.
171. Fletcher O, Johnson N, Orr N, Hosking FJ, Gibson LJ, Walker K, Zelenika D, Gut I, Heath S, Palles C *et al*: **Novel Breast Cancer Susceptibility Locus at 9q31.2: Results of a Genome-Wide Association Study**. *Journal of the National Cancer Institute* 2011, **103**(5):425-435.
172. Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, Brook MN, Orr N, Rhie SK, Riboli E, Feigelson HS *et al*: **Genome-wide association studies identify four ER negative-specific breast cancer risk loci**. *Nat Genet* 2013, **45**(4):392-398.
173. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, Dicks E, Dennis J, Wang Q, Humphreys MK, Luccarini C *et al*: **Genome-wide association analysis identifies three new breast cancer susceptibility loci**. *Nat Genet* 2012, **44**(3):312-318.
174. Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, Millikan RC, Wang X, Ademuyiwa F, Ahmed S, Ambrosone CB *et al*: **A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer**. *Nat Genet* 2011, **43**(12):1210-1214.
175. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A *et al*: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer**. *Nat Genet* 2007, **39**(7):870-874.
176. Li J, Humphreys K, Darabi H, Rosin G, Hannelius U, Heikkinen T, Aittomaki K, Blomqvist C, Pharoah P, Dunning A *et al*: **A genome-wide association scan on estrogen receptor-negative breast cancer**. *Breast Cancer Research* 2010, **12**(6):R93.
177. Li J, Humphreys K, Heikkinen T, Aittomäki K, Blomqvist C, Pharoah P, Dunning A, Ahmed S, Hooning M, Martens J *et al*: **A combined analysis of genome-wide association studies in breast cancer**. *Breast Cancer Res Treat* 2011, **126**(3):717-727.
178. Mavaddat N, Dunning AM, Ponder BAJ, Easton DF, Pharoah PD: **Common Genetic Variation in Candidate Genes and Susceptibility to Subtypes of**

Breast Cancer. *Cancer Epidemiology Biomarkers & Prevention* 2009, **18**(1):255-259.

179. Sehrawat B, Sridharan M, Ghosh S, Robson P, Cass C, Mackey J, Greiner R, Damaraju S: **Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility**. *Human Genetics* 2011, **130**(4):529-537.
180. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A *et al*: **Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer**. *Nat Genet* 2007, **39**(7):865-869.
181. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, Jakobsdottir M, Bergthorsson JT, Gudmundsson J, Aben KK *et al*: **Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer**. *Nat Genet* 2008, **40**(6):703-706.
182. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS *et al*: **Genome-wide association study identifies five new breast cancer susceptibility loci**. *Nat Genet* 2010, **42**(6):504-507.
183. Cai Q, Long J, Lu W, Qu S, Wen W, Kang D, Lee J-Y, Chen K, Shen H, Shen C-Y *et al*: **Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium**. *Human Molecular Genetics* 2011, **20**(24):4991-4999.
184. Chen F, Chen G, Stram D, Millikan R, Ambrosone C, John E, Bernstein L, Zheng W, Palmer J, Hu J *et al*: **A genome-wide association study of breast cancer in women of African ancestry**. *Human Genetics* 2013, **132**(1):39-48.
185. Long J, Cai Q, Shu X-O, Qu S, Li C, Zheng Y, Gu K, Wang W, Xiang Y-B, Cheng J *et al*: **Identification of a Functional Genetic Variant at 16q12.1 for Breast Cancer Risk: Results from the Asia Breast Cancer Consortium**. *PLoS Genet* 2010, **6**(6):e1001002.
186. Long J, Cai Q, Sung H, Shi J, Zhang B, Choi J-Y, Wen W, Delahanty RJ, Lu W, Gao Y-T *et al*: **Genome-Wide Association Study in East Asians Identifies Novel Susceptibility Loci for Breast Cancer**. *PLoS Genet* 2012, **8**(2):e1002532.
187. Zheng W, Long J, Gao Y-T, Li C, Zheng Y, Xiang Y-B, Wen W, Levy S, Deming SL, Haines JL *et al*: **Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1**. *Nat Genet* 2009, **41**(3):324-328.

LIST OF ABBRVIATIONS

3C- chromosome conformation capture

95% CI- 95% confidence interval

ACI- August Copenhagen Irish

ARE- antioxidant response element

ATM- ataxia telangiectasia mutated

ATR- ataxia telangiectasia mutated rad3-related

BAC- bacterial artificial chromosome

Bcl-2- B-cell lymphoma 2

Bcl-X1- Bcl-2 associated X protein

BN- Brown Norway

BRCA1/2- breast cancer 1/2, early onset

BRIP1-BRCA1 interacting protein

BSA- bovine serum albumin

cAMP- cyclic adenosine monophosphate

CDC2- cell division cycle protein 2

CDH1- cadherin 1

ChIP- chromatin immunoprecipitation

CNV- copy number variation

COP- Copenhagen

CRISPR- clustered regularly interspaced short palindromic repeats

CTF/NF-1 domain- CCAAT box-binding transcription factor/ Nuclear factor -1 DNA binding domain

DES- diethylstilbestrol

DMBA- 7,12-dimethylbenz[a]anthracene

DMEM/ F12- Dulbecco's Modified Eagle Medium

DNA- deoxyribonucleic acid

DTT- dithiothreitol

DZF- domain in DSRM or zinc finger C2H2 domain containing proteins

E2- 17β- estradiol

EMCA-estrogen induced mammary cancer loci

EMSA- electrophoretic mobility shift assay

ER- estrogen receptor

FBS- fetal bovine serum

FGFR2- fibroblast growth factor receptor 2

Frmpd1- FERM And PDZ Domain Containing 1

GFP- green fluorescent protein

GPBP1- GC- rich promoter binding protein 1

GWAS- genome- wide association study

HBSS- Hanks' balanced salt solution

IL6ST-interleukin 6 signal transducer

Il8- interleukin 8

ILF2- interleukin enhancer binding factor 2

ILF3- interleukin enhancer binding factor 3

INDEL- insertion or deletion

LD block- linkage disequilibrium block

LOD score- logarithm (base10) of odds score

MAP3K1- mitogen activated protein kinase kinase kinase 1

Mcs- mammary cancer susceptibility locus

Mcsm: mammary carcinoma susceptibility modifier

Mcsta: mammary cancer susceptibility tumor aggressiveness

Mcstm- mammaryv cancer susceptibility tumor multiplicity

MEC- mammary epithelial cell

MIER3- mesoderm induction early response protein 3

MMTV- mouse mammary tumor virus long terminal repeat promoter

MRI- magnetic resonance imaging

NAT2- N-acetyltransferase 2

NAV3- Neuron navigator 3

NFAT- nuclear factor of activated T-cells

NFIC- nuclear factor (eythroid derived)- like 2

NTN4- netrin4

PAH- polycyclic aromatic hydrocarbons

PALB2- partner and localizer of BRCA2

PBS- phosphate buffered saline

PCB- polychlorinated biphenyls

P:C:I- phenol: choloroform: isoamylalcohol (25:24:1)

PCR- polymerase chain reaction

PPP1R12A or MYPT1- Myosin Phosphatase- Targeting Subunit 1

PR- progesterone receptor

PRE- progesterone response element

PTEN- Phosphatase and tensin homolog

QTL- quantitative trait locus

RACE- rapid amplification of cDNA ends

RNA- ribonucleic acid

RPMI- Roswell Park Memorial Institute medium

RR- relative risk

SAP domain- SAF-A/B, Acinus and PIAS domain

SERM- selective estrogen receptor modulators

SETD9- SET domain containing 9

sgRNA- single guide RNA

SNP- single nucleotide polymorphism

STK11-serine/ threonine kinase 11

TAP- tobacco acid phosphatase

TP53- tumor protein 53

TSPAN- tetraspanin

TSS- transcription start site

UTR- untranslated region

VEGF- vascular endothelial growth factor

WF- Wistar Furth

WKy- Wistar- Kyoto

ZDHHC17 or HIP-14-Huntingtin interacting protein

APPENDIX

PERMISSION TO USE PUBLISHED WORK

For research article: Sanders J, Samuelson D: **Significant overlap between human genome-wide association-study nominated breast cancer risk alleles and rat mammary cancer susceptibility loci**. *Breast Cancer Research* 2014, **16**(1):R14.

Permission from Breast Cancer Research:

"Dear Dr Sanders,
Thank you for your email. As your publication was a research article and open access this is absolutely fine. As long as the article is correctly cited in your dissertation then no further permission is required.
If you have any further questions regarding this please do not hesitate to contact me.
Kind regards,
Laura Anstee"

CURRICULUM VITAE

Jennifer Sanders

E-mail: jennifersanders001@gmail.com
Address: 703 Fountain Avenue
Louisville, KY 40222
Phone: 502-777-3019

EDUCATION

University of Louisville, Louisville, KY **August 2009-June 2014**
PhD, Biochemistry and Molecular Biology, GPA 3.9
Mentor: David Samuelson, PhD
Dissertation title: Genetic and Mechanistic Analysis of Rat Mammary Cancer Susceptibility

University of Louisville, Louisville, KY **August 2005- May 2009**
Bachelor of Science in Chemistry with concentration in Biochemistry
Minor in Mathematics, GPA 3.9

POSITIONS

- 08/2010-12/2010 Teaching Assistant, University of Louisville
- 08/2009- Graduate Research Assistant, University of Louisville

HONORS AND AWARDS

Honors

- Won first place for poster presentation at University of Louisville Department of Biochemistry and Molecular Biology Retreat 2013
- Selected to give oral presentation at Cincinnati Children's Hospital Postdoctoral Symposium 2013
- Selected to give oral presentation at Cold Spring Harbor Meeting on Rat Genomics and Models 2011
- Selected to give oral presentation at University of Louisville Department of Biochemistry and Molecular Biology Retreat 2011
- CGeMM Travel Award Recipient 2011
- Research Committee Travel Award Recipient 2011
- Recipient of the National Scholars Award 2005-2009

- 3 time recipient of the Etscorn Scholar Award 2006-2009
- B.S. awarded with magna cum laude
- Graduated as an Honors Scholar
- Member of Woodcock Honors Society
- Member of Phi Kappa Phi Honors Society
- Member of Alpha Epsilon Delta National Honorary Pre-Health Fraternity

Awards

- **Integrated Programs in Biological Sciences Fellowship** from 8/2009 to 7/2011
- **T32-ES011564**
  Dr. Hein (PI) 7/2012- 6/2013 NIEHS
  National Institute of Health Sciences-funded Training Program in Environmental Health Sciences
  Role: Pre-doctoral trainee
- 1st place poster presentation at Biochemistry and Molecular Biology Retreat 2013
- CGeMM Travel Award Recipient 2011
- Research Committee Travel Award Recipient 2011

ABSTRACTS AND PRESENTATIONS

Oral Presentations

- Samuelson DJ, Sanders J, Xu X, Vaughn D. Identification of rat *mammary carcinoma susceptibility-1b* (*Mcs1b*) QTL candidate genes and elements. Abstract for presentation. Rat Genomics and Models Cold Spring Harbor Laboratory Conference, December 7-10, 2011, Cold Spring Harbor, New York.
- Sanders J, Xu X, Kemper AF, Kalbfleisch T, Samuelson DJ. *A074-SNP-17* is a Candidate Rat Mammary Cancer Susceptibility SNP: Analysis of a M ammary Cancer QTL. Abstract for presentation. Cincinnati Children's Hospital Medical Center Postdoctoral Research Symposium 2013, September 6, 2013, Cincinnati, Ohio

Posters

- Sanders J, Xu X, Kemper AF, Kalbfleisch T, Samuelson DJ. *A074-SNP-17* is a Candidate Rat Mammary Cancer Susceptibility SNP: Analysis of a Mammary Cancer QTL. Abstract for poster presentation. James Brown Cancer Center Retreat 2013, October 25, 2013, Louisville, Kentucky
- Sanders J, Xu X, Kemper AF, Kalbfleisch T, Samuelson DJ. *A074-SNP-17* is a Candidate Rat Mammary Cancer Susceptibility SNP: Analysis of a Mammary Cancer QTL. Abstract for poster presentation. Research! Louisville 2013, September 24, 2013, Louisville, Kentucky
- Sanders J, Xu X, Kemper AF, Kalbfleisch T, Samuelson DJ. *A074-SNP-17* is a Candidate Rat Mammary Cancer Susceptibility SNP: Analysis of a Mammary Cancer QTL. Abstract for poster presentation. Biochemistry and Molecular Biology Retreat 2013, August 23, 2013, Louisville, Kentucky
- Samuelson DJ, Sanders J, Xu X. Rat Mammary carcinoma susceptibility-1b single-nucleotide-variant A074-SNV-17 is a candidate Mcs1b quantitative trait

nucleotide. Poster Presentation, Complex Trait Community 12th Annual Meeting, Madison, Wisconsin, May 28-31, 2013

- Sanders J, Xu X, Kemper AF, Kalbfleisch T, Samuelson DJ. Identification and Mechanistic analysis of rat *mammary carcinoma susceptibility-1b* genetic elements. Abstract for presentation. Biochemistry and Molecular Biology Department Student Recruitment, February 15 2013, Louisville, Kentucky
- Sanders J, Xu X, Kemper AF, Kalbfleisch T, Samuelson DJ. Identification and Mechanistic analysis of rat *mammary carcinoma susceptibility-1b* genetic elements. Abstract for presentation. Rat Genomics and Models Cold Spring Harbor Laboratory Conference, December 7-10, 2011, Cold Spring Harbor, New York
- Sanders J, Xu X, Kemper AF, Kalbfleisch T, Samuelson DJ. Identification and Mechanistic analysis of rat *mammary carcinoma susceptibility-1b* genetic elements. Abstract for poster presentation. James Graham Brown Cancer Retreat, Oct 28, 2011, Louisville, Kentucky.
- Sanders J, Xu X, Kemper AF, Kalbfleisch T, Samuelson DJ. Identification and Mechanistic analysis of rat *mammary carcinoma susceptibility-1b* genetic elements. Abstract for poster presentation. Research! Louisville, October 10-15, 2011, Louisville, Kentucky.
- Sanders J, Xu X, Kemper AF, Kalbfleisch T, Samuelson DJ. Mechanistic analysis of rat *mammary carcinoma susceptibility-1b* genetic elements. Abstract for talk. Biochemistry and Molecular Biology Retreat 2011, August 26, 2011, Louisville, Kentucky.
- Samuelson DJ, Xu X, denDekker AD, Sanders J, Kemper AF, Kalbfleisch T. Congenic Mapping and Functional Characterization of a Rat Mammary Carcinoma Susceptibility QTL*, Mcs1b*, Identifies *MAP3K1* and *MIER3* as Candidate Breast Cancer Susceptibility Genes. Poster Presentation, Quantitative Genetics & Genomics Gordon Research Conference, February 20-25, 2011, Galveston, Texas
- Sanders J. 2009. Analyzing Inpatient Data for Hepatitis Patients. Abstract for poster presentation, International Society for Pharmacoeconomics and Outcomes Research Conference 2009.

PUBLICATIONS

Research articles

- Sanders J, Samuelson D: Significant overlap between human genome-wide association-study nominated breast cancer risk alleles and rat mammary cancer susceptibility loci. *Breast Cancer Research* 2014, 16(1):R14

- Sanders J, Haag JD, Samuelson DJ (2011) Physical Confirmation and Mapping of Overlapping Rat Mammary Carcinoma Susceptibility QTLs, *Mcs2* and *Mcs6*. PLoS ONE 6(5): e19891. doi:10.1371/journal.pone.0019891

- Devapatla B, Sanders J, Samuelson DJ (2012) Genetically determined inflammatory-response related cytokine and chemokine transcript profiles between mammary carcinoma resistant and susceptible rat strains. Cytokine 59: 223-227

Books

- Co-author of: Cerrito, P.B. (2009). A Casebook on Pediatric Disease (Bentham Science Publishers).