Identification of Germline Alterations in the Mad-Homology 2 (MH2) Domain of SMAD3 and SMAD4 in Breast Cancer Susceptibility

By

Eric Tram

A thesis submitted in conformity with the requirements for the degree of Master of Science
Department of Laboratory Medicine & Pathobiology
University of Toronto

© Copyright by Eric Tram 2010
Identification of Germline Alterations in the Mad-Homology 2 (MH2) Domain of SMAD3 and SMAD4 in Breast Cancer Susceptibility

Eric Tram
Master of Science
Department of Laboratory Medicine And Pathobiology
University of Toronto
2010

Abstract
A common feature in many neoplastic cells is that mutations in SMADs, intermediates of TGF-β signaling, are known to contribute to the loss of sensitivity to the anti-tumour effects of TGF-β. At present it is not known whether germline alterations in the two key signaling intermediates SMAD3 and SMAD4 are involved in breast cancer predisposition. Here, we undertook mutation analysis of the highly conserved Mad-Homology 2 (MH2) domains of both genes in 408 non-BRCA1/BRCA2 breast cancer cases and 710 controls recruited by the Ontario Familial Breast Cancer Registry (OFBCR). We found no SMAD3 coding variants but identified two novel SMAD4 coding alterations c.1478G>A and c.1701A>G. Nucleotide diversity estimation show alterations in the SMAD3 coding region but not SMAD4 to be less than expected in both cases and controls suggesting SMAD3 function could be important for tumourigenesis. While the variants did not induce aberrant splicing, real-time PCR analysis showed that breast cancer cases were associated with significantly elevated SMAD3 and SMAD4 germline expression relative to population controls. Among these was SMAD4 c.1478G>A from a familial breast cancer case had 5-fold increase and represents a potential breast cancer susceptibility alteration. Microarray gene expression data also support the tumour over-expression of both genes relative to normal tissues. Thus the aberrant germline expression observed may provide novel insight into the roles of these genes in the predisposition of disease.
Acknowledgements

I would like to express a most heartfelt thanks to my supervisor, Dr. Hilmi Ozcelik, for his guidance and support. My graduate training has been a journey to be remembered and at the end of it I feel very privileged to have had such an enriching learning experience. Thank you for being here for me at every step of the way.

I would also like to thank my committee members Dr. L Briollais, Dr. P Hamel, and Dr. J Wrana for their patience and understanding as well as their expert knowledge.

To the past and present members of the Ozcelik lab thank you for your friendships and, in more times than I can remember, for sharing your expertise. I would particularly like to thank Lynda Doughty for her technical help with the dHPLC and Hong Li, Susan Lau, Priscilla Chan, and Dr. Irada for helping me with the myriad of issues I had during my time here.

Finally I would like to acknowledge funding support during my Master’s program through OSOTF awarded by the SLRI and the tremendous support from Dr. Elsholtz and Ferzeen of the LMP department.
# Table of Contents

Abstract

Acknowledgement

Table of Contents

List of Tables

List of Figures

List of Abbreviations

## 1.0 INTRODUCTION

1.1 Prevalence and risk of breast cancer

1.2 Genetic predisposition to breast cancer

1.3 Mechanisms of sporadic breast cancer

1.4 Current clinical diagnosis and management of breast cancer

1.5 The TGF-β signalling pathway in mammary development

1.6 SMAD3 and SMAD4 as candidate breast tumour suppressors

1.7 The SMAD3 and SMAD4 genes

1.8 Structure of the SMAD3 and SMAD4 transcription factor

1.9 Function of SMAD3 and SMAD4

1.10 Inactivating mutations in the TGF-β signalling pathway in breast cancer

1.11 Mutation analyses of SMAD3

1.12 Mutation analyses of SMAD4

1.13 The Mad-Homology 2 (MH2) domain as a somatic mutation hot spot in cancer
2.0 HYPOTHESIS AND OBJECTIVES

2.1 Rationale
2.2 Hypothesis
2.3 Objectives
2.4 Significance

3.0 Materials and Methods

3.1 Discovery of novel polymorphisms by DHPLC Analysis
   3.1.1 Selection and clinical characteristics of the study population
   3.1.2 PCR amplification of exon/intron-exon junctions of the MH2 domain
   3.1.3 Determination of optimal amplicon melting temperature for DHPLC

3.2 Nucleotide diversity estimation

3.3 In silico characterization of identified SNPs
   3.3.1 Evolutionary conservation analysis
   3.3.2 Alternative Splice Analyzer for variants promoting splice defects
   3.3.3 FASTsnp analysis for variants affecting Exonic Splicing Enhancers

3.4 In vitro characterization of the effect of genetic polymorphisms on mRNA splicing
   3.4.1 Preparation of cDNA
   3.4.2 PCR (RT-PCR) amplification of the coding exons of the MH2 domain

3.5 Quantitative Real Time PCR for SNP effect on germline expression
   3.5.1 Generation of standard curves for real-time quantitative PCR
   3.5.2 Statistical Analysis

3.6 Catalogue of somatic breast cancer mutations in SMAD3 and SMAD4

3.7. SMAD3 and SMAD4 expression in primary tumour
4.0 Results

4.1 DHPLC optimization and screening elution profiles

4.2 Mutation analysis of the Mad-Homology 2 domain of SMAD3 and SMAD4

4.3 Nucleotide diversity estimations

4.4 In silico characterization of identified SNPs using bioinformatic tools

4.5 In vitro analysis of SNP impact on mRNA splicing

4.6 Standard curve generations for SMAD3 and SMAD4 qPCR primers

4.7 Effect of SNP on germline expression

4.8 Breast cancer mutations in SMAD3 and SMAD4 retrieved from COSMIC

4.9 SMAD3 and SMAD4 expression in primary tumours relative to normal tissues

5.0 Discussion

5.1 Novel germline variants detected in SMAD3 and SMAD4 in breast cancer

5.1.1 Distribution of variants in cases and controls

5.1.2 Nucleotide diversity relative to expected values

5.2 In silico and in vitro characterization of variant’s impact on mRNA splicing

5.2.1 Creation of cryptic splice sites

5.2.2 Abolition of Exonic Enhancer Elements (ESE)

5.2.3 Effect of variants on proper transcript splicing

5.3 Impact of variants on germline gene expressions

5.3.1 SMAD3 germline expression levels

5.3.2 c.1478G>A influence on SMAD4 germline expression

5.4 Characterization of expression level differences between tumour and normal tissue

5.5 Somatic mutation of SMAD3 and SMAD4 in primary breast carcinoma
5.6 Summary 65
5.7 Future direction 67
6.0 References 70
List of Tables

Table 1 Principle breast cancer risk factor and effects 3
Table 2 Summary of mutation analyses of SMAD3 and SMAD4 in breast cancer 16
Table 3 Criteria defining probands (all invasive breast cancer) at increased genetic risk 24
Table 4 Primer sequences and optimized PCR and DHPLC running conditions 26
Table 5 Clinical characteristics of 37 mRNA from cases or controls 32
Table 6 Primer sequences and optimized RT-PCR conditions 33
Table 7 Novel breast cancer germline variants detected in SMAD3 and SMAD4 40
Table 8 Features of the nucleotide diversity comparison studies 41
Table 9 Bioinformatic predictions of effect on mRNA splicing 45
Table 10 Quantitative RT-PCR analysis for SMAD3 germline variants 49
Table 11 Quantitative RT-PCR analysis for SMAD4 germline variants 51
List of Figures

Figure 1 Diagram of the TGF-β signalling pathway 9
Figure 2 Structure of SMAD3 13
Figure 3 Structure of SMAD4 13
Figure 4 Overview of experimental approach 23
Figure 5 Age-Frequency matched case-control population from OFBCR 25
Figure 6 Denaturing High Pressure Liquid Chromatography (DHPLC) SNP detection 28
Figure 7 Illustration of the key motifs involved in splicing 29
Figure 8 Representative experimental DHPLC elution outputs 38
Figure 9 Nucleotide diversity comparison between this study and 3 major studies 43
Figure 10 Reverse transcription PCR analysis for rare variants of SMAD3 & SMAD4 46
Figure 11 Real-time quantitative PCR standard curve and amplification plots 47
Figure 12 Statistical analyses of SMAD3 germline expressions by qRT-PCR 48
Figure 13 Statistical analyses of SMAD4 germline expressions by qRT-PCR 50
Figure 14 Somatic mutations from the COSMIC database 53
Figure 15 SMAD3 and SMAD4 expressions in tumour relative to normal tissues 54
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT1</td>
<td>RAC-alpha serine/threonine-protein kinase</td>
</tr>
<tr>
<td>ATM</td>
<td>Ataxia telangiectasia mutated</td>
</tr>
<tr>
<td>BMP</td>
<td>Bone morphogenic protein</td>
</tr>
<tr>
<td>BRCA1/2</td>
<td>Breast cancer type 1 / 2 susceptibility protein</td>
</tr>
<tr>
<td>BRIP1</td>
<td>BRCA1-interacting c-terminal helicase 1</td>
</tr>
<tr>
<td>CBP</td>
<td>CREB (cAMP response element-binding)-Binding Protein</td>
</tr>
<tr>
<td>CHEK2</td>
<td>CHK2 Checkpoint homologue</td>
</tr>
<tr>
<td>COSMIC</td>
<td>Catalogue of somatic mutation in cancer</td>
</tr>
<tr>
<td>DPHLC</td>
<td>Denaturing high pressure liquid chromatography</td>
</tr>
<tr>
<td>DTT</td>
<td>Dithiothreitol</td>
</tr>
<tr>
<td>EGFR</td>
<td>Epidermal growth factor receptor</td>
</tr>
<tr>
<td>ESE/ESS</td>
<td>Exonic splicing enhancer/silencer</td>
</tr>
<tr>
<td>FASTsnp</td>
<td>Functional analysis and selection tool for SNP</td>
</tr>
<tr>
<td>FDR</td>
<td>First degree relative</td>
</tr>
<tr>
<td>HER2/Neu</td>
<td>Human epidermal growth factor receptor 2</td>
</tr>
<tr>
<td>ISE/ISS</td>
<td>Intronic splicing enhancer/silencer</td>
</tr>
<tr>
<td>LOH</td>
<td>Loss of heterozygosity</td>
</tr>
<tr>
<td>MCR</td>
<td>Mutation cluster region</td>
</tr>
<tr>
<td>MH1/2</td>
<td>Mad homology 1 / 2</td>
</tr>
<tr>
<td>MLPA</td>
<td>Multiplex ligation-dependent probe amplification</td>
</tr>
<tr>
<td>MMP</td>
<td>Matrix metalloproteinase</td>
</tr>
<tr>
<td>NBS1</td>
<td>Nijmegen breakage syndrome 1</td>
</tr>
<tr>
<td>OFBCR</td>
<td>Ontario familial breast cancer registry</td>
</tr>
<tr>
<td>PALB2</td>
<td>Partner and localizer of BRCA2</td>
</tr>
<tr>
<td>PolyPhen</td>
<td>Prediction of functional effect of human non-synonymous SNP</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>PTEN</td>
<td>Phosphatase and tensin homologue</td>
</tr>
<tr>
<td>QPCR</td>
<td>Quantitative real time polymerase chain reaction</td>
</tr>
<tr>
<td>RAD50</td>
<td>DNA repair protein RAD50</td>
</tr>
<tr>
<td>RB1</td>
<td>Retinoblastoma protein 1</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>Reverse transcription PCR</td>
</tr>
<tr>
<td>SAD</td>
<td>SMAD activation domain</td>
</tr>
<tr>
<td>SARA</td>
<td>SMAD anchor for receptor activation</td>
</tr>
<tr>
<td>SC35</td>
<td>RRM-containing SR protein</td>
</tr>
<tr>
<td>SF2/ASF</td>
<td>Splicing factor 2 / alternative splicing factor</td>
</tr>
<tr>
<td>SMAD</td>
<td>Homologue of Sma- and mothers against decapentaplegic</td>
</tr>
<tr>
<td>SRp55</td>
<td>Regulator of calcitonin/CGRP alternative RNA splicing</td>
</tr>
<tr>
<td>SIFT</td>
<td>Sort Intolerant from tolerant</td>
</tr>
<tr>
<td>SBE</td>
<td>SMAD binding element</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>RAS</td>
<td>Harvey rat sarcoma viral oncogene homolog</td>
</tr>
<tr>
<td>SMAD4//DPC4</td>
<td>SMAD4/Deleted in pancreatic cancer 4</td>
</tr>
<tr>
<td>TGF-β</td>
<td>Transforming growth factor beta</td>
</tr>
<tr>
<td>TGFR</td>
<td>Transforming growth factor receptor</td>
</tr>
<tr>
<td>TIE</td>
<td>TGF-beta 1 inhibitory element</td>
</tr>
<tr>
<td>TP53</td>
<td>Tumour protein 53</td>
</tr>
<tr>
<td>TUSC5</td>
<td>Tumour suppressor candidate 5</td>
</tr>
<tr>
<td>VEGF</td>
<td>Vascular endothelial growth factor</td>
</tr>
</tbody>
</table>
Chapter 1
Introduction
1.1 Prevalence and risk of breast cancer

Breast cancer is the most common female malignancy in the world as estimated by the International Agency for Research on cancer. Globally, approximately 22% of all new cancer diagnoses in women can be attributable to breast cancer. This amounts to almost 10% of all new cancer cases when men and women are combined which is quite significant as the occurrence of breast cancer in men is low. Breast cancer-related death is responsible for 7% of the more than 7.6 million cancer mortalities around the world.

In Canada, breast cancer is the most common and number one diagnosed cancer in women, both before and after menopause, with an estimated 22,000 new diagnoses every year. For 2007, more than 5,300 Canadian women died of the disease, second only to lung cancer (8,900). It is estimated that one in nine women will be diagnosed with, and 1 in 27 will die of breast cancer in their lifetime (Canadian Cancer Society/National Cancer Institute of Canada: Canadian Cancer Statistics 2004, Toronto, Canada, 2004; (special topic).

The exact cause of breast cancer is still unknown, but several key factors are able to increase the risk of developing disease such as age, country of residence, menarche and menopause, lifestyle, hereditary factors, and exposure to radiation. A list of main breast cancer risk factors are summarized in Table 1 [1]. Country of residence plays an important role in determining to what extent environmental factors such as diet, lifestyle and radiation exposure can influence overall cancer risk. Factors such as age of menarche and menopause are thought to modulate risk by altering the level of exposure to hormones such as estrogen. For example, an early age of menarche is thought to prolong exposure to estrogen thus increasing breast cancer risk whereas having multiple live births is thought to decrease risk as estrogen production is depressed during the course of pregnancy and breast feeding.

Although incidence remains high, due in part to limited opportunities for primary prevention, breast cancer mortality rates in Canada have been declining since 1993 due to advances in adjuvant therapy and greater participation in screening programs in Canada and other developed countries. However, there are still a significant number of women diagnosed with breast cancer that die from it (www.cancer.ca/ccs/internet/standard/0,2939,3172_14435__langId-en,00.html).
On average, 437 Canadian women are diagnosed with breast cancer weekly and about 104 Canadian women will die of breast cancer each week. Therefore, an important aim to ease the burden of breast cancer on the Canadian health care system is to increase research to identify novel genetic factors to facilitate early detection so women at high risk can make informed decisions on the treatment and management of disease.

Table 1: Principle breast cancer risk factor and effects. The figure was obtained from a review by Singeltary et al., 2003 [1] and references within.
1.2 Genetic predisposition to breast cancer

Breast cancer is a complex disease with genetic and environmental factors modulating risk. Familial breast cancer represents 10-20% of all breast cancers and a strong family history of breast cancer is a significant risk factor. In recent years, understanding of the genetic predisposition to breast cancer has greatly expanded.

Segregation analysis of breast cancer led to the postulation of the existence of an autosomal dominant breast cancer alleles [2]. Linkage of breast cancer to 17q21 was established in 1990 [3] and a few years later the breast cancer, early onset 1 (BRCA1) susceptibility allele was discovered by positional cloning [4]. Subsequent linkage analysis and positional cloning led to the mapping and discovery of a second breast cancer susceptibility allele, BRCA2 [5; 6]. The risks conferred by mutations in BRCA1 and BRCA2 varied depending on the selection of cases studied. Early studies of large cancer families suggested the risk of breast cancer by age 70 could be as high as 87% (95% Confidence Interval = 72%-95%) for BRCA1 and 84% (CI: 43%-95%) for mutation carriers of BRCA2 [7; 8]. For large families, mutations in BRCA1 and BRCA2 are responsible for up to two-thirds of site-specific breast cancers (≥ four cases) but much less so for smaller family clusters [8]. In a population based study on breast cancer cases unselected for family history the risks are lower, 65% (CI: 51%-75%) for BRCA1 and 45% (CI: 33%-54%) for BRCA2 [9]. In the United Kingdom, the estimated population frequency of mutations in both genes is approximately 1/1000 per gene and increase the risk up to 15% to 20% of the cases in familial breast cancer [10; 11; 12]. Therefore, BRCA1 and BRCA2 have been established as the most important high penetrance breast cancer susceptibility alleles [11] in familial cancers.

The functions of BRCA1 and BRCA2 have been well-studied as they regulate DNA repair, transcription, and chromatin remodeling through interactions with other proteins involved in DNA repair, cell cycle, transcriptional regulation and signal transduction [13], all pathways that are heavily disrupted in tumourigenesis. However, in recent years it has become increasing evident that BRCA1/BRCA2 mutations cannot explain all cases of familial breast cancer. About 70% of breast cancers occurring among families with fewer than four or five cases and do not have ovarian or male breast cancer are attributed to genes other than BRCA1 and BRCA2 [8]. This lead to mutation screening of genes functionally related to BRCA1 and BRCA2. These
included CHEK2, ATM, NBS1, RAD50, BRIP1, and PALB2 involved in the homologous recombination DNA repair pathway. Germline mutations in these genes are rare and confer an intermediate risk (~2 fold) of breast cancer [14]. Including PTEN and TP53 it is estimated that mutations in these ten genes altogether account for only 50% of familial breast cancers [8; 14] suggesting that there exists additional undiscovered breast cancer susceptibility alleles. These yet to be discovered susceptibility alleles are not likely to be high penetrant alleles like BRCA1/2 but of intermediate penetrance. It is also possible that they may be low penetrant commonly occurring single nucleotide polymorphisms (SNPs) that act together synergistically with other susceptibility alleles or environmental factors to increase breast cancer risk.

1.3 Mechanisms of sporadic breast cancer

Familial breast cancer has been studied extensively over the past two decades but the majority (30-70%) of breast cancers do not have an inherited component and occur as a woman acquires somatic mutations throughout life.

The genes which undergo sporadic mutation are mostly the tumour suppressor genes. Negative regulators of the cell cycle including TP53 and RB1 are often inactivated by mutations in 15-45% and 20% of breast cancer cases, respectively [15; 16; 17] and oncogenes such as ras, myc, and neu are often mutated in many cancers including breast cancers. While BRCA1 mutations are rare in sporadic cancers, reduced expression or improper localization of BRCA1 were hypothesized to be important in non-familial breast cancers. Hypermethylation of the BRCA1 promoter region has been shown to occur in approximately 10-15% of primary breast tumours [18]. Recently, it was shown that the membrane serine/threonine protein kinase Ba (AKT1) repressed BRCA1-mediated homologous recombination through cytoplasmic retention of BRCA1 and RAD51 proteins resulting in a BRCA1-deficient–like phenotype [19]. As upregulation of this gene is observed in 38-54% of sporadic breast cancer [20], this suggests that AKT1 could be responsible for up to 50% sporadic breast cancers and provide an impetus to study additional interacting partners of BRCA1.

Another major characteristic of breast cancer is the acquisition of the effect of extracellular mitogenic growth signals by the tumour. Growth receptors such as TGFR, HER2/neu, ErbB-3 and ErbB4 are often the target of overexpression by the tumour. For example, EGFR and
HER2/neu are often overexpressed in as high as 20-30% of breast cancers [21; 22] or in the case of HER2/neu mutated to become a constitutively active form of the receptor [23]. Additional targets that are often modified in sporadic breast cancers include overexpression of anti-apoptotic genes such as bcl-2, which is seen in approximately 30-45% of breast tumours [24], decreasing expression in cell-cell adhesion molecules such as E-cadherin in 60% of primary breast cancers [25], and the upregulation of pro-angiogenic factors such as VEGF and downregulation of pro-metastatic factors such as matrix metalloproteinases (MMPs) contribute to the survival and metastatic potential of the tumour.

1.4 Current clinical diagnosis and management of breast cancer

The factors underlying genetic susceptibility to breast cancer continue to remain largely unknown. To date, the prediction of breast cancer susceptibility in high-risk women has been dependent upon genetic testing for the presence of high-penetrant mutations, which deleteriously impact the function of BRCA1 and BRCA2 and recently other genes including TP53, CHEK2, ATM, RAD50, RAD51 and NBS1 [14; 26; 27]. A blood sample is drawn in a laboratory, doctor's office, hospital, or clinic and then analyzed by a laboratory that specializes in genetic tests. The test result can be obtained in a few weeks or longer.

A health care professional is recommended to discuss the results and often involves a risk assessment based on the individual’s family medical history and the implications of a positive, negative or ambiguous result. An important caveat, however, is that these mutations are rare and account for only for 10-20% of familial, 5-10% of all breast cancers, and 1-2% of all women in the population [28]. Furthermore as most genetic studies on the risk conferred by BRCA1 and BRCA2 mutations have been performed on large families with multiple members affected by breast cancer and may be influenced by genetic or environmental factors, an estimation of breast and/or ovarian cancer risk associated with deleterious mutations based on these large families may not accurately correspond to the risk for mutation carriers of BRCA1 and BRCA2 in the general population. Therefore, a negative result does not necessarily represent a “true negative” because mutations may be missed by the screening method but more importantly a deleterious mutation is likely to occur in genes other than BRCA1 and BRCA2. Another issue that continues
to complicate the interpretation of results is if the individual tests negative for known deleterious mutations but harbor missense variants of unknown risk. This has been estimated to occur in about 10% of women screened according to one study [29] and are often labeled “ambiguous” and do not provide useful information.

For women who test positive for a deleterious mutation, however, there are many options to proceed, including surveillance methods such as mammography and clinical breast exams. With careful surveillance, breast cancers can be detected early and treated successfully. Risk avoidance, which modifies lifestyles to reduce risk factors and prophylactic surgeries such as bilateral prophylactic mastectomy are also options that can be undertaken. However, research suggests that prophylactic surgeries may differ in their efficacy in protecting carriers of BRCA1 and BRCA2 mutations from disease [30]. The use of chemoprevention methods such as the drug tamoxifen, an estrogen receptor antagonist, has been shown in numerous clinical studies to reduce risk of developing breast cancer in BRCA1 and BRCA2 [31; 32] as well as reduce the chances of breast cancer occurring in the contralateral breast during initial treatment [33].

Thus, there is a clear need to identify novel breast cancer susceptibility genes to expand the number of genes to be tested to provide a more comprehensive analysis to supplement the current genetic tests so women at high risk can make an informed choice as to which course of action is the safest and most appropriate.
1.5 The TGF-β signalling pathway in mammary development and carcinogenesis

An important class of signalling molecules in the development of breast cancer is the transforming growth factor-β (TGF-β) family of polypeptide growth factors. They regulate pertinent cellular processes, including cell division, differentiation, motility, adhesion, and death, in virtually all tissues (Reviewed in [34]). TGF-β inhibits epithelial cell cycle progression in the lobules and ducts of the mammary gland and consequently controls epithelial cell proliferation and regression during mammary gland development as well as during the pre- and post-lactation period in the adult gland.

There are three TGF-β (TGF-β1, TGF-β2, TGF-β3) isoforms that are each encoded by different gene and expressed according to tissue-specific and developmentally specific fashion. TGF-β1 is the most universally and abundantly expressed isoform and is used in most of the studies conducted on the signalling pathway. TGF-β is secreted into the extracellular matrix as a latent protein complex bound to a latency-associated protein and one of the four isoforms of latent TGF-β binding protein. The signalling cascade responsible for the antiproliferative response of TGF-β is initiated when ligand-activated TGF-β binds to three high affinity cell surface receptors: the type I TGF-β receptor (TβRI), type II TGF-β receptor (TβRII), and type III TGFβ receptor (TβRIII, or also called betaglycan). Where expressed, TβRIII is the most abundant receptor and is most often responsible for binding the TGF-β ligand and transferring it to the signalling receptors TβRI and TβRII [35].

A general mechanism for the signalling cascade has been demonstrated (Figure 1) [36; 37]. TGF-β binds to either TβR-III and is transferred to TβR-II, or bind to TβR-II directly. TβR-I and TβR-II contain specific pair of type I and type II receptor serine/threonine kinases. Ligand binding to TβR-II leads to the transphosphorylation and activation of TβR-I by the TβR-II. Activated TβR-I then phosphorylates a specific subset of SMAD proteins, SMAD2 and SMAD3. These are termed receptor-activated SMADs (R-SMADs) because they are phosphorylated by TβR-I. SMAD2/3 then associates with SMAD4 which is a common partner for all of the receptor-activated SMADs and the complex translocates into the nucleus. Once in the nucleus, SMAD proteins form functional transcription complexes in association with DNA-binding factors, coactivators, or corepressors [38] to regulate the transcription of a variety of TGF-β responsive
genes. TGF-β signalling can be regulated by the level and duration of TGF-β receptor activation via nucleocytoplasmic shuttling of SMADs [39]. Therefore, TβRI, TβRII, SMAD2, SMAD3, and SMAD4 comprise the core SMAD-dependent TGF-β signalling pathway.

Figure 1: Overview of the TGF-β signaling pathway. This figure is from Massague et al. [69]. Activated TβR-I phosphorylates SMAD3. SMAD2/3 then associates with SMAD4 and the complex translocates into the nucleus. Once in the nucleus, SMAD proteins form functional transcription complexes in association with DNA binding factors to effect TGFβ mediated growth inhibition.
TGF-β achieves growth inhibition by arresting cells in the G1 phase of the cell cycle resulting in induction of apoptosis or terminal differentiation [40]. TGF-β achieves G1 arrest by inducing a hypophosphorylated pRB tumour suppressor that bind and sequester the E2F transcription factor through the transcriptional up-regulation of the cyclin-dependent kinase inhibitors p15\textsuperscript{Ink4B} [41], p21\textsuperscript{cip1/waf1} [42], as well as the downregulation of cdc25A (cyclin-dependent kinase activating phosphatase) and the downregulation of proto-oncogene c-Myc [43].

In various stages of breast cancer, TGF-β appears to play a dual role as both tumour suppressor and oncogene. In the early stages of disease, its antiproliferative properties inhibit the outgrowth of carcinomas \textit{in situ}. In transgenic mouse models, tumour formation induced by the overexpression of TGF-α and administration of a chemical carcinogen can be markedly reduced by the concomitant over-expression of TGF-β1 [44]. In later stages of disease, many malignant breast cancers become insensitive to the effects of TGF-β-induced growth inhibition via mechanisms such as downregulation of the TGF-β2 receptor. In such instances, TGF-β is thought to promote tumour progression by enhancing tumour cell motility and invasiveness [45; 46] and the ability for the primary tumour to metastasize via SMAD-independent MAPK signalling pathway [46; 47; 48]. Therefore, the core components involved in TGF-β signalling, if inactivated, may result in the loss of anti-proliferative responsiveness and can either predispose to or cause cancer progression.

1.6 SMAD3 and SMAD4 as candidate tumour suppressors in breast cancer susceptibility

Tumour suppressors known to be somatically inactivated in breast cancer are attractive targets especially for the non-familial form of breast cancer. Two genes that are of interest in clinical genetics of breast cancer are SMAD3 and SMAD4, the key signal transduction proteins of the TGF-β pathway.

SMAD3 located on chromosome 15q21 is a candidate tumour suppressor which has been implicated in several cancer types. The loss of SMAD3 expression and function is detected in gastric, colorectal cancers and acute T-cell lymphoblastic leukemia [49; 50; 51; 52]. The SMAD3 locus has also been shown to undergo loss of heterozygosity (LOH) in the form of frequent allelic loss in breast cancer and allelic imbalance in advanced breast carcinomas [53].
addition SMAD3 has been shown to counteract BRCA1-mediated DNA repair [54] and its MH2 domain has recently been shown to associate with BRCA1 during oxidative stress response [55]. As the majority of the known breast cancer susceptibility proteins (i.e. TP53, CHEK2, ATM, RAD50, NBS1, BRIP1/BACH1, PTEN and PALB2) [14] are involved in direct protein-protein interactions with BRCA1, this suggest that SMAD3 may also be involved in breast cancer susceptibility.

SMAD4/DPC4 tumour suppressor gene on 18q21 was found mutated or deleted in half of all human pancreatic carcinomas [56]. In addition, LOH or loss of expression of SMAD4 has been shown to be important for the progression of gastric [57], cervical [58] and colorectal [59] cancers. In breast cancer, LOH analysis indicates that the 18q21 region is a frequently lost region [60; 61; 62] and homozygous deletion of SMAD4 is associated with a small percentage of invasive ductal carcinomas of breast [63].

1.7 The SMAD3 and SMAD4 genes

In total, eight SMAD proteins are encoded in the human genome [34]. Of these five of the mammalian SMADs, SMAD1, SMAD2, SMAD3, SMAD5, and SMAD8, are referred to as receptor-regulated SMADs (R-SMADs) as they are substrates for the TGF-β family of receptors. SMAD1, SMAD5 and SMAD8 are the main substrates of BMP and anti-Muellerian receptors while SMAD2 and SMAD3 in particular are principle substrates for the TGF-β, activin, and Nodal receptors. SMAD4, commonly referred to as a Co-SMAD, is a common partner for all mammalian SMADs.

SMAD3 consists of 9 exons separated by introns ranging in size from 162 base pairs (bp) up to 98.5 kilo base pairs (kb) [64]. The MH1 domain is essentially encoded by exons 1 and 2, the linker domain by exons 3–5 and the MH2 domain by exons 6–8.

SMAD4 was mapped to chromosome 18q21 and contains 11 exons and is predicted to have a 552 amino acid coding sequence [65]. It is the human homolog of the Drosophila melanogaster
sma-4 but the protein sequence also has similarities to that of the Mad (mothers against dpp) protein and to the *Caenorhabditis elegans* Mad homologs sma-2, sma-3 [66].

### 1.8 Structure of the SMAD3 and SMAD4 transcription factors

SMAD proteins are generally ~500 amino acids in length and possess two globular domains connected by a linker region [67] (Figure 2, 3). The N-terminal domain is known as the Mad-Homology 1 (MH1) domain and is conserved across all R-SMADs as well as SMAD4. This is a DNA-binding domain stabilized by a tightly bound zinc atom. A β-hairpin (HP; two anti-parallel short β-strands separated by a linker loop) that is highly conserved in R-SMADs and SMAD4 [68] inserts into the major groove of DNA and binds three nucleotides of the 5’ GTCT 3’ motifs within SMAD binding element (SBE) of SMAD-responsive gene promoters. As the β-hairpin is highly conserved across all SMADs, specificity of response is often determined through high-affinity binding through association with different DNA binding co-factors with the R-SMAD/SMAD4 complex at the SBE of target gene promoters. This assures high affinity and specificity of the TGF-β response in a variety of cell types [69].

The linker region following the MH1 domain is a flexible segment which is divergent between the SMADs. In R-SMADs, the linker region possesses PY motifs that serve as binding sites for the WW domain on Smurfs (SMAD ubiquitination-related factor), ubiquitin ligases, and phosphorylation sites for MAPK, CDK, and other classes of protein kinases [70]. In SMAD4, this region includes a nuclear export signal (NES).

The C-terminal Mad-Homology 2 domain (MH2) is highly conserved across all SMAD proteins and is one of the most versatile protein-interacting domains in signal transduction. The C-terminal Ser-X-Ser motif of SMAD3 is phosphorylated by the activated receptor. In SMAD3 and other R-SMADs, a pocket consisting of basic residues and a specific L3 loop interacts with the phosphorylated region of the activated receptor. In SMAD4, the same basic pocket interacts with the phosphorylated tail of R-SMADs. In addition, a contiguous “hydrophobic corridor” on the surface of the MH2 domain mediates interaction with a variety of factors pertinent to function
including cytoplasmic retention proteins (e.g. SARA), nucleoporins, and DNA-binding co-factors [70]. Lastly, a SMAD4 activation domain (SAD) is a proline rich region of 48 amino acids found in the overlapping region of the linker region and the hydrophobic corridor of the SMAD4 MH2 domain and is responsible for interactions with transcriptional repressors and activators including p300 [71].

**Figure 2: Structure of SMAD3.** The MH1 domain (AA26-134) is highly conserved across R-SMADs and SMAD4. A β-hairpin contacts SMAD-Binding Elements (SBE) in target promoters. The variable linker (AA135-225) in between harbours phosphorylation sites for kinases including CDK and MAPK and PY motif recognized by SMURFs. The MH2 domain (AA226-403), contain a pocket consisting of basic residues and a specific L3 loop interacts with the phosphorylated region of the activated receptor. The hydrophobic corridor is pertinent for many important protein-protein interactions. The C-terminal Ser-X-Ser motif is phosphorylated by the activated receptor.

**Figure 3: Structure of SMAD4.** The MH1 domain (AA31-140) shares high homology with SMAD3 and other R-SMADs. SMAD4 import involved direct contacts with nucleoporins, whereas export involves binding of CRM1 to the NES motif found in the linker region. SMAD4 activation domain (SAD) overlapping the linker region and the hydrophobic corridor of the SMAD4 MH2 domain. The SMAD4 MH2 domain (AA321-552) possesses a basic pocket which recognizes the phosphorylated tail of R-SMADs for hetero-complex formation.
1.8 Functions of SMAD3 and SMAD4

The SMAD3/SMAD4 oligomer is a transcriptionally active complex. Biochemical analysis of complexes formed with SMADs with acidic residue substitution for the C-terminal serine residues suggests that they consist of two SMAD3 and one SMAD4 heterotrimer [72; 73]. X-ray crystallography further provided evidence showing that the complex formed by the MH2 domain of SMAD3 and SMAD4 is made up of two SMAD3 subunits and one SMAD4 subunit [74]. The SMAD transcriptional complex formation in vivo, on the other hand, appears to be more complicated as different target genes and/or cofactors associated with the complex and determine whether heterodimers and heterotrimers are formed [75]. A SMAD3-SMAD4 heterodimer and an unknown co-factor, for example, is involved in the targeting to the JunB promoter while a SMAD2-SMAD4 heterotrimer bound to FoxH1 target the complex to the Mix2 promoter [75]. It appears, therefore, that formation of SMAD heterodimers or heterotrimers will depend on factors such as the ratio of SMAD2 and SMAD3, presence of DNA binding partners in the complex, and the number of SMAD-binding elements (SBE) in the target promoter.

The SMAD transcriptional complex targets specific gene promoters for activation or repression by recruiting transcriptional co-activators and co-repressors. The most well-characterized SMAD coactivators are p300 and CBP [69]. Experiments using the adenovirus E1a protein showed that its expression inhibited SMAD3/4 mediated responses but the inhibition was not using observed in a E1a mutant lacking the p300/CBP-binding domain [76; 77]. The interaction between SMAD3 and SMAD4 with p300 and CBP can be mapped to the MH2 domains. The identification of histone deacetylases (HDACs) which can also form parts of the SMAD nuclear complexes and the fact that p300/CBP possess the ability to acetylate N-terminal histone tails [78; 79] suggest that SMADs themselves could function by remodelling the chromatin template.

Many genes are also transcriptionally repressed by the SMAD3/SMAD4 complex. The repression of the c-myc gene under the effect of the TGFβ response was the first example [80]. The inhibitory effect of TGFβ on c-myc expression is rapid and occurs in most of the cell types that are under the growth inhibition effect of TGFβ. A TGF-beta1 inhibitory element (TIE) within the c-Myc promoter is required for its transcriptional repression by TGFβ in epithelial cells by binding to a SMAD3/SMAD4 complex together with a E2F4/5 and DP1 heterodimer [81;
The binding of the corepressor p107 through contacts with both E2F4/5 and SMAD3 makes this a very specific complex to allow the downregulation of c-Myc [81]. With microarray-based transcriptome assays, it is estimated that at least one-quarter of all TGFβ gene responses in mammalian cells are repression-based responses [83; 84].

The TGFβ signalling pathway controls many critical processes during development including cell proliferation, differentiation, migration, and apoptosis. The TGFβ signalling pathways acts in tandem with other signalling pathway and it is now clear from the structure and function of SMAD proteins that they act as an important locale for the integration of inputs from many different sources.

1.9 Inactivating mutations in the TGF-β signalling pathway in breast cancer

Mutations in the main components of the TGF-β signalling pathway can directly lead to the loss of the anti-proliferative responses in tumour cells. The TβR-II receptor is inactivated by mutations in most gastric and colon cancers with microsatellite instabilities [85; 86]. In breast cancer, mutations in TβR-II have not been found in primary human breast carcinomas [87; 88] or breast carcinoma cell lines [89]. TβR-I, on the other hand, has been found to be homozygous deleted in a small number of biliary and pancreatic carcinomas [90]. Furthermore a three residue deletion from a nine alanine stretch found in colorectal and cervical carcinomas have an increased risk to develop these cancers [91; 92]. In breast cancer, a S387Y mutation found in TβR-I from 2 of 31 (6%) primary carcinomas and 5 of 12 lymph node metastases (42%) reduces its ability to signal and was postulated to be important in breast cancer progression [93]. This observation suggests that there may also be deleterious mutations in the key mediators of TGF-β signalling, namely SMAD3 and SMAD4.
1.10. Mutation analyses of SMAD3 in breast cancer

Based on existing literature, there is a lack of evidence for SMAD3 deleterious germline and somatic mutations in breast cancer. A summary of SMAD3/SMAD4 analyses in breast cancer shows that the number of samples screened is limited. Several studies failed to show inactivating somatic mutations in colon, lung, pancreas, and breast cancers among others [94; 95; 96] but the conclusions are based on mutation screening of less than 25 tumour samples in most of the cancers assessed. Germline analysis has only been performed for SMAD3 in 21 cases of juvenile polyposis syndrome and the authors did not report any deleterious mutations [97]. It was only recently that a heterozygous missense mutation (R373H) was found in colorectal cancer cell line, SNU-769A. This first reported inactivating mutation in SMAD3 was localized to the MH2 domain [98]. However, there have been no studies on whether germline mutations in SMAD3 are present in breast cancer and what their impact may be.

Table 2: Summary of mutation analyses of SMAD3 and SMAD4 in breast cancer. *Note A SMAD4 homozygous deletion was detected in the COLO-205, HT-29 (Colorectal), IGROV1 (ovarian) lines but none was observed in breast cancer NCI-60 lines.
1.11 Mutation analyses of SMAD4 in breast cancer

Somatic mutation analyses for SMAD4 have also been carried out extensively in various cancer types and inactivating mutations are found with the highest frequency in pancreatic and colon carcinomas. SMAD4/DPC4 was originally found to be a tumour suppressor gene on chromosome 18q21 that is deleted or mutated in almost half of human pancreatic carcinomas [56]. Somatic mutations have been found in cancers including leukemia [99], lung [100], biliary [101], colorectal [102; 103; 104; 105; 106], juvenile polyposis [107; 108; 109], cervical [110], ovarian [111], and less frequently in other cancer types [112; 113; 114; 115]. Germline mutations in SMAD4 reported in juvenile polyposis resulted in a SMAD4 protein with a defective C-terminus MH2 domain [116] and were linked to colon cancer susceptibility [117; 118].

Mutations of SMAD4 in breast cancer are thought to be rare and only homozygous deletions are found in a small portion of the breast cancer cell lines and invasive ductal carcinomas (IDC) [63; 99; 119]. However, the number of samples screened for breast cancer so far remains limited (Table 2) as they are based on 30 breast cancer cell lines or fewer. In addition, germline analysis has not yet been investigated in SMAD4 and thus is a research area that warrants further investigation.

1.12 The Mad-Homology 2 (MH2) domain as a somatic mutation hot spot in cancer

The MH2 domain serves a central role in receptor recognition, transactivation, interaction with transcription factors, co-activators and co-repressors, and homo and hetero-oligomerization among SMADs [34]. In SMAD4, the MH2’s importance is further exemplified by being a mutational hotspot for missense mutations in pancreatic and colorectal cancers, for which the mutation spectrums have been well characterized [120]. In particular a mutation clustering region (MCR), spanning codons 330-370 corresponding to the L2 loop of the loop-helix region responsible for the heterocomplex formation with other SMAD proteins and translocation into the nucleus to effect growth inhibition, has been described [120]. There also appear to be variations in the location of missense mutations depending on the human tumour type. In pancreatic cancers, missense mutations tend to occur in the MH1 or downstream of the MCR in
the MH2 domain. In colorectal or small bowel tumours, on the other hand, missense mutations occur predominantly within the MCR. These observations indicate that the MH2 domain is likely to be the site of novel missense mutations in other cancer types.
Chapter 2
Hypothesis and Objectives
2.1 Rationale
The chromosomal 15q21 region on which SMAD3 is found is a target of allelic imbalance in early breast carcinomas which becomes increasingly frequent in cancer progression [53]. Similarly the 18q21 locus containing SMAD4 is frequently lost in breast cancers and a cluster of breaks in the 18q21 encompassing SMAD4 was recently shown in breast carcinoma cell lines to be associated with minimum copy number [121]. The role of mutations in SMAD3 and SMAD4 in breast cancer is not well elucidated and current studies are based on the screening of 30 breast cancer cell lines or less. Given the importance of the Mad-Homology 2 (MH2) domain in SMAD3 and SMAD4 function as well as being an apparent mutation hot spot it is therefore of interest to identify deleterious germline alterations in this region that could potentially influence breast cancer susceptibility.

2.2 Hypothesis
SMAD3 and SMAD4 represent potential familial breast cancer susceptibility genes. Any inactivating germline mutations disrupting the protein-interacting domains or splicing mutations causing aberrant splicing or altered expression are likely function in breast cancer development and therefore represent novel genetic risk factors.

2.3 Objectives
I. Discovery of germline alterations in familial and non-familial breast cancer cases and non-cancer population controls
II. Quantitative comparison between observed and established mutation rates
III. *In silico* characterization of genetic alterations on protein structure and splicing
IV. *In vitro* functional analyses of the genetic alterations on SMAD3 and SMAD4 splicing and germline expression
2.4 Significance

This study is the first germline mutation analysis in breast cancer cases and controls for SMAD3 and SMAD4 and will serve to expand our understanding the mutation spectrum of the two genes. In addition, results from this study will provide insight as to whether or not the two genes are involved in breast cancer susceptibility. Finally determining whether these genes are altered at the germline level would provide insight as to whether the existence of somatic mutations could provide a “second hit” to promote tumourigenesis and thus would be an important indicator of whether or not more extensive somatic mutation screening of the two genes is warranted.
Chapter 3
Materials and Methods
3.1 Discovery of novel polymorphism by DHPLC analysis

To gain insight into the role of SMAD3 and SMAD4 mutations in breast cancer we performed a genetic screen for the presence of germline genetic alterations within the Mad-homology 2 protein interaction domain sequences of both genes in familial and non-familial breast cancer cases as well as age, gender and ethnicity matched healthy controls, from the Ontario population recruited by the Ontario Familial Breast Cancer Registry (OFBCR). The results were then analyzed by both in silico and in vitro methods to validate the impact of the alterations. An overview of the experimental design for the mutation screen and subsequent analyses is summarized in Figure 4.

**Figure 4: Experimental overview.** This study was divided into a discovery phase (DHPLC and complementary sequencing). The alterations found in the first phase are then validated for potential impact on protein structure and splicing.
3.1.1 Selection and clinical characteristics of the study population

Genomic DNA was prepared from blood lymphocytes from 408 breast cancer cases and 710 non-cancer population controls from the Ontario Familial Breast Cancer Registry (OFBCR), which recruited newly diagnosed cases of invasive breast cancer without BRCA1 and BRCA2 mutations identified through the population-based cancer registry in the province of Ontario [122; 123]. Collection and characteristics of the cases and controls are described in Table 3.

<table>
<thead>
<tr>
<th>Table 3: Criteria defining probands at increased genetic risk of breast cancer. The above criteria are obtained from Knight et al [122]. All probands were diagnosed with invasive breast cancer and are intended to capture wide patterns of breast cancer among families.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) One first degree relative with breast or ovarian cancer</td>
</tr>
<tr>
<td>2) Two second degree relatives with breast or ovarian cancer</td>
</tr>
<tr>
<td>3) One second degree or ≥ one third degree relative with breast cancer at age ≤35 or ovarian cancer at age ≤60</td>
</tr>
<tr>
<td>4) One second degree or ≥ one third degree relative with male breast cancer</td>
</tr>
<tr>
<td>5) One second degree or ≥ one third degree relative with multiple breast cancer primaries</td>
</tr>
<tr>
<td>6) One second degree or ≥ one third degree relative with breast and ovarian cancer</td>
</tr>
<tr>
<td>7) Family has three first degree relatives with any combination of breast, ovarian, colon, prostate, or pancreatic cancer or sarcoma with at least one diagnosis ≤50</td>
</tr>
<tr>
<td>8) Proband is male</td>
</tr>
<tr>
<td>9) Proband diagnosed with breast or ovarian cancer at age ≤35</td>
</tr>
<tr>
<td>10) Proband diagnosed with both breast and ovarian cancer or multiple breast primaries</td>
</tr>
<tr>
<td>11) Proband is Ashkenazi Jewish</td>
</tr>
</tbody>
</table>
Cases were classified as familial breast cancer probands if they met certain high-risk criteria, which includes women with at least one first-degree relative with breast and/or ovarian cancer, with second degree relatives with breast and ovarian cancer, or with additional cancers (e.g. prostate, pancreatic etc) in the first or second degree relatives and who diagnosed at age <36, with multiple breast and/or ovarian primaries, or Ashkenazi Jewish background where breast cancer rate occurs in 2.5% of the population (173 cases).

The breast cancer cases that did not meet the familial criteria listed above were classified as non-familial (sporadic) breast cancer cases (n=235), which represented the older patients with no family history of breast cancer. All the breast cancer cases (n=408) included in this study were previously tested negative for BRCA1/2 mutations.

Lists of randomly selected, listed, residential telephone numbers for the province of Ontario were used to identify female non-cancer population controls (710 controls). They were frequency-matched to female case probands based on their expected 5-year age distribution and ethnicity (64% response rate). The registry sample consist of about 90% Caucasian women and healthy female population controls which were all below 55 years of age. To address concerns regarding population stratification, we have matched the cases and controls by age and ethnicity (Figure 5).

Figure 5: Age-frequency matched case-control population from OFBCR. Blue bars indicate breast cancer cases and green bars present population controls and are matched evenly by 5-year age frequency.
3.1.2 PCR amplification of exon/intron-exon junctions of the MH2 domain

Polymerase Chain Reaction (PCR) was used to amplify the coding exons and exon-intron boundaries of the MH2 domains of both SMAD3 (exons 7, 8, 9) and SMAD4 (exons 10, 11, 12, 13). Thermocycling conditions and PCR primer sequences are summarized in Table 4. PCR was carried out in 50μl volume containing 10ng of genomic DNA, 1xPCR Gold buffer, 25ng of each 10mmol/l primer, 2.5 Unit of Taq DNA polymerase (AmplitaqGold; Perkin-Elmer, Branchbury, NJ, USA). Thermocycling was carried out in a Bio-Rad Dyad thermocycler (Conditions, Table 4) and the quality of PCR products was evaluated on 1.5% agarose gels. Furthermore to ensure proper formation of homo and hetero-duplexes for subsequent DHPLC analysis, PCR products were denatured again at 95°C for 3 minutes and re-natured for 30 minutes by decreasing temperature from 95°C to 65°C.

Table 4: Primer sequences and optimized PCR and DHPLC running conditions. The primers were designed based on genomic sequences of SMAD3 (GenBank accession NC_000015.8) and SMAD4 (GenBank accession NC_000018.8).
3.1.3 Determination of optimal amplicon melting temperature for DHPLC

The PCR amplicons were subject to analysis using denaturing High-Performance Liquid Chromatography (dHPLC). DHPLC is a PCR-based method where PCR fragments can be studied under semi-denaturing conditions using a capillary injection system [124]. Fragment separation is achieved as the PCR amplicon is run through a separation cartridge which binds dsDNA and release it as the helix of the DNA is unwound. The DNA is then eluted from column as increasing concentration of acetonitrile flows across the matrix. Heteroduplexes formed in hetero samples elute prior to homoduplexes thus appearing as two or more peaks in electropherograms and can be easily distinguished from samples without alterations (Figure 6).

DHPLC has been reported to be a highly sensitive technique with close to 100% detection rate while single-strand conformation polymorphism (SSCP), one of the most frequently used sequencing technique, was only 94% [125]. Given that it is fully automated and can run 3x96 well plates at a time, it is a suitable platform for the detection of rare and common variants.

The optimal melting temperature for all PCR amplicons was calculated using the dHPLC Melt Program (http://insertion.stanford.edu/melt.html) and amplicons from breast cancer cell lines (MDA-MB453, MDA-MB468, TD47) were used to optimize the running conditions to enhance mutation detection sensitivity on the Transgenomic WAVE 4500HT (Transgenomic Inc., Omaha., NE, USA).

Approximately 10ng samples corresponding to breast cancer cases and population controls were analyzed based on the optimized conditions. Samples with elution profiles characteristic of heteroduplexes were identified using the Navigator 1.7.0 Software. As an internal control, a fraction of case and control samples were duplicated across our study population to ensure accuracy of the results. All samples, showing heteroduplex profiles, were purified by SAP/ExoI and the direct sequencing was performed by The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada. Mutation Surveyor DNA Variant Analysis software Version 3.20 (Soft Genetics) was used to evaluate the sequencing results.
Figure 6: Denaturing High Pressure Liquid Chromatography (DHPLC) SNP detection. This system utilizes rapid fragment separation and re-annealing of the strands to achieve quantification of single- and double-stranded nucleic acid fragments, allowing for mutation detection and SNP discovery. The bottom left panel shows a homozygote elution profile with one peak while bottom right panel represents a heterozygous elution profile with double peaks.
3.2 Nucleotide diversity estimation

Nucleotide diversity and its standard deviation were calculated under the assumptions of an infinite site neutral allele model:

$$\theta = k / aL \quad \text{and} \quad s(\theta) = \sqrt{a\theta L + b(\theta L)^2 / aL}$$

$$a = \sum_{i=2}^{n} \frac{1}{i-1} \quad b = \sum_{i=2}^{n} \frac{1}{(i-1)^2}$$

where $K$ is the number of SNPs identified in a genomic length, $L$ base pairs and $n$ is the number of alleles analyzed [126].

3.3 In silico characterization of identified SNPs

The majority of the variants found in this study were intragenic with the exception of three variants found in the exon region of SMAD4. Therefore, a series of bioinformatic tools were used in order to predict whether any of the variants would have a functional outcome. Due to the fact that most of the variants were not located in consensus splicing motifs, we tested whether any variants could be creating cryptic splice sites or causing aberrant splicing through the abolition of Exonic Splicing Enhancers (ESE) (Figure 7).

Figure 7: Illustration of the key motifs involved in splicing. The consensus splice sites 5’AG and 3’GU are highly conserved. The polypyrimidine tract P(y)n promotes assembly of the spliceosome (comprised of small nuclear riboproteins U1-U6). The branch site (a) is the site of the lariat formation. Regulators of splicing efficiency include ESE/ESS in exons (regulated by SR protein, and hnRNPs) or in intron (ISE/ISS) the latter can be close to the exon or deep in the intron (Adapted from Pagani et al., [139]).
3.3.1 Evolutionary conservation analysis

The sole identified missense (non-synonymous) variant was assessed for impact on protein function by evolutionary conservation analysis using SIFT [127]. SIFT is a multiple sequence alignment tool that presumes changes of amino acids with important roles in the protein family would be deleterious to protein function as they tend to be evolutionarily conserved. Structural impact was assessed by PolyPhen (http://genetics.bwh.harvard.edu/). PolyPhen is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein. To standardize the predictions made by these two tools, where applicable, we have annotated the “affecting protein function” prediction of SIFT and both the “probably damaging” and “possibly damaging” predictions of PolyPhen as “damaging” in this report. Similarly, the “tolerated” prediction of SIFT and the “benign” prediction of PolyPhen are collectively annotated as “benign”.

3.3.2 Alternative Splice Analyzer (ASSA) analysis for variants promoting splice defects

In the case of intronic variants, predictions for the alteration of consensus donor sites, acceptor sites and branch points as well as creation of novel cryptic sites was carried out by measuring 5’ and 3’ splice site scores using Automated Splice Site Analyses (ASSA) (http://splice.uwo.ca) [129]. ASSA has been shown to be as robust [130] as other prevalent splice predictors NNSplice [131], SpliceSiteFinder, and MaxEntScan for the purpose of our analysis. All in silico splicing analysis tools were run at default threshold values and a comparison between the outputs for wild type versus variant was documented.

3.3.3 FASTsnp analysis for variants affecting Exonic Splicing Enhancers (ESE)

To evaluate the potential impact of the synonymous variants on Exonic Splicing Enhancers (ESE) in alternative splicing regulation, we utilized FASTsnp (Function Analysis and Selection Tool for Single Nucleotide Polymorphisms; http://fastsnp.ibms.sinica.edu.tw/) [132] a web server that allows users to identify high-risk SNPs according to their phenotypic risks and putative functional effects. We performed a novel SNP analysis by inputting the genomic sequence consisting of the genetic alteration flanked by 15 nucleotides (31 nt) and running the algorithm at default thresholds.
3.4 *In vitro* characterization of the effect of genetic polymorphisms on mRNA splicing

To study the effect of germline intronic variants on splicing, cDNA was prepared from mRNA extracted from 37 cases/controls found in this study to harbour a genetic alteration in either SMAD3 or SMAD4. The study was limited to variants occurring fewer than 6 times in the mutation analysis (<1% mAF). The assay would be able to assess whether any of the variants shown to create a cryptic splice site or lead to the abolition of ESEs would have an impact on the proper splicing of the mRNA transcript. Even though many of the variants did not have a bioinformatic prediction, their impact on splicing was still assessed due to the possibility that they may influence intronic features that are not readily predictable such as intronic Exonic Splicing Enhancer / Silencers (ESE/ESS).

3.4.1 Preparation of cDNA

This assay was carried out based on instructions provided by SuperScript III One-Step RT-PCR System with Platinum *Taq* DNA Polymerase kit (Invitrogen). A complete list of cDNA prepared and their associated variants are listed in Table 5. The reverse-transcription–polymerase-chain-reaction (RT-PCR) assay was carried out with the use of Invitrogen’s SuperScript III One-Step RT-PCR System with Platinum *Taq* DNA Polymerase (Invitrogen) according to the manufacturer’s instructions; 5 μg of RNA was used. A 12 μl of RNA with random and oligo dT primers was incubated for 10 minutes at 70°C. Following this, 8 μl of a reaction mix containing 5X First Strand Buffer, dNTP (10mM), 0.1M DDT and Superscript III was added and incubated for 1 minute at 50°C, 15 minutes at 70°C then cooled to 4°C.
<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Age</th>
<th>Familial *</th>
<th>SMAD3</th>
<th>SMAD4</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>53</td>
<td>OFBCR</td>
<td>IVS8+23A&gt;C</td>
<td>None</td>
</tr>
<tr>
<td>P2</td>
<td>57</td>
<td>None</td>
<td>IVS10+41G&gt;A</td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td>39</td>
<td>OFBCR</td>
<td>IVS8+109A&gt;G</td>
<td></td>
</tr>
<tr>
<td>P4</td>
<td>45</td>
<td>OFBCR/FDR</td>
<td>IVS9+132A&gt;T</td>
<td>None</td>
</tr>
<tr>
<td>P5</td>
<td>44</td>
<td>None</td>
<td>IVS9+118A&gt;G</td>
<td></td>
</tr>
<tr>
<td>P6</td>
<td>51</td>
<td>OFBCR</td>
<td>IVS10-33T&gt;A</td>
<td></td>
</tr>
<tr>
<td>P7</td>
<td>47</td>
<td>None</td>
<td>IVS10+132delA</td>
<td></td>
</tr>
<tr>
<td>P8</td>
<td>N/A</td>
<td>OFBCR/FDR</td>
<td>IVS8+23A&gt;C</td>
<td>None</td>
</tr>
<tr>
<td>P9</td>
<td>N/A</td>
<td>OFBCR</td>
<td>None</td>
<td>c.1478G&gt;A / p.Asp450Asp</td>
</tr>
<tr>
<td>C1</td>
<td>44</td>
<td>None</td>
<td>IVS8+44T&gt;C, IVS9+106del17N</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>44</td>
<td>None</td>
<td>IVS9+68delGAA</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>42</td>
<td>IVS8+23A&gt;C</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>37</td>
<td>None</td>
<td>IVS8+109A&gt;G</td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>49</td>
<td>None</td>
<td>IVS10+132delA</td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>34</td>
<td>IVS8-211C&gt;T</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C7</td>
<td>43</td>
<td>None</td>
<td>IVS10-52A&gt;T</td>
<td></td>
</tr>
<tr>
<td>C8</td>
<td>43</td>
<td>OFBCR/FDR</td>
<td>IVS8+23A&gt;C</td>
<td>None</td>
</tr>
<tr>
<td>C9</td>
<td>54</td>
<td>IVS8+23A&gt;C</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C10</td>
<td>45</td>
<td>IVS6-132C&gt;T</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C11</td>
<td>50</td>
<td>IVS7+69G&gt;C</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C12</td>
<td>52</td>
<td>None</td>
<td>c.1214T&gt;C / p.Phe362Phe **</td>
<td></td>
</tr>
<tr>
<td>C13</td>
<td>47</td>
<td>None</td>
<td>IVS7-121A&gt;C</td>
<td></td>
</tr>
<tr>
<td>C14</td>
<td>50</td>
<td>IVS8+161C&gt;T</td>
<td>IVS8+109A&gt;G</td>
<td></td>
</tr>
<tr>
<td>C15</td>
<td>31</td>
<td>OFBCR/FDR</td>
<td>None</td>
<td>IVS8+109A&gt;G</td>
</tr>
<tr>
<td>C16</td>
<td>35</td>
<td>IVS8+23A&gt;C</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C17</td>
<td>48</td>
<td>IVS8+23A&gt;C</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C18</td>
<td>46</td>
<td>IVS8+48T&gt;G</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C19</td>
<td>43</td>
<td>IVS7+69G&gt;C, IVS8-55A&gt;G</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C20</td>
<td>46</td>
<td>None</td>
<td>IVS11+53A&gt;G</td>
<td></td>
</tr>
<tr>
<td>C21</td>
<td>45</td>
<td>IVS8-170C&gt;T</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C22</td>
<td>67</td>
<td>IVS8+48T&gt;G</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C23</td>
<td>49</td>
<td>None</td>
<td>IVS8-121A&gt;C</td>
<td></td>
</tr>
<tr>
<td>C24</td>
<td>37</td>
<td>None</td>
<td>c.1701A&gt;G / p.Ile525Val</td>
<td></td>
</tr>
<tr>
<td>C25</td>
<td>52</td>
<td>IVS7+69G&gt;C, IVS8-55A&gt;G</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C26</td>
<td>66</td>
<td>IVS7+69G&gt;C</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>C27</td>
<td>60</td>
<td>None</td>
<td>IVS10+132delA</td>
<td></td>
</tr>
<tr>
<td>C28</td>
<td>62</td>
<td>None</td>
<td>IVS10+132delA</td>
<td></td>
</tr>
</tbody>
</table>

**Table 5: Clinical characteristics of 37 mRNA from cases or controls.** P represents a patient and C represents control. The familial categorization is based on the OFBCR and/or FDR (First Degree Relative) which indicates the strength of familial association based on clinical characteristics and patient history. Please note that in ** the variant was found in a case and a control, however, we were unable to obtain the cDNA from the breast cancer case due to sample degradation. The genomic reference coordinates are reported in accordance to the HGVS nomenclature.
3.4.2 PCR (RT-PCR) amplification of the coding exons of the MH2 domain

A total of thirty-seven mRNA samples, 18 for SMAD3 and 19 for SMAD4, were extracted from cases and controls harbouring the rare genetic variants (defined as <5 times) identified in this study. Reverse-transcription polymerase-chain-reactions (RT-PCR) primers targeting the flanking exons of the MH2 domain of SMAD3 (exons 6, 9) and SMAD4 (exons 10, 13) were based on cDNA sequences NM_5902 and NM_005359, respectively. This assay was carried out based on instructions provided by SuperScript III One-Step RT-PCR System with Platinum Taq DNA Polymerase kit (Invitrogen). Conditions and primer sequences are summarized in (Table 6)

The RT-PCR products were separated on a 1.5% agarose and a non-denaturing 8% polyacrylamide gel (29:1) to ensure high resolution of fragments. The identity of the RT-PCR products was confirmed by direct sequencing of the gel-purified DNA. Controls consisted of reaction with the exclusion of the reverse transcription enzyme during RT-reaction, and the exclusion of mRNA template during the RT-reaction.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Exons</th>
<th>Amplicon size (bp)</th>
<th>Annealing Temperature (°C)</th>
<th>Mg2+ (mM)</th>
<th>Forward PCR Primer (5'-3')</th>
<th>Reverse PCR Primer (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMAD3</td>
<td>6 to 9</td>
<td>688</td>
<td>61</td>
<td>15</td>
<td>TAAAGCGCTGCTCTCAATGTCAA</td>
<td>TGCGGCCAAGAGGTAAATGTG1</td>
</tr>
<tr>
<td>SMAD4</td>
<td>10 to 13</td>
<td>519</td>
<td>56</td>
<td>15</td>
<td>TATAAATTGAGCTCTTCTA0000</td>
<td>GACATCCCTACTACGCGTGTA00</td>
</tr>
</tbody>
</table>

Table 6: Primer sequences and optimized RT-PCR conditions. The primers were designed based on cDNA sequences of SMAD3 (GenBank accession NM_5902) and SMAD4 (GenBank accession NM_005359).

3.5 Quantitative Real Time PCR for SNP effect on germline expression

Concurrent with the RT-PCR analysis, we also assessed whether or not germline expression levels are altered. The cDNA from cases and controls that harbored the identified SMAD3 and SMAD4 variants were studied using quantitative real-time PCR. For the real-time PCR analysis of the 18 SMAD3 and 19 SMAD4 variants, the cDNA samples were divided into breast cancers cases with variants (BC-VAR n=3, n=5) and controls with variants (CO-VAR n=12, n=13). Additionally, because the case/control samples harbouring SMAD3 variants were negative for SMAD4 variants and vice versa, each group was used as negative controls for the other to increase the power of the analysis (BC-REF, n=5, n=3; CO-REF, n=13, n=13).
3.5.1 Generation of standard curves

Quantitative Real-time PCR (qPCR) was performed using an ABI 7700 Sequence Detection System (PE Applied-Biosystems) in the presence of SYBR-green in a 30 μl reaction. The PCR conditions were standard (SYBR-Green I core reagent protocol) and all reagents were provided in the SYBR-Green I core reagent kit. PrimerBank [133] was used to obtain quantitative real time-PCR primers for SMAD3 (PrimerBank ID 5174513a2) and SMAD4 (PrimerBank ID 4885457a2). All reactions were run in triplicates and incubated in a 96-well optical plate at 95°C for 10 min, followed by 40 cycles of 95°C for 15s and 60° for 10 min. Standard curves were generated using ten-fold dilutions of pooled cell-line cDNA and used as a reference to extrapolate quantitative information for mRNA targets of unknown concentrations. β-actin (Forward 5’ ATCATGTTTGGACCTTCAA3’, Reverse 5 CATCTCTTGCTCGAAGTCCA 3’) was chosen as a standard reference gene for the assay for normalization.

3.5.2 Statistical analysis

Statistical analysis for t-test, and non-parametric Mann-Whitney test of independent samples were performed using SPSS v.13.0. Statistical significance was assumed at p<0.05 with error bars representing standard deviations (SD).

3.6 Catalogue of somatic breast cancer mutations in SMAD3 and SMAD4

We accessed the COSMIC database (http://www.sanger.ac.uk/genetics/CGP/cosmic/) [134], v44 release a project that catalogues homozygous or heterozygous somatic missense mutations and deletions in various cancer types based on curated research publications. Using this resource we compiled the number of currently known somatic mutations of SMAD3 and SMAD4 in breast, colorectal and pancreatic cancers.

3.7. SMAD3 and SMAD4 expression in primary tumour

To address whether or not SMAD3 and SMAD4 are differentially expressed in normal and breast carcinoma, ArrayExpress data (accession number: E-TABM-276) was downloaded from website (http://www.ebi.ac.uk/gxa/) and used to extract cDNA expression data in five probes (239448_at, 218284_at, 205396_at, s05397x_at, s05398_s_at) for SMAD3 and two (202527_s_at,
1563703_at) for SMAD4 from the Affymetric GeneChip U133 Plus 2.0 arrays [135]. We used microarray gene expression data from breast tumour tissues of 23 patients with invasive ductal carcinoma (IDC), 28 surrounding normal tissues, and 10 samples of normal breast tissue taken from surrounding healthy breast tissue of cancer patients. A 2 independent samples t-test and Levin’s test for the equality of variance was performed on mean expression levels to determine significance.
Chapter 4
Results
4.1 DHPLC optimization and screening elution profiles

The optimal melting temperature for each PCR amplicon was calculated using the dHPLC Melt Program (http://insertion.stanford.edu/melt.html). The theoretical temperatures were used as a benchmark to optimize running conditions in order to enhance mutation detection sensitivity. The elution profiles shown (Figure 8, top) represent pooled DNA from breast cancer cell lines (MDA-MB453, MDA-MB468, TD47). The melting temperature(s) closest to the midpoint of the elution run were chosen for the running condition. A representative elution profile of a mutation analysis screening 96 cases/controls in SMAD4 exon 12 is shown in the middle panel of Figure 8. Upon normalization using the mutation calling feature of the DHPLC Navigator program, the sample harbouring the c.1478G>A (Asp450Asp) mutation can be clearly distinguished from the rest of the samples (Figure 8, bottom).
Figure 8: DHPLC elution profiles. Top: Representative optimization of the SMAD4 exon 10 PCR amplicon to obtain the optimal melting temperature (48-57°C) for separation. Middle: Mutation analysis of 80 samples of the SMAD4 exon 10 amplicon. Bottom: Normalized chromatograms using the mutation calling feature of the DHPLC navigator program. Indicated mutatation are c.1214T>C (Phe362Phe), IVS9-121A>C and IVS10+109A>C found in three distinct case or control.
4.2 Mutation analysis of the Mad-Homology 2 (MH2) domain of SMAD3 and SMAD4

Among the 1118 samples screened for each exon, we have identified a total of 11 and 16 distinct genetic variants in the MH2 domains of SMAD3 and SMAD4, respectively (Table 7). The rate of detecting variants in SMAD3 was 0.25% (1/408) in cases, 0.98% (7/710) in controls and 0.27% (3/1118) in both groups. All the variants in SMAD3 represented intronic variants in the form of single base substitutions and small deletions. The frequency of variants in SMAD4 was found to be 0.98% (4/408) in cases and 1.13% (8/710) in controls, and 0.36% (4/1118) in both cases and controls. We report three coding variants in SMAD4 including c.1214T>C (p.Phe362Phe; rs1801250) found in both cases and controls, and two novel variants c.1478G>A (p.Asp450Asp) found in cases and c.1701A>G (p.Ile525Val) in controls (Table 7).

Table 7: Novel germline variants detected in SMAD3 and SMAD4 in breast cancer. Nucleotide position of the variants was numbered based on genomic sequences of SMAD3 (GenBank accession NC_000015.8) and SMAD4 (GenBank accession NC_000018.8). The cDNA sequences referenced here are NM_5902 (SMAD3) and NM_005359 (SMAD4). The coding variants identified are bolded.
4.3 Nucleotide diversity estimations

Under the neutral theory of molecular evolution and infinite sites model, comparison of sequence diversity can be estimated by the heterozygosity per nucleotide site (π), termed nucleotide diversity, or by the mutation parameter (θ). Both correct for sample size and the length of region screened and are nearly equivalent. In this study we report θ and compared the variant frequencies from our sequencing data to other gene sets including (a) mutation analysis of the coding and adjacent non-coding regions of 106 genes from clinically relevant pathways including cardiovascular, neuropsychiatry, endocrinology in 57 individuals from Cargill et al [126], (b) the 5’ and 3’ UTR, intron, and coding region of 75 candidate genes involved in blood pressure homeostasis in 74 individuals by Halushka et al [136], and (c) rate of polymorphisms in the entire coding regions and 3’ UTR of highly conserved and essential genes involved in DNA replication, transcription and replication from ten Ashbroek et al [137] (Table 8) We use results reported for the European/American subgroups where applicable.

<table>
<thead>
<tr>
<th>Site of Variation</th>
<th>Base Pairs Screened (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic of Study Genes</td>
<td>This Study (SMAD3)</td>
</tr>
<tr>
<td>Coding</td>
<td>1135</td>
</tr>
<tr>
<td>Non-Coding</td>
<td>415</td>
</tr>
<tr>
<td>Total</td>
<td>1550</td>
</tr>
</tbody>
</table>

Table 8: Features of the nucleotide diversity comparison studies. Characteristics of this study as well as the reference studies are summarized. The mutation analysis from this study is normalized according to number of base pairs and individuals screened and compared to three large-scale studies. * Base pair count includes 5’ and 3’ UTR, ** includes 3’ UTR.
The frequency of coding variants of SMAD3 in cases and controls (θ=0) was significantly lower compared to that for SMAD4 (θ=3.99x10^{-4} and 3.71x10^{-4}, respectively) (Figure 9, top). In addition, the absence of coding variants was lower than expected when compared to rates observed by Cargill et al., (θ=5.43x10^{-4}) and Halushka et al., (θ=4.5x10^{-4}) and was closest albeit lower compared the study by Ten Ashbroek et al (θ=2.00x10^{-4}). Interestingly, the θ values for SMAD4 coding variants in cases and controls (θ=3.99x10^{-4} and 3.71x10^{-4}, respectively) is similar to that reported by Cargill et al., and Halushka et al., studies (θ=5.43x10^{-4} and 4.5x10^{-4}, respectively), with standard deviations overlapping, suggesting that SMAD4 alterations appear to be occurring at an expected rate.

Comparing coding and non-coding variants in both cases and controls of SMAD3 and SMAD4 to the expected nucleotide diversity the frequency of the non-coding SMAD3 and SMAD4 variants in from cases and controls was much higher than values reported by Cargill et al, Halushka et al when considering the value reported for the European-American samples, and the 3’UTR region reported by Ten Ashbroek et al. (Figure 9, bottom). However, this may be simply due to the difference in the study design, where larger regions of the non-coding exon-intron boundaries were covered in our study (~150bp) compared to <18bp in the reference studies.
Figure 9: Profile of the nucleotide diversity comparison between this study and 3 major studies. Please note used θ in 3’ UTR for non-coding data for Ten Ashbroek et al study as they did not perform any analysis on intronic sequences. Due to the fact that the Halushka et al study was split into half African ethnicity and half north American of European descent, only the coding/non-coding data from the European/American samples subset is used for the purpose of this comparison. The error bars represent standard deviation values.
4.4 *In silico* characterization of identified SNPs using bioinformatic tools

Intronic variants are increasingly found to be associated with aberrant splicing leading to cancer. Since all the intronic variants identified in this study occur outside the consensus splice donor and acceptor sites, we applied Splice Analyzer to address whether these variants are likely to create cryptic splice sites or deleteriously impact the branch points.

Detailed computational analyses for ASSA are summarized in Table 9. Among 11 intronic variants in SMAD3, only IVS8-55A>G, identified in two population controls, was predicted to abolish a branch site. In SMAD4, 6/13 variants were predicted to affect proper splicing. We found that IVS12-33T>A (P6) (in a familial breast cancer patient), IVS11+126 del7 (C1) and IVS12-52 A>T (C7), and IVS10+109 A>G (in both cases and controls) were all predicted to create cryptic sites.

The remaining two synonymous SMAD4 variants, **c.1478G>A** (p.Asp450Asp) (P9) found in a patient with familial breast cancer and **c.1214T>C** (p.Phe362Phe) (C12 only) from a familial breast cancer case and a healthy control with strong family history of breast cancer, were also predicted to impact the regulation of proper splicing via loss of exonic enhancing motifs. A novel non-synonymous variant **c.1701A>G** (p.Ile525Val) (C24) was found once in the control population. However, it does not reside in a functionally significant domain and does not confer a deleterious impact since SIFT or PolyPhen analyses and ESE-Finder predict no changes to protein structure and splicing regulation, respectively.
Table 9: Detailed summary of bioinformatic predictions for SMAD3 and SMAD4 variants. Only the intronic variants which were predicted by ASSA (top) or the synonymous variants by FASTsnp (bottom) to have a functional impact are listed. Most of the intronic were predicted to create cryptic splice sites while the synonymous variants, despite having a benign prediction from SIFT and Polyphen, are shown to abolish important alternative splicing regulatory motifs (ESEs). Most notably the c.1478G>A / p.450Asp450 variant found only in a breast cancer case.

4.5 *In vitro* analysis of SNP impact on mRNA splicing

We investigated the impact of all variants occurring fewer than 5 times in our population, in addition to the seven which had a functional prediction by bioinformatics, on splicing and expression levels of the mRNA transcript. This allowed for a thorough investigation that encompasses both the effect of the predicted splicing mutants as well as alterations on intronic structures such as ISS/ISE which at present cannot be reliably predicted *in silico* that might otherwise be missed. A total of 37 RNA samples carrying SMAD3 or SMAD4 variants (Table 5) were reverse transcribed into cDNA, which was then used to amplify cDNA segments covering...
multiple exons within the MH2 domains of SMAD3 (exons 6-9) and SMAD4 (exons 10-13). However, the gel electrophoresis of the PCR products showed the absence of aberrantly spliced transcripts in the samples studied (Figure 10).

**Figure 10: Reverse-Transcription PCR analysis of SMAD3 and SMAD4 variants.** Patients are labelled from P1 to P9, whereas the controls are from C1 to C28. The 500kb position on the 1kb ladder is labelled. Top panel: SMAD3 RT-PCR analysis shows no aberrant splicing (Expected size: 688bp). Bottom: SMAD4 RT-PCR analysis shows variants do not result in aberrant splicing (Expected size: 518 bp). Controls for both experiments were as follow: lane 1 (no cDNA template), lane 2 (No RT enzyme during RT-reaction), lane 3 (no mRNA during RT-reaction). Patients and control mRNA samples and the identified variants they harbour were listed in Table 5. Note in both SMAD3 and SMAD4 the non-specific bands were sequenced and found to be due to non-specific primer binding and amplification. Subsequent re-amplification of P5 shows normal amplification.
4.6 Standard curve generations for SMAD3 and SMAD4 qPCR primers

Using cell line DNA, standard curves were generated for SMAD3 and SMAD4 (Figure 11). The R2 values were consistent between the two samples at 0.993562 and 0.994772 for SMAD3 and SMAD4, respectively.

![Standard Curve](image)

![Delta Rn vs Cycle](image)

**Figure 11: Real-time quantitative PCR standard curve and amplification plots**

This is a representative standard curve (top) which was generated using serial dilutions of pooled RKO, HCT116 colon cancer cells, MDA-MB453, MDA-MB468 breast cancer cell lines with Pearson correlation of R2>0.990. cDNA was analyzed in triplicate. The amplification plot (bottom) is an example of a typical reaction including standards and unknowns.
4.7 Effect of SNP on germline expression

When the same cDNA samples were subjected to quantitative real-time PCR (qPCR) (Table 5) SMAD3 expression levels in breast cancers harboring variants (BC-VAR) were significantly higher compared to both control groups (CO-VAR; p=0.038, CO-REF; p=0.035) (Figure 12). Of the three variants in the BC-VAR group, IVS8+23A>C (P8) which showed a 6-fold increase is also found in P1 where expression did not change (Table 10), and is present in several population controls. IVS9+132A>T was only found once in the BC-VAR group and showed a 2.39 fold increase. Given that BC-VAR and BC-REF did not differ significantly (p=0.85) and that SMAD3 BC-REF cases such as P5 had a high increase in expression (>12 fold), the SMAD3 variants of the MH2 domain presented here do not seem to be a strong driving force for the observed change in expression. Nevertheless, mean expression levels were significantly higher in the breast cancer (BC) versus control (CO) groups (p=0.02), supporting the notion that altered SMAD3 germline expression could be an important factor in breast cancer.
Figure 12: SMAD3 expression analysis by Quantitative Real-Time PCR (qPCR): The mean expressions of four groups of mRNA analyzed are compared. Statistical significance was determined by Mann-Whitney test of independence with error bars representing standard deviation (SD). The breast cancer cases harbouring variants (BC-VAR) is highly expressed relative to controls groups with (CO-VAR; p=0.038) and without (CO-REF; p=0.035) SMAD3 alterations. Inset: Grouped BC vs. CO statistically significant (p=0.02) showing that SMAD3 expression in breast cancer cases, regardless of having variants, are significantly higher. Line within the box, median; bars above and below the box, 90th and 10th percentiles, respectively.
Table 10: Quantitative RT-PCR analysis of SMAD3. The mRNA from the BCFR repository studied are broken down into four distinct groups: Breast cancer cases and controls with variants (BC-VAR, CO-VAR) and those without variants (BC-REF, CO-REF). The normalized expression ratios from qPCR are listed with ±SD. The familial categorization is based on the BCFR and/or FDR (First Degree Relative) which indicates the strength of familial association based on clinical characteristics and patient history. The intronic variant IVS9+132 A>T (P4) represents a potential breast cancer mutation as it shows double expression levels over baseline. The IVS8+23 A>C (P1, P8) may be a low penetrant mutation as it is also found in normal controls and show normal expression in cases (e.g. P1). Three samples (C11, C20, C21) were removed due to poor mRNA quality. C12, C16 and P9 were not included due to cDNA sample limitation.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Age</th>
<th>Familial</th>
<th>SMAD3 Variant(s)</th>
<th>SMAD3/B2M Ratio</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3</td>
<td>39</td>
<td>OFBCR</td>
<td></td>
<td>0.36 ± 0.02</td>
<td>BC-REF</td>
</tr>
<tr>
<td>P6</td>
<td>51</td>
<td>OFBCR</td>
<td></td>
<td>0.64 ± 0.01</td>
<td>BC-REF</td>
</tr>
<tr>
<td>P7</td>
<td>47</td>
<td></td>
<td></td>
<td>1.94 ± 0.1</td>
<td>BC-REF</td>
</tr>
<tr>
<td>P2</td>
<td>57</td>
<td></td>
<td></td>
<td>3.69 ± 0.1</td>
<td>BC-REF</td>
</tr>
<tr>
<td>P5</td>
<td>44</td>
<td></td>
<td></td>
<td>12.52 ± 1.8</td>
<td>BC-REF</td>
</tr>
<tr>
<td>P1</td>
<td>53</td>
<td>OFBCR</td>
<td>IVS8+23A&gt;C</td>
<td>1.11 ± 0.08</td>
<td>BC-VAR</td>
</tr>
<tr>
<td>P4</td>
<td>45</td>
<td>OFBCR/FDR</td>
<td>IVS9+132 A&gt;T</td>
<td>2.39 ± 0.1</td>
<td>BC-VAR</td>
</tr>
<tr>
<td>P8</td>
<td>N/A</td>
<td>OFBCR/FDR</td>
<td>IVS8+23A&gt;C</td>
<td>6.08 ± 0.7</td>
<td>BC-VAR</td>
</tr>
<tr>
<td>C13</td>
<td>47</td>
<td></td>
<td></td>
<td>0.15±0.002</td>
<td>CO-REF</td>
</tr>
<tr>
<td>C7</td>
<td>43</td>
<td></td>
<td></td>
<td>0.43±0.03</td>
<td>CO-REF</td>
</tr>
<tr>
<td>C5</td>
<td>49</td>
<td></td>
<td></td>
<td>0.47±0.03</td>
<td>CO-REF</td>
</tr>
<tr>
<td>C24</td>
<td>37</td>
<td></td>
<td></td>
<td>0.48±0.02</td>
<td>CO-REF</td>
</tr>
<tr>
<td>C28</td>
<td>62</td>
<td></td>
<td></td>
<td>0.49±0.03</td>
<td>CO-REF</td>
</tr>
<tr>
<td>C23</td>
<td>49</td>
<td></td>
<td></td>
<td>0.5±0.07</td>
<td>CO-REF</td>
</tr>
<tr>
<td>C1</td>
<td>44</td>
<td></td>
<td></td>
<td>0.83±0.1</td>
<td>CO-REF</td>
</tr>
<tr>
<td>C4</td>
<td>37</td>
<td></td>
<td></td>
<td>0.84±0.05</td>
<td>CO-REF</td>
</tr>
<tr>
<td>C15</td>
<td>31</td>
<td>OFBCR/FDR</td>
<td></td>
<td>1.61±0.05</td>
<td>CO-REF</td>
</tr>
<tr>
<td>C27</td>
<td>60</td>
<td></td>
<td></td>
<td>2.05±0.2</td>
<td>CO-REF</td>
</tr>
<tr>
<td>C2</td>
<td>44</td>
<td></td>
<td></td>
<td>2.98±0.2</td>
<td>CO-REF</td>
</tr>
<tr>
<td>C22</td>
<td>67</td>
<td></td>
<td>IVS8+48T&gt;G</td>
<td>0.35±0.05</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C14</td>
<td>50</td>
<td></td>
<td>IVS8+161C&gt;T</td>
<td>0.37±0.008</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C10</td>
<td>45</td>
<td></td>
<td>IVS6-132C&gt;T</td>
<td>0.42±0.03</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C9</td>
<td>54</td>
<td></td>
<td>IVS8+23A&gt;C</td>
<td>0.59±0.05</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C25</td>
<td>52</td>
<td></td>
<td>IVS7+69G&gt;C, IVS8-55 A&gt;G</td>
<td>0.60±0.2</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C18</td>
<td>46</td>
<td></td>
<td>IVS8+48T&gt;G</td>
<td>0.62±0.1</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C26</td>
<td>66</td>
<td></td>
<td>IVS7+69G&gt;C</td>
<td>0.7±0.07</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C3</td>
<td>42</td>
<td></td>
<td>IVS8+23A&gt;C</td>
<td>0.73±0.07</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C6</td>
<td>34</td>
<td></td>
<td>IVS8-211C&gt;T</td>
<td>0.91±0.05</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C19</td>
<td>43</td>
<td>OFBCR/FDR</td>
<td>IVS7+69G&gt;C, IVS8-55 A&gt;G</td>
<td>1.68±0.1</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C8</td>
<td>43</td>
<td></td>
<td>IVS8+23A&gt;C</td>
<td>1.82±0.2</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C17</td>
<td>48</td>
<td></td>
<td>IVS8+23A&gt;C</td>
<td>1.92±0.2</td>
<td>CO-VAR</td>
</tr>
<tr>
<td>C11</td>
<td>50</td>
<td></td>
<td>IVS7+69G&gt;C, IVS8-55 A&gt;G</td>
<td>N/A</td>
<td>Removed</td>
</tr>
<tr>
<td>C20</td>
<td>46</td>
<td></td>
<td></td>
<td>N/A</td>
<td>Removed</td>
</tr>
<tr>
<td>C21</td>
<td>45</td>
<td></td>
<td>IVS8-170C&gt;T</td>
<td>N/A</td>
<td>Removed</td>
</tr>
<tr>
<td>C12</td>
<td>52</td>
<td></td>
<td></td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>C16</td>
<td>35</td>
<td></td>
<td>IVS8+23A&gt;C</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>P9</td>
<td>N/A</td>
<td></td>
<td></td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>
Real time PCR analysis showed that the SMAD4 variants predicted to create cryptic sites or abolish branch sites did not result in aberrant expression patterns, consistent with the negative RT-PCR results. However, the breast cancer group with variants (BC-VAR) but not the reference (BC-REF) exhibited significant up-regulation in expression relative to both control groups (CO-REF; p=0.036, CO-VAR; p=0.037) (Figure 13). The mRNA carrying variants among the BC-VAR group includes P2 (IVS12+41G>A) and P5 (IVS11+118A>G) which showed a 2-fold increase in expression. P9 harboring c.1478G>A (p.Asp450Asp) from a familial breast cancer was the sole variant with a bioinformatic prediction (i.e. abolish ESE motifs) and is associated with a level of high expression (>5 fold) that was not seen in any of the sample studied (Table 11). Therefore, this variant represents a candidate inherited SMAD4 mutation in breast cancer. Grouping the breast cancer cases and controls (BC vs. CO) did not reach statistical significance in expression (p=0.32).
Figure 13: SMAD4 expression analysis by Quantitative Real-Time PCR (qPCR): The mean expressions of four groups of mRNA are compared. Statistical significance determined by Mann-Whitney test of independence with error bars representing standard deviation (SD). The breast cancer cases harbouring variants (BC-VAR) are highly expressed relative to controls groups with (CO-VAR; p=0.038) and without (CO-REF; p=0.035) variants. P9, the familial breast cancer case harbouring the novel c.1478G>A alteration, contributed significantly to the observed aberrant expression. Inset: Grouped BC vs. CO not statistically significant (p=0.32) suggesting that germline variants are responsible for the observed aberrant expression. The upper and lower boundaries of the box indicate 75th and 25th percentiles, respectively. Line within the box, median; bars above and below the box, 90th and 10th percentiles, respectively.
Table 11: Quantitative RT-PCR analysis of SMAD4. The mRNA from the BCFR repository studied are broken down into four distinct groups: Breast cancer cases and controls without variants (BC-REF, CO-REF) and Breast Cancer cases and controls with variants (BC-VAR, CO-VAR). The normalized values from qPCR are listed with ±SD. The familial categorization is based on the OFBCR and/or FDR (First Degree Relative) which indicates the strength of familial association based on clinical characteristics and patient history. The synonymous c.1478G>A possesses a 5-fold expression increase and thus represents a potential breast cancer susceptibility alteration. C11, C20, and C21 were excluded due to mRNA degradation while C26 was excluded due to sample limitation.
4.8 Breast cancer mutations in SMAD3 and SMAD4 retrieved from the COSMIC database

We have surveyed the alterations of SMAD3 and SMAD4 in the Catalogue of Somatic Mutation in Cancer (COSMIC, http://www.sanger.ac.uk/genetics/CGP/cosmic/), a database representing an up-to-date listing of the somatic mutations curated from the literature (last accessed Mar 29, 2010). According to COSMIC no SMAD3 mutations were found in 48 breast tumour samples screened while two homozygous mutations were found in 38 samples screened (data not shown).

In SMAD4 four homozygous deletion mutations were found in 141 unique breast tumour samples (2.8%) screened while in contrast colorectal and pancreatic cancers were better studied with 92 mutations found in 858 samples (10.7%) and 123 mutations from 564 samples (21.8%), respectively (Figure 14, top). Based on the collection of mutations from all tumour types, the MH2 domain is shown to be a mutation hotspot (Figure 14, bottom) as most of the mutations (>60%) are clustered primarily in that region with fewer mutations found in the MH1 and the intervening linker region.
Figure 14: List of SMAD4 somatic mutations in breast cancer retrieved from the COSMIC database. Top: Mutations of SMAD4 in breast cancer is rare (n=6) compared to colorectal (n=104) and pancreatic cancers (n=137). Bottom: The MH2 domain in SMAD4 is a hotspot for somatic mutations in colorectal cancer and pancreatic cancer.
4.9 SMAD3 and SMAD4 expression in primary tumours compared to normal tissues

To gain a better understanding into how SMAD3 and SMAD4 expressions are altered in breast tumours and normal breast tissue, we have utilized the publicly available online tissue expression data (http://www.ebi.ac.uk/gxa/) [41] using a 2 independent samples t-test and Levin’s test for the equality of variance. The mRNA levels were significantly elevated in the tumour tissues compared to normal tissues for 4 of 5 probes (>5-fold average increase, p<0.05) of SMAD3 and one of 2 probes (>10-fold increase, p<0.01) for SMAD4 (Figure 15).

Figure 15: SMAD3 and SMAD4 expressions in tumour relative to normal tissues. Both SMAD3 and SMAD4 are over-expressed in Breast Carcinoma tissues compared to surrounding normal tissue. Significant expression in 4/5 SMAD3 probes and 1/2 SMAD4 probes. **, P<0.01, *, P<0.05
Chapter 5
Discussion
5.1 Novel germline variants detected in SMAD3 and SMAD4 in breast cancer

Breast cancer is among the most prevalent of human cancers. However, the involvement of the TGF-β signaling pathway in its initiation and progression is not as explicit as in pancreatic and colorectal cancers. Indeed, the roles of SMAD3 and SMAD4 differ quite sharply between cancers as well. While both respective loci are subject to loss of heterozygosity and allelic imbalance in breast cancer, germline mutation analyses for these two genes in breast cancer is an area that has not been examined. This may be due, in part, to the observation that SMAD3 is not known to be somatically mutated in almost every cancer while SMAD4 is highly mutated in colorectal and pancreatic tumours but does not appear to be the frequent target of somatic inactivation in other cancers including breast. This may, however, be due to a limitation in sample sizes screened from previous mutation analyses. Therefore, in this study we performed the first germline mutation analysis in 408 breast cancer cases and 710 population controls of the highly conserved MH2 domain in SMAD3 and SMAD4.

5.1.1 Distribution of variants in cases and controls

For the discovery of the variants the dHPLC methodology is robust and has sensitivity and accuracy of 92% to 100% for PCR fragments harboring genetic variations (point mutations, inter/intragenic deletions) 198 to 732bp, respectively [138]. The MH2 domain was targeted for studies because it has been shown to be a SMAD4 mutational hotspot in other cancer types [120] and was the region where the only mutation had been identified to date for SMAD3 [98] and the domain of interaction with BRCA1 [55]. While there may be additional germline alterations in regions not covered in our screen (i.e. MH1 domain, linker region) and that homozygous alterations would be missed by our method, the observation that 90% of all known somatic SMAD4 mutations reported in COSMIC are located in the MH2 domain suggest the number of undetected mutations is expected to be low when analysis is confined to this mutation hotspot but more importantly that a comprehensive screen of the MH2 domain represents the best method to detect novel SMAD3 and SMAD4 mutations.

Therefore we have confined our genetic analyses to the highly conserved Mad-Homology 2 (MH2) domains of both SMAD3 and SMAD4, where we have evaluated 408 cases including both sporadic and familial breast cancer cases, and 710 non-cancer controls obtained from the
population-based OFBCR. This is the first study of germline alterations of SMAD3 and SMAD4 screened using breast cancer and controls samples representing the general population, and thus suitable for addressing the question of interest.

In the present study, 27 distinct variants were identified in SMAD3 and SMAD4, 18 of which had not been previously reported. Approximately, 62.9% (17/27) were found only once in this study, whereas relatively more common variants varied between 1-2% in the population and are found in both cases and controls (Table 7). No coding variants were detected in the MH2 domain of SMAD3 while, in the MH2 domain of SMAD4, we report one previously reported germline alteration c.1214T>C (Phe362Phe) (rs1801250) and two novel coding variants c.1478G>A (Asp450Asp) (P9), and c.1701A>G (p.Ile525Val) (C24) only found once in the breast cancer case and control population, respectively.

5.1.2 Nucleotide diversity relative to expected values

Concurrent with our mutation analysis, we were interested in addressing whether the frequencies of coding and non-coding variants found in this study are more or less than expected. By normalizing the number of identified variants to the base pairs screened in the number of individuals assessed (θ) we compared our result to large scale studies that have established a frequency based on a large number of genes analyzed in the germline. The study by Cargill et al [126] represented the closest approximation as they screened genes from a large number of clinically relevant pathways including cardiovascular, endocrine and neurological. The studies performed by Halushka et al. [136] and Ten Ashbroek et al. [137] contained genes with specific criteria (i.e. blood pressure and essential genes, respectively). While these studies are performed on samples obtained from healthy individuals, for the purpose of this study the frequencies reported nonetheless provide an adequate baseline for which to quantitatively compare our mutation frequencies.

The frequency of alterations in both cases and controls in the coding region of SMAD3 was far less compared to all the three reference studies. This difference is not likely attributable to a discrepancy in sensitivity of detection of germline variants since we find comparable frequencies for non-coding variants for both SMAD3 and SMAD4 (Figure 8, bottom). The fact that nucleotide diversity in the coding region of SMAD3 is very low in both cases and controls
strongly supports the current notion that SMAD3 alteration is very infrequent and that it is under stringent selective pressure where deleterious mutations impeding proper function would also negatively influence tumourigenesis.

Within the coding region of SMAD4, on the other hand, nucleotide diversity estimations indicated that variants in cases and controls appear to occur at a similar, albeit slightly lower rate than the reference samples (Figure 8, top). This demonstrates that SMAD4 is not preferentially mutated in breast cancer but also suggests that, unlike previous observations where no alterations were detected [94], genetic alterations are expected to exist in the MH2 coding region and by increasing the sample size screened the detection rate can be improved.

The non-coding regions studied here have a higher θ compared to the reference studies (Figure 8, bottom). However, it should be noted that both Cargill et al and Halushka et al studies remarked that their non-coding regions are comprised of perigenic sequences (<18 bp from the exon) while our study spans up to 150bp of the intron and thus is more representative of a neutral rate of polymorphism. In fact the study by Cargill et al suggests that the θ for fourfold degenerate sites reported in their study had the highest nucleotide diversity (θ = 9.73 ± 2.46) may most approximate the neutral rate of polymorphism. If we assume this θ to be the neutral rate of polymorphism then what was observed in the non-coding regions of SMAD3 and SMAD4 cases (θ= 13.24 ± 7.02, 7.56 ± 3.36) and controls (θ= 11.24 ± 4, 11.71 ± 4.17) would be in agreement.

**5.2 In silico and in vitro characterization of variant’s impact on mRNA splicing**

The majority of the alterations (24/27) identified in this study were intronic. To date, germline SMAD3 or SMAD4 alterations have not been shown to result in aberrant splicing associated with breast cancer. However, it is recently appreciated that many intronic and synonymous alterations generally thought to be benign have been found to be associated with splicing defects causing cancer [139].
5.2.1 Creation of cryptic splice sites and/or abolition of branch sites

The majority of the variants found in this study were intragenic with the exception of 3 variants found in the exon region of SMAD4. Bioinformatic analysis indicates that the consensus donor and acceptor splice sites are not affected in the intragenic variants but leads to the creation of cryptic splice sites (Table 9). In splicing, cryptic sites are created if the consensus sites are mutated such that an alternate site becomes a stronger candidate for the binding of components of the spliceosome. Another way cryptic sites may arise is through mutations that increase the strength of pseudo-splice sites such that they could bind the spliceosome proteins with equal or greater strength, resulting in an aberrant mRNA transcript including exon skipping or intron inclusion.

Inactivating mutations in SMAD3 from aberrant splicing have not yet been found from screening of germline or tumour-derived DNA for a number of cancers. IVS9+132A>T was the only variant found in the breast cancer cases during our screening. However, analysis by ASSA does not appear to suggest a possible functional effect.

It is interesting to note that IVS8-55 A>G variants which were both found in the non-familial control population were both predicted to have an impact on the splicing apparatus. The branch point is an essential element in the formation of lariat during the excision of the intron and its abolition is typically associated with deregulation of the splicing process. However, in the IVS8-55A>G found only in control population, splice analyzer predicted the abolition of a branch point. Why such an apparent deleterious change occurs only in a variant found in the control population is not clear. However, this suggests the possibility that this may be a pseudo branch site and thus have no functional impact.

Several variants in the SMAD4 MH2 are also predicted to create cryptic splice sites including cryptic donor sites (IVS10+109 A>G, IVS12-52 A>T) and branch points (IVS11+126 delTATATTA, IVS12-33 T>A). For IVS11+126del6N and IVS12-52A>T, it is difficult to ascertain the functional significance of these predictions as both variants do not segregate with disease phenotype and the cryptic sites occur far from their expected locations. IVS10+109 A>G results in a cryptic donor site creation 108bp upstream of the consensus donor type. This variant is found in a familial case and a carrier with familial history but also more predominantly in
sporadic controls therefore the cryptic donor site created by this variant is likely not strong enough to influence the normal splicing machinery.

The novel IVS12-33T>A variant (P6) in SMAD4 was thought to be a candidate disease predisposition mutation as it is rare and segregates with only a familial breast cancer case (Table 7). It was expected to create a cryptic branch point 33 nucleotides upstream of the consensus acceptor site of exon 12. The endogenous branch point is normally the first adenine found 18-40 bp upstream of a splice acceptor site. The A in position 6 of the branch point consensus sequence YNYURAC (Y = C or T; R = G or A; N = any base) is required to form the lariat structure, a mammalian splicing intermediate. This type of alteration is very rare but has been reported in neurofibromatosis 2 (NF2), where a G to A transition resulted in the creation of a cryptic branch point 18bp upstream of the consensus acceptor site and predisposes to hereditary NF2 as it resulted in the creation of an alternatively spliced exon containing 106 bp from the intron [140].

5.2.2 Abolition of Exonic Enhancer Elements (ESE)

Coding variations including synonymous changes have recently gained attention as capable of being tumourigenic mutations. Nucleotide changes abolishing exonic enhancers and silencer motifs (ESE/ESS) are capable of interfering with the proper excision of introns by disrupting recognition by SR proteins important in the regulation and selection of splice sites [139] resulting most frequently in exon skipping.

The I525V missense variant is identified in this study has not been previously reported in SMAD4. However, due to its presence in only the control population and the likelihood of affecting protein function or ESE motifs is small especially given the negative predictions. Thus it not likely associated with breast cancer susceptibility. Both SMAD4 synonymous variants (c.1478G>A / Asp450Asp (P9) from a familial case and c.1214T>C (C12, rs1801250)) were predicted to gain or abolition of ESE motifs for several SR proteins including SC35, Srp55 and SF2/ASF. Such nucleotide changes affecting exonic enhancers and silencer motifs (ESE/ESS) interferes with the proper excision of introns via disrupting recognition by SR proteins that regulates and selects proper splice sites [139] resulting frequently in exon skipping.
5.2.3 Effect of variants on transcript splicing

Reverse-transcription PCR (RT-PCR analysis of the cDNA of all of the rare variants found in this study (regardless of being in a case or control) showed that the variants do not result in an aberrant transcript (Figure 10). This included the sole variant in SMAD3, IVS8-55 A>G (C19, C25), to have a bioinformatic impact was predicted to abolish a branch site. Since no aberrant splicing was detectable by RT-PCR this reinforces the observation that being found in a control suggests that this branch site is a pseudosite.

Furthermore, c.1478G>A (Asp450Asp) (C12) and c.1214T>C (Phe362Phe) (P9), both of which had strong familial association with breast cancer and were predicted to gain or abolish ESE motifs for several SR proteins including SC35, Srp55 and SF2/ASF did not promote aberrant splicing. We were also unable to observe a splicing defect of the transcript caused by the IVS12-33T>A variant (P6) suggesting that either the wildtype branch site is still predominantly selected for formation of the lariat or the mutant transcript could have been eliminated by non-sense mediated mRNA decay. Based on these results, it is reasonable to conclude therefore that overall SMAD3 and SMAD4 splicing defects do not appear to be a likely mechanism for breast cancer susceptibility.

5.3 Impact of variants on germline gene expressions

As allelic expression in red blood cells has been shown to be a viable surrogate for breast tissue for studies on predisposition to breast cancer [141], we have also performed quantitative real time PCR (qPCR) expression analysis in the same sample set. SMAD4, but not SMAD3, possesses variants predicted to create cryptic splice sites as well as abolition of ESE motifs. However, it appears that cryptic sites in general did not result in aberrant splicing or expression whereas a variant predicted to abolish ESEs altered expression but not splicing. Nonetheless, the breast cancer cases in both SMAD3 and SMAD4 had higher germline expression levels compared to the population controls. This has important implications on the role of the genes in breast cancer susceptibility. Interpreting how increased expression of SMAD3 and SMAD4 affect their activities in the cell facilitates distinguishing their roles as tumour suppressors or oncogenes in breast cancer.
5.3.1 SMAD3 germline variants on breast cancer expression levels

There is strong evidence for SMAD3 having tumour suppressor functions as its loss is associated with tumourigenesis in various cancers [49; 50; 51; 52]. However, in this study quantitative real time PCR analysis showed that SMAD3 mRNA from breast cancer cases was actually significantly highly expressed relative to both control groups (Figure 12, BC vs. CO; P<0.05, t-test). In addition, the variants found in the breast cancer cases do not appear to be responsible for the higher expression levels as the same variants possessed variable expression and/or are also found in the controls. Therefore, this observation is likely attributable to regulatory factors beyond the MH2 domain. These results together with the lack of inactivating mutations from this study suggest that SMAD3 is not functioning as a tumour suppressor in breast cancer.

5.3.2 SMAD4 germline variants on breast cancer expression levels

Loss of expression and allelic imbalance at the SMAD4 locus has been shown to promote carcinogenesis of ovarian, colorectal and gastric cancers [57; 111; 142]. Overall, we did not find SMAD4 cases to be differently expressed compared to controls. However, the germline variants found in breast cancer cases (BC-VAR) appear to contribute to a significantly elevated expression level (Figure 13).

Two intronic variants found in breast cancer cases in SMAD4 (P2, IVS12+41G>A; P5, IVS11+118A>G) showed a 2-fold increase in expression. Although these variants were not associated with a functional prediction, we postulate that they could influence expression without creating aberrant splicing patterns if cis-regulators of splicing known as intronic splicing enhancers and silencers (ISE/ISS) are affected [139]. These elements lie within 100nt of the consensus splice sites or deep in the intron and can act across a great distance and therefore are currently not feasible to accurately predict via bioinformatics. The observation of exonic variants such as c.1478G>A associated with altered expression levels strengthens the idea that variants affecting ISE/ISS may share a similar mechanism as their exonic counterparts.

Importantly, our qPCR data highlights a potential value for evaluating variants that affect ESE/ESS for abnormal expression even if they do not influence splicing. The novel c.1478G>A / Asp450Asp variant in exon 12 of SMAD4 found in a hereditary breast cancer case was predicted
to gain a binding site for SRp55 and abolish motifs for SF2/ASF and SC35. No aberrant splicing was detected but interestingly the breast cancer case (P9) harbouring this variant had a significant increase expression by almost 5-fold indicating that the full length transcript is preferentially produced. Therefore this variant represents a potential inherited mutation as it is functional and found in 0.24% of cases (1/408) matching the definition for a rare mutation.

5.4 Expression level differences between tumour and normal tissues

The abnormally high levels of germline expression as well as statistically significant overexpression of SMAD3 in invasive ductal carcinoma (IDC) compared to normal tissues in 4/5 (80%) of probes analyzed (Figure 15) raises the possibility that epistatic interactions of SMAD3 may contribute to the oncogenic activities of TGF-β. For example, SMAD3 has been shown to counteract BRCA1-dependent DNA repair in response to DNA-damaging agents and overexpression of SMAD3 decreases BRCA1-dependent cell survival [54]. Thus high levels of SMAD3 expression may mimic a BRCA1-deficient phenotype. Furthermore, the aberrant expression agrees with the allelic imbalance often associated with the 15q locus in breast cancer [53]. Future studies into mechanisms of expression regulation including mutation analysis in the promoter region and 5’UTR (untranslated region) and methylation analysis will be essential in addressing the mechanism of SMAD3 expression deregulation.

In SMAD4, recent reports showed that mRNA levels and protein expression are decreased in breast cancer relative to normal tissues [54, 55]. However, this loss is not significantly correlated with tumour size, metastases, nodal status, histological grade, histological type, or estrogen receptor expression. There was a trend toward longer survival times in patients with SMAD4 negative tumours [54]. A loss of expression is also correlated with decrease in axillary lymph node metastasis [55]. Therefore it is possible that any variants increasing SMAD4 expression would promote tumourigenesis. Indeed, our analysis of SMAD4 expression indicated that at least 1/2 (50%) of the probes analyzed are significantly higher (p<0.01) in invasive ductal carcinoma compared to normal tissues. Interestingly, this likely reflects a very important feature of the TGF-β signalling pathway where it functions as a tumour-suppressor in early cancer development but becomes progressively proto-oncogenic in later stages. The role of SMAD4 in breast cancer to bone metastasis has recently been documented in mouse xenograph models.
where genetic depletion of SMAD4 demonstrated that it is essential for the induction of IL-11, a key gene in the formation of osteolytic bone metastases [143]. In addition, another group has shown, using a xenograph mouse model with a conditional control of SMAD4 expression, that disruption of SMAD4 expression prior to the development of metastasis can significantly reduce the osteolytic bone metastasis potential of breast cancer [144]. Following the establishment of bone lesions, however, reduction of SMAD4 expression becomes less effective at influencing progression of metastasis. This showed that SMAD4 is important in the development of osteolytic bone metastases but not in maintenance or progression and agrees with the levels of over-expression observed in our analysis.

5.5 Somatic mutation of SMAD3 and SMAD4 in primary breast carcinoma

In this study we utilized the resources of the COSMIC database to summarize mutation analyses of SMAD3 and SMAD4 in terms of number of mutations found and sample size screened (Figure 14). In particular, we were interested in determining whether or not the paucity of mutations in SMAD3 was due to limitations in sample sizes. In the case of SMAD4, it was of interest whether or not the samples screened would account for the relative infrequency of mutations in breast compared to colorectal and pancreatic cancers.

COSMIC confirms the infrequency of the mutations in SMAD3 breast tumour samples, but we noted that the study sizes (n=48) are too small to provide conclusive evidence that SMAD3 mutations do not exist in breast cancer. The mutation analyses presented in this study provide some insight into this question as it would appear that SMAD3 is devoid of deleterious germline mutation (coding or non-coding) in breast cancer and is supported by the nucleotide diversity estimation. This is in support of the apparent lack of somatic mutations in SMAD3 and suggests that this is not likely due to a limited sample size screened. Therefore, further screening of the MH1 and linker region of SMAD3 is unlikely to uncover any variants of significant interest.

SMAD4 mutations were found in 2.8% (2 homozygous deletion mutations) of unique breast cancer tumour samples (n=141) and are relatively lower than 10.7% in colorectal and 21.8% in pancreatic cancers but data for these two cancers are based on far greater sample sizes screened (n=858, n=564, respectively). Therefore it is clear that from a mutation point of view SMAD4 appears to be better studied in pancreatic and colorectal cancers compared to breast cancer and
thus raises the possibility that there may be screening bias as opposed to SMAD4 not functioning as tumour suppressor genes involved in tumourigenesis.

Our germline results provide support in that increasing the number of samples screened would increase the power to detect novel coding variants as evidenced by the discovery of the novel missense variant c.1701A>G (Ile525Val) and nucleotide diversity estimations. Furthermore, all synonymous mutations should also be evaluated for any potential impact on splicing and expression alteration. Therefore our conclusions indicate that SMAD4 is a good candidate for whole gene germline and somatic mutation analysis with increased sample sizes covering the MH1, linker domains.

5.6 Summary

BRCA1 and BRCA2 are the most prominent examples of breast cancer susceptibility genes. However, there remains a great need to identify additional susceptibility genes as it has become increasingly evident that BRCA1/BRCA2 mutations cannot explain all cases of familial breast cancer. About 70% of breast cancers occurring among families with fewer than four or five cases and do not have ovarian or male breast cancer are attributed to genes other than BRCA1 and BRCA2 [8]. The functionally related genes CHEK2, ATM, NBS1, RAD50, BRIP1, and PALB2 of the DNA repair pathway harbours germline mutations that are rare and confer an intermediate risk (~2 fold) of breast cancer [14]. Including PTEN and TP53 it is estimated that mutations in these ten genes altogether account for only 50% of familial breast cancers [8; 14] suggesting that there exist additional undiscovered breast cancer susceptibility alleles.

Two candidate genes that are of interest in clinical genetics of breast cancer are SMAD3 and SMAD4, the key signaling transduction proteins of the Transforming Growth Factor-β (TGF-β) pathway. The loci on which they are found are frequently lost in breast cancer but their roles in breast cancer have not been well studied. Currently, screens for somatic mutations in these two genes have not yet been thoroughly investigated as the sample sizes remain limited and germline analysis has not yet been undertaken.

Therefore, we have performed the first case-control study in the MH2 domain of both genes. Given that this region is a mutation hotspot in pancreatic and colorectal cancers, we
hypothesized that a comprehensive screening of this region represents the best way to discover novel mutations.

The germline analysis and nucleotide diversity estimation support the paucity of somatic breast cancer mutations in SMAD3. Therefore, the rare mutation rate observed is not likely due to limitation in samples screened. In contrast, screening more tumour samples would increase the power to detect more variants in SMAD4 since it was shown to be mutated at a neutral rate but is unlikely to yield high penetrance point mutations that would inactivate both alleles. Therefore, it is likely that most mutations in SMAD4 would be rare and of low to medium penetrance.

The majority of the variants found in this study were intronic in nature. However, RT-PCR did not show that they influence splicing even though the splice analyzer program (ASSA) had predicted several variants to create cryptic splice sites and abolition of branch sites and exonic splicing enhancers (ESE). Together with the lack of inactivating germline nor somatic mutations, SMAD3 and SMAD4 are not functioning as tumour suppressors in breast cancer.

Importantly, this study has demonstrated that expression levels of SMAD3 and SMAD4 are important factors in breast cancer but have different consequences. For SMAD3, due to counteracting BRCA-1 mediated DNA repair, overexpression is likely to contribute to susceptibility by mimicking a BRCA-1 deficient phenotype. However, since this was not due to any of the variants found in the study, alternate mechanisms must be investigated. For SMAD4 a novel germline variant, c.1478G>A, contributed to abnormal overexpression. Increasing SMAD4 germline expression is unlikely to predispose to breast cancer due to its important role as a tumour suppressor, suggesting that SMAD4 is not involved in susceptibility. However, as tumourigenesis develops the cell becomes increasingly desensitized to the anti-proliferative effects of TGF-β but retains sensitivity for its oncogenic properties. Therefore, such germline variant could represent potentially valuable prognostic markers as SMAD4 is an important mediator in the development of osteolytic bone metastasis in late cancer stages. Therefore, future research to explore the mechanisms of disregulation of SMAD4 expressions and characterization of the c.1478G>A mutation effect on splicing efficiency, its frequency in breast cancer cases, will be essential in determining its value as a breast cancer prognostic marker.
5.7 Future direction

The findings of this study elucidated the presence of germline variants in the highly conserved Mad-Homology 2 (MH2) domain of SMAD3 and SMAD4. As well, this study evaluated whether or not germline variants are able to affect mRNA splicing as well as expression levels. However, it is still unclear whether or not germline variants exist in the conserved MH1 and linker domains of SMAD3 and SMAD4.

The results from this study suggest that deleterious mutations affecting SMAD3 function are not likely to be found since intact SMAD3 function appear to be pertinent to tumourigenesis. Therefore, it is unlikely that additional screening in the MH1 or linker region would yield novel mutations. However, real-time PCR and microarray data showed that SMAD3 was over-expressed in the germline and tumour breast cancer cases, respectively, relative to controls (or normal tissue). Interestingly, it has also been shown that knockdown of SMAD3 in MDA-MB-231 resulted in prolonged latency and delayed growth of bone metastasis, in contrast to SMAD2 knock-down which was more aggressive in phenotype [143]. Therefore, immunohistochemical studies measuring SMAD3 breast tumour expression relative to normal tissues will be necessary in validating the significance of this observation.

While the MH2 domain is a hotspot for SMAD4 mutations in many cancer types, mutations have been found in the MH1 and linker region, albeit not as frequently. Our case-control study was able to detect two novel coding alterations with at least one that had a potential functional impact (c.1478G>A). Along with nucleotide diversity analysis, low to medium penetrance mutations are likely to be found throughout the gene. A recent survey of the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP) indicated that missense mutations may be found in the linker region as evidenced by the recent addition of c.190C>G (rs61751988), c.229G>T (rs75667697) and c.294G>A (rs7238500). While the SMAD4 linker region is not as conserved as the MH1 or MH2 domains, it is nevertheless the site of important motifs such as the nuclear export signal (NES). Furthermore, alternatively spliced tumour-associated SMAD4 isoforms, lacking portions of the linker region have been detected in thyroid tumours [144] and it is possible that similar mechanisms could be at work in breast cancers. Future mutation analysis in the MH1 and linker
regions using DHPLC and direct sequencing in the OFBCR cases and population controls would likely yield additional coding and non-coding variants and expand upon the mutation spectrum of SMAD4 in breast cancer.

Homozygous deletions were one major class of alterations that this study was unable to capture due to the limitation of the DHPLC to detect only heterozygous variants. SMAD4 deletions have been observed in a small number of invasive ductal carcinomas (IDC) [63]. Interestingly, based on our preliminary germline mutation screening of 15 breast cancer cases, selected from the OFBCR population, using Multiplex Ligation-dependent Probe Amplification (MLPA) technology, we have shown that exon 1 and exon 2 corresponding to the SMAD4 MH1 domain was deleted in 1/15 (6.67%) of the cases analyzed. MLPA is a multiplex PCR method which is capable of detecting abnormal copy numbers of up to 50 different genomic DNA or RNA sequences [145]. These results suggest that homozygous deletion may be a potential mechanism of SMAD4 inactivation in breast cancer. At a rate of 6.67%, such homozygous deletions may represent highly penetrant mutations and thus requires further investigation in a pilot 408 breast cancer case population using MLPA.

One SMAD4 variant emerged from this study as being a potential breast cancer susceptibility allele. The c.1478G>A resulting in the synonymous substitution Asp450Asp was not predicted to have a deleterious impact on the domain structure. However, this variant was predicted to alter several exonic splicing enhancers (ESE), regulators that are critical to the efficiency of splicing. While no aberrant splicing was associated with this variant, real time PCR showed that SMAD4 was very highly expressed relative to other breast cancer cases as well as the population controls.

This alteration occurs at a rate of 0.25% (1/408) but as it is a synonymous variant and that the DHPLC has a ~1% failure rate, additional genotyping studies with a larger sample size is required to determine its association with breast cancer. Future studies utilizing the Taqman genotyping technology with expanded breast cancer cases (n=1228) versus controls (n=710) will provide conclusive evidence that c.1478G>A is a bona fide breast cancer susceptibility mutation.

In addition, c.1478G>A as well as several intronic variants were associated with increased SMAD4 expression. C.1478G>A was predicted to affect exonic splicing enhancer (ESE), a class of exonic splicing regulators important in splicing efficiency by promoting exon usage. Similarly,
we postulated that intronic variants which showed increased expression may be affecting intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs) within the upstream and downstream introns which can also influence use of associated splice sites. To determine whether or not these allelic variants have an effect on splicing efficiency, future studies making use of the minigene constructs [146] containing the cis-element of interest will be necessary to determine whether or not these sites represent exonic and intronic elements that enhance splicing. Minigene constructs are an important tool for the identification and in vivo analysis of such cis-acting regulatory elements and trans-acting factors that are pertinent in splicing efficiency. By transiently transfecting the construct into a cells with high transfection efficiency (e.g. COS, HeLa), the expression of the minigene pre-mRNA is a convenient assay in assessing loss-of-function or gain-of-function analyses for cis-elements and trans-acting factors affecting regulation of splicing. In this way, each variant can be assessed based whether or not they are affecting an exonic or intronic element that enhance splicing as well as whether or not they have an effect on splicing efficiency.
CHAPTER 6 REFERENCES


