2010

Molecular Characterization of Ductal Carcinoma In Situ: Pilot Studies

Neil Bipinchandra Desai
Yale University

Molecular Characterization of Progression in Ductal Carcinoma *In Situ*: Pilot Studies

A Thesis Submitted to the

Yale University School of Medicine

in Partial Fulfillment of the Requirements of the

Degree of Doctor of Medicine

by

Neil Desai

2010

Abstract

MOLECULAR CHARACTERIZATION OF PROGRESSION IN DUCTAL CARCINOMA IN SITU: PILOT STUDIES

Neil B. Desai,[1] Arun Gopinath,[1] Sebastian Szpakowski,[1] Rajini Haraksingh,[3] Veerle Bossuyt,[1] Michael O. Krauthammer,[1] David Tuck,[1] Donald R. Lannin,[2] and David F. Stern.[1] [1]Department of Pathology, Yale University, School of Medicine, New Haven, CT. [2]Department of Surgery, Yale University, School of Medicine. [3]Department of Genetics, Stanford University, Palo Alto, CA.

Ductal carcinoma *in situ* (DCIS) is thought directly to precede invasive breast cancer (IBC). Screening mammography has driven the incidence of this key precursor lesion to >65,000 cases per year. However, little is known about the factors controlling the natural history or risk for recurrence following treatment of a particular patient's DCIS. Though the heterogeneity of the disease is well established, no histologic or demographic criteria have been able to stratify DCIS for treatment. We hypothesize that at initial diagnosis there exist biologically distinct subsets of DCIS with associated prognoses that may be recognized by molecular markers. Molecular approaches have been limited by technical design issues related to the types of tissue available for analysis, namely degraded formalin-fixed paraffin embedded (FFPE) specimens and small core biopsy samples. However, new technologies promise to overcome these issues. In the first phase of our investigation, we aimed a) to pilot feasibility studies on the use of FFPE DCIS for molecular analyses including gene expression microarray and b) to pilot feasibility study of selective, high throughput sequencing through the use of "exon capture" on small input material that simulated expected DCIS core biopsy amounts. The results of this work offer specific technical guidelines for the molecular study of DCIS.

Moreover, they have enabled the initiation of the second phase of this study, which aims to assess molecular profiles of DCIS recurrence and progression.

Acknowledgements

This work was funded by a grant from the Susan G. Komen Foundation and the thesis candidate's research fellowship from the Howard Hughes Medical Institute. It was performed with the mentorship and guidance of the candidate's principal investigator David F. Stern, Ph.D and the members of his laboratory, including in no order: Maureen Gilmore-Hebert Ph.D, Alexandra Teixeira, Christina Zito Ph.D, Kathleen Wilson, Jerrica Breindel, Kathryn Tworkoski, Kerry Williams, and Rahul Dalal.

Notable collaborators involved in the project were David Rimm M.D. and his laboratory including Arun Gopinath M.D. and Seema Agarwal Ph.D, Fatteneh Tavassoli M.D., Veerle Bossuyt M.D., Lyndsay Harris M.D. and her laboratory including Kyle Halligan and Kimberly Lezon-Geyda Ph.D., Donald Lannin M.D., Director of the Breast Cancer Center at Yale, Ruth Halaban Ph.D, Mario Sznol M.D., Michael Krauthammer M.D., Ph.D and his lab including Sebastian Szpakowski, James B. Hicks Ph.D and his laboratory at Cold Spring Harbor , and David Tuck M.D and his laboratory including Vince Schulz. Other collaborators were Michael Snyder Ph.D and his lab now at Stanford University including Rajini Haraksingh.

This work is also thankful for the support of the Office of Student Research at Yale University School of Medicine, including Dr. John Forrest, Donna Carranzo, and Mae Geter.

Finally, this thesis candidate's work on this study and otherwise is as always indebted to those relatives, especially parents, and friends who have supported and enabled his goals.

Table of Contents

Introduction

Contextual Overview

The study of tumor development, tumorigenesis, offers to elucidate the basic biologic mechanisms controlling the clinical characteristics of neoplasms and their precursors. The importance of this work becomes clear when considering the prognostic tools and therapeutic advances made possible by knowledge of the development of invasive disease: Histopathologic classification of tumor grades informs staging and underwrites prognosis and treatment discussions. Efforts to define imaging characteristics of developing lesions allow detection of neoplastic processes when they are often more effectively managed. A new generation of biologic and immunologic agents promises a future of personalized treatments largely based on better understanding of the expansion of tumor lineages and the epigenetic changes that drive them.

Many of these insights draw from efforts at modeling the order and behavior of a spectrum of lesions that populate a particular oncologic process. A prominent example is adenocarcinoma of the colon wherein a series of precursor lesions have been paired with characteristic molecular and clinical changes. This model and its understanding that select pathways of tumorigenesis showed reliance on vascular endothelial growth factor (VEGF) receptor signaling led to targeted treatment selection with anti-VEGF antibodies such as bevacizumab.

Similar modeling has proven particularly useful in the study of breast cancer, a heterogeneous group of diseases with multiple lineages and a large spectrum of pre-invasive lesions. Histopathologic characterization and empirical clinical investigation have been supplemented by molecular approaches in efforts to appropriately tailor studies

and treatments to each subtype. The earliest 'targeted' therapies in oncology capitalized on these efforts, including the selection of estrogen receptor (ER) positive patients for the adjuvant antagonistic hormonal therapy tamoxifen and trastuzumab usage in HER2/neu-overexpressing tumors.[1] Continuing efforts to extend this strategy focus on genetic alterations in key genes, global expression profile patterns, and gene copy number.[2] Despite the increased incidence of overall disease, a combination of screening mammography and treatment advances has contributed to a decline in mortality in the Western World.[3]

At the same time, the very success of screening has introduced uncertainty regarding treatment selection for patients who increasingly present with pre-invasive disease. As attention has turned to these lesions, there has been renewed focus on the crucial stages of tumorigenesis that dictate whether a lesion will become invasive and require treatment or not. In this context, Ductal Carcinoma *In Situ* (DCIS), a pre-invasive breast lesion, has figured prominently. The importance of DCIS, its role in breast tumorigenesis, and the gaps in knowledge surrounding its management will be reviewed here as a rationale for the study conducted. Furthermore, the technical advances that allow such investigation will also be described.

Biology of Ductal Carcinoma In Situ

DCIS is defined as a clonal proliferation of malignant appearing cells confined to the lumen of a mammary duct without evidence of penetration through the epithelial basement membrane.[4,5] DCIS is considered an immediate precursor lesion to invasive breast cancer (IBC)[6] with some suggesting it specifically gives rise to Invasive Ductal Carcinoma (IDC), which accounts for 76% of invasive breast cancers annually in the

United States.[7] Traditional models of tumorigenesis in ductal breast tumors have described a linear progression of lesions beginning with benign proliferative changes and culminating in DCIS that corresponds to an accumulation of genetic alterations (**Figure 1**).[8] Despite a large body of work on invasive breast disease, the initiation and culmination of the tumorigenesis process remains poorly defined. Nonetheless, multiple lines of investigation strongly support DCIS as a fundamental evolution of tumor biology in this progression due to a) its similarity to invasive disease and b) its distinction from less advanced precursor lesions.

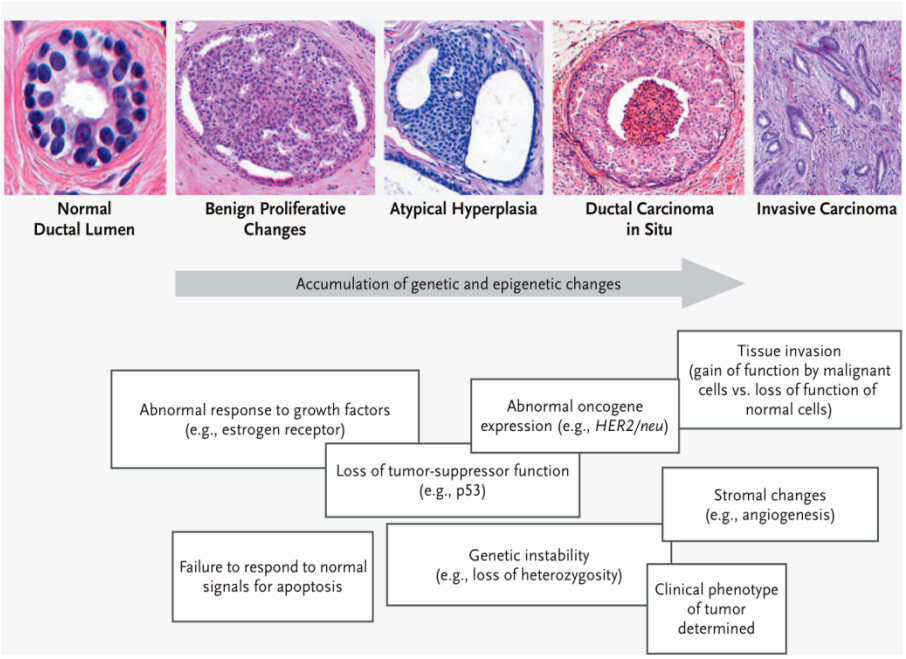


Figure 1. Histological development of ductal carcinoma *in situ*. The transformation of breast tissue is accompanied by an accumulation of molecular, cellular, and pathological changes. The majority of these appear to have occurred by the time ductal carcinoma in situ has arisen. Figure taken from Burstein *et al*..

DCIS Closely Resembles Invasive Disease on a Molecular Level

Most invasive ductal breast disease is thought to arise from DCIS in a clonal, evolutionary manner. Evidence for this relationship has been drawn from the conservation of mutations and chromosomal changes in this progression. Loss of heterozygosity (LOH) analysis first established this comparison on the basis of shared allelic imbalances between DCIS and synchronous, adjacent invasive disease. [9,10] The

development of comparative genomic hybridization (CGH) allowed investigators subsequently to deepen this evidence with whole genome DNA copy number data.[11-13] The high incidence of mutations such as in *p53*,[14] over-expression of the proto-oncogene *HER2/neu*,[15] and loss of estrogen receptor (ER) expression[4] in DCIS similarly mirrored the findings in associated invasive disease. Indeed, no mutations unique to invasive disease have been identified yet on comparison to DCIS. Gene expression profiling likewise has demonstrated highly similar patterns in coincident invasive disease and DCIS.[16-20]

The heterogeneity of DCIS also parallels invasive disease. Histological study has differentiated DCIS into subtypes seen in invasive disease based on similar markers, such as hormone receptors and cytoskeletal proteins.[21] On a molecular level, low- and high-grade DCIS lesions have been correlated with distinguishing genetic alterations much akin to low- and high-grade invasive disease. While 75% of high grade lesions lose ER expression and two-thirds harbor *p53* mutations or HER2/neu over-expression, 90% of low-grade lesions preserve ER expression while less than 20% over-express HER2/neu or contain *p53* mutations.[22] Low-grade lesions further are characterized frequently by chromosomal loss at 16q and gain at 1q in contrast with local amplifications in high-grade lesions at 11q13 (*CCND1*) and 17q12 (*ERBB2*).[12,23] Moreover, it seems these different populations of DCIS represent independent pathways of genetic evolution to IDC. Specifically, low-grade lesions are apparent direct precursors to IDC without requirement for evolution to high-grade DCIS first. Preservation of genetic changes and histological observation of synchronous, adjacent lesions argue that low-grade DCIS give rise to more differentiated IDC, whereas high-grade DCIS often gives rise to grade III

IDC.[23-25] The heterogeneity of DCIS thus suggests it should be analyzed and managed clinically with greater resolution than currently is done. Nonetheless, all of these DCIS lesions seem to represent a stage of tumorigenesis at which most of the molecular changes that define invasive disease are already present.[4]

DCIS is Molecularly Distinct from Less Advanced Precursors

In contrast, the development of DCIS seems to be marked by distinctive clinical, histopathologic, and biologic features when compared with other presumptive precursor lesions such as Atypical Ductal Hyperplasia (ADH). CGH analyses of such lesions, including ADH,[26-29] demonstrated that the copy number imbalance profiles of these lesions differ from invasive disease or DCIS. DCIS seemed to harbor more widespread changes.[11,12] Further, work showing a stepwise accumulation of global LOH from 0% in normal breast tissue to 35-40% in ADH to >70% in DCIS suggested a linear order to these lesions culminating in DCIS.[13,30-32] Key gene alterations common to invasive disease and DCIS, such as in *p53* tumor suppressor mutation or *HER2/neu* proto-oncogene over-expression, are rarely observed in ADH or other early proliferative lesions.[14,15] Finally, the most dramatic gene expression pattern changes of tumorigenesis seemed to occur during the transition from normal tissue to DCIS.[16-18]

Thus, most evidence suggests that DCIS evolves from other precursor lesions, namely ADH, and has accumulated most of the molecular changes of invasive disease to which it gives rise. In this sense, DCIS has been studied thoroughly for its significance in breast cancer tumorigenesis. At the same time, DCIS is a heterogeneous disease with distinct populations likely undergoing transitions to invasive disease with differing

latency and frequency. The factors affecting this latter transition remain poorly defined and constitute the most important information about its clinical management.

Models of Progression of DCIS to Invasive Disease

While past work on DCIS has been unable to consistently differentiate DCIS from paired IBC, it is strongly suspected that a final series of events in the lesion drives the transition to invasiveness. The traditional progression hypothesis holds that the epithelial cells making up the lesion evolve based on such events, which are subtle and/or difficult to detect by previous study designs. However, notable inconsistencies in this 'linear' model have led to alternative theories that will be mentioned briefly here. These focus on a) the tumor microenvironment and b) stem cell populations.

Microenvironment Theory

The most obvious challenge to the traditional, 'linear' model of DCIS evolution is the lack of evidence thus far for genetic change in the epithelial cells of the lesion during progression. This problem is often attributed to the sub-optimal study design in the past caused by limitations on technical ability and access to 'pure' DCIS tissue. As an alternative, interesting new evidence suggests a role of tumor microenvironment in DCIS transition to IBC. Changes in DCIS are specifically documented in myoepithelial cells (MECs) and stromal cells, including fibroblasts and myofibroblasts. MECs are important in normal breast duct development and physiology, providing natural tumor suppressor functions.[33,34] In contrast to 'normal' MECs, DCIS-associated MECs show down-regulation of genes involved in normal duct function, while showing up-regulation of transcripts that support epithelial cell proliferation, migration, invasion, and stromal angiogenesis.[35-37] Further, DCIS-associated MECs show distinct epigenetic[38] and

immunophenotypic changes.[39] Similarly, stroma associated with DCIS has been associated with changes seen in invasive disease.[40-42] However, little work has been done on following the microenvironment of DCIS as progression occurs, limiting interpretation to correlation rather than causation.

A Common Stem Cell Progenitor for DCIS and IBC?

Another alternative hypothesis for explaining DCIS to IBC transition is borne out of the contradiction between the predictions of the 'linear' model and the results of mathematical modeling. One notable study assessed four different models of progression for fit of clinical observations to expected frequencies of co-occurrence of DCIS and IDC of different grades.[43] The 'linear' model, as well as the offshoot 'branched' and 'non-linear' models, all assumed IBC arose from DCIS as traditionally thought. In contrast, a fourth 'parallel' model described DCIS and IDC diverging from a common progenitor cell and developing through different grades together (**Figure 2**). The authors found most

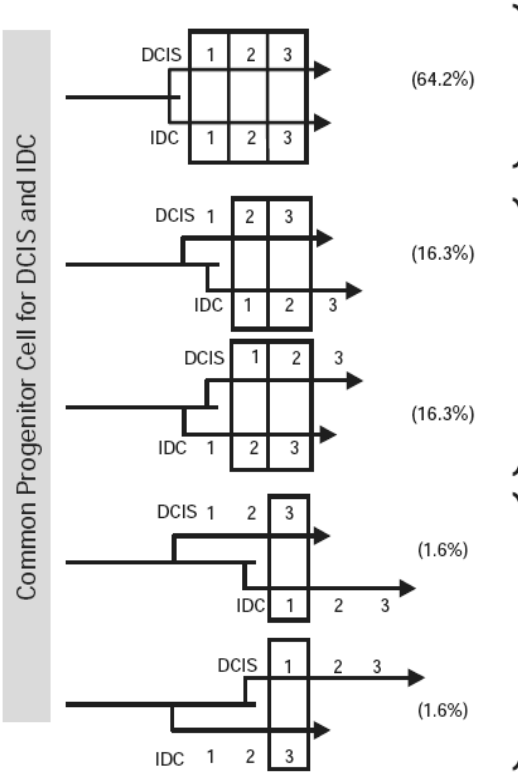


Fig 2. The Sontag-Axelrod 'parallel' model of progression of ductal carcinoma *in situ* (DCIS) and invasive ductal carcinoma (IDC). This model holds that the two diseases diverge from a common progenitor cell and progress through different grades in parallel. The figure shows pairs of lineages of progression for DCIS and IDC through their respective histologic grades. The order in which DCIS and IDC begin their progression varies among the 5 listed sub-groups with their calculated percentages noted. The basis of this model overall lies in matching of mathematic modeling to clinically observed frequencies of grades of DCIS and IDC in co-existing disease. The projections of the parallel model outlined here matched most strongly with the data according to the authors. Figure adapted from Kuerer *et al*..

robust performance in the 'parallel' model. The results suggest the possibility that DCIS and IDC sprout from a common progenitor and that the critical molecular events driving clinical behavior of disease take place before DCIS morphology is manifest.[6] Moreover, this theory would provocatively imply low utility for the standard of care treatment of DCIS to prevent IDC. [6]

However, like the traditional linear model and microenvironment theory, the 'parallel' model relies on key missing data. In the linear model, this would be some genetic alteration associated with DCIS progression to IBC. For the microenvironment theory, a similar need for distinguishing factors in the microenvironment of DCIS that progresses to IBC has yet to be established. In the 'parallel' progression model, progenitor stem-like cells have been tentatively explored but none have been consistently validated. Thus, no biologic model yet exists to illuminate the mechanisms of progression in DCIS.

Clinical Features and Treatment Issues

The lack of biologic characterization of the ductal carcinoma *in situ* (DCIS) to invasive ductal carcinoma (IDC) transition underlies a lack of a precise clinical characterization. This gap in knowledge has inhibited refinement of disease management in a heterogeneous disease that continues to be treated with relatively homogenous measures. Nonetheless, in response to the large epidemiologic impact of the disease today, robust empirical clinical investigation has produced diagnostic and management guidelines with positive effect on mortality and morbidity. Current clinical management and the key prognostic dilemma of DCIS to IBC prediction will be reviewed here as background for the clinical significance of this study.

Diagnosis

Diagnosis of DCIS historically relied on gross palpation of a mass or note of secondary effects such as nipple discharge or Paget's disease of the nipple.[4] Several advances in diagnosis in imaging, biopsy technique, and histologic assessment have enhanced sensitivity for early detection, though not without their own limitations.

Most notably, the introduction of widespread mammography in the 1980's has resulted in a >10x increase in the diagnosis of the disease in the U.S..[44] Whereas only 4800 cases were diagnosed in 1983, roughly 64,000 cases are diagnosed annually today.[45] DCIS now represents 15-25% of newly diagnosed breast carcinomas,[44,45] increased from 1-2% pre-mammography.[46] In terms of sensitivity, 90% of DCIS cases diagnosed as suspicious calcifications on mammography.[47] The specificity of this test is limited regarding the extent of involvement in multifocal disease[48] and in those lesions without significant necrosis or calcifications.

Stereotactic core needle biopsy is the tissue sampling modality of choice for suspicious lesions. Direct biopsy has drawbacks, which are compensated for in management guidelines. Namely, core needle diagnosis of atypical ductal hyperplasia often prompts wider, surgical excision for more extensive pathological assessment due to a 10-50% risk of concurrent invasive or *in situ* disease.[49,50] Biopsy diagnosis of DCIS likewise shows a 10-15% incidence of concurrent invasive disease[50,51] that increases with histologic grade. This further is one of the justifications for full resection.

Following biopsy, histologic analysis primarily assesses for any invasive disease and aids in specification of the few characteristics that are known to affect DCIS

behavior. Close or involved margin status, presence of comedonecrosis, and high-grade are most consistently linked with increased risk of recurrence after resection.[4,6]

Treatment

Once the diagnosis of pure DCIS is made, the goal of treatment is to remove the neoplastic tissue and prevent recurrence. A series of large studies has produced robust data on the dominant treatment strategies that include excision +/- radiation.[6,52] Currently, the standard of therapy is breast conserving surgery (BCS) followed by irradiation. However, significant variation in treatment still exists with a mix of therapies used including mastectomy with or without reconstruction, breast conserving surgery (BCS), and BCS with adjuvant radiotherapy[53] with 10-year local recurrence rates of 1%, 30%, and 10%, respectively.[52]

Historically, mastectomy was first employed with near complete cure rates. The finding that BCS offered identical survival to modified radical mastectomy in invasive breast cancer[54] prompted concern that mastectomy may be overtreatment for some DCIS patients, particularly in that increasing cohort of cases with small lesions detected by mammography. While no direct randomized controlled comparison of mastectomy and BCS for DCIS exists, indirect, retrospective data from surgical trials[55] and treatment registries[56-58] indicate that mastectomy provides a locoregional recurrence benefit but no improved overall survival or rare disease-specific mortality,[56,59] BCS is the most common surgical component of therapy today.

Adjuvant whole-breast radiotherapy is often added to address the increased locoregional recurrence risk in BCS. This combination therapy has been evaluated in 3 large randomized controlled trials in both the U.S. through the National Surgical

Adjuvant Breast and Bowel Project (NSABP) and Europe through the European Organization for Research and Treatment of Cancer (EORTC) and U.K. Coordinating Committee on Cancer Research (UKCC). NSABP B-17 showed a 58% reduction of locoregional recurrence at 12 years in the adjuvant radiotherapy group vs. BCS alone.[60] However, overall survival was not significantly affected. The EORTC and UKCC trials showed an identical pattern of benefit.[61,62]

Irradiation comes with its own morbidity as well as time commitment. Moreover, in some subgroups of patients with DCIS, such as those with no histological high-risk factors for recurrence or of very old age, the absolute locoregional recurrence risk reduction seen in the adjuvant radiation trials was small.[63] In response, some investigators have attempted predict which groups of patients may omit radiation. Notably, the Van Nuys Prognostic Index (VNPI) used a system of risk classification based on grade, width of margin, and size of lesion in which the lowest risk category would receive excision alone.[64] However, neither this index nor any other predictive scoring system based on histology has successfully validated on prospective study.[65-68] The reasons for this failure include the initial 'validation' of the index on retrospective sets, the use of samples across a long period of time that saw an evolution in treatment and diagnostic modalities, and poorly reproducible classification methodology.[69] More fundamentally, the predictive index was unable to address the lack of understanding about the biology of the transition to invasiveness that would allow a more rigorous method for classifying heterogeneous DCIS. Thus, adjuvant radiation remains standard of care following BCS.

Hormonal therapy has also been evaluated following successes in the treatment of invasive disease. 5-year adjuvant tamoxifen in patients treated with BCS and irradiation

was studied in NSABP B-24[60,70] and the UKCC trial[61] with conflicting results. While initial analyses on this discrepancy suggested benefit was isolated to ER+ groups,[71,72] others focused on younger patients (<50 years old).[6] In either case, absolute risk reduction is considered marginal, and the side effects of increased thromboembolic events, endometrial cancer risk, and menopausal symptoms are particularly undesirable in the target group of young women.[73] Like with adjuvant irradiation, development of more specific treatment stratification strategies for hormonal therapy is limited by a lack of biological discrimination of DCIS.

Recurrence

Locoregional recurrence refers to ipsilateral breast lesions found subsequent to treatment of an initial malignant breast lesion. Half manifest as invasive disease in DCIS.[74] The similarity of pathologic[75] and genetic features[76,77] to the index lesion suggest that recurrences most often arise from residual microscopic disease. Clearly, even with various treatment options available, some patients with biologically more aggressive lesions are being under-treated. The result is that these patients often must undergo salvage mastectomy with sentinel lymph node biopsy.[78]

Prognostic Challenges

Thus, despite improved outcomes and reduced morbidity from empirical clinical investigation, it is clear that the treatment decision process at initial diagnosis of DCIS remains muddled by a lack of prognostic ability. Multiple histologic and demographic risk factors for recurrence have been described. Yet, the most publicized approach to tailoring treatment to a predictive index based on such risk factors, the VNPI, has failed repeatedly as described. No criteria exist to answer the fundamental question facing

DCIS management: which of these cases would progress to invasive disease without treatment and thus truly require treatment?

Study of the natural history of DCIS in an attempt to answer these questions is difficult. Standard management calls for full excision of the lesion. Even in those patients opting to defer care, the biopsy diagnosing the lesion often also removes it. The most useful data on rates of progression come from lesions that were initially misdiagnosed as ADH. The most thorough review suggests that these lesions progress to IDC at a rate of 10-53% over a period of 10 or more years.[79] This is a wide estimate that is plagued by the skew towards low-grade lesions that are more often mistaken for ADH.

Without better characterization of sub-group specific prognosis within DCIS, a homogeneous treatment approach will remain. The lesions treated today are fundamentally different in size and possibly biology from those initially used pre-mammography to justify treatment, On the other hand, even with standard of care, a small fraction of cases will recur systemically and cause death. Thus, some populations likely receive over-treatment and others under-treatment.

Molecular Approaches to Prognostic Prediction

With histologic and demographic risk factors failing to translate into clinical predictive ability, attention has returned towards developing a better biologic understanding of the transition to invasiveness in DCIS. Ideally, this would involve the study of DCIS lesions as they evolve with comparisons before and after the onset of invasiveness. The major molecular investigations in this area have been hampered by technical limitations of design with both retrospective and prospective studies. Recent

improvements in technology have introduced possible solutions for these issues and serve as the methodology background for the study.

Past molecular work on DCIS evolution has often focused on snapshot comparisons of DCIS of various grades and associations. For instance, Adeyinka *et al*.[19] supplemented earlier CGH work by demonstrating different grades and morphologies of DCIS could be discerned by gene expression analysis. Scheutz *et al*.[80] and Ma *et al*.[18] conducted patient matched comparisons of synchronous DCIS and IBC. These studies have advanced the use of limited RNA in microarrays and solidified laser microdissection (LCM) as a method of purifying neoplastic epithelial cells in DCIS from contaminating stroma. However, they analyze IBC associated DCIS which are not representative of the 'pure' DCIS that is the focus of the prognostic dilemmas described. None of these studies have made the critical comparisons of a given DCIS specimen before and after invasiveness arises and to DCIS that did not give rise to invasive disease. This line of investigation alone can properly assess the factors determining why some lesions progress and others remain stable as *in situ* lesions for long periods of time.

Prospective Methodology Limitations and Advances

Prospective studies are the most obvious setting for such a design. However, it is difficult to access the necessary DCIS tissue at initial resection because standard of care requires the entire specimen to be examined for evidence of invasive disease and margin status, both of which are critical to treatment decisions. This inhibits the ability to make comparisons to subsequent recurrences. However, the increased sensitivity of molecular analyses, especially gene expression arrays, has offered the potential of banking mRNA or expression data from core biopsy samples. Indeed, this has been accomplished peri-

operatively in invasive disease albeit on more voluminous samples than DCIS usually offers.[81] Nonetheless, given the estimated median latency of DCIS to IDC progression near a decade in length, comparative investigations of patient-matched samples will require much time for completion. Thus, they are usually best used as validation sets for candidate markers developed from retrospective data, as suggested by the experience with the failed Van Nuys Prognostic Index, or as an adjunct to a retrospective study.

Retrospective Methodology Limitations and Advances

Retrospective work has the benefit of clinical follow up data on recurrence and outcome matched to the samples to be analyzed. Further, unlike prospective trials, they can be done relatively quickly if existing tissue and data banks are identified. Molecular analysis of these samples is challenging given the alterations to nucleic acids and proteins in the storage process, which usually produces with a formalin-fixed, paraffin embedded (FFPE) sample.

Proteins become cross-linked, allowing only *in situ* methods based on immunohistochemistry or immunoflourescence. These methods are limited by inability to quantify protein levels, high requirement of rare tissue in the case of archival DCIS, and high labor requirement when scaled to large studies. However, tissue microarrays (TMAs) have been employed with increasing frequency to allow mass examination of samples across multiple markers with small amounts of tissue. The development of automated quantitative analysis (AQUA™) by Dr. David Rimm at Yale University School of Medicine[82] has expanded on this technology to allow quantification of protein levels and statistical clustering of samples by these results.

DNA and RNA are both fragmented and chemically modified by the heating and treatment of the fixation and embedding process in archival specimens. FFPE RNA is especially vulnerable due to endogenous ribonucleases and was long considered unusable. Like with protein analysis, RNA harvesting and analysis has evolved, and FFPE tissue has been used for expression analysis by both real time polymerase chain reaction (RT-PCR) and microarray.[83-85] The factors most affecting successful use seem to be time to fixation, time of storage, amount of tissue used, and extraction technique.[86] Even with these optimized, past microarray studies with FFPE samples in DCIS have been forced to rely on amplification of RNA, specifically via T7-primers, to generate requisite template amounts. This method results in attachment of a T7 sequence to amplified RNA (aRNA). Unfortunately, these sequences on the aRNAs have the potential to subsequently hybridize to complementary motifs on microarray probes, resulting in non-specific signal. This has been seen to occur in up to 1-9% of probes on some arrays.[87] New microarray technology has attempted address these and other problems in using FFPE RNA. In our study we will use one such approach developed by Illumina, the cDNA mediated Annealing, Selection, extension and Ligation (DASL) array. DASL uses random priming to form cDNA followed by oligo hybridization and advertises higher sensitivity for low RNA input. It thus avoids both amplification bias and dependence on an intact poly-A tail for oligo-d(T) priming. Further, DASL probes recognize small (<50 base pair) regions within genes and produce uniform 100 base pair amplicons, which should reduce bias against transcripts more prone to degradation. With these and other features, the use of unamplified RNA in global expression analysis is now potentially feasible.[83]

Though DNA is sturdier, it too requires special adjustments for use in larger scale studies such as comparative genomic hybridization. The high quality DNA required of current CGH analyses in particular have limited past work on allelic imbalances to a combination of FISH and select LOH analysis, which lacks high resolution and scalability.[86] Our collaborator Dr. Jim Hicks at Cold Spring Harbor is pioneering techniques for the use of FFPE DNA in CGH.

Mass Mutation Screening

Notable advances have been made in the field of mutation re-sequencing which are especially valuable in setting of the challenges facing DCIS work. As a neoplastic process, DCIS evolution is driven by inactivating mutations in tumor suppressors and activating mutations in proto-oncogenes, which select for growth advantage.

Recently, landmark studies using unbiased sequencing of breast and colon cancer genomes revealed a number of novel candidate "cancer genes," most of which had no previous connection to malignant diseases.[88,89] These studies have the potential to reveal new therapeutic targets and prognostic markers. Moreover, they have been followed up with attempts to replicate the scale of sequencing using new high-throughput technologies that promise cost-effective, personalized mutation sequencing.[90,91] Still, available sequencing platforms are unable at present to provide coverage of all known genes. Development of an interim strategy focuses on directing the high-throughput platform sequencing to high priority targets, namely gene exons. In this method, a high-density microarray is used to hybridize and thus '*capture'* coding exons of interest for sequencing. This has described with some success in experimental runs on breast and colorectal carcinomas.[92]

The study of DCIS transition is particularly suited to high-throughput mutation re-sequencing analyses since a) little tissue and thus DNA is available for analysis, b) though a genetic alteration almost certainly drives the acquirement of invasiveness in DCIS, no known causative mutations have yet been consistently identified, and c) extension of large scale mutation profiling to a common lesion like DCIS needs to be done in a cost conscious manner. Only prospective, fresh tissue samples can be utilized due to technical requirements. Even with prospective specimens, core biopsies often are all that can be sacrificed from diagnostic pathology on the specimen. In this setting, new methods of growing primary lines briefly before molecular analyses have been introduced, such as the mammosphere.[93,94]

Summary of Rationale

Widespread screening mammography has led both to earlier detection of breast cancer and increased diagnosis of precursor lesions to invasive disease. DCIS is the most prominent of these for its large epidemiological footprint and its biological importance as a direct precursor to IBC. Though DCIS in itself is not capable of metastasis, the increased risk for invasive disease that it incurs has prompted empirical treatment measures consisting of surgery, radiation, plus/minus hormonal therapy. However, it is also known that not all DCIS will progress to invasive disease. Moreover, current therapies do not come without their own morbidities and costs.

Thus, great effort has gone towards producing predictive methods for assessing which DCIS lesions will go onto become invasive and which will not following observation or therapy. Unfortunately, no histological or demographic criteria have been able to affect treatment stratification. This is paired with a lack of fundamental biologic

understanding of DCIS evolution. In this setting, molecular approaches to characterization of the transition become attractive for offering a more in depth analysis than histologic or demographic correlations can allow.

Objectives

Hypothesis

There are biologically distinct populations of DCIS that possess different prognoses for recurrence and progression to invasiveness and which can be distinguished at the time of diagnosis based on molecular markers. Comparative study of DCIS lesions as they evolve may elucidate these markers and provide predictive tools for clinicians treating DCIS. The utilization and validation of newly developed technologies may facilitate such study.

Design

In testing the hypothesis, the overall work encapsulating this study divides into two phases. As previously discussed, 'ideal' research designs for study of DCIS recurrence have been limited previously by technical capabilities. Thus, the first phase, presented here, is designed to pilot new molecular assays for use on the type and amount of DCIS specimen material expected to be available for analysis. The target parameters for assays and tissue type are guided by the second phase of the study in which the actual analysis of DCIS heterogeneity occurs as tentatively projected here:

- Compare molecular profiles of archival DCIS that did recur to DCIS that did not recur.
- Compare archival DCIS that recurred as DCIS to DCIS that recurred as IBC.
- Compare newly diagnosed DCIS without associated invasion to DCIS with associated invasion.
- Compare DCIS and IBC in newly diagnosed cases with adjacent, synchronous DCIS and IBC.

These proposed goals call for the ability to conduct studies on two categories of tissue: retrospective and prospective. Retrospectively, molecular assays able to utilize FFPE material are needed for application to archival DCIS specimens excised by lumpectomy and paired with clinical follow up. This allows comparison of DCIS that did not recur with DCIS that does recur as a whole and by type: invasive or *in situ*. Further, since LCM will be used to selectively gather neoplastic epithelium from rare archival DCIS specimens, the techniques will need to be able to use very small amounts of substrate. Prospectively, assays capable of using small amounts of fresh material from core biopsy samples of DCIS are needed. Thus, our aims for this study are as follows:

Aim 1: Pilot molecular analyses including microarray gene expression assay on archival LCM FFPE DCIS specimens. These will be paired with comparative genomic hybridization and tissue microarray pilots in FFPE being performed by Dr. Jim Hicks and Dr. David Rimm, respectively.

Aim 2: Pilot selective high throughput sequencing technology using exon capture for planned application to fresh core biopsy specimens of DCIS and IBC.

Methods

Statement of Involvement in Experiment and Design

The study was a multi-group collaborative at Yale University School of Medicine. Each section within the methods will indicate where investigators aside from the thesis candidate either performed or will perform work. The candidate specifically played a primary role in assessing the parameters for technical pilots and designing them. This involved developing and adapting protocols for the use of archival tissue for nucleic acid extraction and preparation and microarray gene expression analysis and for the use of fresh material for high-throughput capture sequencing.

Case Selection

Procurement of specimens was overseen by Dr. Donald Lannin, Director of the Yale Breast Cancer Center, from patients treated there and from its tissue bank. Informed consent was obtained from patients at the time of tissue retrieval by resection or core biopsy. Archival specimens were initially prepared and stored as formalin-fixed, paraffin (FFPE) embedded blocks according to standard pathological protocols in use at the Center. Blocks selected by Dr. Lannin for use in the retrospective cohorts were stripped of identifying information. Diagnosis and histological features were verified by Dr. Veerle Bossuyt, a breast pathologist in the Department of Pathology (Yale University School of Medicine). The first cohort of FFPE blocks obtained for **Aim 1** pilot work consisted of DCIS specimens with synchronous IDC. For the follow up studies making use of this work, 50 specimens of pure DCIS without recurrence after treatment with lumpectomy and/or radiation were first identified. Subsequently, 25 specimens of DCIS that recurred as *in situ* disease and 25 specimens of DCIS that recurred with invasive

disease were identified as controls. These controls were matched by Dr. Bossuyt to invasive specimens in pairs according to decade of diagnosis, age (within 5 years), tumor size (<2cm or >2cm), histological grade according to the DIN system, presence or absence of comedonecrosis, margin width, and treatment (+/- adjuvant radiation). Quality control on the specimens included review by Dr. Lannin of the cases selected for completeness of diagnostic workup and surgical approach. Further, the specimens used were extracted from a single institution repository and thus were all subjected to a consistent management approach such as the involved clinical oncology team and storage protocols.

Prospective pilot trials for high-throughput sequencing utilized primary cell lines derived from primary lesions of malignant melanomas in conjunction with a study on metastatic melanoma. These lines were generated by the laboratory of Dr. Ruth Halaban and based on pathological specimens obtained through Dr. Mario Sznol in the Department of Medical Oncology at Yale University School of Medicine (Yale-New Haven Hospital). Identifying information was removed prior to work.

Archival Specimen Preparation

Certain analyses required extraction of nucleic acids from the specimens. In these cases, archival FFPE specimens of DCIS involved in **Aims 1** were cut by Research Histology at Yale University School of Medicine into serial 5-μm or 10-μm sections using a microtome, which was treated with RNAseZap® (Ambion, Austin, TX) and rinsed with RNAse-free water. Deparaffinization of slices was conducted immediately prior to their use with storage in nitrogen chambers before this time. Using index H&E stained slides for discrimination for each block, Dr. Arun Gopinath in the Department of Pathology

performed laser microdissection (LCM) to isolate invasive, *in situ*, and normal cell populations as needed with the Leica LMD7000 (Leica Microsystems, Wetzlar, Germany). LCM use in breast cancer can reduce contaminating cell content to 0.6%.[95] The dissected tissue were collected in 40µl of proteinase K digestion buffer (Ambion) used directly or stored at -80°C in accordance with data showing these approaches as ideal for minimizing nucleic acid degradation.[84]

FFPE Nucleic Acid Extraction

This study sought to apply advances in FFPE molecular analysis described in the introduction to the study of DCIS, which presents a particular challenge due to the small amounts of tissue for analysis available from an *in situ* lesion after LCM purification. Archival specimens that had been microdissected were put through a proprietary extraction process, the RecoverAll® Total Nucleic Acid Isolation Kit for FFPE (AM1975, Ambion), in order to obtain both RNA and DNA simultaneously from serial slices of the same case. The modified protocol employed began with standard protease digestion (proteinase K) followed by high-salt washes on a silicon bead column and nuclease digestion of the unwanted nucleic acid component (i.e. DNAse for RNA extraction). 60-µL nuclease-free water was used to elute the RNA or DNA into nuclease-free tubes. The recovered nucleic acid was then vacuum dried and re-dissolved in 5µL of nuclease-free water. Quantification was accomplished by spectrometry on the NanoDrop-1000 (Thermo Scientific, Wilmington, DE) per manufacturer protocol. RNA samples were stored at -80°C, and DNA samples were stored at -20°C. One aspect of the pilot investigations employed a heating protocol, specifically 5 minutes at 70°C, to remove chemical modifications (mono-methyolol: $–CH_2OH$) from RNA.[96] RNA integrity was

quality tested by Bioanalyzer 2100 system (Agilent, Palo Alto, CA) at the Keck facilities. RT-PCR of (glyceraldehyde 3-phosphate dehydrogenase) GAPDH, as described below, was conducted prior to use in molecular analyses as a further quality check.

RT-PCR

Experiment specific standard inputs, typically 100ng, of total RNA were reverse transcribed into cDNA libraries using the iScript system (Bio-Rad, Hercules, CA), which contains both random and oligo-T primers suited to degraded FFPE RNA. Standard aliquots of 1:10 dilutions of these cDNA were analyzed by RT-PCR using an iCycler (Bio-Rad) system according to manufacturer's instructions. All probes used were TaqMan gene expression assays (Applied Biosystems Inc., Foster City, CA), which are proprietary 5' nuclease assays with standard annealing temperature of 60°C (**Appendix 1).** Relative quantification of gene expression was performed in triplicate reactions normalized to GAPDH (Hs99999905_m1, Applied Biosystems Inc.). Negative controls consisted of no RNA input into reverse transcription as a control against genomic DNA contamination on extraction and no cDNA input. Positive controls were derived from both MCF7 cell line RNA and DNA.

DASL Oligonucleotide Microarray

The 96-well whole transcriptome DASL microarray by Illumina (San Diego, CA) was the chosen platform for its high sensitivity requiring small input amount and specific probe design for degraded FFPE RNA. All RNA used was functionally tested by RT-PCR of GAPDH (Hs99999905_m1, Applied Biosystems Inc.) prior to usage with Ct value <35 used as a threshold for microarray application. In the study pilots, deliberate amounts of input RNA were loaded according to stated design in optimizing the array for our work.

For ongoing and future analyses, 450ng of RNA are loaded in 5 μL volumes. This benefitted from the aid of Kyle Halligan and Dr. Kimberly Lezon-Geyda in the laboratory of Dr. Lindsay Harris with whom we shared arrays in our pilots. Each array is run using the Illumina Golden Gate Assay Protocol at the W.M. Keck Facility, a biotechnology core associated with Yale University School of Medicine, under the supervision of its microarray resource director Dr. Shrikant Mane. Images scanned by BeadArray Reader (Illumina) at the Keck are visually inspected and evaluated using the associated BeadScan software for image processing and intensity data extraction; BeadStudio (Illumina) enables export of this data. The lumiR 2.8.0 (Open Source, Pan Du, Chicago, IL) package software is employed by Drs. Tuck and Schulz to analyze the data. Probe IDs were mapped to genes using lumiHumanAll annotation package. Background was adjusted to force all values to be positive, log2 transformed, and quantile normalized. Unsupervised, hierarchical clustering is performed on all DCIS specimens to discover a) if they cluster together vs. invasive specimens in pilot work and b) if they cluster together within same case replicates in the pilots. For this, all expression values were standardized to have a constant mean and scaled root-mean-square. Only probes of variance >0.1 were analyzed. The Euclidian distance matrix was then calculated, and clustering analysis was performed using the average agglomeration method. Heat maps were produced with default R Euclidean clustering while sample relation dendogram was produced using the lumi R default Multi-Dimensional Scaling algorithm. Supervised clustering will be performed in the follow up experiments to assess for differentially expressed gene transcripts in the proposed comparison cohorts.

FFPE DNA Sequencing

FFPE DNA is amplified via Sanger PCR method with Invitrogen (Carlsbad, CA) primers for selected gene exons stated in each experiment with annealing temperatures empirically assessed by gradient protocols (**Appendix 2)**. PCR was performed using the PfuUltra HF polymerase and kit (formerly Stratagene, now Agilent). After verification of amplicon generation by gel electrophoresis, amplified DNA is purified using a DNA PCR Purification Kit (Qiagen, Valencia, CA) and sequenced at the Keck Facility on Applied Biosystems 3730 capillary instruments using fluorescently-labelled dideoxynucleotides (Big Dye Terminations) and Taq FS DNA Polymerase in a thermal cycling protocol. Electrophoretic data is returned standard file sequence for analysis on 4Peaks software (Mekentosj, Aalsmeer, The Netherlands).

Exon Capture High-Throughput Sequencing

Design: The pilot work aimed to use hybridization for capture of exons from selected genes of interest. The gene list for this design was generated based on a review of several databases that have identified genes of interest in cancer generally and in breast cancer specifically that are known to harbor mutations. Genes only having alterations in amplification or expression, such as Epidermal Growth Factor Receptor or Estrogen Receptor, were not included as the capture sequencing only detects sequence alterations within a gene copy. The selected 874 genes were taken from the Sanger Welcome Database (http://www.sanger.ac.uk/genetics/CGP), the protein kinase family,[97,98] the tyrosine phosphatases,[99] the lipid kinases,[100] the Ras family,[101] the cancer candidate genes in the seminal work of Wood *et al*.,[89] and several associated labs (**Appendix 3**). Primary sequence for these genes was extracted from Build 36.1 of the NCBI's genome annotation using the UCSC Table Browser (http://genome.ucsc.edu); this track on the

Table Browser is also known as hg18 and released on March 2006. Exons from these sequences less than 250 bases in length were buffered in both the 5' and 3' directions to 250 base pairs minimum for use as targets for probe design. The locations of these sequences were submitted to the tiling array design team at Roche NimbleGen (Madison, WI) for generation of overlapping 60mer oligo probes on a 385k feature array; the total coverage was 4.86 Mbp with probe spacing minimized. The sequence search and alignment hash algorithm (SSAHA) was used to generate unique probes allowing up to 5 indels for improved specificity of capture. The specifics of the probe design process beyond these including masking and consolidation algorithms is proprietary to NimbleGen but is based on their technical reports.[92] However, post-hoc analysis using ELAND-extended program assessed quality of design through measurement of number of probes per targeted RefSeq transcript, number of probes per targeted exon, and uniqueness of probes as measured by number of transcripts targeted by a probe

Capture: High quality gDNA from primary cell lines (YUMINE from melanoma associated fibroblasts, YUHEF from melanoma) was obtained from the laboratory of Dr. Ruth Halaban through ethanol precipitation with quality checks by Nano-Drop A260/280 ratio, gel electrophoresis showing high molecular weight bands, and RT-PCR of GAPDH. The following steps in preparation and hybridization of gDNA to array were aided by Rajini Haraksingh and performed according to the NimbleGen Array User's Guide for Sequence Capture Array Delivery 1.0 with added quality control steps as noted. 20μg of DNA for each sample was randomly fragmented through sonication using a Diagenode Biorupter (Sparta, NJ) to a median size of ~500 base pairs, which was confirmed visually by gel electrophoresis and Bioanalyzer. Blunt end generation with the

Klenow fragment of T4 DNA polymerase (NEB) and phosphorylation with T4 polynucleotide kinase (NEB) followed during the 'polishing' reaction. gSel3 oligo (5' – CTC GAG AAT TCT GGA TCC TC – 3') was used to create linkers which were attached to the DNA fragments. Small fragments were removed according to manufacturer protocol with Dynal (NimbleGen) and Bioanalyzer again used to confirm. After confirmation of appropriate yield with LM-PCR, hybridization was carried out per manufacturer protocol on a NimbleGen Hybridization System (MAUI) with active mixing in 1x hybridization buffer for ~64 hours at 42°C. The array was then subjected to washing 3 times each with Stringent Wash Buffer 1x (NimbleGen) and Wash Buffers I, II, and III (NimbleGen). Hybridized, or 'captured,' gDNA was then eluted with 2x 425μL of water at 95°C. The samples were then lyophilized and resuspended in water. Linker-mediated PCR (LM-PCR) was then performed comparing pre- and post-capture samples. These samples were further evaluated by RT-PCR as a preliminary surrogate of efficiency of capture.

Sequencing: Illumina's Solexa platform (San Diego, CA) was used to generate sequence data on the captured DNA. This platform relies on fluorophore-coded synthesis of DNA templates that are bound to a solid phase support. Synthesis from each fragment is enabled by adjacent oligonucleotide primers, which are anchored across the support. Sequences of up to 35 base pairs (23 base pairs after excluding primer sequence) can be produced from each fragment's 'colony' in a massively parallel manner. This information is then interpreted using bioinformatics to align individual reads to the genome and generate continuous sequence information.

Sequencing requires compatible linkers known as 1G adaptors to be affixed to the ends of the eluted, 'capture' fragments of mean size 200 base pairs for ideal efficiency. The capture process, however, was most effective at 500 base pair sizes and involved the use of non-compatible gSel3 linkers. Sonication with the Biorupter again was employed to generate smaller fragments with linker-free ends. These free ends were blunted and phosphorylated with the Klenow fragment of T4 DNA polymerase (NEB) and T4 polynucleotide kinase (NEB), respectively, with subsequent 3' adenylation with Klenow fragment per Illumina protocol. With each step, DNA was purified with QIAquick PCR Purification Kit and Protocol (Qiagen). LigaFast (Promega #M8221) was used with the Illumina adaptor mix to rapidly affix linkers to the adenylated ends. The fragments were gel purified on a 2% Invitrogen gel (#G5018-02) with a Qiagen Gel Extraction Kit and then amplified with PCR using a Phusion DNA polymerase kit (Finnzymes) and adaptor-appropriate primers 1.1 and 2.1 (Illumina). After spin column purification of the PCR product (Qiagen), the captured DNA was quantified by NanoDrop and inputted at 10ng/μL into a single Solexa flow cell lane. Flow cell hybridization, priming, and pyro-sequencing by base incorporation were carried out by the laboratory of Dr. Michael Snyder on an Illumina 1G analyzer.

Read mapping and coverage analysis: Project collaborators Dr. Michael Krauthammer and Sebastian Szpakowski, a Ph.D candidate, performed analyses of the sequencing data as generally described here. Probe sequence targets for the NimbleGen array were mapped to the human genome's hg18 annotation using BLAT. This was done to assess the genomic coverage of the custom microarray. Subsequently, the reads generated by the Solexa (Illumina) sequencer runs were processed for quality and uniqueness using the

ELAND-extended analysis pipeline. Only the high quality reads that had no mismatches and uniquely mapped to the human genome were retained for further analysis. The BLAT and ELAND results were then aligned, allowing assessment of the number of reads per probe. A "false image" of the microarray was generated where the reads-per-probe count could be matched to physical locations on the arrays and thereby allow a quality check across its surface. Quantitative and visual assessment of any such 'smudging' effect was allowed by the R package SmudgeKit, which generates contours according to signal intensity and highlights outlier regions representative of potential spatial artifact. The alignment of probe sequence and reads generated estimation of several metrics of efficacy of array design and capture sequencing. The most important include coverage for the selected gene exons and specificity. This was assessed in terms of reads associated with a) the target exon regions and b) each base pair of the target exons. Correlation of read count with exon length was also measured given the number of probes likely increased with targeted sequence length. Specificity was calculated by comparison of the reads covering intended targets and the total number of reads generated uniquely mapping to the genome. In contrast to other studies, no calculation was made of number of reads mapping to regions near targeted sequence that might represent positive capture without target sequence generation. This conceptually would measure the captured DNA fragments inputted into sequencing which contain internal targeted regions allowing faithful hybridization but which are also located away from a 1G compatible linker.[92]

Results

Overview

The goal of this study was a) to pilot new technologies for novel investigations of tumorigenesis in DCIS and cancer generally and b) to apply these technologies to molecular characterization of DCIS progression to invasive disease. Due to limitations of time and the need to present the latter data as a completed whole for coherency, only the pilot data is available here though the methodology for all work intended was described. These pilots broadly can be classified into those directed at enabling molecular analyses of archival FFPE tissue and those directed at extending sequencing abilities from small prospective specimens.

Molecular analysis of archival tissues

Archival tissue, predominately stored as FFPE blocks, are well suited to tumorigenesis studies in DCIS given their associated clinical follow up and the difficulty of access to prospective tissue given diagnostic requirements. Here we describe pilot studies on the use of nucleic acids from FFPE DCIS samples across varying ages in regards to extraction, enhancement protocols, exon sequencing, and expression analyses.

Tissue selection, LCM, and nucleic acid extraction

Pilot DCIS specimens of varying grades and histologic features were selected from the same tissue bank as the planned study (Yale Breast Cancer Center, New Haven, CT). All tissue was obtained following lumpectomy without prior systemic treatment. A total of 8 FFPE specimens of DCIS were selected, including 2 from the same patient to allow for assessment of assay reproducibility across specimens (**Table 1**). All came from patients with co-existing DCIS and IDC that were hormone receptor negative but not *HER2/neu*

amplified. After confirmation of diagnosis and histologic characteristics by a breast pathologist, each specimen was sliced into 10 μm sections and underwent laser microdissection for neoplastic epithelium.

Table 1. DCIS Case Description. Cases are organized and named according to patient source and age of storage. **Cases 1.1 and 1.2 were derived from separate blocks of the same patient tumor.*

Patient	Code	Block Age (yrs)	DASL Array Samples
1	1.1*	2	N1, N2, N3
1	1.2*	2	N4
2	2	2	N5
3	3	3	-
4	4	9	-
5	5	19	-
6	6	19	-
7	7	29	-

This approach and nucleic acid extraction protocols were first established on non-study tissue. The pilot study tissues were utilized to describe the relationship of extraction yield to tissue input (**Figure 3**). For this assessment, individual extractions of nucleic acids were paired with data from laser mirodissection on the total surface area of ducts collected. 117 RNA extractions, 29 DNA extractions, and 9 simultaneous extractions of RNA and DNA from microdissected material from a single side were used. Simultaneous extraction produced unreliable, low yields and were excluded from analysis (data not shown). Ratios of the RNA and DNA amounts extracted to the surface area collected were then calculated for each specimen based on initial index slices. Yields varied significantly between specimens from 2007 and 1990 (**Figure 3c)**. However, yields also seemed to vary between other specimens of the same approximate age, suggesting age was not the only factor affecting yields. For instance, the specimen from patient 3 produced far less RNA per area dissected than specimens from patient 1 or 2, which were

only a year 'younger.' The nucleic acid yield of a specimen was thus extrapolated from initial extractions on its tissue block, guiding further dissection as needed. This approach was important for preserving the rest of the specimen, which would ideally be used for *in situ* protein expression analysis by TMA. Spectroscopy based quantification (NanoDrop) also allowed initial quality assessment of all nucleic acids by A260/A280 ratio. Further, all RNA was assessed by Bioanalyzer with a RNA integrity number (RIN) >1.9 chosen as a conservative minimum threshold for further use.

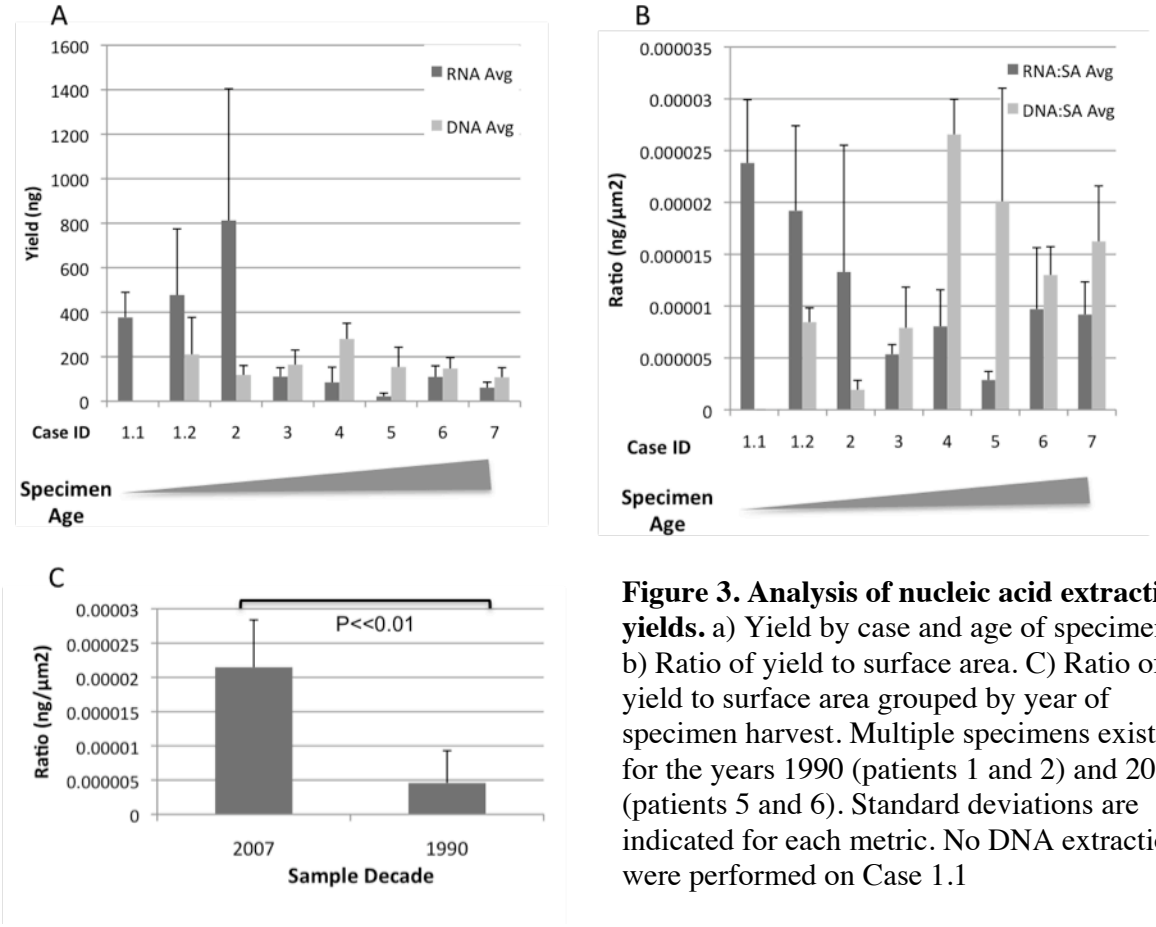


Figure 3. Analysis of nucleic acid extraction yields. a) Yield by case and age of specimen. b) Ratio of yield to surface area. C) Ratio of yield to surface area grouped by year of specimen harvest. Multiple specimens existed for the years 1990 (patients 1 and 2) and 2007 (patients 5 and 6). Standard deviations are indicated for each metric. No DNA extractions were performed on Case 1.1

RNA Expression Analysis

The goal of the pilot for RNA expression analysis was to validate a reproducible genome wide expression analysis platform using unamplified, FFPE RNA from DCIS neoplastic

epithelium. After initial quantification and quality assessment, qualifying RNA extracts from the pilot study specimens were subjected to a heating protocol, which aimed to remove chemical modifications from the storage process. A representative example of the benefit of this heating protocol on signal detection in RT-PCR, a surrogate for signal detection on microarray, is shown in **Figure 4** . RNA samples were then assessed for signal detection by RT-PCR of a standard probe (GAPDH Hs99999905_m1, ABI). A Ct value <35 was required for use on microarray. A selection of samples that had passed the quality parameters of spectroscopy, Bioanalyzer, and RT-PCR then were used to pilot the use of unamplified FFPE RNA on the DASL (Illumina) whole genome microarray.

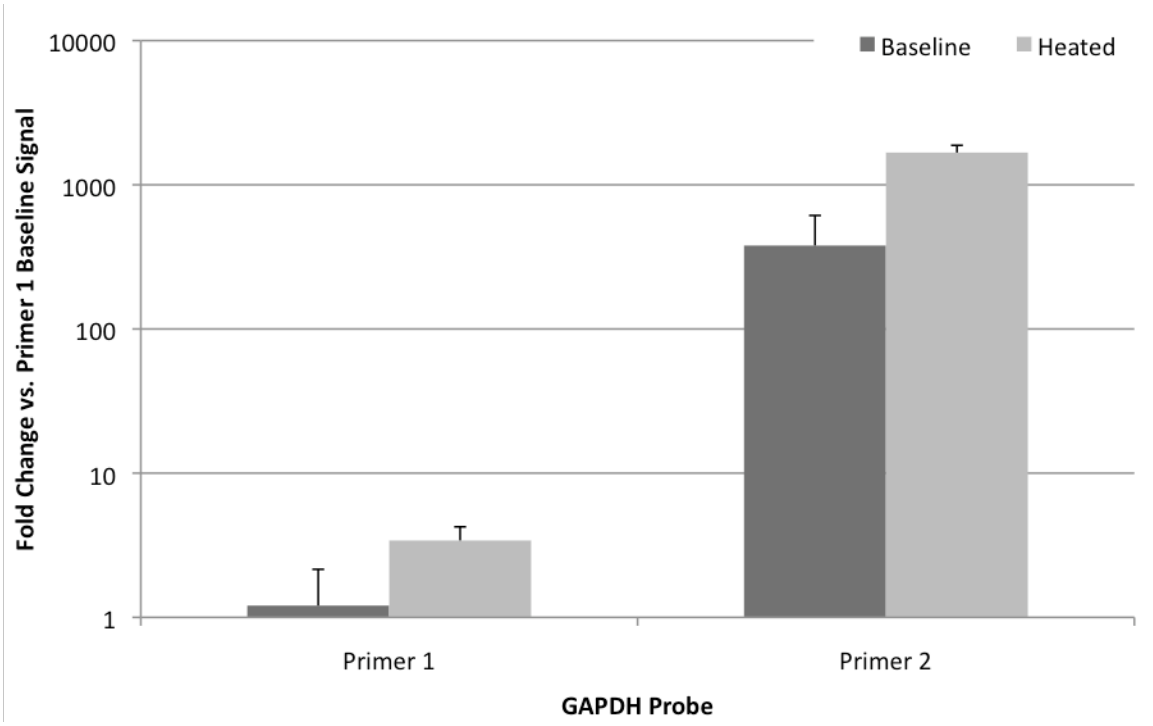


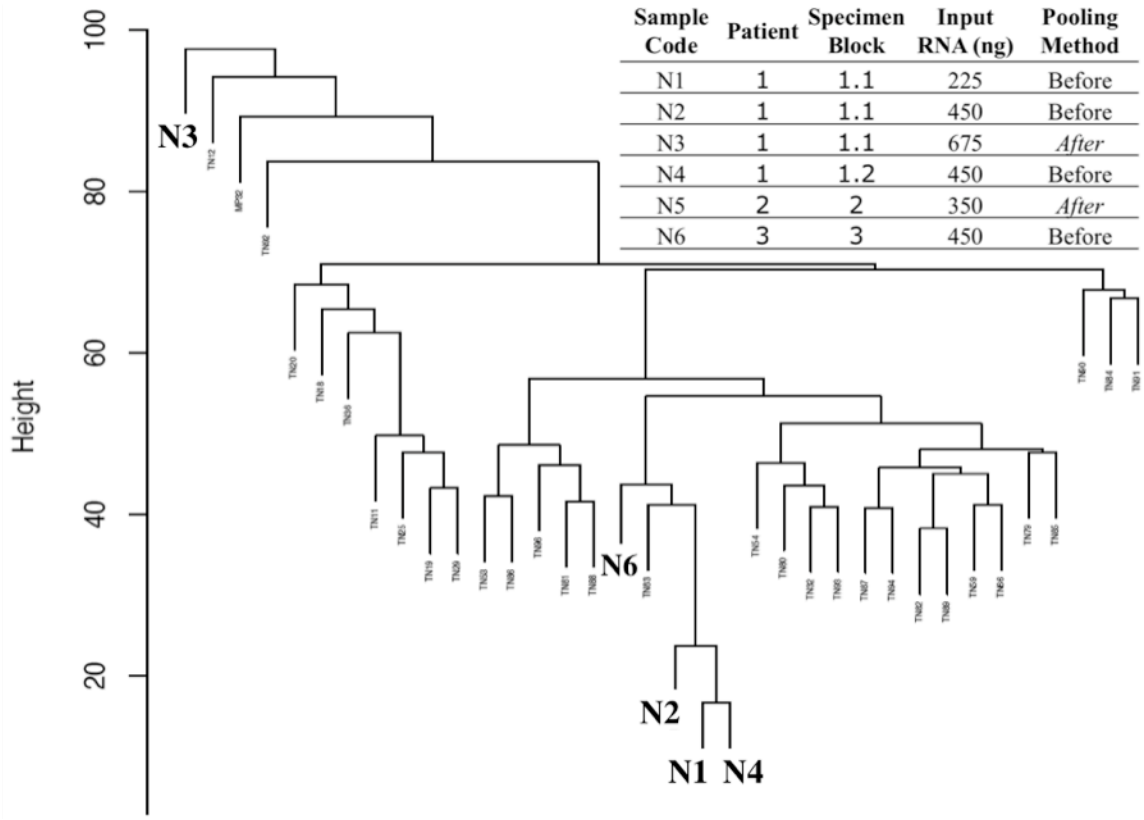
Figure 4. Enhancement of baseline expression signal in RT-PCR by RNA heating protocol. Representative examples of enhancement here are seen across two different probes for GAPDH cDNA. Experiments were done in triplicate and shown as fold change versus primer 1 baseline signal for GAPDH at a constant 54ng RNA input with identical cDNA aliquots into RT-PCR. Standard deviations are shown.

Several variables were assessed by comparison in the pilot design. First, input RNA amount was varied to assess for a) the minimum required amount for reliable signal

detection and b) whether signal data was influenced by input and not compensated for by internal array controls as advertised. Though the indicated minimum for RNA input on the DASL arrays used was 50 ng, it was unknown how the degradation level of FFPE RNA affected this requirement. At the same time, the minimum RNA required was likely to not always come from a laser microdissected single slide. Thus, two different methods for pooling extractions from separate slides of a tissue block were assessed for best signal detection. One involved combining samples prior to the extraction protocol, while the second combined specimens after separate extraction. Lastly, the use of extractions from separate blocks of the same specimen assessed for reproducibility of expression data on this platform. As a positive control, the samples were run alongside RNA extractions from FFPE cores of recent (2009) invasive breast disease that had successfully produced data in the past. This control also provided a set of tumors with an array of expression profiles within which the pilot DCIS samples could undergo clustering analysis. Unsupervised hierarchical clustering was generated by lumiR default Multi-Dimensional Scaling algorithm to produce a dendogram of the interrelatedness of the DCIS samples and the invasive tumors run (**Figure 5**).

First, these results demonstrated that unamplified, FFPE derived RNA from DCIS specimens could be used to generate whole genome expression data. Secondly, these results seemed reproducible. 3 of the 4 samples from the same patient (patient 1), N1, 2, and 4, which came from different blocks (1.1 and 1.2), grouped together closely. The outlier sample, N3, differed in preparation, using both a higher input of RNA (650 ng) and a contrasting *post*-extraction pooling method. N5, like N3, was RNA pooled post-extraction. Both N3 (**Figure 6**) and N5 had a low signal detection rate, though only N3

surpassed the threshold for inclusion in analysis. N5 did not meet the threshold signal detection and was excluded from clustering dendograms. The detectable DCIS samples, excepting N3, grouped together, distinct from the vast majority of the invasive specimens. Only one invasive specimen grouped with the DCIS sub-cluster. N6, from patient 3, did however lie outside the patient 1 sub-group of N1, 2, and 4 as expected given the different tumors of their origin.



Sample Code	Patient	Specimen Block	Input RNA (ng)	Pooling Method
N1	1	1.1	225	Before
N2	1	1.1	450	Before
N3	1	1.1	675	*After*
N4	1	1.2	450	Before
N5	2	2	350	*After*
N6	3	3	450	Before

Figure 5. Hierarchical clustering dendogram of gene expression in FFPE DCIS and invasive specimens on DASL (Illumina) whole genome microarray. A table indicates the coding of the DCIS samples used, of which all but N5 produced threshold signal data. Samples beginning with 'N' represent DCIS specimens run alongside all other named invasive specimens; these are enlarged for visualization. Clustering was performed as described based on an 'intrinsic' gene set of 8,751 genes with standard deviation relative to mean > 0.1.

Since the pilot DASL was aimed primarily at evaluating feasibility of producing reproducible signal and clustering results, it was not deemed necessary to mass validate

gene transcript expression profiles by RT-PCR. However, select validations were performed with a representative example of this effort demonstrating high levels of ErbB2 transcript by RT-PCR in agreement with the array data (**Figure 7**).

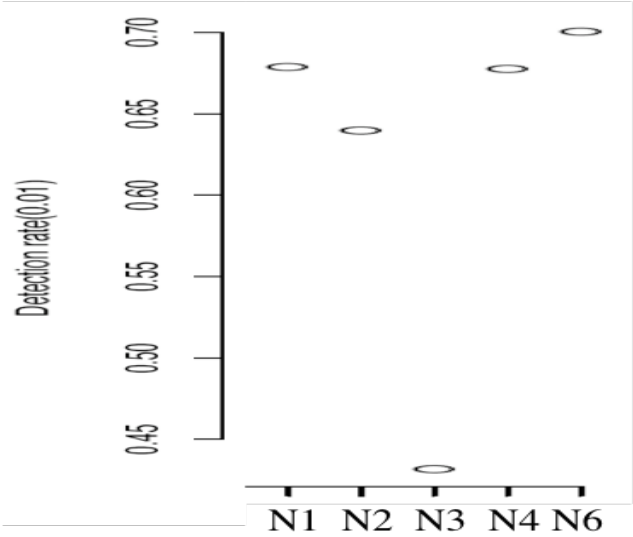


Figure 6. Signal detection rate on DASL microarray gene expression analysis of DCIS specimens. N5 samples is excluded as it did not meet the detection rate threshold. N3 is noted to be just above the 0.4 threshold. Detection rate is relative to the total number of probes on the array.

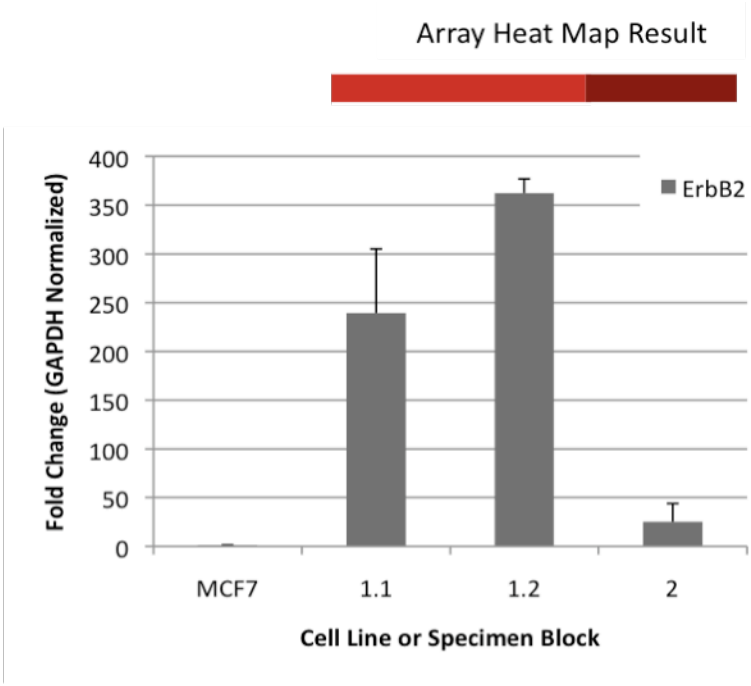


Figure 7. Representative validation of array results with RT-PCR using the *ERBB2* transcript probe. RT-PCR was carried out using identical inputs of RNA and cDNA in triplicate and shown as fold change normalized to GAPDH. Results are aligned with heat map representation of results from array against control MCF7 line with known low *ERBB2* expression.

DNA mutation re-sequencing by PCR pilot

DNA is considered more resistant to degradation by the formalin fixation and paraffin embedding process in archival specimen storage. Nonetheless, concern over the

feasibility and validity of the sequencing results from these specimens prompted a pilot study. The pilot specifically assessed the length of exon reads routinely possible with FFPE DNA derived from laser microdissected DCIS. Primer sets (Invitrogen) for ErbB4 (**Appendix 2**), which had been previously validated and used in the laboratory, were used. Amplicons of 186 base pairs from exon 20 of ErbB4 were routinely sequenced by Sanger PCR method and verified on by BLAT (UCSC Genome Browser). The redundant use of both forward and reverse strand primers allowed high quality, reproducible, and minimal error sequencing throughout the amplicon (data not shown due). Use of a single strand primer tended, in contrast, to produce uncertainty at the end of a read corresponding to the 5' or 3' end of exon.

Selective high-throughput sequencing

Prospective studies will require validated techniques for maximizing molecular analysis of small amounts of tissue. Selective mass sequencing from small inputs has been suggested as a possibility in the literature.[92] The pilot studies here describe efforts to validate a custom array based capture method for selective high-throughput sequencing of breast cancer and melanoma related genes (**Appendix 3**).

Design

Design of the capture array was performed on a NimbleGen 385K feature array, as described in the methodology, by Sebastian Szpakowski and Dr. Michael Krauthammer. Quality assessment of probe design was conducted using ELAND-extended software across several metrics shown in histogram format (**Figure 8**). The microarray contains 381,034 probes designed for human genes. Based on the overlap of BLAT mapping of probe sequences and known gene transcript sequences (RefSeq transcripts), 372,143

probes were found to overlap with at least one Refseq transcript. This includes probes that align to a transcript ideally or with 1 mismatch. Overall the probes were mapped to 2,273 Refseq transcripts and their 22,492 exons. Measurement of specificity by the number of transcripts sampled by each probe revealed that the majority of probes aligned to a unique transcript region. A significant number also aligned to more than one transcript, however, few aligned to more than 2. Coverage of target regions was measured by the number of probes assigned to each transcript, which varied widely. The majority of transcript features contained at least several probes. Within a targeted transcript, most exons contained several probes.

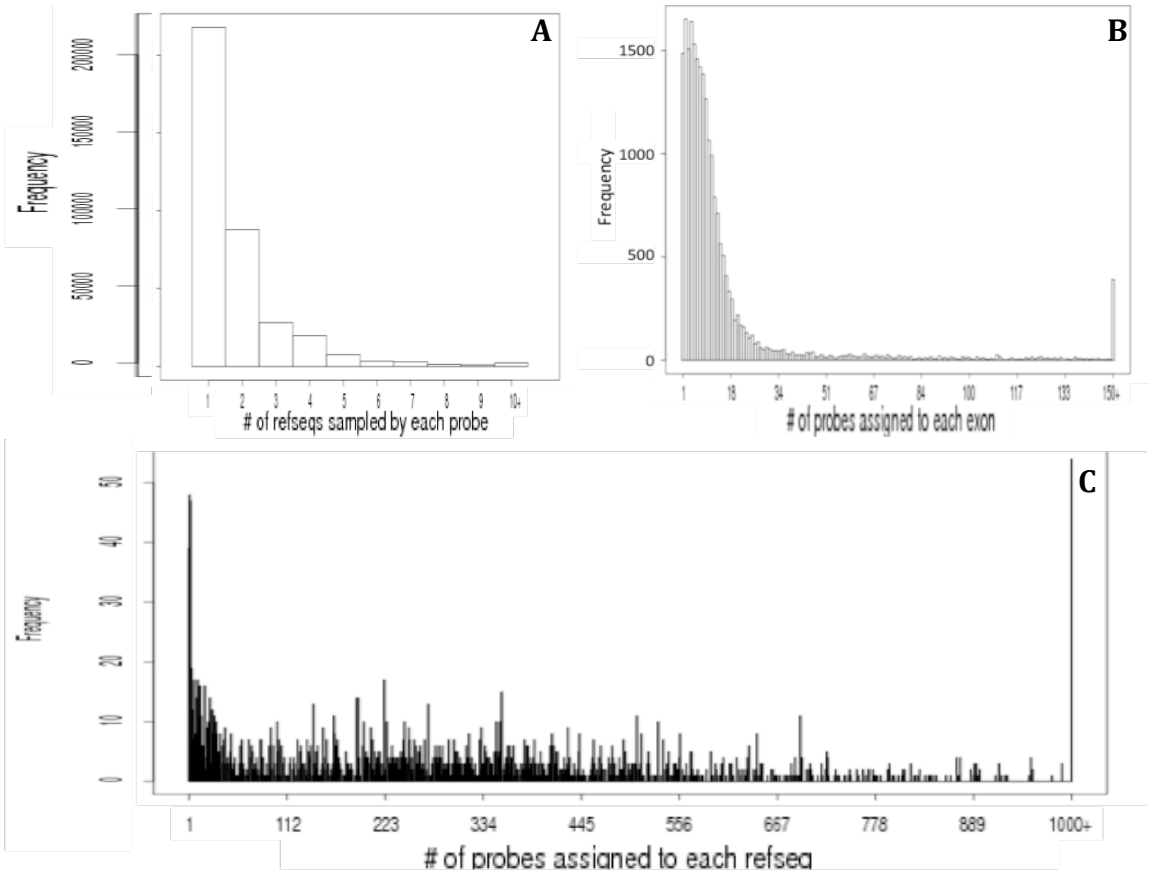


Figure 8. Design analysis by ELAND-extended of capture array probes. All analysis done by Krauthammer lab of capture sequencing performed by thesis candidate and collaborators as indicated. a) Number of transcripts in RefSeq database sampled by probe based on sequence. b) Number of probes assigned to each targeted exon. c) Number of probes assigned to each target RefSeq gene transcript.

Capture

Fragmented DNA from two melanoma lines, YUMINE and YUHEF, were hybridized to the custom array and the post-capture DNA sequenced. Capture efficiency across the arrays was first assessed for performance by alignment of sequence reads to the physical location of corresponding targeting probes on the array. This allowed visual inspection for differences in signal intensity, representing sequence reads, according to a probe's location on the array that might be suggestive of a 'smudge' artifact during the capture. Contour plots of signal intensity across the arrays demonstrated sufficiently homoegeneous capture to rule out major smudge artifacts (**Figure 9**). However, comparison of the "false images" of signal intensity according to physical probe location indicate that YUMINE had on average lower signals than YUHEF (data not shown).

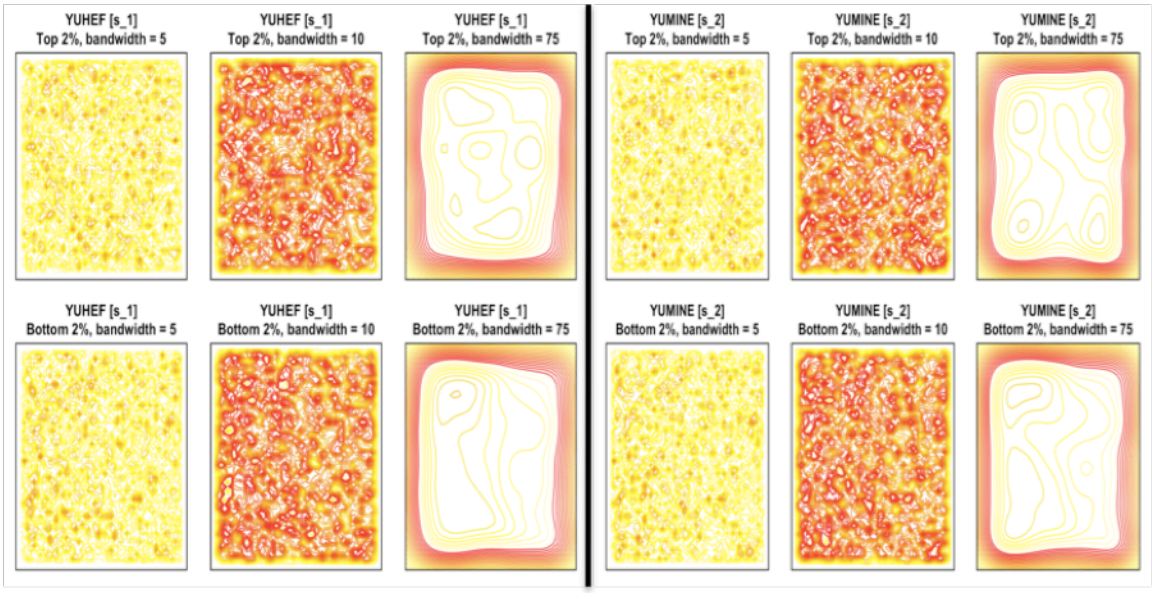


Figure 9. Contour plot of sequence signal across virtual surface of each array. R package SmudgeKit was used by Krauthammer laboratory as noted to generate plots of the top and bottom 2% of signals. Different bandwidths for kernel density reveal no evidence of uneven distribution of signals, or 'smudging,' due to technical artifacts. YUHEF and YUMINE plots are divided by a line while top 2% plots are divided from the bottom 2% plots by rows.

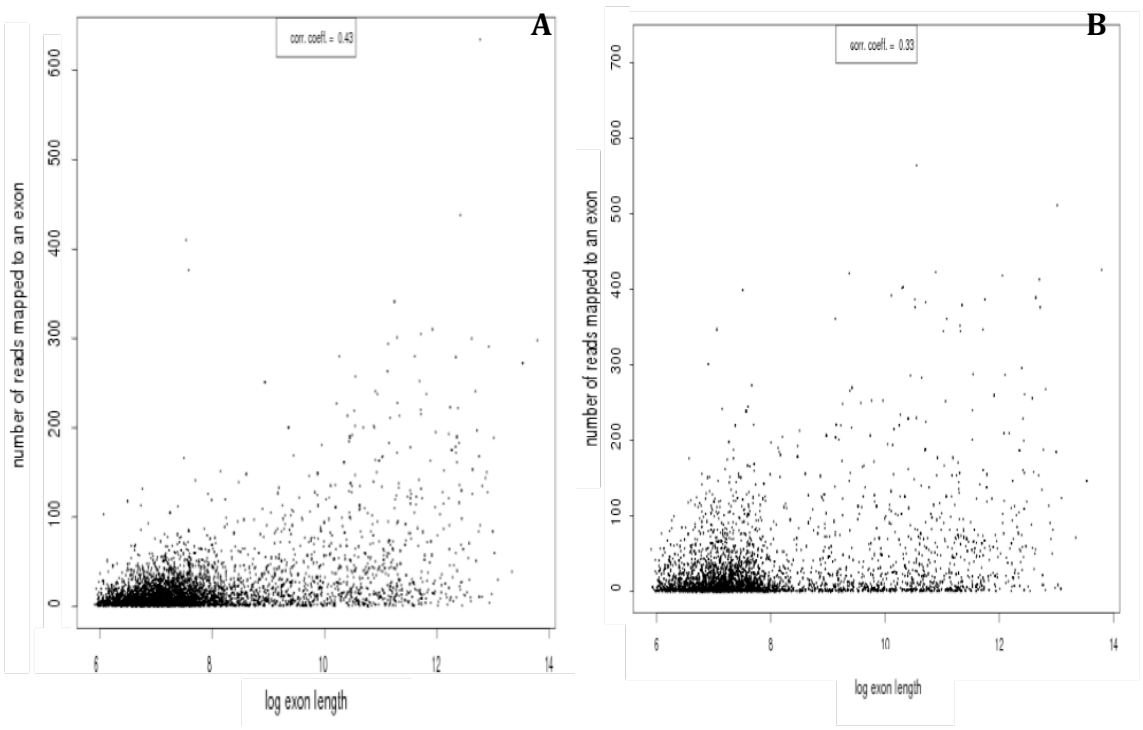


Figure 10. Relationship between exon length and read count. A) YUHEF. B) YUMINE.

Further characterization of the capture focused on specificity and sensitivity. While the microarray targeted 22,492 exons, the sequence reads aligned to 7,297 of these exons (32.4%) in the YUHEF cell line and to 5,549 of these exons (24.7%) in the YUMINE cell line. The relationship between number of reads per exon and exon length was characterized as well (**Figure 10**). However, the association was weak with a correlation coefficient in the YUHEF capture of 0.43 and in the YUMINE capture of 0.33. Analysis of the relationship of number of probes designed per exon to the number of reads obtained from sequencing yielded higher correlations of 0.52 and 0.4 for YUHEF and YUMINE, respectively. In addition to the 22,492 targeted exons, the YUHEF capture sampled 20,448 additional exons, and the YUMINE capture sampled 7,399 additional exons. This non-specific sampling was shared amongst the two captures in only 1,334 exons. Specificity of reads thus was calculated as 26.3% in YUHEF and 42.9% in YUMINE. When sequencing coverage is applied to whole genes, 223,204

unique sequence reads align to 828 targeted genes (94.7% of the full 874 gene set) in the YUHEF capture, and 203,965 reads aligned to 815 genes (93.2%) in YUMINE capture. Overall, YUHEF's 1,842,511 reads and YUMINE's 1,539,151 reads aligned to a total of 14,790 genes each. While the average read coverage for the 14,790 known human genes was 99.8 reads/gene in YUHEF and 107.4 reads/gene in YUMINE, the average read count for the specifically targeted 874 genes was higher at 261.3 reads/gene and 243.4 reads/gene for YUHEF and YUMINE, respectively.

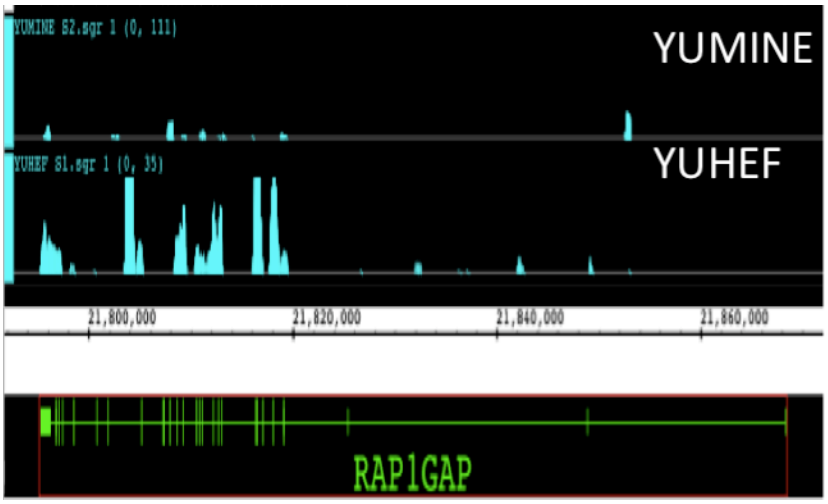


Figure 11. Visual representation of example of capture sequencing using *RAP1GAP* gene exons. Integrated Genome Browser was used to align sequence 'hits' from the YUMINE and YUHEF captures to genomic coordinates. Comparison can thus be made to the coordinates of target exons (green), here showing poor capture in YUMINE relative to YHUEF.

Visual representations of successful capture examples were generated using the Integrated Genome Browser (**Figure 11**). A list of 'indel' and single nucleotide sequence alterations was produced by comparison of sequences with multiple reads to BLAT aligned regions (data not shown). Validation by Sanger PCR sequencing is pending. Furthermore, comparison to single nucleotide polymorphism (SNP) databanks has not yet been performed; thus, these candidate mutation lists are unrefined. More important again is the feasibility endpoint of this pilot, which was achieved.

Discussion

The study of tumorigenesis requires information on the evolution of a lesion *in vivo*, which has presented barriers to effective molecular analysis of progression in diseases like DCIS. Prospective studies incur large time investments and are inhibited by the difficulty with accessing fresh tissue due to diagnostic needs. Various cross-sectional and retrospective approaches have been utilized in an attempt to circumvent prospective study. However, each thus far contained confounding design issues due to technical limitations in studying archival tissue. Mainly, none have been able to properly evaluate a given lesion's propensity for recurrence or progression, which is the key prognostic question at time of diagnosis. The goals of the overall study were to pilot technologies that would overcome limitations in past approaches to study of DCIS with the intent of applying them to the investigation of DCIS progression as laid out in the objectives. Our pilot work marked technical advances in a) the use of retrospective FFPE tissue for molecular analyses and b) the use of small amounts of DNA input for selective high-throughput gene re-sequencing. Further, this work is paired with information from pilot studies from collaborators on CGH and quantitative TMAs in FFPE material that will allow comprehensive analysis of archival and prospective DCIS specimens. Our results will be discussed here in terms of methodology recommendations for ongoing future study objectives and implications for tumorigenesis study generally.

Archival tissue analysis in DCIS

The main body of material that would allow improved retrospective cohort designs, archival FFPE tissue with clinical follow up, has long been considered inaccessible to molecular characterization. In our pilot work on archival DCIS samples,

we demonstrated that laser microdissected neoplastic DCIS epithelium could be used successfully in microarray gene expression analysis and Sanger PCR based exon resequencing.

The microarray pilot assessed feasibility, input amounts, RNA pooling methodology, and reproducibility of results across duplicates from different specimen blocks of the same patient. Comparison of signal data from our DCIS samples to signal data of the invasive specimens indicates the majority of our samples produced interpretable data. Conversely, our 'failed' sample, N5, has the same poor signal qualities, namely low detection rate (**Figure 6**), of poor performance invasive specimens. On hierarchal clustering, the grouping of most DCIS specimens together distinct from the unrelated invasive tumors is consistent with expectations given biologic differences. More importantly, the close grouping of 75% of the 4 samples from patient 1 regardless of tissue block source indicates that expression data was reproducible within a tumor. N3, the outlier sample from patient 1, may represent a limitation in the technology's ability to produce reproducible data or a statistical variance given the low number of samples used. However, the poor signal detection rate of N3 suggests that the gene expression differences might be due to poorer performance from different preparation of the sample. Specifically, N3 was prepared with higher RNA input (650 ng) and a post-extraction method of pooling RNA from separate tissue slide extractions. N1 and N2, which represent differing RNA inputs from the same block of patient 1, nonetheless grouped together closely and shared pre-extraction RNA pooling methodology. Thus, it is more likely that N3's failure to group with either the rest of the DCIS specimens or other samples from patient 1 is due to a different RNA pooling method rather than true gene

expression differences. Indeed, the only other sample prepared with post-extraction pooling was N5, which failed to meet signal detection threshold. In light of this correlation, the failure of N5 and the aberrant expression data of N3 perhaps can be explained by the higher salt concentration in these samples. This would result from combination of separate extracts each containing a set amount of salt residue after high salt column isolation. This contrasts with pre-extraction pooling in which the combined tissues only are subjected to high salt washes once.

The sum of these results and their implications thus suggests that a) laser microdissected neoplastic epithelium from DCIS FFPE can be used in the DASL microarray to produce expression data, b) the input RNA amount should be kept standard at either 450 ng or 225 ng, c) laser microdissected material for RNA isolation from a specimen should be pooled before extraction, and d) replicates are useful for assessing data reliability. The limitations of this work lie primarily in the low numbers used for comparisons. However, the pilots incur substantial costs due to pathologist time, technological expense of microarrays, and use of precious archival material. Given the endpoint of the pilot was to establish basic feasibility and a workflow, we thus deemed the uncertainty due to low statistical powering of the work to be acceptable.

In contrast to previous work, we were able to avoid amplification of FFPE RNA prior to oligomer hybridization on the microarray. Instead, a combination of improved extraction technique, heating protocol for chemical modification removal[96], and a new microarray platform was employed. Importantly, the DASL array hybridizes transcripts prior to amplification of cDNA.[83,86] The reason for this approach is evidence that standard T7 based linear amplification of RNA arguably introduces bias into data.[87] Conceptually,

this bias is likely even more significant in degraded, chemically modified RNA from FFPE specimens that are crucial to DCIS retrospective analyses. Like any manipulation of such RNA, T7 primer ligation is potentially biased towards certain transcripts which are better or worse substrates.

More broadly, the success of establishing a methodology workflow and platform for DCIS FFPE microarray based expression analysis allows the study's key retrospective investigations regarding recurrence and progression to go forward. Previous molecular studies on DCIS progression have used DCIS specimens co-existing with invasive disease in snapshot analyses as surrogates of progression.[18,80] Part of the reason for this is to minimize inter-tumor variability background in assessing differences between DCIS and invasive disease. However, the main reason for doing so is the technical inability to utilize FFPE specimens for their analyses. These approaches have found few consistent differences between invasive disease and DCIS. Indeed, the two studies that analyzed FFPE DCIS after LCM isolation of neoplastic epithelium contain a marginal overlap of only 4 genes that are significant for differential expression. We hypothesize that study of pure DCIS specimens will reflect a more sensitive assay for detecting biologically distinct populations with different fates that might yield candidate markers of progression and/or recurrence.

Nonetheless, in proceeding with our retrospective profiling studies, we will take note of concerns raised by our predecessors. Schuetz *et al*.[80] has argued insightfully that the use of different microarray platforms likely contributes to the lack of overlap between gene candidate lists between their study and other studies on DCIS expression profiles. This same concern has plagued the comparison of data sets in gene expression profiling

for invasive breast cancer.[102,103] Often, the "molecular signatures" generated by profiling studies exhibit little overlap and fail to validate on prospective analysis. Thus, we will take care to focus on the use of a single microarray platform we have validated for LCM FFPE RNA in DCIS and to validate any putative markers on both independent retrospective and prospective sets of DCIS.

Less nuanced were our findings on the use of LCM FFPE DNA from DCIS for selected exon re-sequencing. Given our success on ~200 base pair reads, we may seek to perform limited sequencing analysis on candidate markers of DCIS progression and/or recurrence generated by the expression profiling studies.

Selective high-throughput sequencing

Prospective study of DCIS is made difficult technically by the diagnostic requirements on the excised tissue for ruling out invasive disease and for characterizing the histology. At the current time, a core biopsy of tissue after lumpectomy represents the total substrate of tissue available for analysis. In our study, we adapted the principles of exon capture-based selective sequencing to the need to extract large amounts of sequence data from small quantities of DNA that would be available from such core biopsies.

Our custom array enriched for exons of genes implicated in breast cancer and melanoma (**Appendix 3**). This design both tested potential probes sets for future DCIS work and made any data gained from the pilot useful given the primary melanoma lines used were under active investigation. The use of the YUHEF and YUMINE primary melanoma cell lines in our pilot work simulated extractions from mammosphere primary lines of DCIS that are being developed in collaborator labs as means for enabling

prospective analysis of small biopsy samples in tumors. Specifically, these melanoma lines were early generation cultures from recently excised melanomas.

Our results confirmed the use of array-based enrichment for relatively selective sequencing on a Solexa (Illumina) platform. They also justify further development of the assay for use on our proposed prospective cohorts in the subsequent study phase. Nonetheless, the analysis of efficiency of our approach nonetheless revealed several areas for improvement. Foremost are improving the sensitivity of capture to increase the number of targeted exon reads and improving the specificity to eliminate wasted reads on non-target sequence. In comparison to technical paper reports of target exon coverage of 78-99% depending on technique,[92] our efficiencies (24.7% and 32.4% in YUMINE and YUHEF, respectively) were very low.

Multiple limitations of experimental methodology rather than technical capability likely explain this discrepancy. First, our custom array was built around 60mer probes from NimbleGen rather than the 100mer capture array normally used for its larger and more specific hybridization potential. This produced a lower cost pilot array that also had the possibility of translating into a more cost efficient clinical application. Secondly, the sequence reads only come from 23 base pair end fragments of captured DNA segments that had a Solexa 1G-compatabile linker attached. This linker is a primer sequence attached to input DNA that allows priming of the flourophore based synthesis reactions that make up Solexa sequencing. As described in the methods, hybridization was carried out with non-compatible linkers on larger fragments that are optimal for capture. This was followed by repeat fragmentation to create smaller fragments ideal for Solexa sequencing and free ends for attachment of Solexa 1G-compatabile linkers. Thus, only a

fraction of captured DNA had a free end with the appropriate linker for sequencing. This resulted in lower efficiency of selective sequencing overall than efficiency of target fragment capture itself.

The large reduction in sensitivity and specificity likely due to these modifications will be addressed going forward in the continuing studies. Firstly, 100mer capture arrays have been procured with the probes validated on the pilot. Secondly, future capture hybridizations will be tailored to the planned sequencing platform according to fragment length and linker selection. In the case of Roche 454 sequencing,[104] which advertises faithful reads of 200-400 base pairs, hybridization and sequencing can occur using fragment sizes ideal for capture (400-500 base pairs). However, due to cost issues, Solexa is currently more viable. Solexa sequencing will require 1G compatible linkers to be attached to DNA fragments of <200 base pairs which will be used for capture and sequencing. Based on previous technical literature, the cost to capture efficiency of using shorter fragments should be more than compensated for by the increased sequencing efficiency.[92]

Our pilot work did also produce sequence information on the melanoma lines we analyzed. Analysis of the sequence aligned to the genome generated a list of indels and mismatches. These data were not validated by Sanger sequencing. Moreover, it is likely many of these alterations represent SNPs or other changes not associated with tumor behavior. These analyses, however, were not the endpoint of the study, which instead focused on assessing capture efficiency rather than the fidelity of a proven sequencing platform.[105] However, in future work, candidate mutations will be validated in such a manner. In summary, exploration of hybridization-based capture for selective high

throughput sequencing resulted in discrete recommendations for our future work on prospective DCIS analysis from core biopsy samples.

References

1. Brenton J, Carey L, Ahmed A, Caldas C. Molecular classification and molecular forecasting of breast cancer: ready for clinical application? J Clin Oncol 2005;23:7350-60.
2. Caldas C, Aparicio S. The molecular outlook. Nature 2002;415:484-5.
3. Peto R, Boreham J, Clarke M, Davies C, Beral V. UK and USA breast cancer deaths down 25% in year 2000 at ages 20-69 years. Lancet 2000;355:1822.
4. Burstein H, Polyak K, Wong J, Lester S, Kaelin C. Ductal carcinoma in situ of the breast. N Engl J Med 2004;350:1430-41.
5. Lakhani S, Collins N, Stratton M, Sloane J. Atypical ductal hyperplasia of the breast: clonal proliferation with loss of heterozygosity on chromosomes 16q and 17p. J Clin Pathol 1995;48:611-5.
6. Kuerer HM, Albarracin CT, Yang WT, et al. Ductal carcinoma in situ: state of the science and roadmap to advance the field. J Clin Oncol 2009;27:279-88.
7. Li C, Uribe D, Daling J. Clinical characteristics of different histologic types of breast cancer. Br J Cancer 2005;93:1046-52.
8. Allred D, Mohsin S, Fuqua S. Histological and biological evolution of human premalignant breast disease. Endocr Relat Cancer 2001;8:47-61.
9. Radford D, Phillips N, Fair K, Ritter J, Holt M, Donis-Keller H. Allelic loss and the progression of breast cancer. Cancer Res 1995;55:5180-3.
10. Stratton M, Collins N, Lakhani S, Sloane J. Loss of heterozygosity in ductal carcinoma in situ of the breast. J Pathol 1995;175:195-201.
11. Moore E, Magee H, Coyne J, Gorey T, Dervan P. Widespread chromosomal abnormalities in high-grade ductal carcinoma in situ of the breast. Comparative genomic hybridization study of pure high-grade DCIS. J Pathol 1999;187:403-9.
12. Buerger H, Otterbach F, Simon R, et al. Comparative genomic hybridization of ductal carcinoma in situ of the breast-evidence of multiple genetic pathways. J Pathol 1999;187:396-402.
13. Aubele M, Cummings M, Mattis A, et al. Accumulation of chromosomal imbalances from intraductal proliferative lesions to adjacent in situ and invasive ductal breast cancer. Diagn Mol Pathol 2000;9:14-9.
14. Rudas M, Neumayer R, Gnant M, Mittelböck M, Jakesz R, Reiner A. p53 protein expression, cell proliferation and steroid hormone receptors in ductal and lobular in situ carcinomas of the breast. Eur J Cancer 1997;33:39-44.
15. Allred D, Clark G, Molina R, et al. Overexpression of HER-2/neu and its relationship with other prognostic factors change during the progression of in situ to invasive breast cancer. Hum Pathol 1992;23:974-9.
16. Porter D, Krop I, Nasser S, et al. A SAGE (serial analysis of gene expression) view of breast tumor progression. Cancer Res 2001;61:5697-702.
17. Porter D, Lahti-Domenici J, Keshaviah A, et al. Molecular markers in ductal carcinoma in situ of the breast. Mol Cancer Res 2003;1:362-75.
18. Ma X, Salunga R, Tuggle J, et al. Gene expression profiles of human breast cancer progression. Proc Natl Acad Sci U S A 2003;100:5974-9.

19. Adeyinka A, Emberley E, Niu Y, et al. Analysis of gene expression in ductal carcinoma in situ of the breast. Clin Cancer Res 2002;8:3788-95.
20. Seth A, Kitching R, Landberg G, Xu J, Zubovits J, Burger A. Gene expression profiling of ductal carcinomas in situ and invasive breast tumors. Anticancer Res;23:2043-51.
21. Tamimi R, Baer H, Marotti J, et al. Comparison of molecular phenotypes of ductal carcinoma in situ and invasive breast cancer. Breast Cancer Res 2008;10:R67.
22. Allred D. Biologic characteristics of ductal carcinoma in situ. In: Silverman M, ed. Ductal carcinoma in situ of the breast. 2nd ed. Philidelphia: Lippincott, Williams and Wilkins; 2002:37-48.
23. Buerger H, Otterbach F, Simon R, et al. Different genetic pathways in the evolution of invasive breast cancer are associated with distinct morphological subtypes. J Pathol 1999;189:521-6.
24. Jeffrey S, Pollack J. The diagnosis and management of pre-invasive breast disease: promise of new technologies in understanding pre-invasive breast lesions. Breast Cancer Res 2003;5:320-8.
25. Buerger H, Mommers E, Littmann R, et al. Ductal invasive G2 and G3 carcinomas of the breast are the end stages of at least two different lines of genetic evolution. J Pathol 2001;194:165-70.
26. Werner M, Mattis A, Aubele M, et al. 20q13.2 amplification in intraductal hyperplasia adjacent to in situ and invasive ductal carcinoma of the breast. Virchows Arch 1999;435:469-72.
27. Gong G, DeVries S, Chew K, Cha I, Ljung B, Waldman F. Genetic changes in paired atypical and usual ductal hyperplasia of the breast by comparative genomic hybridization. Clin Cancer Res 2001;7:2410-4.
28. Boecker W, Buerger H, Schmitz K, et al. Ductal epithelial proliferations of the breast: a biological continuum? Comparative genomic hybridization and high-molecular-weight cytokeratin expression patterns. J Pathol 2001;195:415-21.
29. Jones C, Merrett S, Thomas V, Barker T, Lakhani S. Comparative genomic hybridization analysis of bilateral hyperplasia of usual type of the breast. J Pathol 2003;199:152-6.
30. O'Connell P, Pekkel V, Fuqua S, Osborne C, Allred D. Molecular genetic studies of early breast cancer evolution. Breast Cancer Res Treat 1994;32:5-12.
31. O'Connell P, Pekkel V, Fuqua S, Osborne C, Clark G, Allred D. Analysis of loss of heterozygosity in 399 premalignant breast lesions at 15 genetic loci. J Natl Cancer Inst 1998;90:697-703.
32. Farabegoli F, Champeme M, Bieche I, et al. Genetic pathways in the evolution of breast ductal carcinoma in situ. J Pathol 2002;196:280-6.
33. Lakhani S, O'Hare M. The mammary myoepithelial cell--Cinderella or ugly sister? Breast Cancer Res 2001;3:1-4.
34. Barsky S, Karlin N. Mechanisms of disease: breast tumor pathogenesis and the role of the myoepithelial cell. Nat Clin Pract Oncol 2006;3:138-51.
35. Polyak K, Hu M. Do myoepithelial cells hold the key for breast tumor progression? J Mammary Gland Biol Neoplasia 2005;10:231-47.
36. Allinen M, Beroukhim R, Cai L, et al. Molecular characterization of the tumor microenvironment in breast cancer. Cancer Cell 2004;6:17-32.

37. Adriance M, Inman J, Petersen O, Bissell M. Myoepithelial cells: good fences make good neighbors. Breast Cancer Res 2005;7:190-7.
38. Hu M, Yao J, Cai L, et al. Distinct epigenetic changes in the stromal cells of breast cancers. Nat Genet 2005;37:899-905.
39. Hilson J, Schnitt S, Collins L. Phenotypic alterations in ductal carcinoma in situ-associated myoepithelial cells: biologic and diagnostic implications. Am J Surg Pathol 2009;33:227-32.
40. Guidi A, Fischer L, Harris J, Schnitt S. Microvessel density and distribution in ductal carcinoma in situ of the breast. J Natl Cancer Inst 1994;86:614-9.
41. Kleer C, Bloushtain-Qimron N, Chen Y, et al. Epithelial and stromal cathepsin K and CXCL14 expression in breast tumor progression. Clin Cancer Res 2008;14:5357-67.
42. Brown L, Guidi A, Schnitt S, et al. Vascular stroma formation in carcinoma in situ, invasive carcinoma, and metastatic carcinoma of the breast. Clin Cancer Res 1999;5:1041-56.
43. Sontag L, Axelrod D. Evaluation of pathways for progression of heterogeneous breast tumors. J Theor Biol 2005;232:179-89.
44. Ernster V, Barclay J, Kerlikowske K, Grady D, Henderson C. Incidence of and treatment for ductal carcinoma in situ of the breast. JAMA 1996;275:913-8.
45. Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun M. Cancer statistics, 2007. CA Cancer J Clin;57:43-66.
46. Rosner D, Bedwani R, Vana J, Baker H, Murphy G. Noninvasive breast carcinoma: results of a national survey by the American College of Surgeons. Ann Surg 1980;192:139-47.
47. Dershaw D, Abramson A, Kinne D. Ductal carcinoma in situ: mammographic findings and clinical implications. Radiology 1989;170:411-5.
48. Holland R, Hendriks J, Vebeek A, Mravunac M, Schuurmans Stekhoven J. Extent, distribution, and mammographic/histological correlations of breast ductal carcinoma in situ. Lancet 1990;335:519-22.
49. Gilles R, Zafrani B, Guinebretière J, et al. Ductal carcinoma in situ: MR imaging-histopathologic correlation. Radiology 1995;196:415-9.
50. Morakkabati-Spitz N, Leutner C, Schild H, Traeber F, Kuhl C. Diagnostic usefulness of segmental and linear enhancement in dynamic breast MRI. Eur Radiol 2005;15:2010-7.
51. Liberman L, Morris E, Dershaw D, Abramson A, Tan L. Ductal enhancement on MR imaging of the breast. AJR Am J Roentgenol 2003;181:519-25.
52. Boughey J, Gonzalez R, Bonner E, Kuerer H. Current treatment and clinical trial developments for ductal carcinoma in situ of the breast. Oncologist 2007;12:1276-87.
53. Smith G, Smith B, Haffty B. Rationalization and regionalization of treatment for ductal carcinoma in situ of the breast. Int J Radiat Oncol Biol Phys 2006;65:1397-403.
54. Fisher B, Bauer M, Margolese R, et al. Five-year results of a randomized clinical trial comparing total mastectomy and segmental mastectomy with or without radiation in the treatment of breast cancer. N Engl J Med 1985;312:665-73.

55. Fisher E, Leeming R, Anderson S, Redmond C, Fisher B. Conservative management of intraductal carcinoma (DCIS) of the breast. Collaborating NSABP investigators. J Surg Oncol 1991;47:139-47.
56. Ernster V, Barclay J, Kerlikowske K, Wilkie H, Ballard-Barbash R. Mortality among women with ductal carcinoma in situ of the breast in the population-based surveillance, epidemiology and end results program. Arch Intern Med 2000;160:953-8.
57. Cutuli B, Cohen-Solal-Le Nir C, De Lafontan B, et al. Ductal carcinoma in situ of the breast results of conservative and radical treatments in 716 patients. Eur J Cancer 2001;37:2365-72.
58. Silverstein M, Barth A, Poller D, et al. Ten-year results comparing mastectomy to excision and radiation therapy for ductal carcinoma in situ of the breast. Eur J Cancer 1995;31A:1425-7.
59. Solin L, Kurtz J, Fourquet A, et al. Fifteen-year results of breast-conserving surgery and definitive breast irradiation for the treatment of ductal carcinoma in situ of the breast. J Clin Oncol 1996;14:754-63.
60. Fisher B, Land S, Mamounas E, Dignam J, Fisher E, Wolmark N. Prevention of invasive breast cancer in women with ductal carcinoma in situ: an update of the National Surgical Adjuvant Breast and Bowel Project experience. Semin Oncol 2001;28:400-18.
61. Houghton J, George W, Cuzick J, Duggan C, Fentiman I, Spittle M. Radiotherapy and tamoxifen in women with completely excised ductal carcinoma in situ of the breast in the UK, Australia, and New Zealand: randomised controlled trial. Lancet 2003;362:95-102.
62. Julien J, Bijker N, Fentiman I, et al. Radiotherapy in breast-conserving treatment for ductal carcinoma in situ: first results of the EORTC randomised phase III trial 10853. EORTC Breast Cancer Cooperative Group and EORTC Radiotherapy Group. Lancet 2000;355:528-33.
63. Practice guideline for the management of ductal carcinoma in-situ of the breast (DCIS). J Am Coll Surg 2007;205:145-61.
64. I S. Conservation therapy of DCIS without radiation. Breast Disease 1996;9:27-36.
65. Boland G, Chan K, Knox W, Roberts S, Bundred N. Value of the Van Nuys Prognostic Index in prediction of recurrence of ductal carcinoma in situ after breast-conserving surgery. Br J Surg 2003;90:426-32.
66. Silverstein M, Lagios M, Groshen S, et al. The influence of margin width on local control of ductal carcinoma in situ of the breast. N Engl J Med 1999;340:1455-61.
67. Wong J, Kaelin C, Troyan S, et al. Prospective study of wide excision alone for ductal carcinoma in situ of the breast. J Clin Oncol 2006;24:1031-6.
68. Rampaul R VP, Pinder S, et al. Wide local excision with 10mm clearance without radiotherapy for DCIS. Eur J Surg Oncol 2001;27.
69. Schnitt S, Harris J, Smith B. Developing a prognostic index for ductal carcinoma in situ of the breast. Are we there yet? Cancer 1996;77:2189-92.

70. Fisher B, Dignam J, Wolmark N, et al. Tamoxifen in treatment of intraductal breast cancer: National Surgical Adjuvant Breast and Bowel Project B-24 randomised controlled trial. Lancet 1999;353:1993-2000.
71. Vogel V, Costantino J, Wickerham D, et al. Effects of tamoxifen vs raloxifene on the risk of developing invasive breast cancer and other disease outcomes: the NSABP Study of Tamoxifen and Raloxifene (STAR) P-2 trial. JAMA 2006;295:2727-41.
72. D A. What is hormone responsive? In: Breast Cancer Symposium September 7-8, 2007; 2007 September 7-8; San Francisco, CA: American Society of Clinical Oncology; 2007.
73. Yen T, Hunt K, Mirza N, et al. Physician recommendations regarding tamoxifen and patient utilization of tamoxifen after surgery for ductal carcinoma in situ. Cancer 2004;100:942-9.
74. Boyages J, Delaney G, Taylor R. Predictors of local recurrence after treatment of ductal carcinoma in situ: a meta-analysis. Cancer 1999;85:616-28.
75. Fisher E, Sass R, Fisher B, Wickerham L, Paik S. Pathologic findings from the National Surgical Adjuvant Breast Project (protocol 6). I. Intraductal carcinoma (DCIS). Cancer 1986;57:197-208.
76. Waldman F, DeVries S, Chew K, Moore Dn, Kerlikowske K, Ljung B. Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences. J Natl Cancer Inst 2000;92:313-20.
77. Bijker N, Peterse J, Duchateau L, et al. Histological type and marker expression of the primary tumour compared with its local recurrence after breast-conserving therapy for ductal carcinoma in situ. Br J Cancer 2001;84:539-44.
78. Virnig B, Tuttle T, Shamliyan T, Kane R. Ductal Carcinoma In Situ of the Breast: A Systematic Review of Incidence, Treatment, and Outcomes. J Natl Cancer Inst 2010.
79. Erbas B, Provenzano E, Armes J, Gertig D. The natural history of ductal carcinoma in situ of the breast: a review. Breast Cancer Res Treat 2006;97:135-44.
80. Schuetz C, Bonin M, Clare S, et al. Progression-specific genes identified by expression profiling of matched ductal carcinomas in situ and invasive breast tumors, combining laser capture microdissection and oligonucleotide microarray analysis. Cancer Res 2006;66:5278-86.
81. Ellis M, Davis N, Coop A, et al. Development and validation of a method for using breast core needle biopsies for gene expression microarray analyses. Clin Cancer Res 2002;8:1155-66.
82. Moeder C, Giltnane J, Moulis S, Rimm D. Quantitative, fluorescence-based in-situ assessment of protein expression. Methods Mol Biol 2009;520:163-75.
83. Bibikova M, Talantov D, Chudin E, et al. Quantitative gene expression profiling in formalin-fixed, paraffin-embedded tissues using universal bead arrays. Am J Pathol 2004;165:1799-807.
84. Coudry R, Meireles S, Stoyanova R, et al. Successful application of microarray technology to microdissected formalin-fixed, paraffin-embedded tissue. J Mol Diagn 2007;9:70-9.

85. Karsten S, Van Deerlin V, Sabatti C, Gill L, Geschwind D. An evaluation of tyramide signal amplification and archived fixed and frozen tissue in microarray gene expression analysis. Nucleic Acids Res 2002;30:E4.
86. Paik S, Kim C, Song Y, Kim W. Technology insight: Application of molecular techniques to formalin-fixed paraffin-embedded tissues from breast cancer. Nat Clin Pract Oncol 2005;2:246-54.
87. Kerkhoven R, Sie D, Nieuwland M, et al. The T7-primer is a source of experimental bias and introduces variability between microarray platforms. PLoS One 2008;3:e1980.
88. Sjoblom T, Jones S, Wood LD, et al. The consensus coding sequences of human breast and colorectal cancers. Science 2006;314:268-74.
89. Wood LD, Parsons DW, Jones S, et al. The genomic landscapes of human breast and colorectal cancers. Science 2007;318:1108-13.
90. Sjöblom T. Systematic analyses of the cancer genome: lessons learned from sequencing most of the annotated human protein-coding genes. Curr Opin Oncol 2008;20:66-71.
91. Volik S, Zhao S, Chin K, et al. End-sequence profiling: sequence-based analysis of aberrant genomes. Proc Natl Acad Sci U S A 2003;100:7696-701.
92. Hodges E, Xuan Z, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007;39:1522-7.
93. Farnie G, Clarke RB. Mammary stem cells and breast cancer--role of Notch signalling. Stem Cell Rev 2007;3:169-75.
94. Farnie G, Clarke RB, Spence K, et al. Novel cell culture technique for primary ductal carcinoma in situ: role of Notch and epidermal growth factor receptor signaling pathways. J Natl Cancer Inst 2007;99:616-27.
95. Nishidate T, Katagiri T, Lin M, et al. Genome-wide gene-expression profiles of breast-cancer cells purified with laser microbeam microdissection: identification of genes associated with progression and metastasis. Int J Oncol 2004;25:797-819.
96. Masuda N, Ohnishi T, Kawamoto S, Monden M, Okubo K. Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples. Nucleic Acids Res 1999;27:4436-43.
97. Stephens P, Edkins S, Davies H, et al. A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nat Genet 2005;37:590-2.
98. Greenman C, Stephens P, Smith R, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446:153-8.
99. Wang Z, Shen D, Parsons D, et al. Mutational analysis of the tyrosine phosphatome in colorectal cancers. Science 2004;304:1164-6.
100. Samuels Y, Wang Z, Bardelli A, et al. High frequency of mutations of the PIK3CA gene in human cancers. Science 2004;304:554.
101. Colicelli J. Human RAS superfamily proteins and related GTPases. Sci STKE 2004;2004:RE13.
102. Järvinen A, Hautaniemi S, Edgren H, et al. Are data from different gene expression microarray platforms comparable? Genomics 2004;83:1164-8.
103. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 2005;21:171-8.

104. Droege M, Hill B. The Genome Sequencer FLX System--longer reads, more applications, straight forward bioinformatics and more complete data sets. J Biotechnol 2008;136:3-10.
105. Fox S, Filichkin S, Mockler TC. Applications of ultra-high-throughput sequencing. Methods Mol Biol 2009;553:79-108.

Appendix

Appendix 1. TaqMan Gene Expression Assays with reference ID (Applied Biosystems, Inc.). All assays run under protocol supplied with common annealing temperature of 60°C. IDs ending in -s1 refer to amplicon targets within an exon which will detect gDNA as well as cDNA. -m1 suffixes refer to amplicon targets spanning introns thus detecting only cDNA and not gDNA. -g1 refer to targets spanning small introns which may still detect gDNA.

Gene	Reference ID
ARMCX2	Hs01932946_s1
CHEK2	Hs01007278_g1
EGFR1	Hs00152928_m1
ERBB2	Hs01001598_g1
FGFR	Hs00259959_s1
GAPDH	Hs00266705_g1
GAPDH	Hs99999905_m1
IDH1	Hs01855675_s1
IGF1R	Hs00541255_s1
PTEN	Hs00829813_s1
TP53	Hs00153349_m1

Appendix 2. Invitrogen primers used for DNA sequencing. Primers include a 454 linker allowing Sanger pyrosequencing at an affiliated core facility as described.

Gene	Exon	Amplicon Length (base pairs)	Tm*	Annealing T*
ERBB4	18	123	67.3,66.9	54.9
ERBB4	19	99	69.3,68.8	
ERBB4	20	186	68.7, 69.0	54.9
ERBB4	21	156	68.3, 69.2	54.9
ERBB4	22	76	68.1, 67.4	52.4
ERBB4	23	147	65.0, 66.9	55.3

**All temperatures in celsius.*

Appendix 3. Gene list used for exon capture design. Transcripts employed are listed according to RefSeq and/or CCDS accession number. UCSC Table Browser was used to generate exon coordinates based on these inputs. 874 genes are listed here derived from melanoma and breast cancer gene lists from the literature as described.

Gene	***CCDS/Accession***	***Name/Accession***
ABCA3	NM_001089	ATP-binding cassette, sub-family A (ABC1), member 3
ABCB10	NM_012089	ATP-binding cassette, sub-family B (MDR/TAP), member 10
ABCB8	NM_007188	ATP-binding cassette, sub-family B (MDR/TAP), member 8
ABI1	NM_005470	spectrin SH3 domain binding protein 1
ABP1	NM_001091	
ACADM	NM_001127328	acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain
ACAN	NM_001135	formerly AGC1
ACCS	NM_032592	1-aminocyclopropane-1-carboxylate synthase ; formerly PHACS
ACP1	NM_004300	Acid phosphatase 1, soluble
ACSL6	NM_001009185	fatty-acid-coenzyme A ligase, long-chain 6; FACL6
ADAM12	NM_003474	ADAM metallopeptidase domain 12
ADRBK2	NM_005160	
AFF1	NM_005935	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 2 (AF4)
AFF3	NM_002285	lymphoid nuclear protein related to AF4; formerly LAF4
AFF4	NM_014423	ALL1 fused gene from 5q31 (formerly AF5q31)
AGT	NM_000029	
AIM1	NM_001624	
Alk	NM_004304	
ALPK2	NM_052947	HAK
ALPK3	NM_020778	MIDORI
ALS2CL	NM_147129	ALS2 C-terminal like
AMFR	NM_001144	autocrine motility factor receptor
ARAF	NM_001654	
ARFGAP1	NM_175609	RPA human homolog
ARFGEF2	NM_006420	
ARHGAP25	NM_001007231	
ARHGAP26	NM_015071	Rho GTPase activating protein 26;Formerly known as GRAF or GTPase regulator associated with focal adhesion kinase pp125(FAK)
ARHGEF12	NM_015313	RHO guanine nucleotide exchange factor (GEF) 12 (LARG)
ARHGEF4	NM_032995	Rho guanine nucleotide exchange factor (GEF) 4
ARL4C	NM_025144	LAK
ARNTL	NM_001030273	Aryl hydrocarbon receptor nuclear translocator-like; formerly BMAL1
ASL	NM_001024943	argininosuccinate lyase
ATF1	NM_005171	activating transcription factor 1
ATIC	NM_004044	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase
ATN1	NM_001940	
ATP8B1	NM_005603	ATPase, Class I, type 8B, member 1
ATR	NM_001184	
AURKA	NM_198433	
AURKB	NM_004217	
AURKC	NM_001015878	
AXL	NM_001699	
BAI1	NM_001702	Brain-specific angiogenesis inhibitor 1

BAT2	NM_080686	
BAX	NM_138764	BCL2-associated X protein
BCL10	NM_003921	B-cell CLL/lymphoma 10
BCL11A	NM_022893	B-cell CLL/lymphoma 11A
BCL11B	NM_138576	B-cell CLL/lymphoma 11B (CTIP2)
BCL2	NM_000633	B-cell CLL/lymphoma 2
BCL3	NM_005178	B-cell CLL/lymphoma 3
BCL6	NM_001706	B-cell CLL/lymphoma 6
BCL7A	NM_020993	B-cell CLL/lymphoma 7A
BCL9	NM_004326	B-cell CLL/lymphoma 9
BCR	NM_004327	breakpoint cluster region
BGN	NM_001711	biglycan
BIRC3	NM_001165	baculoviral IAP repeat-containing 3
BLK	NM_001715	
BLM	NM_000057	Bloom Syndrome
BMP2K	NM_198892	formerly BIKE
BMPR1A	NM_004329	bone morphogenetic protein receptor, type IA
BMPR1B	NM_001203	
BMPR2	NM_001204	
BRAF	NM_004333	v-raf murine sarcoma viral oncogene homolog B1
BRCA1	NM_007295	breast cancer 1, early onset
BRCA2	NM_000059	familial breast/ovarian cancer gene 2
BRD2	NM_005104	NM_005104
BRD3	NM_007371	NM_007371
BRDT	NM_001726	
BRIP1	NM_032043	BRCA1 interacting protein C-terminal helicase 1
BRSK1	NM_032430	
BTK	NM_000061	
C14orf100	NM_016475	
C15orf55	NM_175741	nuclear protien in testis; formerly NUT
C1orf64	NM_178840	chromosome 1 open reading frame 64; aka MGC24047
C5orf42	NM_023073	FLJ13231
C9ORF96	NM_153710	MGC43306
CACNA1F	NM_005183	
CAMK1	NM_003656	
CAMK1D	NM_020397	
CAMK1G	NM_020439	
CAMK2A	NM_015981	
CAMK2B	NM_001220	
CAMK2G	NM_001222	
CAMK4	NM_001744	
CAMKK2	NM_006549	
CAMKV	NM_024046	MGC8407
CARD11	NM_032415	caspase recruitment domain family, member 11
CASK	NM_003688	
CATSPERB	NM_024764	C14orf161
CCDC6	NM_005436	DNA segment on chromosome 10 (unique) 170, H4 gene (PTC1); D10S170
CCNB1IP1	NM_182851	cyclin B1 interacting protein 1;formerly HEI10 or enhancer of invasion 10 - fused to HMGA2
CCND1	NM_053056	cyclin D1
CCND2	NM_001759	cyclin D2

CCND3	NM_001760	cyclin D3
CCNL1	NM_020307	PRO1073 protein (ALPHA)
CD46	NM_002389	CD46 molecule, complement regulatory protein; formerly MCP
CD93	NM_012072	C1QR1
CD97	NM_078481	
CDC42BPA	NM_014826	
CDC42BPB	NM_006035	
CDC42BPG	NM_017525.1	DMPK2
CDC7	NM_003503	cdc7l1
CDC73	NM_024529	hyperparathyroidism 2 formerly HRPT2
CDH1	NM_004360	cadherin 1, type 1, E-cadherin (epithelial) (ECAD)
CDH10	NM_006727	cadherin 10, type 2 (T2-cadherin)
CDH11	NM_001797	cadherin 11, type 2, OB-cadherin (osteoblast)
CDH20	NM_031891	cadherin 20, type 2
CDK2	NM_001798	
CDK3	NM_001258	
CDK4	NM_000075	
CDK6	NM_001259	
CDK8	NM_001260	
CDKL2	NM_003948	
CDKL3	NM_016508	
CDKL5	NM_003159	
CDKN2A	NM_000077	cyclin-dependent kinase inhibitor 2A (p16(INK4a)) gene
CEBPA	NM_004364	CCAAT/enhancer binding protein (C/EBP), alpha
CENTB1	NM_014716	centaurin, beta 1
CENTG1	NM_001122772	centaurin, gamma 1
CEP110	NM_007018	centrosomal protein 110kda aka cep1
CFP	NM_002621	complement factor properdin, PFC
CHD5	NM_001795	chromodomain helicase DNA binding protein 5
CHEK2	NM_001005735	CHK2 checkpoint homolog (S. pombe)
CHUK	NM_001278	
CIC	NM_015125	capicua homolog
CIITA	NM_000246	MHC class II transactivator; formerly MHC2TA
CIT	NM_007174	
CLCN3	NM_001829	
CLP1	NM_006831	CLP1, cleavage and polyadenylation factor I subunit, homolog (S. cerevisiae); formerly HEAB or ATP_GTP binding protein
CNBP	NM_001127192	zinc finger protein 9 (a cellular retroviral nucleic acid binding protein)
CNNM4	NM_020184	cyclin M4
CNTN3	NM_020872	
CNTN6	NM_014461	contactin 6
COL11A1	NM_080629	collagen, type XI, alpha 1
COL19A1	NM_001858	collagen, type XIX, alpha 1
COL1A1	NM_000088	collagen, type I, alpha 1
COL7A1	NM_000094	collagen, type VII, alpha 1
COX6C	NM_004374	cytochrome c oxidase subunit VIc
CREB1	NM_134442	cAMP responsive element binding protein 1
CREB3L2	NM_194071	cAMP responsive element binding protein 3-like 2
CREBBP	NM_004380	CREB binding protein (CBP)
CRKRS	NM_016507	CRK7

CRTC1	NM_001098482	CREB regulated transcription coactivator 1; formerly MECT1 mucoepidermoid translocated 1
CRY1	NM_004075	Cryptochrome 1 (photolyase-like)
CRY2	NM_021117	Cryptochrome 2 (photolyase-like)
CSF1R	NM_005211	
CSH2	NM_022644	Chorionic somatomammotropin hormone 2
CSK	NM_004383	
CSNK1A1	NM_001892	
CSNK1D	NM_001893	
CSNK1E	NM_001894	
CSPP1	NM_024790	
CUBN	NM_001081	cubilin (intrinsic factor-cobalamin receptor)
CUX1	NM_001913	CUTL1
CXCR7	NM_020311	chemokine orphan receptor 1
CYB5R4	NM_016230	cytochrome b5 reductase 4; formerly ncb5or
CYP1A1	NM_000499	cytochrome P450, family 1, subfamily A, polypeptide 1
DAPK1	NM_004938	
DAPK2	NM_014326	
DAPK3	NM_001348	
DBN1	NM_080881	drebrin 1
DCLK1	AF052152	formerly dcamkl1
DCLK3	NM_033403	formerly dcamkl3
DDB1	NM_001923	Damage-specific DNA binding protein 1, 127kDa
DDB2	NM_000107	damage-specific DNA binding protein 2
DDIT3	NM_004083	DNA-damage-inducible transcript 3
DDR1	NM_013993	
DDR2	NM_006182	
DDX10	NM_004398	DEAD (Asp-Glu-Ala-Asp) box polypeptide 10
DEK	NM_003472	DEK oncogene (DNA binding)
DHH	NM_021044	Desert hedgehog homolog (Drosophila)
DIP2C	NM_014974	DIP2 disco-interacting protein 2 homolog C; aka KIAA0934
DIRAS1	NM_145173	DIRAS family, GTP-binding RAS-like 1
DIRAS2	NM_017594	DIRAS family, GTP-binding RAS-like 2
DMPK	NM_004409	
DNAH9	NM_001372	dynein, axonemal, heavy polypeptide 9
DNAJC24	NM_181706	DPH4, JJJ3 homolog (S. cerevisiae); formerly ZCSL3 zinc finger, CSL-type containing 3
DNASE1L3	NM_004944	deoxyribonuclease I-like 3
DPAGT1	NM_001382	
DPYD	NM_000110	
DUSP7	NM_001947	
DYRK1B	NM_006484	
DYRK2	NM_006482	
DYRK3	NM_003582	
DYRK4	NM_003845	
E2F1	NM_005225	E2F transcription factor 1
EEF2K	NM_013302	
EGFL6	NM_015507	EGF-like-domain, multiple 6
EHMT1	NM_024757	euchromatic histone-lysine N-methyltransferase 1
EIF2AK2	NM_002759	PRKR
EIF2AK4	NM_001013703	

ELN	NM_000501	elastin
EMR1	NM_001974	Egf-like module containing, mucin-like, hormone receptor-like 1
EPHA1	NM_005232	
EPHA10	NM_001004338	
EPHA2	NM_004431	
EPHA4	NM_004438	
EPHA5	NM_004439	
EPHA6	XM_114973	
EPHA7	NM_004440	
EPHA8	NM_020526	
EPHB1	NM_004441	
EPHB2	NM_017449	
EPHB3	NM_004443	
EPHB4	NM_004444	
EPS15	NM_001981	epidermal growth factor receptor pathway substrate 15 (AF1p)
ERAS	NM_181532	ES cell expressed Ras
ERBB2	NM_004448	
ERBB3	NM_001982	
ERBB4	NM_005235	
ERC1	NM_178037	ELKS protein
ERCC2	NM_000400	excision repair cross-complementing rodent repair deficiency, complementation group 2 (xeroderma pigmentosum D)
ERCC3	NM_000122	excision repair cross-complementing rodent repair deficiency, complementation group 3 (xeroderma pigmentosum group B complementing)
ERCC4	NM_005236	excision repair cross-complementing rodent repair deficiency, complementation group 4
ERCC5	NM_000123	excision repair cross-complementing rodent repair deficiency, complementation group 5 (xeroderma pigmentosum, complementation group G (Cockayne syndrome))
ERCC8	NM_000082	Excision repair cross-complementing rodent repair deficiency, complementation group 8; formerly ERCC8
ERG	NM_004449	v-ets erythroblastosis virus E26 oncogene like (avian)
ERGIC3	NM_198398	ERGIC and golgi 3; formerly ERGIC and golgi 3
ERICH1	NM_207332	glutamate-rich 1; formerly LOC157697
ERN1	NM_001433	
ERN2	NM_033266	
ETV1	NM_004956	ets variant gene 1
ETV4	NM_001079675	ets variant gene 4 (E1A enhancer binding protein, E1AF)
ETV5	NM_004454	ets variant gene 5
ETV6	NM_001987	ets variant gene 6 (TEL oncogene)
EVI1	NM_001105078	ecotropic viral integration site 1
EXOC3L	NM_178516	exocyst complex component 3-like ; formerly hypothetical protein LOC283849
EXOC4	NM_021807	exocyst complex component 4; formerly sec8l1
FAM123B	NM_152424	family with sequence similarity 123B formerly WTX
FAM161A	NM_032180	FLJ13305
FAM171B	NM_177454	KIAA1946
FANCA	NM_000135	Fanconi anemia, complementation group A
FANCB	NM_001018113	Fanconi anemia, complementation group B

FANCC	NM_000136	Fanconi anemia, complementation group C
FANCD2	NM_033084	Fanconi anemia, complementation group D2
FANCE	NM_021922	Fanconi anemia, complementation group E
FANCF	NM_022725	Fanconi anemia, complementation group F
FANCG	NM_004629	Fanconi anemia, complementation group G
FANCI	NM_001113378	Fanconi anemia, complementation group I
FANCL	NM_001114636	Fanconi anemia, complementation group L
FANCM	NM_020937	Fanconi anemia, complementation group M
FAS	NM_000043	tumor necrosis factor receptor superfamily, member 6 (FAS) (TNFRSF6)
FASTK	NM_006712	NM_025096
FCRL4	NM_031282	Fc receptor-like 4; formerly IRTA1 immunoglobulin superfamily receptor translocation associated 1
FCRL5	NM_031281	Fc receptor-like 5
FEN1	NM_004111	
FER	NM_005246	
FGFR1	NM_000604	
FGFR1OP	NM_007045	FGFR1 oncogene partner (FOP)
FGFR2	NM_022970	
FGFR3	NM_000142	
FGFR4	NM_022963	
FGR	NM_005248	
FH	NM_000143	fumarate hydratase
FLCN	NM_144997	folliculin, Birt-Hogg-Dube syndrome; formerly BHD
FLJ23356	NM_032237	
FLNA	NM_001456	
FLNB	NM_001457	filamin B, beta
FLT1	NM_002019	
FLT3	NM_004119	
FLT4	NM_002020	
FOXO1	NM_002015	forkhead box O1A (FKHR)
FOXO3	NM_001455	forkhead box O3A
FOXO4	NM_005938	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 7 (AFX1)
FREM1	NM_144966	
FRK	NM_002031	
FSCB	NM_175741	chromosome 14 open reading frame 155
FUS	NM_004960	fusion, derived from t(12;16) malignant liposarcoma
FYN	NM_002037	
GAB1	AK074381	GRB2-associated binding protein 1
GAK	NM_005255	
GALNT17	NM_001034845	
GALNT5	NM_014568	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 5
GEM	NM_005261	GTP binding protein overexpressed in skeletal muscle
GEN1	NM_182625	hypothetical protein FLJ40869
GGA1	NM_013365	golgi associated, gamma adaptin ear containing, ARF binding protein 1
GJD4	NM_153368	CX40.1
GLI1	NM_005269	glioma-associated oncogene homolog 1
GLI2	NM_005270	GLI-Kruppel family member GLI2
GPNMB	NM_001005340	glycoprotein (transmembrane) nmb
GPR64	NM_005756	G protein-coupled receptor 64; formerly HE6

GRIN2D	NM_000836	glutamate receptor, ionotropic, N-methyl D-aspartate 2D
GRK1	NM_002929	RHOK
GRK4	NM_005307	
GRK5	NM_005308	GPRK5
GRK6	NM_002082	
GRK7	NM_139209	
GSN	NM_000177	gelsolin
GUCY2C	NM_004963	
GUCY2D	NM_000180	
GUCY2F	NM_001522	
HCK	NM_002110	
HDAC4	CCDS2529.1	histone deacetylase 4
HDLBP	CCDS2547.1	high density lipoprotein binding protein
HIPK1	NM_152696	
HIPK2	NM_022740	
HIPK3	NM_005734	
HNF1A	NM_000545	HNF1 Homeobox A;transcription factor 1, hepatic; formerly tcf1
HOXA3	NM_153631	homeobox A3
HRAS	NM_176795	v-Ha-ras Harvey rat sarcoma viral oncogene homolog
HSP90AB1	BG393867	heat shock 90kDa protein 1, beta
HSPB8	NM_014365	H11
ICAM5	NM_003259	intercellular adhesion molecule 5, telencephalin
ICK	NM_016513	
IGF1R	NM_000875	
IHH	NM_002181	Indian hedgehog homolog (Drosophila)
IKBKB	NM_001556	
IKZF1	AF432219	zinc finger protein, subfamily 1A, 1 (Ikaros)
IL2	NM_000586	interleukin 2
IL21R	NM_181078	interleukin 21 receptor
INHBE	NM_031479	inhibin, beta E
INPP4A	NM_001566	Inositol polyphosphate-4-phosphatase, type I, 107kDa
INPP4B	NM_003866	Inositol polyphosphate-4-phosphatase, type II, 105kDa
INPP5B	NM_005540	Inositol polyphosphate-5-phosphatase, 75kDa
INPP5E	NM_019892	Inositol polyphosphate-5-phosphatase, 72 kDa
INSR	NM_000208	
INSRR	NM_014215	
IRAK1	NM_001569	
IRAK2	NM_001570	
IRAK3	NM_007199	
IRF4	NM_002460	interferon regulatory factor 4
ITGA9	NM_002207	integrin, alpha 9
ITK	NM_005546	
JAK1	NM_002227	
JAK2	NM_004972	
JAK3	NM_000215	
JARID1B	NM_006618	
JAZF1	NM_175061	juxtaposed with another zinc finger gene 1
JTV1	NM_014413	HRI
KALRN	NM_007064	
KDR	NM_002253	

KDSR	NM_002035	3-ketodihydrosphingosine reductase; follicular lymphoma variant translocation 1 aka FVT1
KEAP1	NM_203500	kelch-like ECH-associated protein 1
KIAA0427	NM_014772	KIAA0427
KIAA0467	NM_015284	
KIAA0664	NM_015229	
KIAA0774	NM_001033602	
KIAA0999	NM_025164	KIAA0999 protein
KIAA1632	NM_020964	KIAA1632
KIF14	NM_014875	
KIT	NM_000222	
KLF6	NM_001300	core promoter element binding protein (KLF6); formerly COPEB
KPNA5	NM_002269	karyopherin alpha 5
KRT73	NM_175068	keratin 6 irs3 formerly k6irs3
KSR1	NM_014238	
KSR2	NM_173598	
KTN1	NM_182926	kinectin 1
LASP1	NM_006148	LIM and SH3 protein 1
LATS1	NM_004690	
LATS2	NM_014572	
LCK	NM_005356	
Lig1	NM_000234	
Lig3	NM_013975	
LIMK1	NM_016735	
LMTK2	NM_014916	Lemur tyrosine kinase 2
LMTK3	NM_001080434	Lemur tyrosine kinase 3
LRBA	NM_006726	LPS-responsive vesicle trafficking, beach and anchor containing
LRRC7	NM_020794	
LRRFIP1	NM_004735	leucine rich repeat (in FLII) interacting protein 1
LRRK1	NM_024652	FLJ23119
LRRK2	XM_058513	
LTK	NM_002344	
LYN	NM_002350	
MACF1	NM_033044	microtubule-actin crosslinking factor 1
MAGEE1	NM_020932	melanoma antigen family E, 1
MAK	NM_005906	
MAMDC4	NM_206920	MAM domain containing 4
MAML2	NM_032427	mastermind-like 2 (Drosophila)
MAP2K4	NM_003010	
MAP2K7	NM_005043	
MAP3K1	XM_042066	
MAP3K10	NM_002446	
MAP3K11	NM_002419	
MAP3K12	NM_006301	
MAP3K13	NM_004721	
MAP3K14	NM_003954	
MAP3K15	NM_001001671	SK681
MAP3K2	NM_006609	
MAP3K3	NM_002401	
MAP3K4	NM_005922	
MAP3K6	NM_004672	mitogen-activated protein kinase kinase kinase 6

MAP3K9	NM_033141	
MAP4K1	NM_007181	
MAP4K3	NM_003618	
MAP4K4	NM_145686	
MAP4K5	NM_006575	
MAPK7	NM_002749	
MAPK8	NM_002750	
MAPK9	NM_002752	
MAPKAPK3	NM_004635	
MAPKBP1	NM_014994	
MARK1	NM_018650	
MARK2	NM_017490	
MARK3	NM_002376	
MARK4	NM_031417	
MAST1	NM_014975	
MAST2	NM_015112	
MAST3	XM_038150	
MAST4	XM_291141	
MASTL	NM_032844	
MATK	NM_139355	
MBD4	NM_003925	
MC1R	NM_002386	Melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor)
MDM2	NM_002392	Mdm2 p53 binding protein homolog (mouse)
MDS1	NM_004991	myelodysplasia syndrome 1
MELK	NM_014791	
MEN1	NM_130803	multiple endocrine neoplasia type 1 gene
MERTK	NM_006343	
MET	NM_000245	
MGC16169	NM_033115	TBCK
MGC42105	NM_153361	
MGMT	NM_002412	
MINK1	NM_015716	
MLH1	NM_000249	E.coli MutL homolog gene
MLKL	NM_152649	
MLL	NM_005933	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila)
MN1	NM_002430	meningioma (disrupted in balanced translocation) 1
MNX1	NM_005515	homeo box HB9
MOS	NM_005372	
MRAS	NM_012219	Muscle RAS oncogene homolog
MRE11A	NM_005590	MRE11 meiotic recombination 11 homolog A
MSH2	NM_000251	mutS homolog 2 (E. coli)
MSH4	NM_002440	MutS homolog 4 (E. coli)
MSH6	NM_000179	mutS homolog 6 (E. coli)
MST1R	NM_002447	
MST4	NM_016542	
MTMR14	NM_001077525	Myotubularin related protein 14
MTMR3	NM_021090	myotubularin related protein 3
MUSK	NM_005592	
MYC	NM_002467	v-myc myelocytomatosis viral oncogene homolog (avian)
MYH1	NM_005963	myosin, heavy polypeptide 1, skeletal muscle, adult

MYLK	NM_053025	
MYLK4	XM_373109	SgK085
MYO19	NM_001033579	MYOHD1
MYO1G	NM_033054	
MYO3A	NM_017433	
MYO3B	NM_138995	
MYOD1	NM_002478	myogenic differentiation 1
MYST3	NM_001099412	MYST histone acetyltransferase (monocytic leukemia) 3;formerly Runxbp2 runt-related transcription factor binding protein 2 (MOZ/ZNF220)
MYST4	NM_012330	MYST histone acetyltransferase (monocytic leukemia) 4 (MORF)
NBN	NM_002485	Nijmegen breakage syndrome 1 (nibrin)
NCOA6	NM_014071	nuclear receptor coactivator 6
NEIL1	NM_024608	Nei endonuclease VIII-like 1 (E. coli)
NEIL2	NM_145043	Nei like 2 (E. coli)
NEIL3	NM_018248	Nei endonuclease VIII-like 3 (E. coli)
NEK1	NM_012224	
NEK10	NM_152534	
NEK11	NM_024800	
NEK4	NM_003157	
NEK6	NM_014397	
NEK7	NM_133494	
NEK8	NM_178170	
NEK9	NM_033116	
NF2	NM_181832	neurofibromatosis type 2 gene
NKIRAS1	NM_020345	NFKB inhibitor interacting Ras-like 1
NKIRAS2	NM_001001349	NFKB inhibitor interacting Ras-like 2
NLE1	NM_001014445	notchless homolog 1, aka FLJ10458
NLK	NM_016231	
NLRP8	NM_176811	NACHT, leucine rich repeat and PYD containing 8; formerly NALP8
NOTCH1	NM_017617	Notch homolog 1, translocation-associated (Drosophila) (TAN1)
NPAT	NM_000051	ataxia telangiectasia mutated, formerly ATM
NPM1	NM_002520	nucleophosmin (nucleolar phosphoprotein B23, numatrin)
NPR1	NM_000906	
NR4A3	NM_173198	nuclear receptor subfamily 4, group A, member 3 (NOR1)
NRAS	NM_002524	neuroblastoma RAS viral (v-ras) oncogene homolog
NRBP1	NM_013392	
NRBP2	NM_178564	
NRCAM	NM_001037132	neuronal cell adhesion molecule
NTRK1	NM_002529	
NTRK2	NM_006180	
NTRK3	NM_002530	
NUAK2	NM_030952	SNARK
NUMA1	NM_006185	nuclear mitotic apparatus protein 1
NUP133	NM_018230	nucleoporin 133kDa
NUP214	NM_005085	nucleoporin 214kDa
OCA2	NM_000275	Oculocutaneous albinism II
OCRL	NM_000276	Oculocerebrorenal syndrome of Lowe
OGG1	NM_016819	8-oxoguanine DNA glycosylase
OSR1	NM_005109	
OTOF	NM_194248	otoferlin
PAK1	NM_002576	

PAK3	NM_002578	
PAK6	NM_020168	
PAK7	NM_020341	
PALB2	NM_024675	partner and localizer of BRCA2
PARG	NM_003631	
PARP1	NM_001618	
PARP2	NM_005484	
PASK	NM_015148	
PATZ1	NM_014323	POZ (BTB) and AT hook containing zinc finger 1; formerly ZNF278
PAX3	NM_181458	paired box gene 3
PAX5	NM_016734	paired box gene 5 (B-cell lineage specific activator protein)
PAX7	NM_002584	paired box gene 7
PAX8	NM_003466	paired box gene 8
PBX1	NM_002585	pre-B-cell leukemia transcription factor 1
PCDHB15	NM_018935	protocadherin beta 15
PCM1	NM_006197	pericentriolar material 1 (PTC4)
PCNA	NM_002592	Proliferating cell nuclear antigen
PCTK2	NM_002595	
PDCD11	NM_014976	
PDGFB	NM_002608	platelet-derived growth factor beta polypeptide (simian sarcoma viral (v-sis) oncogene homolog)
PDGFRA	NM_006206	
PDGFRB	NM_002609	
PDIK1L	NM_152835	
PDK2	NM_002611	
PDK3	NM_005391	
PDPK1	NM_002613	
PDXK	NM_003681	PNK
PER1	NM_002616	period homolog 1 (Drosophila)
PER2	NM_022817	Period homolog 2 (Drosophila)
PER3	NM_016831	Period homolog 3 (Drosophila)
PFTK1	NM_012395	
PHKG1	NM_006213	
PICALM	NM_007166	phosphatidylinositol binding clathrin assembly protein (CALM)
PIK3C2A	NM_002645	CPK, PI3-K-C2A, PI3K-C2alpha
PIK3C2B	NM_002646	C2-PI3K, PI3K-C2beta
PIK3C2G	NM_004570	PI3K-C2-gamma
PIK3C3	NM_002647	Vps34
PIK3CA	NM_006218	p110-alpha
PIK3CB	NM_006219	PIK3C1, p110-beta
PIK3CD	NM_005026	p110-delta
PIK3CG	NM_002649	PI3CG, PI3K-gamma
PIK3R4	NM_014602	
PIM1	NM_002648	
PIM2	NM_006875	
PINK1	NM_032409	
PKDREJ	NM_006071	polycystic kidney disease (polycystin) and REJ (sperm receptor for egg jelly homolog, sea urchin)-like
PKN1	NM_002741	PRKCL1
PKN2	NM_006256	PRKCL2
PKN3	NM_013355	

PLD2	NM_002663	
PLEKHA8	NM_032639	pleckstrin homology domain containing, family A
PLK1	NM_005030	
PLK2	NM_006622	SNK
PLK3	NM_004073	
PML	NM_033238	promyelocytic leukemia
PMS1	NM_000534	PMS1 postmeiotic segregation increased 1 (S. cerevisiae)
PMS2	NM_000535	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)
PNCK	NM_198452	CaMK1b
POLB	NM_002690	
PPAP2B	NM_003713	Phosphatidic acid phosphatase type 2B
PPM1E	NM_014906	protein phosphatase 1E
PRDM16	NM_022114	PR domain containing 16
PRKAA2	NM_006252	
PRKACB	NM_002731	
PRKAR1A	NM_212472	protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1)
PRKCA	NM_002737	
PRKCB	NM_002738	
PRKCG	NM_002739	
PRKCH	NM_006255	
PRKCI	NM_002740	
PRKCQ	NM_006257	
PRKCZ	NM_002744	
PRKD2	NM_016457	
PRKD3	NM_005813	
PRKDC	NM_006904	
PRKG1	NM_006258	
PRKG2	NM_006259	
PRMT1	NM_198319	
PRMT6	NM_018137	
PRPF4B	NM_003913	PRP4 pre-mRNA processing factor 4 homolog B
PRPS1	NM_002764	phosphoribosyl pyrophosphate synthetase 1
PRRX1	NM_006902	paired mesoderm homeo box 1
PRUNE2	NM_015225	KIAA0367
PSIP1	NM_033222	PC4 and SFRS1 interacting protein 2 (LEDGF)
PSKH2	NM_033126	
PTCH1	NM_001083602	Homolog of Drosophila Patched gene
PTEN	NM_000314	phosphatase and tensin homolog gene
PTK2	NM_005607	
PTK2B	NM_004103	
PTK6	NM_005975	
PTK7	NM_002821	
PTPN13	NM_080685	
PTPN14	NM_005401	protein tyrosine phosphatase, non-receptor type 14
PTPN18	NM_014369	
PTPN20B	NM_015605.2	DKFZP566K0524
PTPN3	NM_002829.2	
PTPRF	NM_002840	
PTPRG	NM_002841	
PTPRT	NM_133170	

RABEP1	NM_004703	rabaptin, RAB GTPase binding effector protein 1 (RABPT5)
RAD23B	NM_002874	RAD23 homolog B (S. cerevisiae)
Rad50	NM_005732	
Rad51	NM_002875	
RAD51L1	NM_133509	RAD51-like 1 (S. cerevisiae) (RAD51B)
RAD52	NM_134424	
RAF1	NM_002880	
RAGE	NM_014226	
RALA	NM_005402	V-ral simian leukemia viral oncogene homolog A (ras related)
RALB	NM_002881	V-ral simian leukemia viral oncogene homolog B (ras related; GTP binding protein)
RALBP1	NM_006788	RalA binding protein 1
RAP1A	NM_001010935	RAP1A, member of RAS oncogene family
RAP1B	NM_015646	RAP1B, member of RAS oncogene family
RAP1GAP	NM_002885	RAP1 GTPase activating protein
RAP2A	NM_021033	RAP2A, member of RAS oncogene family
RAP2B	NM_002886	RAP2B, member of RAS oncogene family
RAP2C	NM_021183	RAP2C, member of RAS oncogene family
RAPH1	NM_213589	Ras association (RalGDS/AF-6) and pleckstrin homology domains 1
RASAL2	NM_170692	RAS protein activator like 2
RASD1	NM_016084	RAS, dexamethasone-induced 1
RASD2	NM_014310	RASD family, member 2
RASGRF2	NM_006909	Ras protein-specific guanine nucleotide-releasing factor 2
RASL10A	NM_001007279	RAS-like, family 10, member A
RASL10B	NM_033315	RAS-like, family 10, member B
RASL11A	NM_206827	RAS-like, family 11, member A
RASL11B	NM_023940	RAS-like, family 11, member B
RASL12	NM_016563	RAS-like, family 12
RASSF1	NM_170714	Ras association (RalGDS/AF-6) domain family member 1
RASSF10	NM_001080521	Ras association (RalGDS/AF-6) domain family (N-terminal) member 10
RASSF2	NM_014737	Ras association (RalGDS/AF-6) domain family member 2
RASSF3	NM_178169	Ras association (RalGDS/AF-6) domain family member 3
RASSF4	NM_032023	Ras association (RalGDS/AF-6) domain family member 4
RASSF5	NM_182663	Ras association (RalGDS/AF-6) domain family member 5
RASSF6	NM_201431	Ras association (RalGDS/AF-6) domain family member 6
RASSF7	NM_003475	Ras association (RalGDS/AF-6) domain family (N-terminal) member 7
RASSF8	NM_007211	Ras association (RalGDS/AF-6) domain family (N-terminal) member 8
RASSF9	NM_005447	Ras association (RalGDS/AF-6) domain family (N-terminal) member 9
RB1	NM_000321	retinoblastoma gene
RBL1	NM_002895	Retinoblastoma-like p107
RBL2	NM_005611	Retinoblastoma-like 2 (p130)
RBM15	NM_022768	RNA binding motif protein 15
REC8	NM_001048205	
RECQL4	NM_004260	RecQ protein-like 4
REM1	NM_014012	RAS (RAD and GEM)-like GTP-binding 1
REM2	NM_173527	RAS (RAD and GEM)-like GTP binding 2
RERG	NM_032918	RAS-like, estrogen-regulated, growth inhibitor
RFX2	NM_000635	regulatory factor X, 2
RGL1	NM_015149	ral guanine nucleotide dissociation stimulator-like 1
RHEB	AF148645	Ras homolog enriched in brain
RHEBL1	NM_144593	Ras homolog enriched in brain like 1

RIMS2	NM_014677	
RIOK2	NM_018343	
RIPK1	NM_003804	
RIPK3	NM_006871	
RIPK5	NM_015375	
RIT1	NM_006912	Ras-like without CAAX 1
RIT2	NM_002930	Ras-like without CAAX 2
RNASEL	NM_021133	
RNF219	NM_024546	C13orf7
ROCK1	NM_005406	
ROCK2	NM_004850	
ROR1	NM_005012	
ROR2	NM_004560	
ROS1	NM_002944	
RP1L1	NM_178857	
RPGRIP1	NM_020366	
RPL22		ribosomal protein L22 (EAP)
RPS6KA2	NM_021135	
RPS6KA3	NM_004586	
RPS6KA4	NM_003942	
RPS6KA6	NM_014496	
RPS6KB2	NM_003952	
RPS6KC1	NM_012424	
RRAD	NM_004165	Ras-related associated with diabetes
RRP9	NM_004704	ribosomal RNA processing 9, small subunit (SSU) processome component, homolog (yeast); formerly RNU3IP2 RNA, U3 small nucleolar interacting protein 2
RUNX1	NM_001001890	runt-related transcription factor 1 (AML1)
RYK	NM_001005861	NM_002958
SAR1A	NM_020150	SAR1 gene homolog A (S. cerevisiae)
SBDS	NM_016038	Shwachman-Bodian-Diamond syndrome protein
SBK1	NM_001024401	
SBNO1	NM_018183	sno, strawberry notch homolog 1
SCNN1B	NM_000336	sodium channel, nonvoltage-gated 1, beta
SCYL1	NM_020680	
SCYL2	NM_017988	
SDHB	NM_003000	succinate dehydrogenase complex, subunit B, iron sulfur (Ip)
SDHC	NM_003001	succinate dehydrogenase complex, subunit C, integral membrane protein, 15kDa
SDHD	NM_003002	succinate dehydrogenase complex, subunit D, integral membrane protein
SEMA5B	NM_001031702	semaphorin 5B
SEPHS2	NM_012248	
SEPT6	NM_145799	septin 6
SERPINB1	NM_030666	serpin peptidase inhibitor, clade B (ovalbumin), member 1
SGK1	NM_005627	
SGK2	NM_016276	
SgK269	XM_370878	
SGK3	NM_013257	SGKL
SH3GL1	NM_003025	SH3-domain GRB2-like 1 (EEN)
SHH	NM_000193	Sonic hedgehog homolog (Drosophila)

SIN3B	NM_015260	SIN3 homolog B, transcription regulator (yeast)
SIX4	NM_017420	sine oculis homeobox homolog 4
SKIP	NM_130766	Skeletal muscle and kidney enriched inositol phosphatase
SLC24A4	NM_153646	Solute carrier family 24 (sodium/potassium/calcium exchanger), member 4
SLC24A5	NM_205850	Solute carrier family 24, member 5
SLC44A4	NM_025257	solute carrier family 44, member 4, formerly c6orf29
SLC6A3	NM_001044	solute carrier family 6 (neurotransmitter transporter, dopamine), member 3
SLC8A3	NM_182932	
SLC9A10	NM_183061	
SLC9A2	NM_003048	solute carrier family 9 (sodium/hydrogen exchanger), member 2
SLK	NM_014720	
SMARCB1	NM_003073	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1
SMG1	NM_015092	
SMUG1	NM_014311	
SNF1LK	NM_173354	
SNF1LK2	NM_015191	
SNRK	NM_017719	
SOCS1	NM_003745	suppressor of cytokine signaling 1
SORL1	NM_003105	sortilin-related receptor, L(DLR class) A repeats-containing
SP110	NM_080424	SP110 nuclear body protein
SPECC1	NM_001033553	sperm antigen HCMOGT-1
SPEG	NM_005876	APEG1
SPTAN1	NM_003127	spectrin, alpha, non-erythrocytic 1
SRPK2	NM_003138	
STARD8	NM_014725	START domain containing 8
STIL	NM_001048166	TAL1 (SCL) interrupting locus
STK10	NM_005990	
STK11	NM_000455	
STK16	NM_003691	
STK17B	NM_004226	
STK19	NM_032454	
STK3	NM_006281	
STK31	NM_031414	
STK32A	NM_145001	YANK1
STK32B	NM_018401	
STK33	NM_030906	
STK35	NM_080836	
STK36	NM_015690	
STK38	NM_007271	
STK38L	NM_015000	
STK39	NM_013233	
STYK1	NM_018423	
SUFU	NM_016169	suppressor of fused homolog (Drosophila)
SULF2	NM_018837	sulfatase 2
SYK	NM_003177	
SYNE2	NM_182914	spectrin repeat containing, nuclear envelope 2
SYNJ1	NM_203446	Synaptojanin 1
SYNJ2	NM_003898	Synaptojanin 2
SYT14L	NM_001014372	CHR415SYT
TACC2	NM_206862	

TAF1	NM_138923	
TAF15	NM_003487	TAF15 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 68kDa
TAF1L	NM_153809	
TAOK1	NM_020791	
TAOK2	NM_004783	TAO1
TAOK3	NM_016281	
TBK1	NM_013254	
TBPL1	NM_004865	TBP-like 1
TDG	NM_003211	
Tdp1	NM_018319	
TDRD6	NM_001010870	
TEC	NM_003215	
TECTA	NM_005422	tectorin alpha
TEK	NM_000459	
TERF2	NM_005652	Telomeric repeat binding factor 2
TESK1	NM_006285	
TESK2	NM_007170	
TET1	AL713658	leukemia-associated protein with a CXXC domain
TEX14	NM_031272	
TFE3	NM_006521	transcription factor binding to IGHM enhancer 3
TFEB	NM_007162	transcription factor EB
TFG	NM_006070	TRK-fused gene
TG	NM_003235	
TGFBR1	NM_004612	
THBS3	NM_007112	thrombospondin 3
THOC5	NM_001002878	THO complex 5, formerly c22orf19
TIE1	NM_005424	
TIMELESS	NM_003920	
TLK1	NM_012290	
TLK2	NM_006852	
TLN1	NM_006289	
TMEM123	NM_052932	
TMPRSS2	NM_005656	transmembrane protease, serine 2
TMPRSS6	NM_153609	transmembrane protease, serine 6
TNFRSF17	NM_001192	tumor necrosis factor receptor superfamily, member 17
TNIK	NM_015028	
TNK1	NM_003985	
TNNI3k	NM_015978	
TOP1	NM_003286	topoisomerase (DNA) I
TP63	NM_003722	Tumor protein p63
TP73	NM_005427	Tumor Protein 73
TPM3	NM_152263	tropomyosin 3
TPR	NM_003292	translocated promoter region
TPTE2	NM_199254	Transmembrane phosphoinositide 3-phosphatase and tensin homolog 2
TRIB1	NM_025195	
TRIB3	NM_021158	
TRIM24	NM_003852	TIF1
TRIM33	NM_015906	
TRIO	NM_007118	
TRPM6	NM_017662	

TRPM7	NM_017672	
TRRAP	NM_003496	
TSC1	NM_000368	tuberous sclerosis 1 gene
TSC2	NM_000548	tuberous sclerosis 2 gene
TSHR	NM_000369	thyroid stimulating hormone receptor
TSSK1B	NM_032028	STK22D
TTBK1	NM_032538	
TTBK2	NM_173500	
TTK	NM_003318	
TTL	NM_153712	tubulin tyrosine ligase
TWF1	NM_002822	PTK9
TWF2	NM_007284	PTK9L
TYK2	NM_003331	
TYR	NM_000372	Tyrosinase (oculocutaneous albinism IA)
TYRO3	NM_006293	
UBQLNL	NM_145053	MGC20470
UBR5	NM_015902	EDD1
ULK1	NM_003565	
ULK2	NM_014683	
ULK3	NM_015518	
UNG	NM_003362	
USP1	NM_003368	Ubiquitin specific peptidase 1
USP6	NM_004505	ubiquitin specific peptidase 6 (Tre-2 oncogene)
VEPH1	NM_024621	ventricular zone expressed PH domain homolog 1
VHL	NM_000551	von Hippel-Lindau syndrome gene
VIPR1	NM_004624	Vasoactive intestinal peptide receptor 1
VRK1	NM_003384	
WDR91	NM_014149	HSPC049
WEE1	NM_003390	
wnk1	NM_018979	PRKWNK1
wnk2	NM_006648	PRKWNK2
wnK3	NM_020922	PRKWNK3
wnk4	NM_032387	PRKWNK4
WRN	NM_000553	Werner syndrome (RECQL2)
WT1	NM_024426	Wilms tumour 1 gene
XDH	NM_000379	xanthine dehydrogenase
XIRP1	NM_194293	formerly cmya1; xin actin-binding repeat containing 1
XPA	NM_000380	xeroderma pigmentosum, complementation group A
XPC	NM_004628	xeroderma pigmentosum, complementation group C
XRCC1	NM_006297	
XRCC3	NM_001100119	
YSK4	NM_025052	FLJ23074
ZAP70	NM_001079	
ZBTB16	NM_006006	zinc finger and BTB domain containing 16; zinc finger protein 145 (PLZF); formerly znf145
ZFP64	NM_018197	zinc finger protein 64 homolog
ZFYVE26	NM_015346	zinc finger, FYVE domain containing 26
ZMYM2	NM_006006	Zinc finger, MYM-type 2; formerly ZNF198 zinc finger protein 198
ZMYM4	NM_005095	ZNF262
ZNF318	NM_014345	zinc finger protein 318
ZNF331	NM_018555	zinc finger protein 331

ZNF384	NM_001039916	zinc finger protein 384 (CIZ/NMP4)
ZNF521	NM_015461	zinc finger protein 521
ZNF569	NM_152484	zinc finger protein 569
ZNF646	NM_014699	
ZNF668	NM_024706	zinc finger protein 668, aka FLJ13479
NCKIPSD	NM_016453	SH3 protein interacting with Nck, 90 kDa (ALL1 fused gene from 3p21), formerly AF3p21
GAS7	NM_201433	growth arrest-specific 7
PPARG	NM_138711	peroxisome proliferative activated receptor, gamma
AKAP9	NM_147171	A kinase (PRKA) anchor protein (yotiao) 9
CTNNB1	NM_001904	catenin (cadherin-associated protein), beta 1
EP300	NM_001429	300 kd E1A-Binding protein gene
FSTL3	NM_005860	follistatin-like 3 (secreted glycoprotein)
HIST1H4I	NM_003495	histone 1, H4i (H4FM)
NIN	NM_020921	ninein (GSK3B interacting protein)
NCOA1	NM_147223	nuclear receptor coactivator 1
NCOA2	NM_006540	nuclear receptor coactivator 2 (TIF2)
NCOA4	AK130612	nuclear receptor coactivator 4 - PTC3 (ELE1)
MUTYH	NM_012222	mutY homolog (E. coli)
NSD1	NM_022455	nuclear receptor binding SET domain protein 1
Sumo1		
CHIC2	NM_012110	cysteine-rich hydrophobic domain 2
ELF4	NM_001421	E74-like factor 4 (ets domain transcription factor)
MSI2	NM_170721	musashi homolog 2 (Drosophila)
MSN	NM_002444	moesin
POU2AF1	NM_006235	POU domain, class 2, associating factor 1 (OBF1)
POU5F1	NM_002701	POU domain, class 5, transcription factor 1
RANBP17	NM_022897	RAN binding protein 17
RAP1GDS1	NM_001100426	RAP1, GTP-GDP dissociation stimulator 1
TCEA1	NM_006756	transcription elongation factor A (SII), 1
TCF12	NM_207037	transcription factor 12 (HTF4, helix-loop-helix transcription factors 4)
TCF3	NM_003200	transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
KCNAB2	NM_172130	
AK5	NM_012093	
DPP4	NM_001935	
IHPK3	NM_054111	
SLC26A4	NM_000441	
RGS3	NM_021106	
ARMCX2	NM_177949	
CNNM1	NM_020348	
RAB3D	NM_004283	
CACNA1A	NM_001127221	
DOK5	NM_018431	
TOM1	NM_005488	
TET3	NM_144993	
SOX11	NM_003108	
SHOX2	NM_003030	
NFE2L3	NM_004289	
SLC38A2	NM_018976	
MEIS2	NM_002399	

PIK3IP1	NM_052880	