

THE INTERACTIONS OF DIFFERENTIALLY EXPRESSED GENES
DIRECTLY TARGETED BY WNT/ β -CATENIN SIGNALING:
IMPLICATIONS FOR COLORECTAL CARCINOGENESIS

By

Charles F. Murchison

A thesis presented to

the Department of Medical Informatics and Clinical Epidemiology

and the Oregon Health & Science University School of Medicine

in partial fulfillment of the requirements for the degree of

Masters of Science

December 2010

School of Medicine
Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Thesis of

Charles F. Murchison

*“The Interactions of Differentially Expressed Genes
Directly Targeted by Wnt/ β -catenin Signaling: Implications
for Colorectal Carcinogenesis”*

Has been approved

Thesis Advisor

Committee Member

Committee Member

TABLE OF CONTENTS

LIST OF TABLES.....	v
LIST OF FIGURES.....	vi
ACKNOWLEDGEMENTS.....	vii
ABSTRACT.....	viii

CHAPTER 1 – INTRODUCTION

1.1 COLON CANCER	1
1.2 ADENOMA DEVELOPMENT AND THE WNT SIGNALING PATHWAY.....	4
1.3 MICROARRAYS AND POOLED ANALYSES.....	6
1.4 OTHER HIGH-THROUGHPUT TECHNIQUES.....	9
1.5 A MULTI-METHOD APPROACH.....	12

CHAPTER 2 – METHODS

2.1 LITERATURE SEARCH FOR EXPRESSION STUDIES.....	14
2.2 NORMALIZATION OF CEL FILES TO ABSOLUTE INTENSITY.....	15
2.3 DIFFERENTIAL EXPRESSION.....	18
2.4 CO-EXPRESSION.....	19

2.5 MODULE COMPARISON: PATIENT VS CELL-LINE	19
2.6 WNT TRANSCRIPTION FACTOR BINDING DATA.....	20
2.7 EXPRESSION PATTERNS AND β -CATENIN BINDING	21
2.8 THE PROTEIN-PROTEIN INTERACTION NETWORKS.....	22
2.9 CODINGS FOR THE NETWORK	24
2.10 EVALUATION OF THE DISCOVERY NETWORK.....	26
2.11 NULL NETWORK MODEL – COMPARING OBSERVED AND PERMUTED PROPORTIONS OF METRICS	28

CHAPTER 3 – RESULTS

3.1 LITERATURE SEARCH OF mRNA EXPRESSION STUDIES	32
3.2 NORMALIZATION ASSESSMENT FOR POOLING.....	33
3.3 AGGREGATE STUDY OF MRNA EXPRESSION	33
3.4 FILTERING OF SACO DATA OF β -CATENIN BINDING.....	36
3.5 THE PROTEIN PRODUCT INTERACTION NETWORK.....	36
3.6 NETWORK NODES WITH SINGLE METRICS OF INTEREST	37
3.7 NODES MATCHING MULTIPLE CRITERIA AND A FINAL NETWORK.....	38
3.8 ASSESSMENT OF THE NULL MODEL	40
3.9 ANNOTATION OF FUNCTION FOR IDENTIFIED GENES/NODES	41

CHAPTER 4 – DISCUSSION

4.1 INITIAL ANALYSIS – DIFFERENTIAL EXPRESSION PROFILES AND DIRECT β-CATENIN TARGETING	44
4.2 BIOLOGICAL ANNOTATION – THE 8 FUNCTIONAL ACTIVITY CATEGORIES FOR THE REFINED NETWORK	47
4.3 CELL GROWTH, PROLIFERATION AND REGULATION OF THE CELL CYCLE	48
4.4 PHOSPHORYLATION AND UBIQUITINATION	49
4.5 CELL FATE AND DIFFERENTIATION, AND EPITHELIAL CELL POLARITY	50
4.6 EPIGENETICS AND HISTONE MODIFICATION.....	51
4.7 CELL-CELL INTERACTIONS, MOLECULAR TRANSPORT, INTER-CELLULAR SIGNALING	52
4.8 PARTICIPANTS IN THE CANONICAL WNT PATHWAY – COFACTORS AND RECEPTORS.....	54
4.9 UBIQUITOUSLY EXPRESSED GENES – GENERALIZED FUNCTIONALITY AND MULTIPLE BIOLOGICAL CIRCUMSTANCES.....	56
4.10 EXPRESSION LOCALIZED OUTSIDE THE COLON	57
4.11 OBSERVATIONS ON NODES IN THE CANONICAL WNT PATHWAY.....	59
4.12 FINAL TRENDS.....	60

CHAPTER 5 – CONCLUSIONS

5.1 LIMITATIONS OF THE CURRENT STUDY.....	64
5.2 FUTURE DIRECTIONS	65

5.3 CONTRIBUTIONS TO THE FIELD	66
REFERENCE LIST	69
TABLES.....	83
FIGURES.....	111

LIST OF TABLES

TABLE 1 – STUDIES USED FOR mRNA ANALYSIS	83
TABLE 2 – SIGNIFICANCE COUNTS OF PROBES FROM mRNA ANALYSIS	84
TABLE 3 – CONCORDANT GENES BY DIRECTION OF DIFFERENTIAL EXPRESSION	85
TABLE 4 – TOP 5 CONCORDANTLY UPREGULATED AND DOWNREGULATED GENES	86
TABLE 5 – DIRECT TARGETS OF β -CATENIN SHOWING UPREGULATION AND DOWNREGULATION	88
TABLE 6 – METRICS OF THE INITIAL PROTEIN INTERACTION NETWORK.....	95
TABLE 7 – GENES WITH MRNA OVEREXPRESSION FOUND IN ADENOMAS AND CELL LINES	96
TABLE 8 – GENES WITH MRNA UNDEREXPRESSION FOUND IN ADENOMAS AND CELL LINES	97
TABLE 9 – METRICS OF NODES WITH ALTERED CO-EXPRESSION UNDER ADENOMATOUS CONDITIONS	98
TABLE 10 – METRICS FOR β -CATENIN DIRECT TARGETS AND NEIGHBORS	99
TABLE 11 – METRICS FOR INTERACTION NETWORK HUBS AND HUB NEIGHBORS	101
TABLE 12 – METRICS FOR KNOWN WNT PARTICIPANTS AND NEIGHBORS	105
TABLE 13 – FULL METRICS LISTING FOR THE FINAL 65 NETWORK NODES.....	107
TABLE 14 – STATISTICAL ASSESSMENT OF NULL NETWORK MODEL.....	110

LIST OF FIGURES

FIGURE 1 – COLONIC CRYPT AND ADENOCARCINOMA DEVELOPMENT	111
FIGURE 2 – WNT ACTIVATION OF β -CATENIN DEPENDENT TRANSCRIPTION	113
FIGURE 3 – SERIAL ANALYSIS OF CHROMATIN OCCUPANCY	115
FIGURE 4 – DENSITY GRAPH OF SPOT INTENSITIES FROM mRNA AGGREGATION ANALYSIS.....	117
FIGURE 5 – BOX PLOTS OF INDIVIDUAL ARRAYS COMPARING NORMALIZATIONS.....	119
FIGURE 6 – VENN DIAGRAM FOR DIFFERENTIALLY EXPRESSED GENES BY MODULE GROUPING	123
FIGURE 7 – OVERVIEW OF THE BASE PROTEIN INTERACTION NETWORK.....	125
FIGURE 8 – COLOR CODED LEGEND FOR TABLES 7-12	127
FIGURE 9 – REFINED INTERACTION NETWORK USED FOR VISUALIZATION	129

ACKNOWLEDGEMENTS

This thesis would not have been possible without the guidance and support of my advisor Dr. Shannon McWeeney. Her tireless devotion and formidable intellect are matched only by her enthusiasm for the study of Bioinformatics and the entire discipline benefits from her involvement. Words do little justice to the generosity and insight she has given over the years but thank you, Shannon, for everything.

Equally central in the development and implementation of this project were the members of my thesis advisory committee, Drs. Greg Yochum and Motomi Mori. Such a broad and inter-disciplinary study required a great deal of understanding of numerous disparate fields. Their advice in molecular biology and microarray analysis were indispensable in bringing this thesis to fruition.

Finally, a thank you to the faculty, staff and students in the Department of Medical Informatics and Clinical Epidemiology at OHSU. I can think of no better place to have begun my career in Bioinformatics.

ABSTRACT

Colorectal cancer is one of the most pervasive and deadly diseases in the world today. Although heavily studied with a wealth of available scientific data, little research has been directed at uniting the various biological metrics to test for novel genetic signals and markers that may be otherwise overlooked in singular studies. The current thesis addresses this by utilizing a combined, multi-method analysis to examine biological archetypes in colorectal adenomas by evaluating patterns of differential expression, co-expression disruption, protein product interaction, and transcription factor binding across an available selection of cancerous cell lines, patient colorectal adenomas and normal patient mucosa. First, an aggregate evaluation was used to identify disruptions in co-expression resulting from adenocarcinoma development in addition to differential gene regulation. The resulting gene expression patterns were compared against the chromatin occupancy of β -catenin, a key molecule in the canonical Wnt pathway. The results of this pooled analysis were overlain onto a protein-protein interaction network derived from the Wnt pathway to identify genes of interest matching multiple categories of biological metrics. A number of genes identified in the empirical network were further categorized based on their functional ontologies and examined for their prospective role in either the treatment or detection of adenocarcinomas. In addition to isolating a number of potentially intriguing markers, this study also represents a novel method of disease assessment with applicability in conditions beyond colon cancer.

CHAPTER 1 – INTRODUCTION

1.1 Colon cancer

Few would argue that cancer is the primary medical epidemic facing the world today. According to the World Health Organization, cancer was the leading cause of death across the globe in 2007, accounting for 7.9 million deaths (13%) worldwide [1]. Typically, cancer is accepted as a blanket term, encompassing a spectrum of disease conditions, all united by a common progression of uncontrolled cell growth, possibly in conjunction with invasion to neighboring tissue and metastasis. Other similarities across cancers include a tendency for all uncontrolled growths to begin with abnormalities in the originating tissue's genetic material, whether these abnormalities are inherited, induced by a carcinogen or both [2]. Yet, despite these parallels, not all cancers are created, develop or progress equally. The time-course of the disease, the development of the uncontrolled growth, and potential treatment options are all highly dependent on the cancer's tissue of origin.

Certain cancers by nature, such as mesothelioma of the lung lining, are harder to initially detect and have far fewer treatment options available. Other cancers, with stronger inheritance patterns for example, are more easily found and have much greater treatment efficacy [3]. One cancer in particular fits nicely into the later grouping: colorectal adenocarcinomas or colon cancer. The colon is a unique organ, being one of

the few that has a perpetual supply of undifferentiated stem cells available, specifically for the regeneration of the intestinal epithelium lining [4]. This is a necessity due to the caustic and toxic environment the colon is constantly subjected to.

In order to properly facilitate nutrient absorption across the lining, an extensive surface area is used to maximize possible absorption into the blood stream. To properly fit in the limited space of an organism's torso, the intestinal epithelium folds on itself creating a series of small nubs, villi, as well as multiple divots called intestinal crypts. These crypts are the source of lining replenishment with each representing a cross-sectional snapshot of the developmental progression of epithelial cells [5]. Nascent cells begin at the base of the crypt as undifferentiated stem cells until mitosis is induced. These progenitor cells can undergo one of two types of division. The first, asymmetric division, results in a new partially differentiated cell ready for intestinal function in addition to another stem cell ready for subsequent division. Symmetric division on the other hand results in a matched pair of either partially differentiated cells or a pair of progenitor cell. The unchecked accumulation of the later often results in early adenomas. For those cells properly designated for intestinal activity, final differentiation to full functionality continues as the cell migrates along the crypt [6]. Thus, more efficacious cells are in closer proximity to potential nutrient sources and ready to replace epithelial cells which have reached the end of their effective lifespan. When this process is properly regulated, intestinal activity is smooth and efficient (see figure 1).

With the capacity to duplicate and differentiate as needed, the intestinal wall can be replenished continuously and promote proper digestion. However, this collection of cells, designed to propagate and replace, creates a volatile situation where unhindered proliferation can easily lead to tumorous growths and the development of adenocarcinomas. Commonly, these carcinomas begin as benign adenomatous polyps in the epithelial lining of the colon, which eventually develop into malignant tumors after a certain (and often unknown) latency [7]. Furthermore, many colon cancers show a strong hereditary component, although the exact component differs between disease types. For example, Familial Adenomatous Polyposis involves lesions in the APC or MUTYH tumor suppressor genes, conferring a selective advantage towards the creation of hundreds and thousands of colonic polyps [8]. Comparatively, Hereditary Non-Polyposis Colon Cancer is due to inherited abnormalities in any of five DNA mismatch repair genes, giving Lynch Syndrome far fewer polyps and an increased tendency for the cancer to also express in nearby organs such as the endometrium or other parts of the reproductive system [9].

Despite mild differences in specifics, colon cancer in all its forms is a deadly and pervasive disease. The CDC estimated nearly 140,000 cases of colon cancer were diagnosed in the U.S. alone in 2006, with over 53,000 deaths [10]. The World Health Organization reiterates this deadly prevalence, citing colon cancer as the number two leading cause of cancer related death in the western world, second only to lung cancer [1]. Yet despite these sobering statistics, colorectal adenocarcinomas are also one of

the most treatable types of cancer: polyps can be evaluated before they become malignant with many more viable (although still unpleasant) treatment options than many cancers originating from different organs. Nonetheless, the treatment of colon cancer is still predicated on successful detection, and like all cancers, the earlier the better.

1.2 Adenoma development and the Wnt signaling pathway

The progression of colorectal adenocarcinomas is very well characterized, and despite variations in genetic abnormalities, a number of specific protein interactions and signaling pathways have been implicated in all of its developmental phases (see figure 1); from a normal colon epithelial cell, to a benign adenomatous polyp, to a malignant carcinoma and finally, to a tumor that has undergone metastasis [11]. One particular pathway has been identified whose genetic lesions are commonly found in the determination of normal colon cells into the initial adenomatous polyps: the Wnt signaling pathway (visualized in figure 2). The Wnt pathway is an enormous and complex collection of molecules, proteins and physiological responses all related to extra-cellular signaling due to a ligand involving a Wnt molecule. Although many branches and divergences exist in the pathway, one particular branch is highly involved with adenomatous cell development, referred to as the canonical Wnt pathway [12]. The canonical Wnt pathway primarily regulates the amount of nuclear β -catenin available within a cell. β -catenin is a transcription factor most commonly contained in

the cytoplasm, where it is most often degraded by a holoenzyme before it can migrate into the nucleus. That destruction complex, made up of a number of proteins such as APC and AXIN2, is in turn regulated by Wnt signaling. When Wnt binds to transmembrane proteins in the Frizzled family, key components in the destruction holoenzyme are prevented from associating and subsequent degradation of β -catenin is inhibited. As the transcription factor accumulates in the cytoplasm, it begins to migrate into the nucleus where it associates with a number of DNA associated compounds in the TCF/LEF family already bound to the chromatin. Once associated, these β -catenin/TCF-LEF compounds regulate expression of their affiliate genes. In Wnt, these genes are most commonly involved in embryogenesis, epithelium development and cancer.

Wnt-like activity can be easily simulated through abnormalities to the genes coding for proteins involved in the destruction of β -catenin. By preventing the ability of the complex to poly-ubiquitinate β -catenin, the transcription factor accumulates and enters the nucleus where it alters mRNA expression. A very common mutation, which is also highly hereditary, is to the gene coding for the APC protein [13]. A number of mutations lead to a truncated version of APC, which can still combine into the destruction complex, but cannot adequately bind β -catenin to mark it for degradation. In fact, mutations to the APC gene are one of the hallmark abnormalities involved in Familial Adenomatous Polyposis, as mentioned earlier. Beyond inheritance, the significance of APC is highlighted by the propensity of its mutations to be found in incident cancers as well. Especially critical is the relation between Wnt and colon cancer's developmental

time course. β -catenin signaling through Wnt activity is associated with the earliest possible step in tumorigenesis, from normal mucosa to initial adenoma before the growth is even malignant [14]. By focusing on pathways intrinsically involved with this initial transition step, the potential of genetic and molecular signals to assist in the detection and treatment of colorectal adenocarcinomas will be at its greatest.

1.3 Microarrays and pooled analyses

One of the greatest discoveries to accompany the genomic revolution was the advent of high-throughput methods that allowed testing of upwards of tens of thousands of genetic sequences, all at the same time and in many cases on a single experimental medium. By far the most popular high-throughput technique is the mRNA microarray. The concept of the array is fairly simple: to hybridize corresponding single strands of nucleic acids to look for the presence of sequences of interest. The true utility of the microarray is multiplying this process numerous times, allowing for a quantifiable level of expression for a specific gene and testing this expression for thousands of genes simultaneously [15]. By extracting mRNA from a cell under the specific experimental conditions in question (e.g. from a colorectal adenoma), a genetic profile snapshot can be taken showing the levels of expression for a specified collection of genes. Comparing these results against the expression levels of a cell under control conditions can calculate fold changes in mRNA levels, giving an effective measure of differential gene regulation [16]. As the technology has improved, microarrays have benefited from a

number of advances in extraction methods, quality control of both cell and array, imaging techniques and, most importantly, standardization.

It should be noted that microarrays are not without their own weaknesses. Although many advances have been made, genomic analysis is still a burgeoning field with barely two decades worth of research behind it. One such limitation is cost. Microarrays, in spite of becoming smaller, more condensed and cheaper to fabricate, are not considered inexpensive; major difficulties (especially for smaller laboratories) with technical replicates, and in some cases even biological replicates, are far from uncommon [17]. Additionally, standardization is only a recent phenomenon and inter-study comparisons are still a major hurdle. Differences between arrays often seem innumerable, ranging from more benign situations, such as varying numbers of probe replicates for a given spot, to conditions that fully prevent array comparisons, like evaluating a two-channel array with competitive hybridization against a pair of single-channel arrays showing absolute expression [18]. Even with mass-produced identical arrays, the human component introduces its own sources of study error [19]. Fortunately, judicious use of multi-method techniques and pooling of data and results allows aggregation of disparate studies into viable datasets for joint analysis.

Pooling of data, no matter the underlying design, attempts to address concerns over reduced statistical power, which commonly arises from issues due to small sample sizes. By looking at a common effect measure from studies with similar research hypotheses,

inferences can be drawn to better estimate a true population parameter [20]. Joint measurements allow for generalities to the population to be more easily drawn, effect size parameters from multiple studies to be combined into a single measure, and accountability for confounding variables due to inter-study variability within a larger model [21]. When used properly, a pooled analysis can yield effect parameters that might have otherwise been masked by a single experiment.

These aggregation methods are well established analysis techniques. For example, meta-analysis was first used as a viable process in 1904 by Karl Pearson, the founding father of modern statistics, and has since undergone numerous adjustments and refinements, from the coining of the term 'meta-analysis' by Gene Glass in 1976 to the development of the Cochrane handbook, a gold-standard in meta-analysis techniques. The use of meta-analysis has also been applied to numerous fields including psychotherapy, epidemiology and, as is becoming more and more common, genetic array analysis [22, 23]. Addressing issues such as collective normalization and background subtraction alleviates many of the concerns inherent to inter-study analysis of arrays. Pooling techniques are further facilitated through powerful software packages that have numerous normalization and aggregation methods available, such as the Affy module of the Bioconductor package of R [24]. However, like all multi-study methods, the utility of the pooled genetic analysis is dependent on the strength of the individual studies. Experiments must be as similar as possible with respect to the effect size which was measured and the research question asked . This is especially important

since, by nature, the bias sources for the individual studies are ignored. Ultimately, this highlights the importance of the initial selection of the experiments to be used. By being stringent in selection from the literature, studies that are initially evaluated using poor statistics can be avoided [25]. Despite these concerns, pooling analyses are powerful and robust techniques that can effectively bring to light inferences and conclusions that would have otherwise been lost to shallow statistical power.

1.4 Other high-throughput techniques

Much insight has been gained through mRNA profiling, allowing differentially expressed genes to be identified in a number of biological conditions across a breadth of organisms. Yet the presence of mRNA is only a single facet of the central dogma. Correspondingly, research has also been directed towards developing other methods to analyze genetic and molecular signals such as variations and adaptations of the *in situ* method of chromatin-immunoprecipitation (ChIP) [26].

In its most basic form, immunoprecipitation elucidates and isolates proteins that have been histochemically bound to an antibody (selected specific for the protein) that is additionally tailored to allow for ease of collection and enrichment. When the targeted proteins are also chromatin-bound, the antibody complexes can additionally affix the protein-associated tracts of genomic DNA thus collecting actual binding sites for the protein of interest. After purification, these harvested stretches of nucleic acids can be

evaluated using any number of standard sequence-based high-throughput techniques; for example, hybridization against microarray probes as in the prototypical ChIP-on-Chip methodology [27]. In addition to the obvious utility with microarray screening, another method expands on the tag tabulation aspects of serial analysis to uniquely identify binding loci and quantify transcription factor association using a process called Serial Analysis of Chromatin Occupancy (SACO, see figure 3).

SACO begins much like any standard ChIP experiment, by sonicating chromatin and using antibodies to separate out the desired transcription factor associated nucleotides. After isolation and purification (including the use of multiple endonucleases to remove contamination by PCR purification adapters as well as artifact effects due to undigested DNA) the previously bound DNA sites are ready for sequence-based analysis to evaluate loci affinity to the transcription factor being studied [28]. This analysis process utilizes the same principles as the serial analysis of gene expression by concatenating tags that uniquely identify genes, creating vectors readymade for amplification and sequencing. Tabulation against a sequence database not only identifies ostensible binding sites, but also provides a discrete measure of affinity using tag counts. As such, the collected SACO library gives a great deal of insight towards transcription factor associations, quantifying binding while allowing for the discovery of novel sites. By evaluating chromatin under various experimental conditions, a binding profile for a transcription factor of interest can be developed much like an expression profile is assembled from a standard mRNA serial analysis.

Many *in silico* methods are also in place to help researchers understand the intricacies of various biological conditions, notably the protein-protein interaction (PPI) network [29]. Although a PPI network is developed, visualized and analyzed using software, the data itself comes from laboratory bench results as well as validated citations from the literature and widely accessible protein databases [30]. In this regard, a PPI network is simply another type of aggregate study using the condition of gene product interaction, whatever form it may take, as the joint effect. By evaluating codings upon the nodes and edges in the network, statistical inferences can be drawn ranging from strength of interactions, to levels of connectivity, to the classification of a specific gene/protein as a 'hub' involved in multiple higher order physiological processes [31]. Adjusting the presence of nodes, by either parsing out less important nodes or adding nodes that create new interactions, in turn recreates the interacting topology so that multiple networks can be compared; disruptions in the network due to biological conditions are highlighted and genes/gene products not immediately implicated in the condition of interest through previous laboratory methods can be isolated [32]. A pure expression profile of mRNA is certainly a valuable and viable technique for genetic signal analysis; although, much like the role of mRNA in the central dogma, it is but one of many aspects relating to the genetic and proteomic changes resulting from a physiological effect.

1.5 A multi-method approach

Despite the breadth of testable methods for genetic activity available, only a few studies have attempted to combine multiple, disparate measures together in order to develop more refined and direct models of the genetics of biological states. Some studies have combined differential expression metrics into PPI networks to assess changes in mRNA expression and graph node connectivity. Most studies have focused on specific diseases such as gastric cancer [33] and diabetes [34] although some such as Xu et. al. [35] have attempted to identify multiple candidate disease genes by evaluating the topology of numerous networks with respect to differential expression of disease-associated genes. The study met with a degree of success but was mainly seen as validation studies rather than an attempt to elicit new information regarding genetic markers for their ailments of interest. Camargo et. al. [36] created a *de novo* PPI network based on a specific array study and coded the ensuing network with differential expression with regards to human heart failure. Although details about specific genes and gene products were not discussed in their study, a number of interesting observations were made. One such discovery noted differentially expressed nodes were not highly connected, but the neighbors of these nodes did show high connectivity. These results just begin to emphasize the potential a combined approach could bring to genetic marker analysis. Due to its extensive prevalence, high mortality rate, ease of treatment when properly detected, and an enormous literature collection of biological effects, colorectal adenocarcinomas are an ideal candidate for such a collective analysis.

This study brought the idea of a combined approach further by looking for genes that fit multiple categories of experimental classification from a range of measures including differential expression, direct chromatin binding, and levels of protein product connectivity. To emphasize the role of pathways previously implicated in early adenoma development, literature studies were limited to patient colorectal adenomas albeit in conjunction with cancerous simulating cell lines. These experimental modules were also compared against normal colonic mucosa. Additionally, co-expression was examined within each sample medium to look for disruptions due to changes in transcription during adenomatous development, and related to protein product interactions. The goal of this study was one of discovery, both with regards to identification of genes highly implicated in colon adenocarcinoma development and to assess the validity of a multi-method approach in implicating such genes. Although many individual questions were posited, the primary research question asked was can any novel information related to genes, their protein products, their biological interactions and their regulatory control in colorectal adenocarcinomas be determined using a multi-method approach?

CHAPTER 2 – METHODS

2.1 Literature search for expression studies

The available literature was searched for mRNA expression experiments to be utilized in the pooled study of differential and co-expression. Two experimental groups were considered based on the sample medium of a given array; one module based on derived cancerous cell lines (HCT116 and LS174T specifically [37, 38]) with the other exclusive to human colon adenomatous tissue. Both of these groups were compared against a third collection of data, where normal human colon mucosa was considered as a control group. To appropriately run the meta-analysis, a number of restrictions were placed on the potential studies:

- i. The experiments must have been run using the Affymetrix GeneChip HG-U133 Plus 2.0 array [39]. As a consequence, only a single CDF file was required
- ii. RNA extraction of the samples was done with RNeasy and the probe sets were labeled using the standard reagents, quantities and procedures as described by Affymetrix
- iii. Fluorescent signals from the array were read using a GeneChip Scanner to create CEL files available in their entirety from a public repository e.g. GEO

CEL files of the biological replicates matching these criteria were downloaded and grouped based on the three tissue conditions (cancerous cell/patient adenoma/mucosa control) before being subjected to analysis.

These restriction criteria conferred a number of advantages to the aggregation study, the most notable benefit being the standardization inherent to the Affymetrix array. By focusing on a single probe set (contained in one of the most comprehensive arrays available) there would be little confounding due to missing probes, differing numbers of repeated sequences within spots, or other concerns that could arise by comparing multiple array types against each other. Furthermore, using the CEL intensity files allowed for normalization to be done before aggregation so that the pooled analysis was based on the raw data rather than summary statistics. This limited any statistical inequities that may have arisen within the analysis of an individual study. Finally, any subsequent microarray experiments suggested based on these results can be easily validated since the HG-U133 Plus 2.0 array is so well characterized.

2.2 Normalization of CEL files to absolute intensity

The downloaded CEL files were analyzed using the Affy package of the Bioconductor module of R [40, 41]. Analysis began with normalization and background subtraction. Since the collected studies were taken from a variety of experiments and laboratories,

there existed a potential to have a large amount of inter-study variability introduced, which would need to be removed with a strong normalization technique. Conversely, it could be argued the standardization inherent to the Affymetrix arrays would adequately limit the cross-study variance, requiring a less forceful normalization [42]. To address this question of baseline, three separate normalizations and background subtraction combinations (listed below in increasing strength) were run and compared:

1. Linear scaling using a pre-summarization normalization of MAS 5.0 (utilizing the 16 square MAS 5.0 background correction with and idealized mismatch pair subtraction)
2. The dChip method of invariant set scaling against a median expression reference from the data file (although again utilizing the MAS 5.0 background correction with idealized mismatch subtraction) [43]
3. Quantile normalization via Robust Multi-Array Average (using RMA background correction with no subtraction of mismatch hybridization) [44]

The normalization process was evaluated by looking at the box plots for all arrays within each module. Centering of the IQRs was compared as were spots differentially identified as outliers. These criteria were used to select the least stringent normalization method that adequately centered the expression intensities while still retaining the natural variability of the data.

To assess inter-module variability, the normalization was applied in two fashions: to each array module separately (separate normalization) and to all arrays in all modules at one time (collected normalization). Although the self-contained nature of co-expression analysis necessitates a separate analysis, different requirements exist for appropriate testing of differential gene regulation. Ideally, any differential expression calculations would be done on arrays that have all been normalized together; however, a concern with the collected normalization is that smaller differences at the expression level could be lost among the modules when they are aggregated together. To address this, the scatter-plots for the collected normalization were evaluated to identify excessive scattering, as determined by correlation. Considerations for both methods exist with regards to calculating differential expression. A consequence of the separate normalization would be placing the baseline expression of the various modules at different levels. However, this method also highlights differences based on tissue type that may have been lost with a collected normalization.

Ultimately, the goal of the normalization process was to bring all the tested studies to a common field at the expression index level albeit within reason. Since co-expression analysis is module specific, first thoughts suggested that a collected normalization may be of no real benefit. Although differential expression would ideally be measured with all biological replicates normalized together, excessive data point scattering at the origin

was concerned to have greatly affected any correlation so that the collected normalization could introduce even more variance into the model than already existed.

The summarization metric used was also directed by the normalization method selected. Linear scaling uses MAS 5.0's robust average (average difference could also have been used but is considered a poorer metric), invariant set normalization is summarized with the model based expression-indexes of Li and Wong, and the quantile normalization makes use of the median polish metric. Each measure was given (and was statistically tested) on the \log_2 scale. After normalization and subtraction, an absolute intensity value was then available for each probe spot corresponding to the biological replicates within that module. These absolute expression intensities were subsequently used to test for differential expression between modules and co-expression within a module.

2.3 Differential Expression

The summarized, absolute expression values were used to determine the relative fold-changes in mRNA expression from one module to the next (for all three modules) in a pair-wise fashion. Significantly differentially expressed genes were identified by creating a generalized linear model implemented using the LIMMA package of Bioconductor [45]. As mentioned, all data was pre-processed using the Affy package; therefore, linear model creation was done via normalized summarizations on a \log_2

scale for each gene of each array. Generation of the linear model utilized an empirical Bayesian approach to create a priors distribution from the data itself, with the design matrix organized to keep the individual experimental modules distinct. A pair-wise comparison between all three modules for each gene determined differential expression after selecting for genes using a false discovery rate of 5% [46].

2.4 Co-expression

Co-expression of all gene pairs was determined by correlating the absolute intensities of the biological replicates of a gene pair within a given module, for each module. This gave three Pearson's correlation coefficients for each gene pair, one corresponding to each of the three modules. Changes in co-expression between modules were evaluated for each gene pair to test for disruptions in gene co-expression under disease conditions. Multi-factorial statistical testing was done using one-way ANOVA to compare changes in Pearson's r for each gene pair across the three modules. Multiple tests were controlled for by using a false discovery rate of 5%.

2.5 Module comparison: patient vs cell-line

At this early stage assessment, the expression patterns were compared between the cancerous simulating cell-line module and the adenomatous patient tissue module to test for differences in co-expression and differential expression between the sample

sources. The specific metrics looked for genes that were differentially expressed, or showed disruption in co-expression, in one module but not the other. A small fraction threshold (less than 1%) was put into place to demarcate genes as being shown to have different expression patterns between patient tissue and cell lines to represent progression of colorectal adenocarcinomas. A consequence of the sample modules having a similar pattern of gene regulation was a need for normalization, summarization and statistical analysis to be rerun using cell-line data and patient tissue data aggregated together so that all comparisons would then be considered as “adenomatous/cancerous expression” vs mucosa control. Conversely, differences in fold-changes of expression or co-expression patterns between cell lines and patient tissue would keep the modules separate. As was known in advance, the largest consequence from this would occur during evaluation of the PPI network since separate codings would be required for differential expression and co-expression based on the tissue condition of the module and relative progression of the disease.

2.6 Wnt transcription factor binding data

In addition to the mRNA expression data, transcription factor binding data in colorectal carcinomas was used as a measure of implication in colon tumorigenesis. Due to vast differences in techniques and a general limitation on available data, an aggregation of the transcription factor binding was not run. Instead, a study by Yochum et. al. [47] was utilized which used the previously described SACO technique to quantify the number of

gene sequence tags that were bound to β -catenin. Since the study involved transcription factors involved in the canonical Wnt pathway they provided extra insight into potential genetic markers for colon adenocarcinomas.

A threshold was specifically established to determine if a particular gene can be classified as a “strong, direct target” of β -catenin. This was possible as the SACO approach gives a quantifiable measure for tagged gene loci. Tags that were found to be located close together were grouped into a single cluster giving a specific gene not only a number of tags, but a number of clusters as well. For a gene to be considered a strong, direct target of β -catenin, the gene was required to show at least one cluster with a minimum of three tags. Once a gene matches met this threshold any subsequent evaluation considered all tags for that gene, regardless of cluster associations.

2.7 Expression patterns and β -catenin binding

At this stage, an evaluation determined which genes already fit multiple criteria related to colorectal adenocarcinomas, namely showing either differential expression and/or disrupted coexpression in addition to being a direct transcription factor target. This created a first-stage assessment for comparison against subsequent curations to see if additional information identified any new genes as being relevant to adenocarcinoma development. In its own right, this collection is useful as a list of direct targets of the Wnt pathway differentially expressed under disease conditions.

Before this association could be done, the fraction of unique high-confidence β -catenin target tags that could be mapped to the probe set tags of the HG-U133 array was calculated in order to verify an appropriate collection of affiliated probes could be used for meaningful interpretation. SACO tags that were unable to be mapped were marked for addition to the discovery PPI network. Furthermore, the coverage of the binding data was potentially less than the coverage of the HG-U133 chips used in the expression studies which may have had a further effect on data interpretation. Because of this coverage issue, analysis was limited to those genes that could be adequately mapped from the SACO library.

2.8 The protein-protein interaction networks

To serve as a structural backbone for the expression and binding data, a protein product interaction network was developed using previously determined genes related to Wnt and colorectal adenocarcinoma development. This protein-protein interaction (PPI) network was comprised of gene products known for their involvement with the Wnt pathway as determined by the literature. The intent of this PPI network was as a validated model to highlight the current state of known interactions as they relate to Wnt and colorectal adenocarcinoma development in order to further refine their specific role and timeline. Confirmed genes were identified using Agilent's Literature

Search program along with verified interactions as determined by the cPath database of molecular associations [48].

Beyond its initial utility, this network had potential to be used as a baseline and seed network for a second discovery network designed to be much larger and much more comprehensive. By selecting and expanding on genes identified as notable participants in adenocarcinoma development, this discovery network could identify genes related to the source nodes yet not originally implicated in the canonical Wnt pathway. As would be expected from an array of this size, the topology of the network would be of less use since many nodes potentially have no classification. Unlike the validated network, the purpose of the discovery network would be to start with a large list of potential candidates (namely the added nodes expanded from around the seed proteins) that would be subsequently reduced based on the extensive codings of the nodes.

All networks were visualized and analyzed using the program Cytoscape [49]. Cytoscape allows for easy coding and statistical assessment and also includes modules for the Agilent Literature Search program as well as an interface with the cPath database. The validated network was assembled directly through Cytoscape and its modules, while the discovery network required a Java script to interface with a downloadable version of cPath's dataset to establish the comprehensive collection of interactions.

2.9 Codings for the network

To obtain the full utility of the network, many codings were applied to the various nodes.

1. mRNA expression specific results were applied to the network with mutually exclusive codes for a gene being significantly upregulated in adenomas and/or cell lines, significantly downregulated in adenomas and/or cell lines, or showing no differential expression. Codings were fully refined to show changes in differential expression in either patient adenoma tissue or cancerous cell lines, based on the previous comparison of mRNA expression patterns between the patient tissue and cell line modules. Those nodes found in the literature but not contained with the Affymetrix array were left uncoded.
2. Genes that were found to be β -catenin targets (as per the three tag minimum previously mentioned) as indicated by the chromatin binding study were specially coded as being direct targets of the Wnt transcription factor. Tag counts were included in the coding indicating the total number of tags found in all clusters for the gene in question.

3. The original node population was based solely on some involvement of their protein product with the general Wnt pathway. Using the annotations of the Wnt homepage, additional coding identified specific genes that have been previously determined to be direct targets of the Wnt pathway.

4. Degree of connectivity between protein products was considered at two levels with especially connected nodes as to be marked as being either hubs (5-13 interacting neighbors) or super-hubs (14 or more neighbors). The super-hub node levels were determined heuristically in order to identify the most heavily network associated nodes. Additionally, the interacting partners of the highly connected super-hub nodes were specifically marked.

5. A final coding simply identified a node as being one of a pair of genes that showed disruption in co-expression when compared between the sample modules. For coding purposes of the network, only one co-expression relation was required to be disrupted for the gene in question to flag the node although co-expression disruption hubs were identified for nodes that showed significant changes in co-expression with three or more other genes.

Thresholds for hubs and super-hubs was determined heuristically to prevent any prior self-imposed restrictions that could occur by forcing a discrete number of protein interactions or disrupted co-expression gene partners (e.g. a minimum of 10 interactions to be considered a highly connected node). Since the discovery network is highly comprehensive by design, no issues with nodes being erroneously classified using the heuristic thresholds were anticipated. Since the probe set of the HG-U133 Plus 2.0 array is designed to be genome spanning, it is a reasonable assumption that the most highly connected protein product in the network is an adequate estimate for the most highly connected protein product in the human genome when taken in conjunction with the network context of the Wnt pathway. However, it is worth mentioning the disrupted co-expressing gene pairs metric may not be an especially practical measure since a great deal of biological “distance” exists between transcription of mRNA (co-expression) and associations of translated products (protein interactions). Nonetheless, it may be a viable metric for analysis even within the topographical context of a protein-protein interaction network and was evaluated as such. Finally, the threshold of one disruption for the final coding may seem low, but the question of interest was to identify direct participants of the Wnt pathway that have altered mRNA co-expression. This made a minimum of one a necessity.

2.10 Evaluation of the discovery network

As previously stated, the primary research question asked was, can any novel

information related to genes, their protein products, their biological interactions and their regulatory control in colorectal adenocarcinomas be determined using a multi-method approach? To address this principle question, a number of smaller questions were considered with regards to the evaluation of the discovery network after the nodes were properly coded. These questions included:

- Are highly connected nodes or genes heavily co-expressed in patient adenoma tissue and/or cancerous cell lines also directly targeted by transcription factors in the Wnt pathway? Do these same nodes exhibit differential mRNA expression under disease conditions?
- Are the interacting partners of highly connected nodes directly targeted by β -catenin? Are these same interacting partners differentially expressed in adenomatous conditions?
- Are there any direct targets of β -catenin also showing significant changes in expression patterns that have not been previously implicated in colon adenocarcinoma development?
- Do nodes showing disruption of co-expression display similar pattern changes in expression under disease conditions? Are these partners also likely to show chromatin binding by β -catenin in a pattern related to their expression profiles?

- Are there any implicated nodes that match multiple coding criteria of the network and warrant further examination as markers for colon tumorigenesis or potential treatment targets?

From a medical treatment/detection standpoint this last question is the most critical. A specific emphasis was placed on finding differentially expressed genes in adenocarcinoma tissue, or gene pairs that are disrupted in the same disease conditions, which are also directly targeted by Wnt signaled transcription factors.

The utility of the discovery network was based as much on the comprehensiveness of its extensive coding as it was on its direct interaction topology. The first goal of the network was to identify patterns between differential expression, levels of protein interaction, disruption of gene co-expression in diseased tissue, and being directly targeted by β -catenin, all within the context of a Wnt protein product network. The second goal was to identify specific nodes/genes that fit multiple criteria; especially genes found to be differentially expressed or show co-expression disruption, and are also direct targets of β -catenin. As mentioned, these nodes could further serve as seeds to make a future *de novo* network for comparison to the current validated network.

2.11 Null network model – comparing observed and permuted proportions of metrics

The statistical significance of the full protein interaction framework was determined through multiple random collections of nodes from the network to examine the

uniqueness of the model. This was accomplished by sampling network nodes and comparing the observed and expected proportions of nodes matching the various permutations of criteria. Four label vectors were used, in turn giving each node a binary representation as to whether it was differentially expressed, a direct target of β -catenin, a network hub or super-hub, or displayed disrupted co-expression under disease conditions. Accordingly, this led to a collection of observed proportions, using the combinations of label vectors, and an experimental network against which control models could be subsequently compared via a null hypothesis positing no difference between the observed and expected proportions of the experimental and null networks.

Numerous bootstrap samples were taken from the network to test model significance. Within the experimental network, a specific number of nodes matched at least one of the individual metrics listed above with 15 different combinations possible (4 individual criteria, 6 pairs, 4 triplets and the full set of four labels). For each permutation, an equal fraction of nodes was randomly selected, with replacement, from the experimental network giving a new random set of proportions for the same fifteen metrics combinations. This expected set of criteria was used to test if the proportion of metric matched nodes expected by chance was greater than the experimentally observed proportions. This permutation and tabulation was repeated 50,000 times. Ultimately, the statistical significance of the experimental network was evaluated by testing if the chance expected proportions were larger than the observed proportions for all the examined combinations. The 15 sets of criteria tested were as follows:

The four individual metrics

- Significant differential expression
- Target of β -catenin
- Network hub or super-hub
- Disrupted co-expression

The six pairs of criteria

- Significant differential expression and target of β -catenin
- Significant differential expression and network hub/super-hub
- Significant differential expressed and disrupted coexpression
- Target of β -catenin and network hub/super-hub
- Target of β -catenin and disrupted co-expression
- Network hub/super-hub and disrupted co-expression

The four metric triplets of

- Significant differential expression, target of β -catenin, and network hub/super-hub
- Significant differential expression, target of β -catenin, and disrupted co-expression
- Significant differential expression, network hub/super-hub, and disrupted co-expression
- Target of β -catenin, network hub/super-hub, and disrupted co-expression

- Finally, all four criteria – differential expression, β -catenin targeting, extensive protein interactions and disrupted coexpression under disease conditions

Although some combinations have less biological applicability, this significance testing was designed solely to assess the uniqueness of the network; therefore, comprehensiveness of the tested proportions was deemed of greater value than focusing on specific combinations of metrics that are more contextually relevant to the study.

CHAPTER 3 – RESULTS

3.1 Literature search of mRNA expression studies

Keeping with the previously outlined criteria, ten different source studies from the literature were found with microarray data available for aggregation. Citations with individual sample counts and descriptions can be found in table 1.

When grouped according to module, the original 117 individual arrays gave final totals of 12 from cell lines (9 HCT116 and 3 LS174T), 52 from patient colon adenomas and 53 from normal patient mucosa. Of the patient tissue, 32 arrays from each group were same-patient matched.

Although many of the studies involving cell lines subjected them to experimental conditions, only the unaltered controls were used. This led to far less replicates but a much more appropriate dataset for the mRNA expression analysis. Fortunately, limited replicates of the cell lines were of less concern due to little overall variation in mRNA expression from generation to generation for the diploid lines. Both cell lines carry mutations to their final β -catenin products that prevent them from being marked for phosphorylation. In turn, they are unable to be degraded in the cytoplasm and easily transfect into the nucleus effectively mirroring the cancerous conditions affiliated with

extensive Wnt activation. As a result of their similarity (diploid, mutations to β -catenin, wild-type alleles of p53 and APC), the two cell lines were grouped into a single module for analysis purposes to further enhance the underlying statistical power.

3.2 Normalization assessment for pooling

Of the three normalization procedures tested (MAS 5.0, dChip, RMA), Robust Multi-array Average gave the best overall normalization when compared during exploratory data analysis. Overall, the quantile normalization method of RMA resulted in much closer median intensities when compared across disease conditions and sample sources. For initial analysis, each module was normalized individually and then normalized as a single group to compare intensities. Both density graphs of module intensities as well as box plots of each array showed RMA had the most effective normalization for the disparate sources. It did so while retaining differences in expression due to variation in sample rather than array analysis technique and keeping the same robustness found in the less stringent normalizations. Density curves for the various module comparisons can be found in figure 4 while the box plots of intensities are shown in figure 5.

3.3 Aggregate study of mRNA expression

Given the three distinct sample groups, the linear model used to test mRNA expression significance contained three different comparisons: normal patient mucosa vs

adenomatous patient tissue (N-A), normal patient mucosa vs cancerous cell lines (N-C), and adenomatous patient tissue vs cancerous cell lines (A-C). In addition to the false discovery rate of 5%, an mRNA differential expression threshold of a 1.5x fold change in either direction was applied to indicate those genes with the most extensive alterations in their expression profiles due to the disease conditions. Gene counts are out of the total 54675 probes on the Affymetrix HG-U133 Plus 2.0 array. For the N-A comparison, 5135 probes (2626 upregulated and 2509 downregulated) met the significance and expression thresholds while the N-C comparison identified 16078 probes (8515 upregulated and 7563 downregulated). The A-C comparison found 11610 probes (5782 upregulated and 5828 downregulated) with differential expression between the two disease modules. For further information see table 2. Breakdown counts of those probes showing differential expression in multiple comparisons can be found in the Venn diagram in figure 6.

Despite the isolation of probes by comparison, it was important to further refine the probe set to identify genes showing concordant expression between the N-A and N-C comparisons. This would ensure that identified genes showed the same expression profile across both the adenomatous and cancerous conditions as would be expected for the typical progression of an adenocarcinoma *in vivo*. Probes with discordant expression on the other hand would be expected from genes involved with regulation or other functions outside the realm of adenocarcinoma development and would be of little interest in the current study. In total, 4356 distinct probes with significant changes

in mRNA expression were found to meet the fold-change threshold of 50% in both the N-A as well as the N-C comparison. Of those, 271 (6.2%) showed discordance with differential expression in different directions between the two comparisons and were not considered for the analysis. For individual counts based on direction of concordant differential expression see table 3. The remaining 4085 probes were further organized into six sets based on direction of concordance (N-A and N-C upregulated or downregulated) in addition to the relative expression of the cell lines compared to the adenomatous tissue (A-C upregulated, downregulated or no change in expression between disease states). Listings of the top 5 probes and their fold changes for each of the six groupings can be found in tables 4a and 4b. Even before the application of the transcription factor binding and the protein interaction network backbone, the pooled analysis had already given an excellent comparison of changes in mRNA expression profiles under the colorectal adenocarcinoma conditions across the entire spectrum of development from normal mucosa to benign adenoma to malignant carcinoma.

Coexpression analysis was done by first filtering the total gene set based on the soon to be described protein interaction network. After limiting analyzed genes to those found in the network, 32 disruptions were found across a total of 29 unique genes. These disruptions were based off their alterations in expression when compared between normal mucosa and patient adenomas.

3.4 Filtering of SACO data of β -catenin binding

By applying the criteria of at least 3 tags in a cluster on a gene, 1636 genes were identified as direct targets of β -catenin under the cancerous conditions of the HCT116 cell line. Of those 1636 direct targets, 213 also exhibited changes in differential expression concordantly in the N-A and N-C comparisons of the mRNA analysis. Of the total set, 115 showed concordant upregulation under disease conditions while 98 were downregulated. The complete list of upregulated and downregulated genes, along with their fold changes for each comparison and the total number of SACO tags across all clusters for that gene, can be found in tables 5.a and 5.b respectively. Note that although an increased number of tags probe certainly implies a greater affinity for β -catenin to that specific gene, it is not necessarily a measure of strength of association or propensity for expression of said gene and should not be seen as such. Nonetheless, meeting the tag criteria previously established unquestionably identifies the gene as a target of β -catenin which was the primary factor in this study.

3.5 The protein product interaction network

Initial assembly of the protein-protein interaction network was carried out via Pathway Commons as a point-of-access repository. Specifically, Cancer Cell Map's and NCI's Natural Protein Interactions databases were used to populate the Wnt pathway network. Once initial compilation was finished, the network was comprised of 168

nodes with 554 interactions. With the backbone of the analysis in place, results from the mRNA expression and β -catenin targeting studies were overlain on the network along with appropriate network metrics to identify levels of connectivity. Of the 168 nodes, 111 also had matching probes on the Affymetrix HG-U133 2.0 Plus array while 16 nodes were identified as direct targets of β -catenin by the SACO filtering. With regards to connectivity, 36 nodes qualified as hubs (5-13 neighbor nodes) while 5 nodes were found as super-hubs (14+ neighbors) due to the uniqueness of their neighbor counts. For comparison, 6 hubs (a number greater than the total of super-hubs) had 13 neighbors. Further details on the network can be found in table 6 and an overview image of the populated network can be seen with figure 7.

3.6 Network nodes with single metrics of interest

Although emphasis was on the identification of nodes meeting multiple criteria, it was deemed important to also determine which genes within the network were prominent participants in the role of the Wnt pathway in colorectal adenocarcinoma development simply based on single metric involvement. Towards that end, 20 genes were identified as having upregulation of mRNA expression under the disease conditions (3 in patient adenomas only, 11 in cancerous cell lines only, and 6 upregulated in both conditions, see table 7). With respect to downregulation of expression, 17 additional nodes were found (none unique to adenomas but 12 in cancerous cell lines and 5 in both, see table 8). As stated in the expression study results, 29 nodes were involved in 32 different

coexpression disruptions under adenomatous conditions with 7 nodes showing disruptions with three or more other nodes (see table 9). The 16 SACO targets of β -catenin have already been mentioned, but there were also 30 first-degree neighbor nodes of those targets, identified as potentially informative downstream participants (see table 10). Therefore under similar reasoning, beyond the 35 hubs and 5 super-hubs, 41 first-degree neighbors of the super-hubs were also isolated (see table 11). Finally, the previously validated targets found on the Wnt homepage accounted for 13 nodes with 23 first-degree neighbors of their own (see table 12). Tables 7-12 contain extensive color-coding to aid in tracking nodes across the various combinations of evaluated metrics that were assessed; a descriptive legend of the codings can be found in figure 8.

3.7 Nodes matching multiple criteria and a final network

While the single metrics were of great value, of particular interest were those nodes which fit into numerous different categories. Among the most notable were the 16 direct β -catenin targets of the network. A total of 5 targets were found to be upregulated under disease conditions (N-A only – FHL2, AXIN2; N-C only – FBXW11; both conditions – CCND1) while 2 were downregulated (LRP1, TAX1BP3 - both in the N-C comparison only). Furthermore, 2 targets were also super-hubs (CSNK1A1 and CTBP1) with 8 of the hub neighbors targets themselves (in addition to hubs CSNK1A1 and CTBP1 - FBXW11, FHL2, MACF1, NFATC2, TCF7L2, TLE1). Iterations continued by also observing

the 30 neighbors of β -catenin targets that were additionally upregulated (Total 4; N-A only – 1; N-C only – 2; both conditions – 1), downregulated (4 total; all in the N-C comparison only), super-hubs (all 5 super-hubs were target neighbors) and hub neighbors (21 total).

A similar set of secondary classifications was carried out on the super-hubs and their first-degree neighbors. Although none of the super-hubs were upregulated in patient adenomas or cancerous cell lines, 1 was downregulated (CTNNB1 in the N-C comparison only). As mentioned earlier, CSNK1A1 and CTBP1 were identified as targets of β -catenin and all five super-hubs were neighbors of direct targets. With respect to the 42 super-hub neighbors, 7 nodes were found to be upregulated (N-A only – 2; N-C only – 4; both conditions – 1) while another 7 were downregulated (N-A only – 0; N-C only – 6; both conditions – 1). Direct targeting by β -catenin was found for 8 of the super-hub neighbors with 21 also being neighbors of directly targeted nodes.

The identification of upregulated, downregulated, targeted or hub nodes was overlapped with two more criteria: disruption of coexpression under adenomatous constraints and previous identification as a target of the Wnt pathway as per the Wnt homepage. The collection of disrupted nodes included 5 upregulated genes (N-A only – 1; N-C only – 3; both conditions – 1), 7 downregulated (N-A only – 0; N-C only – 4; both conditions – 3), 4 β -catenin targets and 10 target neighbors, and a single super-hub (AXIN1) along with 8 super-hub neighbors. For the final metrics pairings, the previously

identified Wnt targets included 5 upregulated genes (N-A only – 1; N-C only – 3; both conditions – 1), 3 downregulated genes (all found in the N-C comparison only), 2 β -catenin targets and 4 direct target neighbors. None of the super-hubs were implicated but 5 super-hub neighbors were.

Having determined those genes strongly implicated in Wnt-related development of colorectal adenocarcinomas via multiple categorization, a final network was compiled showing the interactions between all identified nodes. This refinement (representation in figure 9 with node metrics in table 13) was the final network used for visualization of critical nodes. With a selection of pertinent genes, categorization and annotation was straightforward as well as enlightening.

3.8 Assessment of the null model

With the production of the final network complete, the uniqueness of the model was tested by comparing the experimentally observed and randomly collected prediction proportions for nodes meeting the multiple network criteria in tandem. Of the 168 nodes in the final network, 107 displayed at least one of the overlain metrics. Thus, each permutation consisted of 107 nodes randomly collected from the experimental sample space to create a new set of proportions. Specifics on the observed and expected proportions for each of the 15 tested metrics combinations can be found in table 14. As shown, none of the calculated proportions of label combinations were

found to exceed the proportions seen within the experimental network ($p < 0.0001$ for all 15 proportions), showing a significant network representative of the actual patterns of differential expression, transcription factor binding, protein interactions and co-expression disruption found in colorectal adenocarcinomas. In truth, this is the best validation for the study's model since it corroborates the uniqueness of the network. No matter how well-described, enlightening or informative the network is, it can only be of real value if it is shown to be significantly valid and not simply an amalgamation of chance.

3.9 Annotation of function for identified genes/nodes

The final selection of 65 nodes was organized based on annotation and ontologies, with the entire collection easily fitting into one of eight categories of biological activity; the various functional groups and their corresponding genes included:

- **Cell growth, proliferation and life cycle moderation**

CCND1, CDC42, CTBP1, PIN1, PPP2R5D

- **Phosphorylation and ubiquitination**

AXIN1, AXIN2, FBXW11, APC, BTRC, CUL1, SKP1, GSK3B, CSNK1A1, CSNK1D,
MAP3K4, MAP3K7

- **Cell fate, differentiation and polarity**

AES, SOX1, CDX1, MARK2

- **Epigenetics and histone modification**

HDAC1, SALL1, RUVBL1, CREBBP

- **Cell-cell interactions, molecular transport and inter-cellular signaling**

MAGI3, DLG1, GNG2, CDH1, MACF1, RAN, RANBP3, PIAS4

- **Wnt pathway – receptors and cofactors**

DVL1, DVL2, FZD1, FZD3, FZD5, LRP1, LRP6, DKK1, KREMEN2, TLE1

- **Ubiquitously expressed genes**

CTNNB1, TCF4, TCF7L2, MYC, FHL2, NFATC2, PRKCA, TAX1BP3, FRAT1, SMAD3,
SMAD4, ARRB2

- **Expression localized outside the colon**

LRP5, NLK, RUNX2, SFRP2, FZD7, WNT3A, CAMK2D, DAB2, MAP1B, NR5A1

Ultimately, this collection of genes was the final dataset for the study. As a whole it constituted, within the context of a protein interaction network of the Wnt pathway, a collection of genes notable for their involvement in the development and progression of colon adenocarcinomas. By combing multiple metrics including changes to mRNA expression profiles, targeting by the canonical Wnt transcript factor β -catenin, and specifics to the interaction network itself, a compilation of pertinent genes became

available to evaluate for possible roles in the detection and intervention of colorectal cancer.

CHAPTER 4 – DISCUSSION

4.1 Initial analysis – differential expression profiles and direct β -catenin targeting

Even the earliest assessments of the study yielded insight towards potential genes of interest in colorectal adenocarcinoma development. Specifically this includes the overlap of probes showing significant differential expression in the HG-U133 2.0 Plus arrays as well as being identified as direct targets of β -catenin. Even before the application of the protein interaction framework, this initial analysis immediately identified key genes whose expression and binding profiles, alongside their functional activity, could prove them to be efficacious targets in discovery and treatment. Of the most differentially expressed genes shown in tables 4.a and 4.b, five probes in particular stood out among the annotated set.

CDA – 5 SACO tags; N-A FC: -3.09; N-C FC: -2.06. The final protein product for this gene is Cytidine Deaminase, a standard pyrimidine scavenger. Of particular note, the CDA protein also catalyzes deamination of nucleoside analogs such as the therapeutic Ara-C. This becomes especially key since administration of Ara-C has become a common treatment for acute myeloid leukemia with associated drug resistance attributed to CDA activity [50]. Although its presence leads to a

dampening of therapeutic treatments in some cancers, it was interesting to see CDA showing reduced expression under the current disease conditions.

CDH3 – 9 SACO tags; N-A FC: +6.43; N-C FC: +13.12. Leads to the production of Cadherin3, one of the many membrane-bound glycoproteins involved in calcium-dependent cell-cell adhesion. Of particular interest is evidence that cells expressing CDH3 tend to bind together preferentially over parental cell lines showing reduced expression of the gene [51]. Alongside the increased mRNA transcription, this suggests a precedence for cells constitutively expressing CDH3 to affiliate together, creating a direct implication towards the promotion of adenocarcinoma development. Note that multiple cadherin genes exist albeit without the intriguing profile of CDH3. For example, note that CDH1 (PPI network member and known Wnt target) was found to be downregulated in cancerous cell lines (N-A -1.206, N-C -3.115).

KIAA1199 – 17 SACO tags; N-A FC: +26.95; N-C FC: +7.84. One of the more novel probes implicated by this study, KIAA1199 has recently come under scrutiny for its possible involvement in adenoma development, specifically by one of the source studies used for the pooled analysis of mRNA differential expression [52]. In normal tissue it is typically expressed in the cochlea of the ear although its specific function is largely uncharacterized [53]. Nonetheless, KIAA1199 has also been located in intestinal crypts, and in conjunction with its expression and binding

patterns, could be well worth further investigation towards its function in colon carcinogenesis.

MAMDC2 – 7 SACO tags; N-A: -5.71; N-C: -3.42. Final protein product is a member of a protein family characterized by their possession of a MAM (mephrin, A5 antigen, protein tyrosine phosphatase mu) domain. Much like KIAA1199, the specific function of this gene is unknown. Previous experimentation discovered the gene during walks on chromosome 9 while looking for altered genes due to the congenital disorder Kabuki Syndrome [54]. Although MAMDC2 was unaltered for that disease, the differences between a pediatric development disorder and a nascent adenocarcinoma are rather obvious. Given MAMDC2's notable expression and binding profile, it may well be worth further research to determine a precise role in tumorigenesis.

PYY – 5 SACO tags; N-A FC: -29.08; N-C FC: -39.82. Expression leads to the final protein product Peptide Tyrosine Tyrosine. The protein is secreted from the pancreas and regulates both the intestine and duodenum through a combination of inhibition of gastric acid as well as digestive enzymes [55]. While much characterization has been related to appetite regulation (e.g. PYY application and food intake), interestingly, peripheral application has been found to upregulate expression of the proto-oncogene c-fos [56]. A strong inhibition under the presence of β -catenin suggests a critical role for gastric acid and digestive enzymes

in the promotion of a normal mucosa state. Activation of PYY and other related genes may in turn have a role as potential treatment options for adenomas and possibly even cancerous tumors.

Even without the backdrop of the *in silico* Wnt pathway network, a great deal of insight into noteworthy participants in carcinogenesis was readily available from just the pure experimental data.

4.2 Biological annotation – the 8 functional activity categories for the refined network

The final refinement to the network was the inclusion of functional headings to organize the biological activity of the previously identified nodes; instead of being reductive, this created another layer of information for the network. Nonetheless, it was important to design appropriate categories that reflected the typical activities inherent to colorectal disease development but also not restrictively exclusive or limiting. With all the nodes originally united by a common theme of Wnt involvement, a focus towards specific functions which promote tumor formation and progression was considered of greatest importance and can be seen in the first five categories. The final three categories (Wnt participation, ubiquitous expression, and non-intestinal localization) were created to account for genes which, although no less potentially critical, did not readily fall under one of the banners related to adenocarcinoma development.

4.3 Cell growth, proliferation and regulation of the cell cycle

Unquestionably, uncontrolled cell growth and reproduction are hallmarks of carcinoma development in the whole body, including the intestine, making this a rather clear-cut category. Genes involved with cell cycle regulation in particular are especially key points of interest and worthy of further study. This is especially true when they are found to be differentially expressed or translated under adenomatous or cancerous conditions. Of the five genes placed in the group, the most notable is the cyclin CCND1, one of the few nodes identified for both direct β -catenin binding as well as differential expression. Its upregulation (common in both carcinoma and adenoma expression profiles) eminently suggests disruption at the G phase wherein individual cell growth is limited and rapid cell division is encouraged [57].

Although not bound by β -catenin directly, CDC42 is a neighbor of a transcription factor target as well as being downregulated, a combination suggestive of downstream activation. Although the reduced expression seems misleading since CDC42 is a known activator of the cell cycle, it is also involved in assymetrical cell division in epithelium [58]. Considering adenocarcinoma propagation requires symmetrical division to maintain the undifferentiated status of the daughter cells, the reduction in mRNA expression fits nicely given the context. The major phosphatase PPP2R5D also follows an initially deceptive expression profile since, as a negative cell cycle regulator, it would be expected to limit a propensity towards uncontrolled cell reproduction. However, it is

commonly found in the nucleus of mitotic and recently divided cells [59], giving its presence in the disease conditions just as much credence with the rapid division characteristic of adenocarcinomas.

Supplementary downstream activity is even suggested by the classification of CTBP1 as both a super-hub and direct target of β -catenin. Although no changes occur in its specific expression, it is a known regulator of transcription especially during development, one of the times when Wnt activation is highly common. It would not be a stretch to envision a similar regulatory role being played by CTBP1 in adenocarcinomas, with similar indirect changes taking place due to a secondary control.

4.4 Phosphorylation and ubiquitination

This category was designed to emphasize the role of β -catenin since its destruction is preceded by the phosphorylation and ubiquitination required to mark it for degradation. As mentioned, HCT116 cells themselves are characterized by a mutation to one of β -catenin's phosphorylation sites which allows them to mirror the abundance of the transcription factor in adenocarcinomas. Not surprisingly, many of the key participants in the β -catenin marking and destruction complexes were identified in this study. Indicated members of the destruction complex included APC, SKP1 and AXIN2 although only the last one displayed any notable metrics [60]. Still, those metrics included being a super-hub, a direct target of β -catenin and even upregulation of mRNA expression in

both adenomas and carcinomas. In a way, this suggests a reaction by cells as an attempt to stave off a plethora of β -catenin by utilizing its own excess to promote production of an extensive amount of AXIN2, even if the destruction complex is ineffective.

Members of the SCF ubiquitination complex were also found in force, including CUL1, SKP1, FBXW11 and BTRC [61, 62]. The last two in particular showed an increase in expression (perhaps for the same proposed reasons leading to upregulation of AXIN2) with FBXW11 also being a direct target of β -catenin. Finally, it is worth mentioning the super-hub CSNK1A1 which was also a direct target. Much like CTBP1, direct targeting on a super-hub implies a certain amount of downstream activity and secondary regulation. Instead of transcriptional control, CSNK1A1 is an ubiquitously expressed phosphorylating compound that is known to target β -catenin [63]. While the activity CSNK1A1 in adenocarcinomas is not as readily explainable as other nodes, its pervasiveness and clear role in the Wnt pathway make it a prime candidate for further research.

4.5 Cell fate and differentiation, and epithelial cell polarity

With so much of “proper” tumor development reliant on uncontrolled cell duplication, maintenance of an undifferentiated state becomes important for the malignancy to continually divide as well as metastasize into other tissues. In smooth tissues like the

intestine, tumor promotion is further confounded by the tendency of cells to adopt a certain type of polarity in response to their desired growth patterns to accommodate tissue migration and account for creation of an epithelial sheet. All comments about encouraging tumor growth aside, the regulation of cell differentiation and polarity are clearly vital aspects in adenocarcinoma development in the colon.

The downregulation of CDX1 and AES both seem clear considering their involvement in determining cell fate. CDX1 in particular is noted for its role during development, especially in encouraging the differentiation of intestinal tissue [64]. AES has functionality similar to CDX1 in cell fate determination but has been previously identified in developing neurons [65]. Conversely, SOX1 is well known for its role in inhibiting differentiation, although again most commonly localized within neurons during embryogenesis [66]. Specifically, SOX1 is found in self-regenerating neurons that promote cell reproduction while limiting cell determination. Under different environmental conditions this would be fitting with the description of a developing tumor. Although changes in the mRNA profile were only found in patient adenomas, it was still intriguing to identify the increased expression of a negative regulator of cell fate.

4.6 Epigenetics and histone modification

Even though epigenetics is not a commonly cited facet of the Wnt pathway, its

involvement in tumor suppression and regulation in general is heavily characterized and its association with colorectal adenoma development is no exception. Notably, all epigenetic effects appear focused on acetylation based activities rather than alternate modifications such as transcription inhibition through histone methylation or direct, atypical alterations done to specific gene nucleotides. Acetylation of histones has long been known to encourage gene transcription since the bulky electrostatic additions allow access to the DNA contained within chromatin. Within the context of the cancerous cell lines, it was discovered the histone deacetylase (and known expression inhibitor [67]) SALL1 was downregulated while the acetyl transferase RUVBL1 was upregulated. As a promoter of gene expression, RUVBL1 is specifically cited for its promotion of numerous cascade pathways, including Wnt [68]. Despite none of the nodes involved with epigenetics being direct targets of β -catenin, their role in aiding adenocarcinoma development through the Wnt pathway is striking.

4.7 Cell-cell interactions, molecular transport, inter-cellular signaling

Much like polarization and differentiation, cell-cell interactions, signaling and adhesion are distinct hallmarks of epithelium. Despite the loss of typical function during tumor development, tissues can still behave as they would under normal circumstances. When certain functions of a tissue directly impact adenoma or carcinoma growth and propagation there is often a selective advantage towards increasing such activity while limiting inhibition of functions counter to adenocarcinoma development. In the case of

epithelial tissue in the colon, expression related to cell-cell adhesion, mobility along the intestinal wall, and signaling between cells in the epithelium sheets should in turn be selected for so as to encourage the proliferation of tumorous cells in a disease environment. This appears to be the case with MAGI3, DLG1 and CDH1. All three genes were underexpressed in disease conditions, and from the perspective of an adenocarcinoma with good reason. MAGI3 is often expressed at cell-cell junctions and has been identified as an associative protein of PTEN, a known growth suppressor [69]. DLG1 is found at cell-cell junctions also although largely in the scaffolding required for proper neuron development. Instead of associating with a tumor suppressor, DLG1 is a suppressor itself [70]. CDH1 is a member of the same cadherin family as the previously discussed CDH3. Like other cadherins, CDH1 has key role in the motility and adhesion of epithelial sheets. Although downregulated (unlike CDH3) the loss of function of CDH1 has been heavily associated with tumor malignancy and seeing reduced expression in cancerous cell lines is to be expected [71]. Note only one direct target of β -catenin was found in this grouping: MACF1, a microtubule affixing protein at cell junctions [72].

Beyond the importance of the stable coherence of the epithelial sheet, molecular signaling and effector transport have their own role in both the Wnt pathway as well as tumorigenesis. For example, PIAS4 was found to have increased expression appropriate for its role as a sumoylation tether which inhibits LEF1 (a member of the β -catenin destruction complex) while enhancing TCF4 (the co-factor of β -catenin) and subsequently promoting Wnt activity [73]. Changing focus to inside the cell, nuclear

transport of β -catenin is a sometimes understated aspect of Wnt activation yet is no less critical than other steps. RAN and RANBP3 (both members of the family of ras-related nuclear proteins) are both critical in the transport of material into the nucleus, such as transcription factors like β -catenin [74]. Seeing their increased expression in patient adenomas and cancerous cell lines gave even further weight to the role of increased Wnt activation in the development of colorectal cancer.

4.8 Participants in the canonical Wnt pathway – cofactors and receptors

Given the network's backdrop of Wnt pathway involvement, it was fully expected that many genes would be key members in the regulation of β -catenin. Correspondingly, many receptors and associative compounds were identified within the network. Nonetheless, the true benefit of the overlay of multiple metrics was the refinement of the protein interaction network to identify those Wnt participants seemingly related to the fundamentals of adenocarcinoma development. Both families of cell surface receptors of Wnt were represented including FZD1, FZD3, FZD5, LRP1 and LRP6 [75]. Of key note was the downregulation of the sole direct β -catenin target LRP1. This is one of the few cases where a direct target of β -catenin under cancerous conditions was found to be down regulated. In many ways this is suggestive of a negative regulation by a transcription factor noted for activation of mRNA transcription. This is even more interesting since the same inhibition is being placed on a protein whose purpose is to receive (thus galvanizing) Wnt signaling. FZD3 and FZD5 added further layers of intrigue

by being upregulated and downregulated respectively. Although a simpler explanation may exist, such as auto-regulation of FZD5 and LRP1 by β -catenin coupled with a preference towards FZD3 in adenocarcinomas, such a convoluted pattern of expression could easily warrant further investigation.

Additional questions were raised during consideration of the activities of associative proteins like TLE1 and DKK1 [76]. DKK1 associates with KREMEN to form a complex that usurps LRP surface receptors thus limiting Wnt activation [77]. However, DKK1 was found to be overexpressed in disease conditions although this may be due to the often discussed self-regulation. Although there were no changes in the transcription of TLE1, it is a direct target of β -catenin and has been determined prior as a corepressor of β -catenin along with the general TCF family of transcription factors [78]. Although there is no evidence for a direct negative regulation of TLE1 by β -catenin, observing a transcription factor bind to its own corepressor has fascinating implications.

Luckily, other Wnt participants have much more understandable behavior within the pathway's context. Among these are the Dsh homologs DVL1 and DVL2. Both are critical in promoting β -catenin by commandeering AXINs and binding to the cell surface in the presence of Wnt [79]. As a result, the destruction complex is unable to form and degradation of β -catenin is halted. As such, it was no great surprise to see both DVL1 and DVL2 upregulated in the cancerous cell lines. In fact given the rather counterintuitive behavior of other Wnt participants, seeing such an easily explainable

and discrete example of Wnt activation in adenocarcinomas was in many ways encouraging.

4.9 Ubiquitously expressed genes – generalized functionality and multiple biological circumstances

Another example of using the different metrics as a refinement of the protein interaction network was the isolation of specific nodes known for their ubiquitous expression in multiple tissues and numerous biological conditions, yet now are observed for their involvement with adenocarcinoma progression specifically. The most obvious examples were CTNNB1 (the most heavily connected node in the network) and TCF4, the members of the actual transcription activation complex in the canonical Wnt path [80]. While initial expectations would purport an increase in transcription, both genes displayed downregulation under cancerous conditions. Due to the inability of β -catenin to be degraded in the cell lines, it stands to reason there could be an effective saturation of functionality wherein less β -catenin or TCF4 is created simply because continued production no longer serves as beneficial. A similar motivation may describe the downregulation of FRAT1 which traditionally inhibits phosphorylation of β -catenin by binding to GSK3B therefore increasing Wnt activity [81]. If conditions within the cancerous cell lines were such that there were an effective limit on the amount FRAT1 that could disrupt the destruction complex, a reduction in its expression may be expected much like the suggested downregulation of CTNNB1 and TCF4.

The most notable gene within this category shared a similar pattern to LRP1; both being direct targets of β -catenin yet showing decreased mRNA expression under adenocarcinomas conditions. Unlike LRP1, this gene (TAX1BP3) has its explicit function largely uncharacterized aside from its PDZ binding/signaling domain. Evidence has previously linked it to the Rho protein family of repressors yet its role remains ambiguous [82]. On a circumstantial note, the related protein TAX1BP1 has been determined to be involved with inducing apoptosis [83]. With an odd profile as a downregulated target and largely uncharacterized activity, the interactions and functionality of TAX1BP3 stand as a bastion of potential in further understanding the progression of adenocarcinomas.

4.10 Expression localized outside the colon

When first being populated, the initial protein interaction network looked at the numerous compounds involved in the Wnt pathway with little consideration towards biological function, level of expression or localization in the body. With this in mind, it would not be unexpected to observe nodes that are largely expressed outside the colon. This has already been seen with previously identified nodes such as DLG1 and AES, both of which are typically found in neurons. Although this collection of extra-colonic genes may at first appear to be nothing more than false positives and artifacts merely matching multiple metrics due simply to the pro-Wnt environment, they could very

easily represent new conditions and functions found beyond their typical utility and have additional roles in promoting adenocarcinoma development.

One example uses another key participant in neuronal development, MAP1B, which showed an underexpression of mRNA in disease conditions. Traditionally, MAP1B is used for joining microtubules across neurons during embryogenesis and aiding in plasticity. However, overexpression of the gene has been linked with inducing cortical neuron death [84]. Perhaps a similar trend could be applied in intestinal epithelium accounting for the reduced expression that would otherwise be detrimental to the growing adenocarcinoma.

Continuing with ostensibly neuronal genes, the calcium-dependent kinase CAM2KD is frequently located within neurons to aid in long-term potentiation and has a critical role in regulating differentiation in cardiac tissue. Although a battery of Ca^{2+} kinases exist, CAM2KD is singled out since it is also a direct target of β -catenin in colorectal cancer. Further pointing to a possible role in carcinoma development, CAM2KD is also the kinase variant most critical in determining the final phenotype of a number of cardiac diseases as well as being the dominantly expressed variant in multiple cancer cell lines [85]. Although no changes in expression were found under disease conditions, a potential downstream or secondary effect could be a driving force for this highly determinant gene's role in adenocarcinoma progression.

The final notable node outside the colon, DAB2, met many of the descriptive hallmarks of colorectal cancer. The gene is a known messaging component for the SCF ubiquitination complex which normally degrades β -catenin. Furthermore, its reduced expression in cancer is not uncommon. As such, seeing its downregulation in the cancerous cell lines appeared as standard validation. What was surprising was noting that DAB2 is largely found in ovarian epithelium and not in the intestine [86]. Given the activity of DAB2 in both cancer and normal tissue and considering the similarity between ovarian and intestinal epithelial sheets, DAB2 may work just as effectively as a marker for colorectal adenocarcinomas as it does for the ovarian variant.

4.11 Observations on nodes in the canonical Wnt pathway

Of the many trends taken from the organization and annotation, some of the most interesting were the explicit roles and changes wrought on some of the key participants of the canonical Wnt pathway. As mentioned, both CTNNB1 and TCF4 showed downregulation, showing reduced expression as adenocarcinomas develop although this could be explained by functional saturation within the cell lines. Of the three FZD nodes identified, FZD3's increase in transcription could suggest a preference for that specific cell surface receptor over members FZD1 and FZD5. Arguably the most common LRP surface receptors are LRP5 and LRP6 although neither was strongly impacted in the disease conditions; the downregulated target LRP1 is most noted for displaying such an odd coding profile and is a rare node implying negative regulation by

β -catenin. Finally, of all the members of the destruction complex for β -catenin, none were more heavily indicated by the model than AXIN2. When compared to the more constitutively expressed AXIN1, AXIN2 was found to be a direct target of β -catenin, overexpressed in disease conditions, and as a super-hub was one of the most heavily connect nodes in the protein interaction network. Although its increase transcription seems largely auto-regulatory, there is no question regarding the pivotal role AXIN2 plays in moderating adenocarcinomas.

4.12 Final trends

The annotations of functions were the final critical portion in a study uniting mRNA expression profiles, transcription factor binding data, protein interactions and functional ontologies into a descriptive framework detailing the development and progression of colorectal adenocarcinomas. The combination of genes behaving as anticipated alongside genes acting counter to expectations highlighted a number of specific nodes and general trends pertaining to the relation between Wnt and colon cancer and the use of potential genes as targets for treatment or detection.

As consistently mentioned, reproduction of an undifferentiated cell is a hallmark of cancer development. Disruptions in cell cycle regulation due to malfunctions in CCND1 have long been associated with cancer growth but now arise in conjunction with binding by β -catenin. Of practical interest, the maintenance of the undifferentiated state by

expression of SOX1 in neurons may have new applications as a possible marker for early tumorigenesis resulting from an upregulation exclusive to adenomatous tissue.

The possibility of negative regulation by β -catenin, whether direct or otherwise, has been mentioned while discussing genes such as TAX1BP3 and LRP1. Of a similar theme is the targeting of the super-hubs CTBP1 and CSNK1A1. Although the hubs show not direct changes in expression themselves it is often the first-order neighbors of hubs that show the distinct regulatory activities. Even considerations of negative regulation aside, having super-hubs also be direct targets of β -catenin opens much potential towards investigation into the functionality of their interacting protein neighbors. Still, both super-hubs mentioned have typical functions that appear counter towards promoting an adenocarcinomas state: CTBP1 as a known transcription repressor and CSNK1A1 is a common phosphorylating agent. Perhaps some sort of inhibitory regulation would not be so ill-suited for the transcription factor.

Even with the loss of proper function, many genes still take an expected role in combating against the cancerous state. Members of the SCF ubiquitination complex such as BTRC and FBXW11, as well as the oft cited destruction complex participant AXIN2, all showed increased expression ostensibly in an attempt to reduce the excess of β -catenin typical of adenocarcinomas. Whether this was due to specific activation of these components or reduced expression of the other complex members followed by

blanket activation of all participants in β -catenin degradation is just one of many prospective questions arising from this study.

An undeniable benefit in using both patient adenomas as well as cancerous cell lines is the ability to evaluate adenocarcinoma progression across its full spectrum of development. Although genes showing altered mRNA transcription in either patient tissue or the cell lines were valuable in their own right, specific attention was reserved for genes showing differential expression in both disease conditions. Some are more easily described; MAGI3's role in association with the growth suppressor PTEN could easily account for its pervasive reduction in transcription, a pattern mirrored by the epigenetic repressor SALL1. Persistent changes in expression were not limited to downregulation however. From the earliest stages of adenocarcinoma development, RUVBL1 (an epigenetic activator) and RAN (a critical nuclear translocator) are upregulated, a finding well in line with their promotion of Wnt related activity.

With the breadth of insight taken from the full network analysis notwithstanding, arguably the most interesting nodes were those who lacked deep functional ontologies or were localized in non-intestinal tissues yet had now become notable genes of interest in colorectal adenocarcinoma development. CAM2KD, MAP1B and DAB2 are all expressed outside the colon (in cardiac tissue, neurons and ovarian epithelium respectively) yet have functionality and network profiles suggestive of colorectal involvement beyond their native tissues. More intriguing still were those genes not

readily described, not due to conflicting network metrics, but because their functions are still so uncharacterized. TAX1BP3, FHL2 and the non-network probe KIAA1199 are all directly bound by β -catenin and each show significant differential expression yet none are especially described in the literature. Even still, there are some functions that hint at potential roles in adenocarcinomas. As previously described, the downregulated TAX1BP3 contains a PDZ domain with the potential to inhibit Wnt signaling and increased expression of TAX1BP variants is known for inducing cell death. Additionally, FHL2 is a known coactivator in RhoA signaling as well as a negative regulator of the transcriptional repressor E4F1; its overexpression implicates similar roles in colorectal carcinogenesis. Further classification of these genes and their functions in adenocarcinoma development may herald new insights in tumorigenesis that may have been otherwise overlooked, and such a capacity for novel discovery and identification alone substantiate the merits of this study.

CHAPTER 5 – CONCLUSIONS

5.1 Limitations of the current study

Although many captivating results highlighting the many facets and identifying many novel points of regulation in adenocarcinomas, this study was not without its limitations. Efforts to encompass a breadth of development points in the progression from normal intestinal mucosa to adenoma and finally carcinomas were the motivation behind using patient adenomas as well as cancerous cell lines. In spite of the replicability of cell lines, the inclusion of expression data from patient carcinomas would be an easy and very valid addition to the descriptive network metrics. Such an inclusion would also allow for a much more in-depth assessment of changes in expression with respect to progression of adenocarcinomas.

Even with such added information, there is always the underlying concern with the veracity of the studies used for the pooled analysis. Attempts to maintain integrity of the aggregation included the limitation of studies to the use of only a single (although very comprehensive) type of array although this in turn hindered the availability of sample arrays, especially replicates of the cell lines. Even with such safeguards there is still little end-user control of methods used for microarray procedures or tabulation of

the raw expression intensities. Ultimately, there will always be a reliance on the authenticity of the source experiments.

While the use of the protein interaction network was helpful for establishing an underlying framework, it was also restrictive by limiting nodes to genes whose protein products were chosen for their participation in the Wnt pathway, preventing analysis of any prospective involvement of β -catenin in other systems and pathways. Although expansion of the network would certainly add greater depth, there would still be the contextual concern due to the restriction of using protein interactions only and excluding any interactions or relationships between genes at earlier stages of biosequence synthesis. Still, these issues are as much the foundation for future directions as they are acknowledged concerns.

5.2 Future directions

Many of the limitations in fact serve as potential seeds for expanding this study to create an even more comprehensive profile of colorectal adenocarcinomas. As was previously highlighted, applying data on differential expression from patient carcinomas would add a final transcriptional facet, completing an overview of mRNA regulation for colon cancer across all parts of the developmental process. In a similar vein, proteomics data from adenomas and carcinomas could also be overlain to create a broader picture

of gene activation beyond mRNA transcription and even better tie in with the protein product framework of the interaction network.

Instead of simply adding additional metrics, the network itself could also be extended. Certain genes, such as differentially expressed targets of β -catenin, stood out more so than other nodes and would serve as prime candidates as expansion seeds. By including the non-Wnt related neighbors of these seeds and reapplying the evaluated metrics, even more novel and intriguing genes could be identified and further investigated to see what downstream and secondary roles β -catenin plays in colorectal adenocarcinoma development.

5.3 Contributions to the field

No matter the limitations and potential directions, this current study still stands as a valid assessment in the role of many Wnt related genes in adenocarcinomas through a combination of multiple biological metrics. In the basest and most concrete terms, multiple genes were implicated during analysis for their markedly altered behavior under the disease conditions, ranging from differences in mRNA expression profiles to transcription factor binding to dynamics of protein-protein interactions. Clinically, isolation of these noted genes has potential benefit both diagnostically (e.g possible markers to identify the adenomatous condition like the overexpressed direct target CCND1) as well as prospective treatment sites worth investigation for their direct role in

colorectal disease, such as promoting increased expression of CDH1 for possible therapeutic purposes.

As detailed in the potential future directions, the current network and metrics can easily be viewed as simply an early step in the continued assessment of adenocarcinomas genes. The combination of chromatin occupancy, differential expression and protein interaction provides a design ready-made for elaboration and subsequent analysis of genetic marks in colorectal adenomas and carcinomas. Subsequent metrics can easily be added (e.g. expression profiles of carcinomas) for a more refined analysis or the network can be expanded based on seeding by nodes previously implicated during the current study. In either case, this current network represents a starting point for new research as much as it is a stand-alone project.

Beyond identification of potentially interesting adenocarcinomic genes and continued enhancement of the network, this entire process as a whole exists as a robust model for disease analysis with unification from gene activation all the way to final protein product interactions across an entire timeframe of disease development. Although currently specific to the colorectal context, this method of combining multiple results across the entire spectrum of biosequence synthesis has applicability to numerous disease conditions and cellular pathways. The use of metrics beyond the basic standards and removing them from isolation can greatly aid the understanding of disease dynamics and highlight important participants and associations that may

otherwise be missed using only a single profiling method. Ultimately, this is perhaps the greatest contribution by this study: a widely applicable and comprehensive method of disease assessment and evaluation. Beyond any significant utility resulting from this study's gene identification, its relevance with other disease may be its greatest benefit of all.

REFERENCE LIST

1. WHO Media Centre: **World Health Organization - Cancer 2007**
<http://www.who.int/mediacentre/factsheets/fs297/en/index.html>.
2. Croce CM: **Oncogenes and cancer**. *N Engl J Med* 2008, **358**(5):502-511.
3. O'Connell JB, Maggard MA, Ko CY: **Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging**. *J Natl Cancer Inst* 2004, **96**(19):1420-1425.
4. Shaker A, Rubin DC: **Intestinal stem cells and epithelial-mesenchymal interactions in the crypt and stem cell niche**. *Transl Res* 2010, **156**(3):180-187.
5. Yen TH, Wright NA: **The gastrointestinal tract stem cell niche**. *Stem Cell Rev* 2006, **2**(3):203-212.
6. Meinzer HP, Sandblad B, Baur HJ: **Generation-dependent control mechanisms in cell proliferation and differentiation--the power of two**. *Cell Prolif* 1992, **25**(2):125-140.
7. Tahara E: **Molecular biology of gastric cancer**. *World J Surg* 1995, **19**(4):484-488; discussion 489-490.
8. Half E, Bercovich D, Rozen P: **Familial adenomatous polyposis**. *Orphanet J Rare Dis* 2009, **4**:22.

9. Lindor NM: **Familial colorectal cancer type X: the other half of hereditary nonpolyposis colon cancer syndrome.** *Surg Oncol Clin N Am* 2009, **18**(4):637-645.
10. United States Cancer Statistics: **Centers for Disease Control and Prevention - Colorectal Cancer 2007** <http://www.cdc.gov/cancer/colorectal/>.
11. Kinzler KW, Vogelstein B: **Lessons from hereditary colorectal cancer.** *Cell* 1996, **87**(2):159-170.
12. Grigoryan T, Wend P, Klaus A, Birchmeier W: **Deciphering the function of canonical Wnt signals in development and disease: conditional loss- and gain-of-function mutations of beta-catenin in mice.** *Genes Dev* 2008, **22**(17):2308-2341.
13. Smith KJ, Johnson KA, Bryan TM, Hill DE, Markowitz S, Willson JK, Paraskeva C, Petersen GM, Hamilton SR, Vogelstein B *et al*: **The APC gene product in normal and tumor cells.** *Proc Natl Acad Sci U S A* 1993, **90**(7):2846-2850.
14. Fearon ER, Vogelstein B: **A genetic model for colorectal tumorigenesis.** *Cell* 1990, **61**(5):759-767.
15. Butte A: **The use and analysis of microarray data.** *Nat Rev Drug Discov* 2002, **1**(12):951-960.

16. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3**:Article3.
17. Sebastiani P, Gussoni E, Kohane IS, Ramoni MF: **Statistical challenges in functional genomics.** *Stat Sci* 2003, **18**(1):33-60.
18. Carter SL, Eklund AC, Mecham BH, Kohane IS, Szallasi Z: **Redefinition of affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements.** *BMC Bioinformatics* 2005, **6**:15.
19. Nimgaonkar A, Sanoudou D, Butte AJ, Haslett JN, Kunkel LM, Beggs AH, Kohane IS: **Reproducibility of gene expression across generations of Affymetrix microarrays.** *BMC Bioinformatics* 2003, **4**:12.
20. Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, Wetmore DZ, Mariani TJ, Kohane IS, Szallasi Z: **Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements.** *Nucleic Acids Res* 2004, **32**(9):8.
21. Severgnini M, Bicciato S, Mangano E, Scarlatti F, Mezzelani A, Mattioli M, Ghidoni R, Peano C, Bonnal R, Viti F *et al*: **Strategies for comparing gene**

- expression profiles from different microarray platforms: Application to a case-control experiment.** *Anal Biochem* 2006, **353**(1):43-56.
22. English SB, Butte AJ: **Evaluation and integration of 49 genome-wide experiments and the prediction of previously unknown obesity-related genes.** *Bioinformatics* 2007, **23**(21):2910-2917.
23. Wolfe CJ, Kohane IS, Butte AJ: **Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks.** *BMC Bioinformatics* 2005, **6**:10.
24. Irizarry R, Gautier L, Cope L: **An r package for analyses of Affymetrix oligonucleotide arrays.** In: *The Analysis of Gene Expression Data: Methods and Software*. Edited by Parmigiani G, Garrett E, Irizarry R, Zeger S. New York, NY: Springer; 2002.
25. Cahan P, Rovegno F, Mooney D, Newman JC, St Laurent G, McCaffrey TA: **Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization.** *Gene* 2007, **401**(1-2):12-18.
26. Farnham PJ: **Insights from genomic profiling of transcription factors.** *Nat Rev Genet* 2009, **10**(9):605-616.
27. Komashko VM, Acevedo LG, Squazzo SL, Iyengar SS, Rabinovich A, O'Geen H, Green R, Farnham PJ: **Using ChIP-chip technology to reveal common principles**

- of transcriptional repression in normal and cancer cells.** *Genome Res* 2008, **18**(4):521-532.
28. Impey S, McCorkle SR, Cha-Molstad H, Dwyer JM, Yochum GS, Boss JM, McWeeney S, Dunn JJ, Mandel G, Goodman RH: **Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions.** *Cell* 2004, **119**(7):1041-1054.
29. Butte AJ, Kohane IS: **Creation and implications of a phenome-genome network.** *Nat Biotechnol* 2006, **24**(1):55-62.
30. Makino T, Gojobori T: **Evolution of protein-protein interaction network.** *Genome Dyn* 2007, **3**:13-29.
31. Friedman N: **Inferring cellular networks using probabilistic graphical models.** *Science* 2004, **303**(5659):799-805.
32. Wu Z, Zhao X, Chen L: **Identifying responsive functional modules from protein-protein interaction network.** *Mol Cells* 2009, **27**(3):271-277.
33. Aggarwal A, Guo DL, Hoshida Y, Yuen ST, Chu KM, So S, Boussioutas A, Chen X, Bowtell D, Aburatani H *et al*: **Topological and functional discovery in a gene coexpression meta-network of gastric cancer.** *Cancer Res* 2006, **66**(1):232-241.
34. Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, Kohane IS, Kasif S: **Network-based analysis of affected biological processes in type 2 diabetes models.** *PLoS Genet* 2007, **3**(6):958-972.

35. Xu JZ, Li YJ: **Discovering disease-genes by topological features in human protein-protein interaction network.** *Bioinformatics* 2006, **22**(22):2800-2805.
36. Camargo A, Azuaje F: **Linking gene expression and functional network data in human heart failure.** *PLoS ONE* 2007, **2**(12):e1347.
37. Sekine S, Shibata T, Sakamoto M, Hirohashi S: **Target disruption of the mutant beta-catenin gene in colon cancer cell line HCT116: preservation of its malignant phenotype.** *Oncogene* 2002, **21**(38):5906-5911.
38. van de Wetering M, Sancho E, Verweij C, de Lau W, Oving I, Hurlstone A, van der Horn K, Batlle E, Coudreuse D, Haramis AP *et al*: **The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells.** *Cell* 2002, **111**(2):241-250.
39. Affymetrix: **Affymetrix Microarray Suite User Guide**, version 4 edn. Santa Clara, CA: Affymetrix; 1999.
40. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**(10):R80.
41. R DCT **R: A language and environment for statistical computing** 1996
<http://www.r-project.org>.

42. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19**(2):185-193.
43. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc Natl Acad Sci U S A* 2001, **98**(1):31-36.
44. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31**(4):e15.
45. Smyth GK: **Limma: linear models for microarray data.** In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Edited by Gentleman R, Carey V, Dudoit S, Irizarry R. New York, NY: Springer; 2005: 397-420.
46. Farcomeni A: **A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion.** *Stat Methods Med Res* 2008, **17**(4):347-388.
47. Yochum GS, McWeeney S, Rajaraman V, Cleland R, Peters S, Goodman RH: **Serial analysis of chromatin occupancy identifies beta-catenin target genes in colorectal carcinoma cells.** *Proc Natl Acad Sci U S A* 2007, **104**(9):3324-3329.
48. Cerami EG, Bader GD, Gross BE, Sander C: **cPath: open source software for collecting, storing, and querying biological pathways.** *BMC Bioinformatics* 2006, **7**:497.

49. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**(11):2498-2504.
50. Jahns-Streubel G, Reuter C, Auf der Landwehr U, Unterhalt M, Schleyer E, Wormann B, Buchner T, Hiddemann W: **Activity of thymidine kinase and of polymerase alpha as well as activity and gene expression of deoxycytidine deaminase in leukemic blasts are correlated with clinical response in the setting of granulocyte-macrophage colony-stimulating factor-based priming before and during TAD-9 induction therapy in acute myeloid leukemia.** *Blood* 1997, **90**(5):1968-1976.
51. Shimoyama Y, Hirohashi S, Hirano S, Noguchi M, Shimosato Y, Takeichi M, Abe O: **Cadherin cell-adhesion molecules in human epithelial tissues and carcinomas.** *Cancer Res* 1989, **49**(8):2128-2133.
52. Sabates-Bellver J, Van der Flier LG, de Palo M, Cattaneo E, Maake C, Rehrauer H, Laczko E, Kurowski MA, Bujnicki JM, Menigatti M *et al*: **Transcriptome profile of human colorectal adenomas.** *Mol Cancer Res* 2007, **5**(12):1263-1275.
53. Abe S, Usami S, Nakamura Y: **Mutations in the gene encoding KIAA1199 protein, an inner-ear protein expressed in Deiters' cells and the fibrocytes, as the cause of nonsyndromic hearing loss.** *J Hum Genet* 2003, **48**(11):564-570.

54. Kuniba H, Yoshiura K, Kondoh T, Ohashi H, Kurosawa K, Tonoki H, Nagai T, Okamoto N, Kato M, Fukushima Y *et al*: **Molecular karyotyping in 17 patients and mutation screening in 41 patients with Kabuki syndrome.** *J Hum Genet* 2009, **54**(5):304-309.
55. Leiter AB, Toder A, Wolfe HJ, Taylor IL, Cooperman S, Mandel G, Goodman RH: **Peptide YY. Structure of the precursor and expression in exocrine pancreas.** *J Biol Chem* 1987, **262**(27):12984-12988.
56. Batterham RL, Cowley MA, Small CJ, Herzog H, Cohen MA, Dakin CL, Wren AM, Brynes AE, Low MJ, Ghatei MA *et al*: **Gut hormone PYY(3-36) physiologically inhibits food intake.** *Nature* 2002, **418**(6898):650-654.
57. Fu M, Wang C, Li Z, Sakamaki T, Pestell RG: **Minireview: Cyclin D1: normal and abnormal functions.** *Endocrinology* 2004, **145**(12):5439-5447.
58. Etienne-Manneville S, Hall A: **Cdc42 regulates GSK-3beta and adenomatous polyposis coli to control cell polarity.** *Nature* 2003, **421**(6924):753-756.
59. McCright B, Rivers AM, Audlin S, Virshup DM: **The B56 family of protein phosphatase 2A (PP2A) regulatory subunits encodes differentiation-induced phosphoproteins that target PP2A to both nucleus and cytoplasm.** *J Biol Chem* 1996, **271**(36):22081-22089.
60. Liu W, Dong X, Mai M, Seelan RS, Taniguchi K, Krishnadath KK, Halling KC, Cunningham JM, Boardman LA, Qian C *et al*: **Mutations in AXIN2 cause**

colorectal cancer with defective mismatch repair by activating beta-catenin/TCF signalling. *Nat Genet* 2000, **26**(2):146-147.

61. Cenciarelli C, Chiaur DS, Guardavaccaro D, Parks W, Vidal M, Pagano M: **Identification of a family of human F-box proteins.** *Curr Biol* 1999, **9**(20):1177-1179.
62. Winston JT, Strack P, Beer-Romero P, Chu CY, Elledge SJ, Harper JW: **The SCFbeta-TRCP-ubiquitin ligase complex associates specifically with phosphorylated destruction motifs in IkappaBalpha and beta-catenin and stimulates IkappaBalpha ubiquitination in vitro.** *Genes Dev* 1999, **13**(3):270-283.
63. Liu C, Li Y, Semenov M, Han C, Baeg GH, Tan Y, Zhang Z, Lin X, He X: **Control of beta-catenin phosphorylation/degradation by a dual-kinase mechanism.** *Cell* 2002, **108**(6):837-847.
64. Mallo GV, Rechreche H, Frigerio JM, Rocha D, Zweibaum A, Lacasa M, Jordan BR, Dusetti NJ, Dagorn JC, Iovanna JL: **Molecular cloning, sequencing and expression of the mRNA encoding human Cdx1 and Cdx2 homeobox. Down-regulation of Cdx1 and Cdx2 mRNA expression during colorectal carcinogenesis.** *Int J Cancer* 1997, **74**(1):35-44.
65. Tetsuka T, Uranishi H, Imai H, Ono T, Sonta S, Takahashi N, Asamitsu K, Okamoto T: **Inhibition of nuclear factor-kappaB-mediated transcription by association**

- with the amino-terminal enhancer of split, a Groucho-related protein lacking WD40 repeats.** *J Biol Chem* 2000, **275**(6):4383-4390.
66. Bylund M, Andersson E, Novitch BG, Muhr J: **Vertebrate neurogenesis is counteracted by Sox1-3 activity.** *Nat Neurosci* 2003, **6**(11):1162-1168.
67. Netzer C, Rieger L, Brero A, Zhang CD, Hinze M, Kohlhase J, Bohlander SK: **SALL1, the gene mutated in Townes-Brocks syndrome, encodes a transcriptional repressor which interacts with TRF1/PIN2 and localizes to pericentromeric heterochromatin.** *Hum Mol Genet* 2001, **10**(26):3017-3024.
68. Feng Y, Lee N, Fearon ER: **TIP49 regulates beta-catenin-mediated neoplastic transformation and T-cell factor target gene induction via effects on chromatin remodeling.** *Cancer Res* 2003, **63**(24):8726-8734.
69. Wu Y, Dowbenko D, Spencer S, Laura R, Lee J, Gu Q, Lasky LA: **Interaction of the tumor suppressor PTEN/MMAC with a PDZ domain of MAGI3, a novel membrane-associated guanylate kinase.** *J Biol Chem* 2000, **275**(28):21477-21485.
70. Azim AC, Knoll JH, Marfatia SM, Peel DJ, Bryant PJ, Chishti AH: **DLG1: chromosome location of the closest human homologue of the Drosophila discs large tumor suppressor gene.** *Genomics* 1995, **30**(3):613-616.

71. Oda T, Kanai Y, Oyama T, Yoshiura K, Shimoyama Y, Birchmeier W, Sugimura T, Hirohashi S: **E-cadherin gene mutations in human gastric carcinoma cell lines.** *Proc Natl Acad Sci U S A* 1994, **91**(5):1858-1862.
72. Gong TW, Besirli CG, Lomax MI: **MACF1 gene structure: a hybrid of plectin and dystrophin.** *Mamm Genome* 2001, **12**(11):852-861.
73. Yamamoto H, Ihara M, Matsuura Y, Kikuchi A: **Sumoylation is involved in beta-catenin-dependent activation of Tcf-4.** *EMBO J* 2003, **22**(9):2047-2059.
74. Ren M, Drivas G, D'Eustachio P, Rush MG: **Ran/TC4: a small nuclear GTP-binding protein that regulates DNA synthesis.** *J Cell Biol* 1993, **120**(2):313-323.
75. Kirikoshi H, Koike J, Sagara N, Saitoh T, Tokuhara M, Tanaka K, Sekihara H, Hirai M, Katoh M: **Molecular cloning and genomic structure of human frizzled-3 at chromosome 8p21.** *Biochem Biophys Res Commun* 2000, **271**(1):8-14.
76. Liu Y, Dehni G, Purcell KJ, Sokolow J, Carcangiu ML, Artavanis-Tsakonas S, Stifani S: **Epithelial expression and chromosomal location of human TLE genes: implications for notch signaling and neoplasia.** *Genomics* 1996, **31**(1):58-64.
77. Fedi P, Bafico A, Nieto Soria A, Burgess WH, Miki T, Bottaro DP, Kraus MH, Aaronson SA: **Isolation and biochemical characterization of the human Dkk-1 homologue, a novel inhibitor of mammalian Wnt signaling.** *J Biol Chem* 1999, **274**(27):19465-19472.

78. Levanon D, Goldstein RE, Bernstein Y, Tang H, Goldenberg D, Stifani S, Paroush Z, Groner Y: **Transcriptional repression by AML1 and LEF-1 is mediated by the TLE/Groucho corepressors.** *Proc Natl Acad Sci U S A* 1998, **95**(20):11590-11595.
79. Pizzuti A, Amati F, Calabrese G, Mari A, Colosimo A, Silani V, Giardino L, Ratti A, Penso D, Calza L *et al*: **cDNA characterization and chromosomal mapping of two human homologues of the Drosophila dishevelled polarity gene.** *Hum Mol Genet* 1996, **5**(7):953-958.
80. Beildeck ME, Islam M, Shah S, Welsh J, Byers SW: **Control of TCF-4 expression by VDR and vitamin D in the mouse mammary gland and colorectal cancer cell lines.** *PLoS One* 2009, **4**(11):e7872.
81. Saitoh T, Mine T, Katoh M: **Molecular cloning and expression of proto-oncogene FRAT1 in human cancer.** *Int J Oncol* 2002, **20**(4):785-789.
82. Reynaud C, Fabre S, Jalinot P: **The PDZ protein TIP-1 interacts with the Rho effector rhotekin and is involved in Rho signaling to the serum response element.** *J Biol Chem* 2000, **275**(43):33962-33968.
83. De Valck D, Jin DY, Heyninck K, Van de Craen M, Contreras R, Fiers W, Jeang KT, Beyaert R: **The zinc finger protein A20 interacts with a novel anti-apoptotic protein which is cleaved by specific caspases.** *Oncogene* 1999, **18**(29):4182-4190.

84. Allen E, Ding J, Wang W, Pramanik S, Chou J, Yau V, Yang Y: **Gigaxonin-controlled degradation of MAP1B light chain is critical to neuronal survival.** *Nature* 2005, **438**(7065):224-228.
85. Tombes RM, Krystal GW: **Identification of novel human tumor cell-specific CaMK-II variants.** *Biochim Biophys Acta* 1997, **1355**(3):281-292.
86. Mok SC, Chan WY, Wong KK, Cheung KK, Lau CC, Ng SW, Baldini A, Colitti CV, Rock CO, Berkowitz RS: **DOC-2, a candidate tumor suppressor gene in human epithelial ovarian cancer.** *Oncogene* 1998, **16**(18):2381-2387.

Table 1 – Studies used for mRNA expression analysis - citations, sample counts and descriptions

Patient tissue – adenomas and normal mucosa

- Galamb O et al. *Dis Markers*. 2008. 25(1): 1-16. GEO: GSE4183 PMID: 18776587
15 adenoma patients and 8 non-matched controls
A gene expression profile of various colonic diseases including adenomas, carcinomas and irritable bowel syndrome
- Sabates-Bellver J et al. *Mol Cancer Res*. Dec 2007. 5(12): 1263-75. GEO: GSE8671
PMID: 18171984
32 adenoma samples with 32 same-patient matched controls
A wide ranging study examining changes in mRNA expression in adenomatous tissue accompanying carcinoma development
- Galamb O et al. *Cancer Epidemiol Biomarkers Prev*. Oct 2008. 17(10): 2835-45. GEO: GSE10714
PMID: 18843029
5 adenoma patient samples and 3 non-matched controls
Companion study following up their paper in *Disease Markers* attempting to use mRNA profiling for diagnostics
- Hong Y et al. *Clin Cancer Res*. Feb 15 2007. 13(4): 1107-14. GEO: GSE4107 PMID: 17317818
10 non-matched patient controls
Systematic search for novel genetic causes in colon cancer aside beyond FAP and hereditary non-polyposis origins

Cell lines – HCT116

- Goetz S et al. June 2007. 27(12): 4475-87. GEO: GSE8690 PMID: 17420274
1 replicate
Transcriptome profile of six different human cell lines
- Wagner KW et al. *Nat Med*. Sep 2007. 13(9): 1070-7. GEO: GSE8332 PMID: 17767167
2 replicates
Effect of the Apo2L/TRAIL ligand in stimulating cell death in multiple lines
- Ruike Y et al. *J Hum Genet*. 2008. 53(6):515-23. GEO: GSE10021 PMID: 18465083
1 replicate
mRNA and miRNA profiling of various cell lines
- Connolly K et al. *Cancer Chemother Pharmacol*. July 2009. 64(2): 307-16. GEO: GSE11618
PMID: 19034449
2 replicates
Assessment of X-linked inhibition of apoptosis in cancerous cell lines
- Dornan D. Gentech – Oncology Diagnostics. Submitted Mar 14 2008. GEO: GSE10843
3 replicates
Independent submission to GEO database

Cell lines – LS174T

- GlaxoSmithKline – caArray submission caArray ID: woost-00041
3 replicates
Affymetrix HGU-133 Plus 2.0 array submission to the caArray compendium

Table 2 – Significance counts of probes from mRNA analysis by linear model comparison

	BH sig (FDR<0.05)	BH Sig & FC Δ0.5	Upregulated (FC > 1.5)	Downregulated (FC < -1.5)
Normal Mucosa vs Adenomatous Tissue	23465	5135	2626	2509
Normal Mucosa vs Derived Cell Line	34292	16078	8515	7563
Adenomatous Tissue vs Derived Cell Line	28819	11610	5782	5828

Table 3 – concordant genes in the N-A and N-C comparisons by direction of differential expression

	Total number of concordantly expressed genes	Significant differential expression in cell lines over adenomas	No significant difference in mRNA expression between patient adenomas and cell lines	Significant differential expression in adenomas over cell lines
N-A & N-C concordant Downregulated probes	2177	1279	878	20
N-A & N-C concordant Upregulated probes	1908	1022	861	25

Although columns are divided by the direction of the A-C comparisons, all probes showed significant concordant differential expression in the N-A and N-C portions of the linear model.

Table 4.a – top five N-A and N-C concordantly upregulated genes – grouped by A-C comparison

TOP FIVE UPREGULATED PROBES	Top sorted probes with largest diff expr in patient adenomas	Top sorted probes with largest diff expr in cancerous cell lines	Probes with largest differences between cell lines and patient tissue
Fold change in cancerous cell lines significantly greater (N- A < N-C)	PSAT1 (+7.3692) PHLDA1 (+6.779/+5.678) CDH3 (+6.4309) GDF15 (+4.1667) SLC7A5 (+4.8031)	PSAT1 (+58.8235) PHLDA1 (+45.662/+51.020) SLC7A5 (+37.5940) AP1S3 (+25.8398) MSX1 (+25.0627)	SFRS6 (+13.4953) DHRS2 (+12.8205) GAL (+12.0627) AREG (+11.4416) HS6ST2 (+9.5694)
No significant difference in expression between patient tissue and cell lines (N-A == N-C)	hCG_1815491 (+10.8108) KLK10 (+7.6220) LST+3TM12 (+6.9832) MSX2 (+6.8729) WDR72 (+6.1162)	hCG_1815491 (+11.1732) KLK10 (+10.1833) WDR72 (+8.8028) LST+3TM12 (+8.1103) CEP55 (+5.8617)	Non-significant A-C FC range: SULT1C2 (-1.4961) to HAS3 (+1.5477)
Fold change in patient adenomas significantly greater (N-A > N-C)	TCN1 (+28.9017) KIAA1199 (+26.9542) FOXQ1 (+22.0264) FAM148A (+12.0337) SERPINB5 (+10.0604)	KIAA1199 (+7.8370) SERPINB5 (+3.5199) IFGBP2 (+2.7420) PPM1H (+2.5628) FOXQ1 (+2.5031)	TCN1 (-13.3271) FOXQ1 (-8.7943) FAM148A (-6.6210) AXIN2 (-3.8980) IL8 (-3.8713)

Each row contains a unique set of concordantly upregulated genes grouped according to their A-C comparison. The first column corresponds to the top genes when sorted on fold changes of expression in the N-A comparison while the second column is a sort of the same set but for the top genes in the N-C comparison. The final column is a sort showing the greatest changes in differential expression between the two disease conditions based on the A-C comparison. Probes that are bolded were found in both the top five N-A and N-C sorts for that A-C group.

Table 4.b – top five N-A and N-C concordantly downregulated genes – grouped by A-C comparison

TOP FIVE UPREGULATED PROBES	Top sorted probes with largest diff expr in patient adenomas	Top sorted probes with largest diff expr in cancerous cell lines	Probes with largest differences between cell lines and patient tissue
Fold change in cancerous cell lines significantly greater (N- A < N-C)	AQP8 (-38.388) CA1 (-33.995) MS4A12 (-33.258) CLDN8 (-32.4201) CLCA4 (-28.983)	SLC26A3 (-896.891) IGJ (-487.828) CEACAM7 (-453.085) IGL2 (-285.950) FABP1 (-285.011)	IGJ (-243.286) IGL2 (-183.455) Cl5orf48 (-147.225) FABP1 (-130.501) KRT20 (-90.439)
No significant difference in expression between patient tissue and cell lines (N-A == N-C)	GCG (-43.992) PYY (-29.076) ABCG2 (-20.705) GUCA2B (-18.2882) CDKN2B (-16.688)	GCG (-84.252) PYY (-39.821) GUCA2B (-21.784) ABCG2 (-20.7646) SST (-17.943)	Non-significant A-C FC range: GCG (-1.9152) to TMEM200A (+1.6418)
Fold change in patient adenomas significantly greater (N-A > N-C)	CPNE8 (-6.2415) MAMDC2 (-5.709) TBC1D9 (-4.3016) RUNDC3B (-3.8366) CNNM2 (-3.8113)	MAMDC2 (-3.4165) TBC1D9 (-2.2319) CPNE8 (-2.2032) CNNM2 (-2.0780) CDA (-2.0566)	CPNE8 (+3.4060) RUNDC3B (+2.5342) LIFR (+2.3348) TBC1D9 (+1.9272) KIAA1211 (+1.8450)

Format is the same as in table 4.a

Table 5.a – Direct targets of β -catenin also showing concordant upregulation

Entrez ID	Affymetrix ID	Hugo Symbol	N-A FC	N-C FC	A-C FC	Total tags for all clusters
4363	202804_at	ABCC1	2.154	3.299	1.531	81
23305	211207_s_at	ACSL6	2.142	1.628	-1.315	6
306	209369_at	ANXA3	3.574	1.944	-1.838	5
130340	1555731_a_at	AP1S3	2.885	25.814	8.948	11
328	210027_s_at	APEX1	1.646	2.178	1.323	3
56938	223586_at	ARNTL2	2.393	3.277	1.369	5
51008	1554627_a_at	ASCC1	1.714	2.227	1.299	9
23250	230875_s_at	ATP11A	2.406	5.163	2.146	19
8313	222696_at	AXIN2	6.349	1.629	-3.898	8
146712	213589_s_at	B3GNTL1	1.799	2.200	1.223	16
63827	223632_s_at	BCAN	1.620	2.979	1.839	6
699	209642_at	BUB1	2.860	5.472	1.913	4
745	204073_s_at	C11orf9	1.832	2.152	1.175	5
90417	225300_at	C15orf23	1.927	3.225	1.673	3
91300	221764_at	C19orf22	1.600	1.916	1.197	3
790	202715_at	CAD	1.901	3.865	2.034	18
23261	213268_at	CAMTA1	1.500	2.569	1.712	104
595	208712_at	CCND1	2.342	5.910	2.523	6
10849	205264_at	CD3EAP	1.829	4.859	2.656	3
1001	203256_at	CDH3	6.432	13.121	2.040	9
1021	243000_at	CDK6	2.322	1.635	-1.420	12
84303	224462_s_at	CHCHD6	1.511	2.209	1.462	18
63922	226569_s_at	CHTF18	1.603	2.579	1.608	6
9075	223509_at	CLDN2	7.572	2.265	-3.343	8
152189	235099_at	CMTM8	1.885	3.688	1.957	9
51692	225082_at	CPSF3	1.619	2.145	1.325	7
115908	225681_at	CTHRC1	2.340	5.184	2.216	3
51523	222996_s_at	CXXC5	1.777	1.562	-1.138	8
54606	217754_at	DDX56	1.619	2.224	1.374	3
56919	222875_at	DHX33	1.773	2.288	1.290	7
1984	201123_s_at	EIF5A	2.413	11.137	4.616	5
54512	218695_at	EXOSC4	1.604	2.227	1.388	16
2172	210445_at	FABP6	3.338	4.548	1.362	3
55179	220643_s_at	FAIM	2.332	3.708	1.590	3
113115	228069_at	FAM54A	2.166	4.088	1.887	4

Table 5.a continued

Entrez ID	Affymetrix ID	Hugo Symbol	N-A FC	N-C FC	A-C FC	Total tags for all clusters
2175	203805_s_at	FANCA	1.546	2.990	1.934	22
2237	204767_s_at	FEN1	2.173	5.868	2.701	3
2272	206492_at	FHIT	1.622	2.200	1.356	74
63979	222843_at	FIGNL1	2.073	2.678	1.292	3
79583	64900_at	FLJ22167	1.944	3.252	1.673	4
79581	222155_s_at	GPR172A	1.678	2.818	1.680	20
2887	209409_at	GRB10	2.013	3.809	1.892	13
57822	232116_at	GRHL3	1.722	1.628	-1.058	26
9569	218412_s_at	GTF2IRD1	3.062	2.777	-1.103	34
3159	206074_s_at	HMGA1	2.045	6.817	3.334	5
9456	213793_s_at	HOMER1	1.639	7.568	4.616	10
2537	204415_at	IFI6	3.100	2.328	-1.331	3
9466	222062_at	IL27RA	1.857	3.754	2.022	4
3609	217805_at	ILF3	1.642	4.057	2.471	6
3615	201892_s_at	IMPDH2	2.016	2.703	1.341	5
55705	217885_at	IPO9	1.534	1.705	1.112	3
3656	231779_at	IRAK2	1.682	1.760	1.046	20
9270	203336_s_at	ITGB1BP1	1.563	1.843	1.179	4
11015	204017_at	KDEL3	1.900	1.532	-1.240	18
57214	212942_s_at	KIAA1199	26.990	7.836	-3.444	17
55425	220171_x_at	KIAA1704	1.578	1.971	1.249	3
64147	231319_x_at	KIF9	1.903	1.601	-1.189	6
8270	219061_s_at	LAGE3	1.575	2.209	1.403	3
84823	216952_s_at	LMNB2	1.709	4.102	2.400	14
7804	205282_at	LRP8	2.215	8.920	4.026	12
57128	218561_s_at	LYRM4	1.826	2.543	1.393	10
51025	218969_at	Magmas	1.521	2.338	1.537	3
90411	212246_at	MCFD2	1.653	3.198	1.934	8
4199	204058_at	ME1	3.080	7.684	2.495	3
79828	1554667_s_at	METTL8	1.816	2.982	1.642	3
64979	224331_s_at	MRPL36	1.568	1.839	1.173	6
124540	225238_at	MSI2	1.621	1.623	1.001	420
10232	204885_s_at	MSLN	3.336	2.917	-1.143	3
9112	211783_s_at	MTA1	1.604	6.074	3.787	22
4522	202309_at	MTHFD1	1.826	2.522	1.381	5
25902	225520_at	MTHFD1L	4.219	16.690	3.956	22

Table 5.a continued

Entrez ID	Affymetrix ID	Hugo Symbol	N-A FC	N-C FC	A-C FC	Total tags for all clusters
140838	228073_at	NANP	1.903	2.863	1.505	4
10529	207279_s_at	NEBL	2.422	2.362	-1.025	14
4833	212739_s_at	NME4	1.851	1.839	-1.006	4
10360	205129_at	NPM3	2.019	8.963	4.440	12
26747	205134_s_at	NUFIP1	1.544	2.737	1.772	4
4957	225617_at	ODF2	1.869	3.716	1.989	5
26031	209626_s_at	OSBPL3	1.910	3.457	1.810	38
78990	219369_s_at	OTUB2	1.846	2.261	1.225	3
10606	201013_s_at	PAICS	2.045	4.301	2.103	4
55795	225149_at	PCID2	1.501	1.675	1.116	6
5238	210041_s_at	PGM3	1.553	1.887	1.215	6
26227	201397_at	PHGDH	1.786	9.771	5.472	6
9487	213889_at	PIGL	1.525	2.391	1.567	8
58473	209504_s_at	PLEKHB1	2.432	2.199	-1.106	4
10733	204886_at	PLK4	1.534	2.435	1.588	3
57048	56197_at	PLSCR3	1.539	2.449	1.592	3
5471	209433_s_at	PPAT	1.944	6.458	3.322	4
10465	204228_at	PPIH	1.536	1.802	1.173	4
5493	203407_at	PPL	1.772	2.686	1.515	7
10848	218849_s_at	PPP1R13L	1.591	2.089	1.312	4
6240	201476_s_at	RRM1	1.828	2.107	1.153	3
6461	1557458_s_at	SHB	1.647	1.553	-1.061	23
65244	218324_s_at	SPATS2	1.553	3.161	2.036	11
125058	228488_at	TBC1D16	1.703	1.827	1.073	13
6904	229192_s_at	TBCD	1.758	2.705	1.538	102
7003	224955_at	TEAD1	1.662	2.071	1.246	10
7027	212330_at	TFDP1	1.504	2.885	1.918	20
54929	43977_at	TMEM161A	1.506	2.241	1.489	7
252839	222987_s_at	TMEM9	1.948	1.817	-1.072	3
8797	231775_at	TNFRSF10A	1.741	2.669	1.533	7
116447	225802_at	TOP1MT	1.907	2.663	1.397	7
10131	201391_at	TRAP1	2.033	3.209	1.579	7
10155	200990_at	TRIM28	1.822	3.955	2.171	3
9319	204033_at	TRIP13	3.180	6.634	2.086	6
7205	209129_at	TRIP6	2.675	2.537	-1.055	7
80746	219581_at	TSEN2	1.549	3.121	2.015	6
150465	224896_s_at	TTL	1.696	4.011	2.364	6

Table 5.a continued

Entrez ID	Affymetrix ID	Hugo Symbol	N-A FC	N-C FC	A-C FC	Total tags for all clusters
7328	221962_s_at	UBE2H	1.620	2.454	1.515	7
54578	232654_s_at	UGT1A6	2.540	1.552	-1.637	7
29128	225655_at	UHRF1	2.067	6.674	3.229	6
89891	224715_at	WDR34	1.582	2.246	1.420	3
25886	226355_at	WDR51A	1.518	1.705	1.124	48
84858	227195_at	ZNF503	1.825	3.447	1.889	4
7625	205881_at	ZNF74	1.555	2.493	1.603	3

Table 5.b – Direct targets of β -catenin also showing concordant downregulation

Entrez ID	Affymetrix ID	Hugo Symbol	N-A FC	N-C FC	A-C FC	Total tags for all clusters
22848	205434_s_at	AAK1	-1.938	-1.737	1.116	6
10351	204719_at	ABCA8	-19.310	-29.886	-1.548	6
115	204497_at	ADCY9	-2.795	-2.715	1.029	13
3899	227198_at	AFF3	-2.782	-3.258	-1.171	25
408	222912_at	ARRB1	-1.743	-2.790	-1.600	11
51676	227915_at	ASB2	-1.712	-3.125	-1.825	7
64115	225372_at	C10orf54	-1.503	-2.319	-1.543	8
81563	223126_s_at	C1orf21	-1.776	-4.384	-2.469	10
718	217767_at	C3	-2.088	-6.263	-3.000	5
51719	217873_at	CAB39	-1.555	-2.312	-1.487	5
817	228555_at	CAMK2D	-1.558	-1.715	-1.100	30
824	208683_at	CAPN2	-1.773	-2.231	-1.258	9
726	226292_at	CAPN5	-1.653	-6.772	-4.098	8
930	206398_s_at	CD19	-1.677	-1.946	-1.160	4
978	205627_at	CDA	-3.090	-2.057	1.503	5
10428	203166_at	CFDP1	-1.806	-2.268	-1.256	5
1113	204697_s_at	CHGA	-7.265	-9.880	-1.360	4
54102	227742_at	CLIC6	-2.721	-4.816	-1.770	5
79789	225757_s_at	CLMN	-1.711	-3.855	-2.253	11
1191	222043_at	CLU	-2.116	-1.790	1.182	6
134147	227522_at	CMBL	-2.095	-1.864	1.124	5
54805	1554522_at	CNNM2	-3.811	-2.078	1.834	9
23242	213050_at	COBL	-1.745	-1.810	-1.037	16
27147	53991_at	DENND2A	-1.758	-5.526	-3.143	7
1756	203881_s_at	DMD	-1.549	-5.910	-3.815	5
667	212254_s_at	DST	-1.614	-2.544	-1.576	15
1946	233814_at	EFNA5	-1.621	-3.565	-2.200	11
2034	200878_at	EPAS1	-1.698	-7.314	-4.307	7
83641	223059_s_at	FAM107B	-1.918	-2.857	-1.490	9
2192	202995_s_at	FBLN1	-3.787	-3.819	-1.008	10
114907	225803_at	FBXO32	-2.057	-3.825	-1.859	12
2534	210105_s_at	FYN	-1.649	-1.831	-1.111	13
51228	226177_at	GLTP	-2.294	-2.337	-1.019	5

Table 5.b continued

Entrez ID	Affymetrix ID	Hugo Symbol	N-A FC	N-C FC	A-C FC	Total tags for all clusters
3603	209827_s_at	IL16	-1.959	-2.908	-1.485	8
7850	205403_at	IL1R2	-3.332	-42.775	-12.838	4
3570	205945_at	IL6R	-4.249	-4.609	-1.085	10
81618	221004_s_at	ITM2C	-2.076	-7.314	-3.523	3
221895	225798_at	JAZF1	-1.908	-1.656	1.153	15
3778	221584_s_at	KCNMA1	-1.578	-6.505	-4.122	32
9764	204546_at	KIAA0513	-1.707	-2.736	-1.602	30
687	203542_s_at	KLF9	-1.766	-2.140	-1.211	4
3990	206606_at	LIPC	-1.562	-1.613	-1.032	9
200879	235871_at	LIPH	-1.672	-3.304	-1.976	10
4026	202822_at	LPP	-1.578	-2.696	-1.709	17
7851	209373_at	MALL	-3.959	-2.930	1.351	9
256691	228885_at	MAMDC2	-5.709	-3.416	1.671	7
57134	218918_at	MAN1C1	-1.852	-2.469	-1.333	24
196410	227055_at	METTL7B	-1.537	-2.145	-1.396	16
51237	221286_s_at	MGC29506	-1.540	-4.317	-2.804	3
80168	207491_at	MOGAT2	-1.540	-2.915	-1.893	4
253827	225782_at	MSRB3	-2.427	-1.560	1.556	14
10398	201058_s_at	MYL9	-1.708	-2.670	-1.564	3
4684	212843_at	NCAM1	-1.515	-1.604	-1.059	5
8648	209106_at	NCOA1	-1.702	-2.425	-1.424	8
10397	200632_s_at	NDRG1	-1.852	-3.245	-1.753	11
23327	212448_at	NEDD4L	-1.752	-2.914	-1.663	17
23114	213438_at	NFASC	-1.629	-2.632	-1.616	11
79840	219418_at	NHEJ1	-1.778	-2.269	-1.276	8
197358	236295_s_at	NLRC3	-2.061	-2.837	-1.376	6
283298	217525_at	OLFML1	-1.987	-2.359	-1.187	3
11252	201651_s_at	PACSIN2	-1.627	-2.316	-1.424	19
11240	1554384_at	PADI2	-1.852	-2.136	-1.154	5
54852	242871_at	PAQR5	-3.410	-2.508	1.360	8
5570	223551_at	PKIB	-8.029	-6.089	1.319	7
5354	210198_s_at	PLP1	-3.970	-4.224	-1.064	3
5467	37152_at	PPARD	-1.955	-2.093	-1.070	11
5592	228396_at	PRKG1	-1.594	-2.846	-1.785	15
5652	202525_at	PRSS8	-1.738	-4.892	-2.815	5
754	200677_at	PTTG1IP	-1.505	-1.995	-1.326	7
5697	207080_s_at	PYY	-29.076	-39.821	-1.370	5

Table 5.b continued

Entrez ID	Affymetrix ID	Hugo Symbol	N-A FC	N-C FC	A-C FC	Total tags for all clusters
5919	209496_at	RARRES2	-2.392	-11.451	-4.788	12
83937	226436_at	RASSF4	-1.812	-3.572	-1.971	5
83593	223322_at	RASSF5	-1.881	-2.242	-1.192	5
27303	238447_at	RBMS3	-2.645	-7.358	-2.782	9
92241	225763_at	RCSD1	-2.393	-5.872	-2.454	5
6256	202449_s_at	RXRA	-1.506	-1.778	-1.180	15
6398	213716_s_at	SECTM1	-4.106	-4.460	-1.086	6
153769	243582_at	SH3RF2	-1.520	-4.922	-3.238	13
5003	206097_at	SLC22A18AS	-2.486	-3.263	-1.313	7
6584	205074_at	SLC22A5	-2.716	-3.680	-1.355	6
114134	227176_at	SLC2A13	-2.485	-4.344	-1.748	9
219855	238638_at	SLC37A2	-1.684	-2.213	-1.314	5
340024	231021_at	SLC6A19	-2.993	-2.766	1.082	6
6548	209453_at	SLC9A1	-1.561	-2.049	-1.312	14
6550	207212_at	SLC9A3	-1.658	-1.821	-1.099	3
55512	219695_at	SMPD3	-1.654	-4.227	-2.556	24
57522	1554473_at	SRGAP1	-1.566	-2.173	-1.388	9
6480	201998_at	ST6GAL1	-1.532	-2.889	-1.885	9
55959	224724_at	SULF2	-1.570	-4.089	-2.605	10
11346	202796_at	SYNPO	-1.952	-3.373	-1.728	6
7077	231579_s_at	TIMP2	-1.993	-2.127	-1.067	13
57458	226489_at	TMCC3	-3.227	-6.279	-1.946	7
64759	217853_at	TNS3	-1.685	-5.604	-3.326	28
1831	208763_s_at	TSC22D3	-1.819	-1.509	1.205	9
10194	223282_at	TSHZ1	-1.640	-2.339	-1.427	7
219699	226899_at	UNC5B	-1.779	-1.652	1.077	57
50853	209950_s_at	VILL	-1.647	-13.564	-8.236	14
7433	205019_s_at	VIPR1	-2.550	-7.213	-2.829	12

Table 6 – metrics of the initial protein interaction network

Number of connected components	11
Number of individual protein product nodes	168
Total number of edges in the network	554
Number of node pairs with multiple edges	120
Isolated nodes with no edges	11
Average number of neighbors per node	4.774
Network density	0.029
Network centralization	0.329
Network hubs (5-13 neighbor nodes)	36
Superhubs (14 or more neighbor nodes)	5
Number of nodes that are direct targets of β-catenin	16
Number of nodes with HGU133 Plus 2.0 array mRNA spots	111

Table 7 – genes with mRNA overexpression found in adenomas and cell lines

7.a: upregulated in both adenomas and cell lines

Joint A/C gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
CCND1	2.3425	5.9104	6	5	4	X
CDC2	1.5025	2.0282		1		
CDC25C	1.7695	1.8137		1		
FZD3	1.8176	5.9420		2		
RAN	1.8697	3.1093		1		
RUVBL1	2.2555	4.5135		2		

7.b upregulated in patient adenomas only

N-A only gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
AXIN2	2.4377	1.4908	8	5	2	X
FHL2	1.8514	1.3280	8	1		
SOX1	1.6872	1.0039		1		

7.c: upregulated in cancerous cell lines only

N-C only gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
BRD7	1.0125	1.5980		0		
BTRC	1.1259	2.0770		6	1	X
DKK1	1.3230	40.6236		4		X
DVL1	1.2027	2.3971		1		
DVL2	1.1105	1.8868		6		
FBXW11	1.1695	1.9334	8	2		
PIAS4	1.3826	1.8332		2	1	
PPP2R5D	1.1345	1.5430		5		
RANBP3	1.1076	1.5524		3		
TBP	1.1232	1.5815		1		
WNT6	1.1061	1.8539		2	1	

Table 8 – genes with mRNA underexpression found in adenomas and cell lines

8.a: downregulated in both adenomas and cell lines

Joint A/C gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
FZD5	-1.6837	-6.4483		2	1	
JUP	-1.5145	-1.9822		1		
MAGI3	-1.6431	-3.2192		4	1	
MAP1B	-2.2633	-1.5526		1	3	
SALL1	-1.5466	-3.3709		1		

8.b: downregulated in cancerous cell lines only

N-C only gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
AES	1.0724	-1.5903		13	2	
CDC42	-1.0832	-10.8582		5		
CDH1	-1.2061	-3.1153		1	1	X
CDX1	-1.1338	-5.8053		1		X
CTNNB1	-1.0976	-1.5156		59		
DAB2	1.2593	-2.6291		1	1	
DLG1	-1.3821	-3.5461		0		
DLG4	-1.1270	-1.7271		1		
FRAT1	-1.0422	-1.9366		3		
LRP1	-1.0276	-1.6999	54	1		
TAX1BP3	-1.3520	-3.3596	4		1	
TCF4	-1.1998	-1.6468		13		X

Table 9 – information on network nodes with altered co-expression under adenomatous conditions

Disrupted gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
BTRC	1.1259	2.0770		6	1	X
CAMK2A	-1.0581	1.1944		6	1	
CDH1	-1.2061	-3.1153		1	1	X
CSNK1D	-1.0655	-1.1329		1	1	
DAAM1	-1.0519	-1.0862		2	1	
DAB2	1.2593	-2.6291		1	1	
FZD5	-1.6837	-6.4483		2	1	
GSK3B	1.3509	-1.2529		9	1	
GSK3B	1.3509	-1.2529		13	1	
MAGI3	-1.6431	-3.2192		4	1	
PIAS4	1.3826	1.8332		2	1	
SENP2	1.2014	1.4892		3	1	
TAX1BP3	-1.3520	-3.3596	4	0	1	
WNT6	1.1061	1.8539		2	1	
AES	1.0724	-1.5903		13	2	
AXIN1	1.1479	1.0531		25	2	
AXIN2	2.4377	1.4908	8	5	2	X
CAMK2D	-1.1189	-1.0669	30	5	2	
HDAC1	1.0633	1.0283		7	2	
PRKCB	-1.0757	1.2911		1	2	
PRKCG	1.0348	1.1938		1	2	
SFRP2	-1.2827	-1.3071		0	2	X
SUMO1	1.0365	1.3626		1	2	
MAP1B	-2.2633	-1.5526		1	3	
MAP3K7	1.1263	1.2797		6	3	
ARRB2	-1.1101	1.6196		2	4	
CCND1	2.3425	5.9104	6	5	4	X
SKP1	1.1149	1.0038		5	5	
PIN1	1.1830	1.3079		1	6	
GNG2	-1.1714	1.0498		4	8	

Table 10.a – metrics tables for direct targets of β -catenin

β-catenin gene nodes	N-A FC	N-C FC	β-catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
TAX1BP3	-1.3520	-3.3596	4	0	1	
CCND1	2.3425	5.9104	6	5	4	X
CSNK1A1	-1.0112	-1.4370	6	15		
AXIN2	2.4377	1.4908	8	5	2	X
FBXW11	1.1695	1.9334	8	2		
FHL2	1.8514	1.3280	8	1		
TLE1	1.4090	-1.1476	9	13		
CTBP1	1.0317	1.4166	10	14		
MARK2	-1.2043	-1.0294	10	0		
NFATC2	-1.0518	1.0457	13	4		
CAMK2D	-1.1189	-1.0669	30	5	2	
MACF1	-1.0026	1.2452	52	6		
LRP1	-1.0276	-1.6999	54	1		
SMAD3	-1.3340	1.0263	63	1		
PRKCA	1.2687	-1.1077	78	1		
TCF7L2	-1.1514	-1.4934	238	9		

Table 10.b – informative listings for first-order protein interaction neighbors of β -catenin targets

β -cat neighb gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
AES	1.0724	-1.5903		13	2	
APC	-1.3493	-1.2140		19		
AXIN1	1.1479	1.0531		25	2	
CAMK2A	-1.0581	1.1944		6	1	
CAMK2B	-1.0130	1.0752		5		
CAMK2G	-1.1167	1.2379		5		
CDC42	-1.0832	-10.8582		5		
CREBBP	-1.0374	-1.0701		13		
CSNK2A1	1.0403	1.1668		9		
CTNNB1	-1.0976	-1.5156		59		
DVL1	-	-		8		
DVL2	1.1105	1.8868		6		
EP300	-1.0963	1.2065		2		
FZD1	-1.3124	-1.1081		11		
GSK3B	1.3509	-1.2529		13	1	
HDAC1	1.0633	1.0283		7	2	
MYC	-1.0075	-1.1048		5		X
NLK	1.1906	1.2528		6		
PIAS4	1.3826	1.8332		2	1	
SENP2	1.2014	1.4892		3	1	
SMAD4	-1.0045	-1.1107		9		
SUMO1	1.0365	1.3626		1	2	
TCF4	-1.1998	-1.6468		13		X

β -catenin targets as target neighbors

AXIN2	2.4377	1.4908	8	5	2	X
CCND1	2.3425	5.9104	6	5	4	X
CSNK1A1	-1.0112	-1.4370	6	15		
CTBP1	1.0317	1.4166	10	14		
NFATC2	-1.0518	1.0457	13	4		
TCF7L2	-1.1514	-1.4934	238	9		
TLE1	1.4090	-1.1476	9	13		

Table 11.a – metrics for heavily connected protein interaction hubs (5-13 neighbor nodes)

Hub gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
AXIN2	2.4377	1.4908	8	5	2	X
CAMK2B	-1.0130	1.0752		5		
CAMK2D	-1.1189	-1.0669	30	5	2	
CAMK2G	-1.1167	1.2379		5		
CCND1	2.3425	5.9104	6	5	4	X
CDC42	-1.0832	-10.8582		5		
CUL1	1.0039	1.3766		5		
GNAS	1.0064	1.1315		5		
KREMEN2	1.0111	1.3491		5		
MYC	-1.0075	-1.1048		5		X
NLK	-	-		5		
PPP2R5D	1.1345	1.5430		5		
SKP1	1.1149	1.0038		5	5	
BTRC	1.1259	2.0770		6	1	X
CAMK2A	-1.0581	1.1944		6	1	
CSNK2A2	1.2909	1.3838		6		
DVL2	1.1105	1.8868		6		
MACF1	-1.0026	1.2452	52	6		
MAP3K7	1.1263	1.2797		6	3	
NLK	1.1906	1.2528		6		
PPP2CA	1.0510	1.3172		6		
HDAC1	1.0633	1.0283		7	2	
DVL1	-	-		8		
DVL3	1.0520	1.2337		8		
CSNK2A1	1.0403	1.1668		9		
GSK3B	1.3509	-1.2529		9	1	
SMAD4	-1.0045	-1.1107		9		
TCF7L2	-1.1514	-1.4934	238	9		
WNT1	-1.0628	-1.0441		9		
FZD1	-1.3124	-1.1081		11		
AES	1.0724	-1.5903		13	2	
CREBBP	-1.0374	-1.0701		13		
GSK3B	1.3509	-1.2529		13	1	
LRP6	-1.0039	1.2464		13		
TCF4	-1.1998	-1.6468		13		X
TLE1	1.4090	-1.1476	9	13		

Table 11.b – metrics for the most connected network super- hubs (14 or more neighbor nodes)

Super-hub gene nodes	N-A FC	N-C FC	β-catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
CTBP1	1.0317	1.4166	10	14		
CSNK1A1	-1.0112	-1.4370	6	15		
APC	-1.3493	-1.2140		19		
AXIN1	1.1479	1.0531		25	2	
CTNNB1	-1.0976	-1.5156		59		

Table 11.c – gene listings for the first-order neighbors of the protein interaction network super-hubs

S.H. neighb gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
AES	1.0724	-1.5903		13	2	
APC	-1.3493	-1.2140		19		
AXIN1	1.1479	1.0531		25	2	
BCL9	1.2254	1.4153		2		
BTRC	1.1259	2.0770		6	1	X
CDC42	-1.0832	-10.8582		5		
CDH1	-1.2061	-3.1153		1	1	X
CREBBP	-1.0374	-1.0701		13		
CSNK1A1	-1.0112	-1.4370	6	15		
CSNK1D	-1.0655	-1.1329		1	1	
CSNK2A1	1.0403	1.1668		9		
CTBP1	1.0317	1.4166	10	14		
CTNNB1	-1.0976	-1.5156		59		
DVL1	-	-		8		
DVL3	1.0520	1.2337		8		
EP300	-1.0963	1.2065		2		
FBXW11	1.1695	1.9334	8	2		
FHL2	1.8514	1.3280	8	1		
FRAT1	-1.0422	-1.9366		3		
FZD1	-1.3124	-1.1081		11		
GSK3B	1.3509	-1.2529		13	1	
GSK3B	1.3509	-1.2529		9	1	
HDAC1	1.0633	1.0283		7	2	
LEF1	1.7077	-2.4090		4		X
LRP6	-1.0039	1.2464		13		
MACF1	-1.0026	1.2452	52	6		
MAP3K4	1.0809	1.3205		2		
NFATC2	-1.0518	1.0457	13	4		
NLK	-	-		5		
NR5A1	-1.0583	-1.1433		1		

Table 11.c continued

S.H. neighb gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
PIN1	1.1830	1.3079		1	6	
PPP2CA	1.0510	1.3172		6		
PPP2R5B	-1.0745	-1.0278		1		
PPP2R5D	1.1345	1.5430		5		
RANBP3	1.1076	1.5524		3		
RUNX2	-1.0474	-1.1022		4		X
RUVBL1	2.2555	4.5135		2		
SALL1	-1.5466	-3.3709		1		
SMAD4	-1.0045	-1.1107		9		
SOX1	1.6872	1.0039		1		
TCF4	-1.1998	-1.6468		13		X
TCF7L2	-1.1514	-1.4934	238	9		
TFAP2A	-1.0213	1.2072		3		
TLE1	1.4090	-1.1476	9	13		
WNT1	-1.0628	-1.0441		9		

Table 12.a – network metrics for genes previously identified direct targets of the Wnt pathway

Wnt target gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
AXIN2	2.4377	1.4908	8	5	2	X
BTRC	1.1259	2.0770		6	1	X
CCND1	2.3425	5.9104	6	5	4	X
CDH1	-1.2061	-3.1153		1	1	X
CDX1	-1.1338	-5.8053		1		X
DKK1	1.3230	40.6236		4		X
FZD7	1.3782	1.4405		1		X
LEF1	1.7077	-2.4090		4		X
MYC	-1.0075	-1.1048		5		X
RUNX2	-1.0474	-1.1022		4		X
SFRP2	-1.2827	-1.3071		0	2	X
TCF4	-1.1998	-1.6468		13		X
WNT3A	-	-		1		X

Table 12.b – gene listings for the first-order neighbors of validated Wnt targets

Wnt neighb gene nodes	N-A FC	N-C FC	β -catenin tags	PPI neighbor count	Co-expression disruptions	Wnt target
AES	1.0724	-1.5903		13	2	
AXIN1	1.1479	1.0531		25	2	
AXIN2	2.4377	1.4908	8	5	2	X
CCND1	2.3425	5.9104	6	5	4	X
CDX1	-1.1338	-5.8053		1		X
CREBBP	-1.0374	-1.0701		13		
CTBP1	1.0317	1.4166	10	14		
CTNNB1	-1.0976	-1.5156		59		
CUL1	1.0039	1.3766		5		
DVL2	1.1105	1.8868		6		
FZD1	-1.3124	-1.1081		11		
HDAC1	1.0633	1.0283		7	2	
KREMEN2	1.0111	1.3491		5		
LEF1	1.7077	-2.4090		4		X
LRP5	1.0325	-1.0025		1		
LRP6	-1.0039	1.2464		13		
MAGI3	-1.6431	-3.2192		4	1	
MAP3K4	1.0809	1.3205		2		
MYC	-1.0075	-1.1048		5		X
NLK	1.1906	1.2528		6		
SKP1	1.1149	1.0038		5	5	
SMAD4	-1.0045	-1.1107		9		
TLE1	1.4090	-1.1476	9	13		

Table 13 – full metrics listing for the final set of 65 network nodes

Gene of interest	N-A upreg	N-C upreg	N-A downreg	N-C downreg	β -catenin binding	β -catenin neighbor	Wnt target	Wnt target neighbor	PPI Superhub	Superhub neighbor	Disrupted coexpression	Disruption neighbor
AES				-1.5903		X		X			2	
APC						X			19			
ARRB2											4	
AXIN1						X		X	25		2	
AXIN2	2.4378				8	X	X	X			2	X
BTRC		2.0768					X				1	
CAMK2D					30						2	
CCND1	2.3425	5.9102			6		X	X			4	X
CDC42				-10.858		X						
CDH1				-3.1153			X			X	1	
CDX1				-5.8053			X	X				
CREBBP						X		X				
CSNK1A1					6				15			X
CSNK1D										X	1	
CTBP1					10			X	14			X
CTNNB1				-1.5156				X	59			X
CUL1								X				
DAB2				-2.6291							1	X
DKK1		40.650					X					
DLG1				-3.5461								
DVL1		2.3969				X						
DVL2		1.8868				X		X				X

Table 13 continued

Gene of interest	N-A upreg	N-C upreg	N-A downreg	N-C downreg	β -catenin binding	β -catenin neighbor	Wnt target	Wnt target neighbor	PPI Superhub	Superhub neighbor	Disrupted coexpression	Disruption neighbor
FBXW11		1.9335			8							
FHL2	1.8515				8							
FRAT1				-1.9366						X		
FZD1						X		X				
FZD3	1.8175	5.9418										X
FZD5			-1.6837	-6.4483							1	
FZD7							X					
GNG2											8	
GSK3B										X	1	
HDAC1						X		X			2	
KREMEN2								X				
LRP1				-1.6999	54							
LRP5								X				
LRP6								X				
MACF1					52							X
MAGI3			-1.6431	-3.2192				X			1	
MAP1B			-2.2633	-1.5526							3	
MAP3K4						X		X				
MAP3K7											3	
MARK2					10							
MYC							X	X				

Table 13 continued

Gene of interest	N-A upreg	N-C upreg	N-A downreg	N-C downreg	β -catenin binding	β -catenin neighbor	Wnt target	Wnt target neighbor	PPI Superhub	Superhub neighbor	Disrupted coexpression	Disruption neighbor
NFATC2					13	X						
NLK						X		X				
NR5A1										X		
PIAS4		1.8332				X					1	
PIN1											6	
PP2R5D		1.5430								X		
PRKCA					78							
RAN	1.8695	3.1095										
RANBP3		1.5523								X		
RUNX2							X					
RUVBL1	2.2553	4.5126										
SALL1			-1.5466	-3.3709								
SFRP2							X				2	
SKP1								X			4	
SMAD3					63							
SMAD4						X		X				
SOX1	1.6872									X		
TAX1BP3				-3.3596	4						1	
TCF4				-1.6468		X	X					X
TCF7L2					238	X						X
TLE1					9	X		X				X
WNT3A							X					

Table 14 – statistical assessment of network uniqueness using a null model 110

Type of comparison based on combined metrics of interest	Obs prop	Mean exp prop	StErr exp prop	Exp 95 CI L	Exp 95 CI U	Odds ratio
Differential expression (DE)	0.238095	0.151580	9.41E-05	0.151396	0.151765	1.7491
Targeting by β -catenin (β -cat)	0.095238	0.060714	4.79E-05	0.060620	0.060808	1.6285
Network hub or super-hub (Hub)	0.375000	0.238932	1.10E-04	0.238716	0.239148	1.9112
Disrupted coexpression (Coex)	0.178571	0.113647	0.000237	0.113183	0.114112	1.6955
DE + β -cat	0.035714	0.022742	4.77E-06	0.022733	0.022751	1.5915
DE + Hub	0.053571	0.034185	6.01E-05	0.034067	0.034303	1.5992
DE + Coex	0.077381	0.049222	9.93E-05	0.049027	0.049416	1.6201
β -cat + Hub	0.047619	0.030355	1.04E-04	0.030151	0.030558	1.5972
β -cat + Coex	0.023810	0.015175	1.20E-05	0.015152	0.015199	1.5828
Hub + Coex	0.071429	0.045466	8.95E-05	0.045291	0.045641	1.6150
DE + β -cat + Hub	0.011905	0.007563	4.60E-05	0.007473	0.007653	1.5810
DE + β -cat + Coex	0.017857	0.011373	2.90E-05	0.011316	0.011430	1.5805
DE + Hub + Coex	0.023810	0.015148	6.54E-05	0.015019	0.015276	1.5858
β -cat + Hub + Coex	0.017857	0.011365	2.90E-05	0.011308	0.011422	1.5816
DE + β -cat + Hub + Coex	0.011905	0.007563	4.60E-05	0.007473	0.007653	1.5810

Figure 1

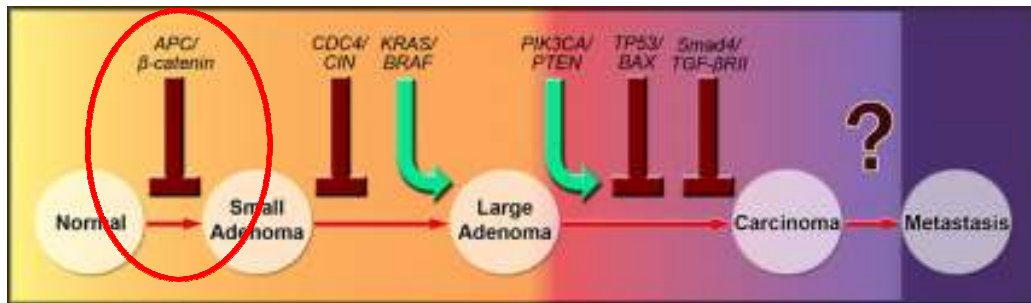
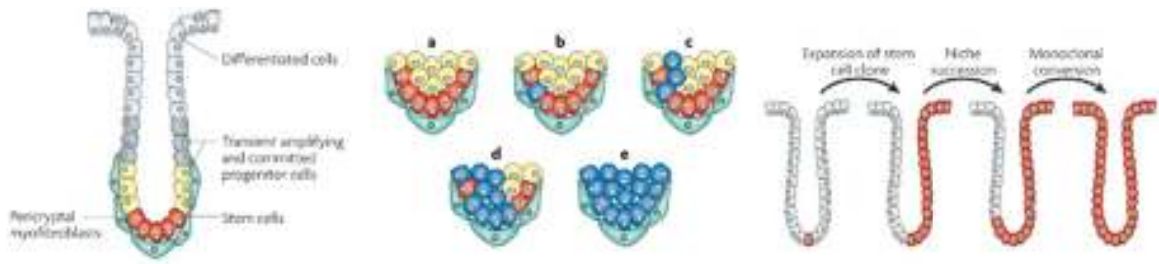


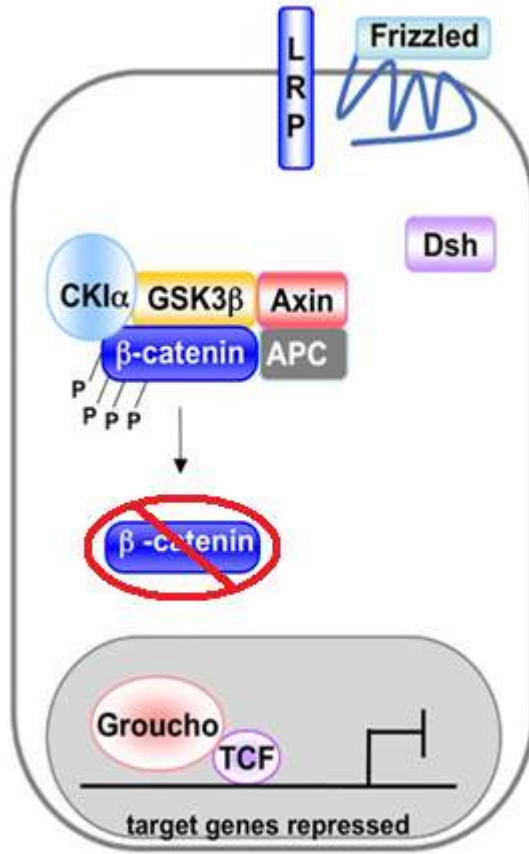
Figure 1

The colonic crypt and adenocarcinoma development. As shown, motility along the intestinal villi mirrors the development of lining cells from undifferentiated stem cells to fully functional epithelial sheets. Proper functioning is reliant on asymmetrical division of the stem cells otherwise a plethora of stem cells can lead to the formation of an early adenoma. Without suitable regulation in place, the crypt is quickly overwhelmed by the aberrant growth leading to the beginnings of an adenocarcinoma. Colorectal cancer in particular is defined by its early loss of function in the Wnt pathway as the adenoma is first forming making it an obvious point of interest for study.

Adapted from: Humpries et. al. *Nat Reviews Cancer*. 2008; **8**, 415-424 and Jones S et. al. *Proc Nat Acad Sci*. 2008; **105**(11): 4283-8.

Figure 2

Absence of Wnt:



Presence of Wnt:

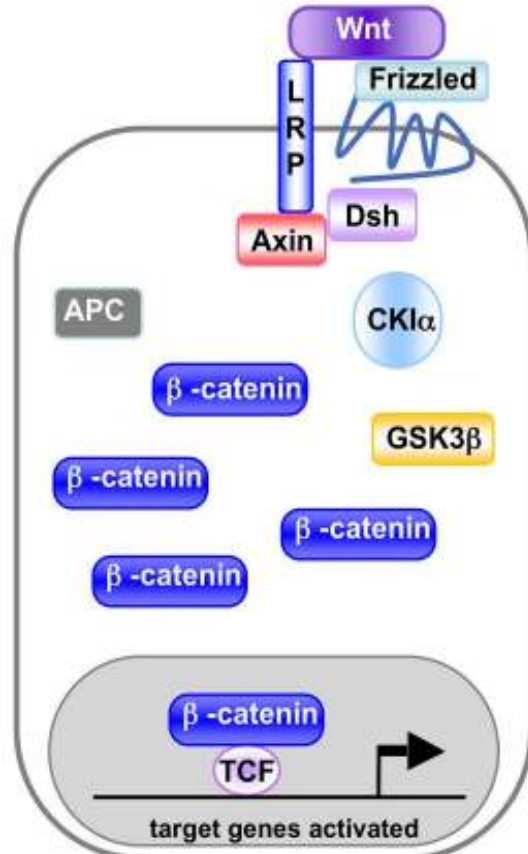


Figure 2

Wnt activation of β -catenin dependent transcription. In the absence of Wnt, cellular levels of β -catenin are limited by the destruction complex which degrades the compound while still in the cytoplasm. Conversely, Wnt signalling prevents the formation of the destruction complex through cell surface receptor binding to Axin. As β -catenin accumulates in the cytoplasm it begins to translocate into the nucleus, consequently activating numerous genes associated with biological functions such as embryogenesis and epithelial cell replenishment as well as cancer development.

Adapted from: Eisenmann. WormBook 2005.

Figure 3

Serial analysis of chromatin occupancy combines chromatin-immunoprecipitation with SAGE based sequencing techniques to create a library of transcription factor binding sites and sequences alongside tag counts for the chromatin-bound protein under investigation. Detailed above are the enrichment and purification techniques used on the collected protein-associated genomic DNA to prepare the binding sites for vector cloning, amplification and sequencing.

Adapted from: Impey S et. al. *Cell* 2004, **119**(7):1041-1054.

Figure 4

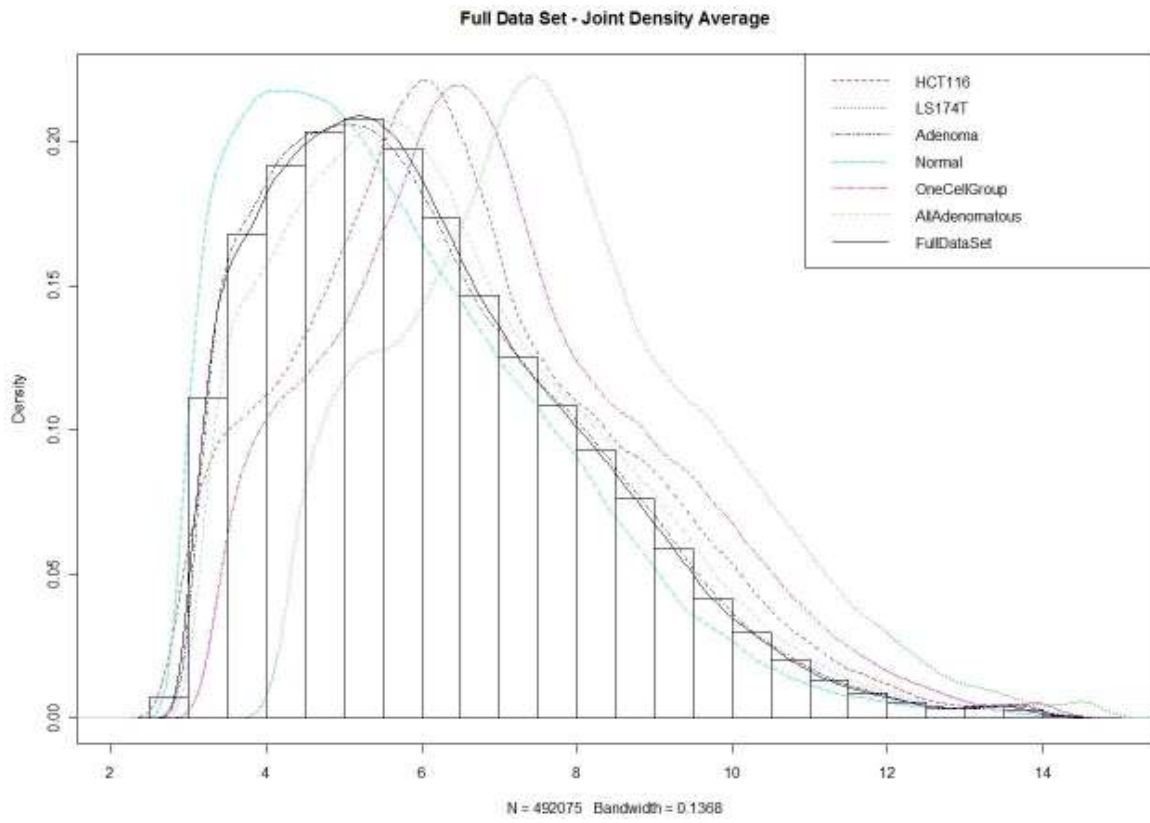
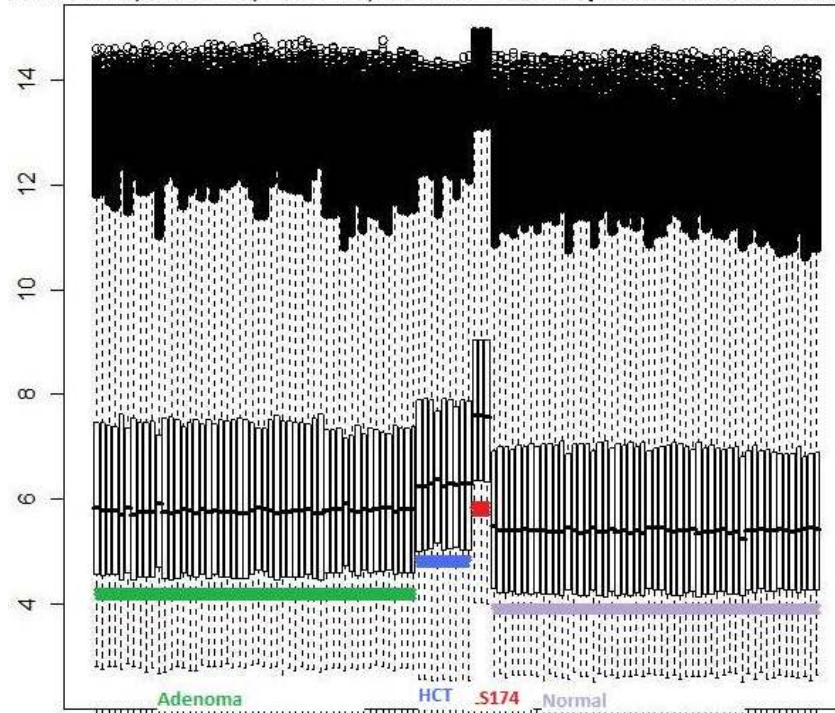


Figure 4

Density graph of spot intensities from the mRNA aggregation analysis. Taken from the exploratory data analysis, multiple density curves are overlain. Each of the four modules densities are shown individually in addition to the density curve for the normalization for the full set of utilized arrays. Partial combinations such as OneCellGroup (the pairing of the HCT116 and LS174T profiles) and AllAdenomatous (both cell lines combined with the patient adenoma data) are also included for comparison purposes.

Figure 5a

Adenoma, HCT116, LS174T, Normal - Full Separate Normalization



Adenoma, HCT116-LS174T Joint, Normal - Separate Normalization

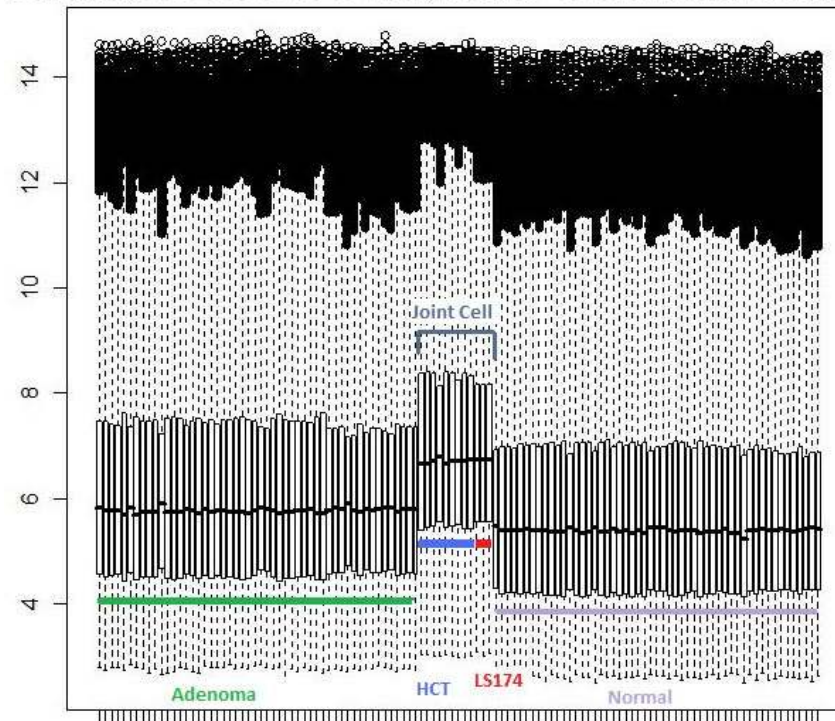


Figure 5a

Box graphs for the exploratory data analysis of the mRNA pooled study. Shown are the separate module normalizations with the expectation of the lower image displaying the effects of normalization when both cell lines are grouped together. This is not a standard method of normalization and was used predominately for illustration purposes.

Figure 5b

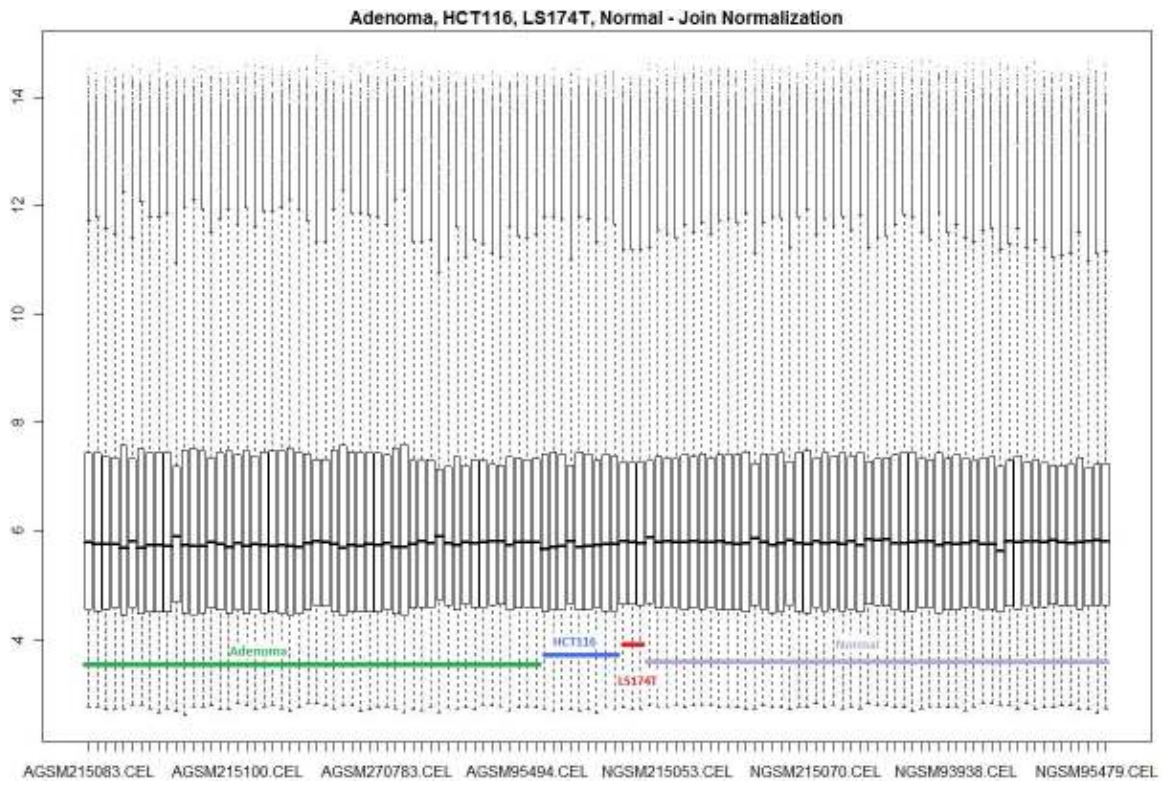


Figure 5b

The box graphs for the intensities of each array after RMA normalization combining all four functional modules. As is traditional for mRNA expression studies, this was the normalization used for the aggregate analysis.

Figure 6

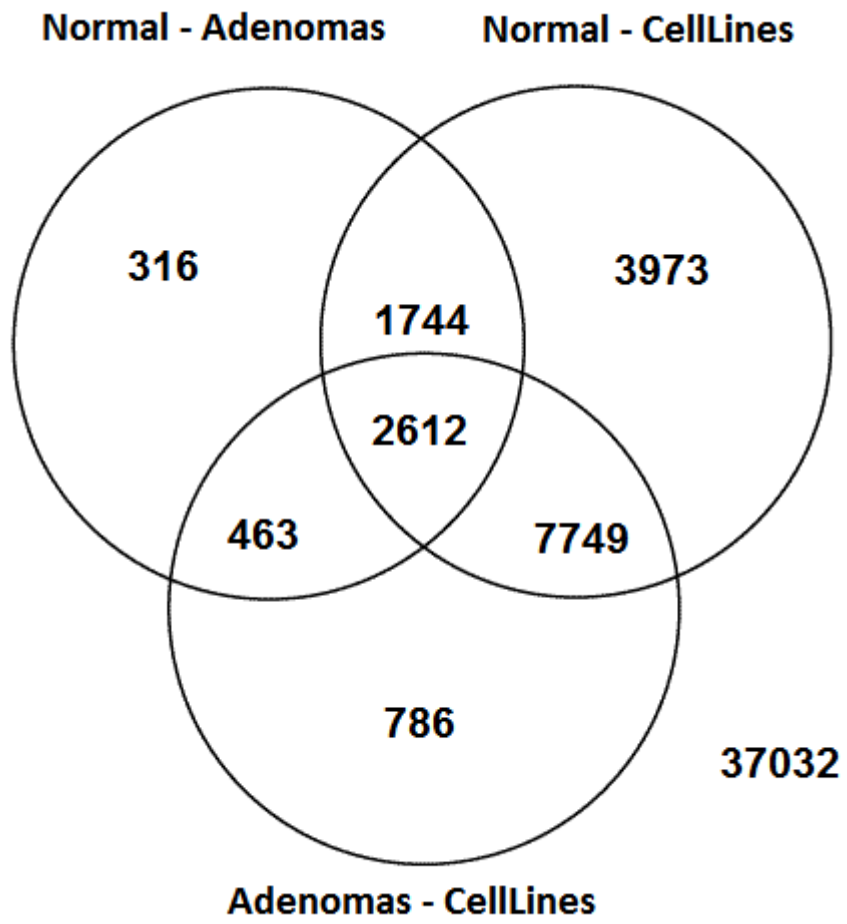


Figure 6

Venn diagram listing the number of probes identified by the linear model during the mRNA analysis as having significant differential expression (FDR 5%) meeting the 1.5x fold-change threshold.

Figure 7

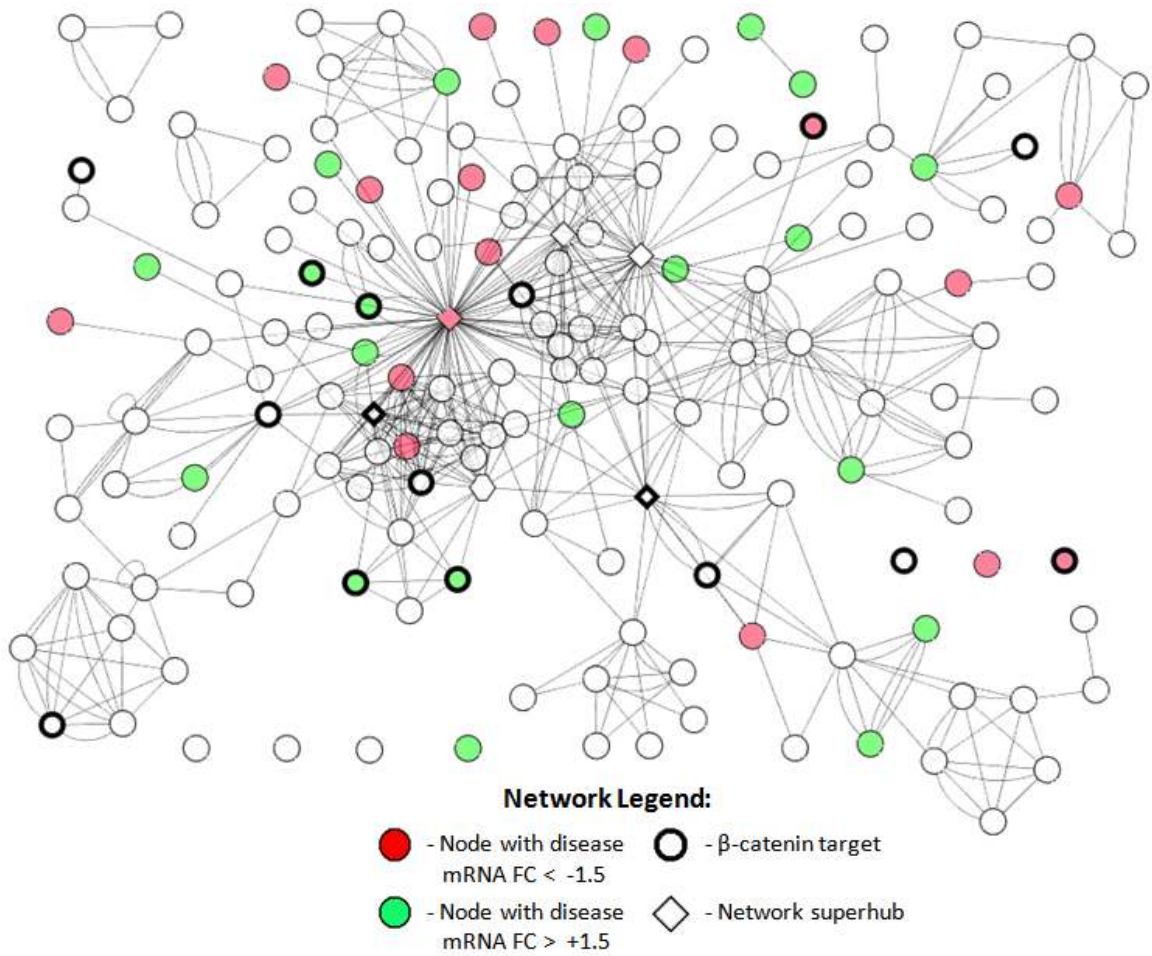


Figure 7

Overview of the base protein product interaction network. Indicated are the 39 genes showing significant mRNA expression changes in patient adenomas or cell lines, the 16 targets of β -catenin and the 5 super-hubs.

Figure 8




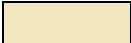

-  - Upregulated transcription in disease conditions
-  - Downregulated transcription in disease conditions
-  - Chromatin binding of β -catenin by SACO tags
-  - Three or more co-expression disruptions in disease conditions
-  - Previously validated Wnt pathway target from the Wnt homepage

Figure 8

Legend of the color-coding of the five metrics found in tables 7-12. Highlighting was done to assist in tracking various nodes through the multiple metrics combinations that were evaluated.

Figure 9

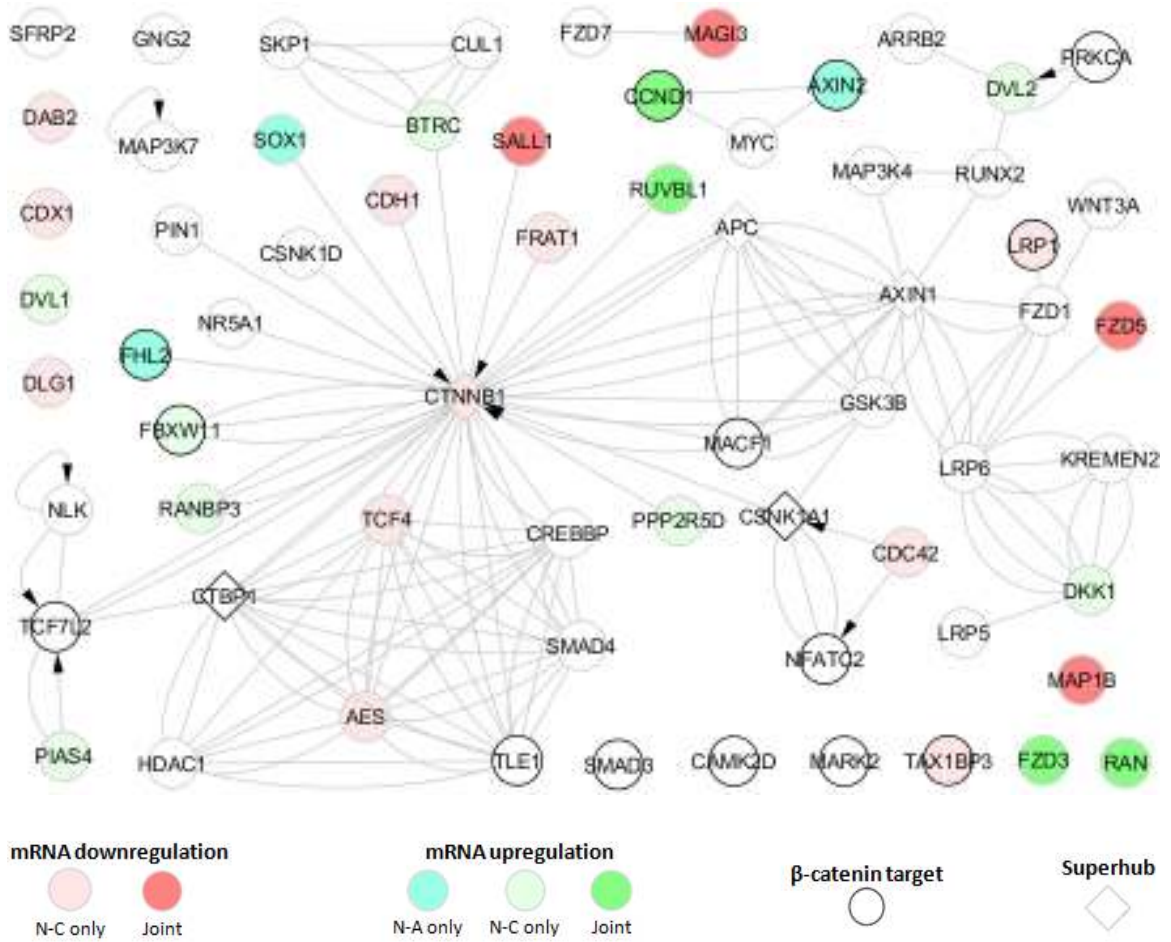


Figure 9

The refined network used for visualization; based on the 65 nodes identified as meeting multiple categorizations for the criteria including: differential expression, direct targeting by β -catenin or being a neighbor of a target, being a super-hub or super-hub neighbor, disruption in coexpression under adenomatous conditions, or previous identification for involvement with the Wnt pathway.