

DEFINING NOVEL CANCER GENES, SAGE TAGS AND CO-REGULATED
REGIONS OF THE HUMAN GENOME EVOLVING FROM THE SEARCH AND
IDENTIFICATION OF NOVEL WNT/TCF/B-CATENIN TARGETS

by

Erşen Kavak

B.S., Molecular Biology and Genetics, Bogaziçi University, 2004

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Molecular Biology and Genetics
Bogaziçi University
2009

Babama...

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor and mentor Prof. Ahmet Koman for teaching me critical thinking and how to work in the laboratory. I also would like to thank him for being patient, understanding and supportive all the time.

I would like to mention and thank my wife's unreferenced contribution to this thesis. She has always not only supported me emotionally but also helped me finding my way in several seemingly no way out situations.

I would also like to acknowledge Assoc. Prof. Rickard Sandberg's significant scientific contribution to the second part of the thesis as well as his financial support (January 2009 – May 2009) and friendship during production of this work.

I would also like to thank Prof. Monica Nister for recruiting me to Sweden and funding me during a part of this thesis work (March 2008 - December 2008).

Assoc. Prof. Durişehvar Özer Ünal and Assist. Prof. Necla Birgül always helped me and supported me with a lot with several different things.

Tuncay Şeker, good scientist of the future, helped me a lot with several experiments throughout the thesis.

I would also like to acknowledge the contribution of Nuri Öztürk, Ayaz Najafov and Tolga Aslan with some of the experiments performed in the first part of the thesis.

I also would like to thank my dear friends from Boğaziçi University; Demet Candaş, Nadire Duru, Kader Çavuşoğlu, H. Emre Etlioğlu, Yeşim Özmen, Adil Doğanay Duru, Sibel Uğur, Tunç Öztunç and Birdal Bilir for being my friends and supporting me.

I also would like to thank my friends from Karolinska Institute for making Sweden a warmer place; Christer Ericsson, Inga Nazeranko, Tomadher Elsir, Ulrica Westermarck, Anna Eriksson and Rona Strawbridge.

Finally, I would like to thank my mom, my dad and all my family for their constant love and support throughout my life and this thesis.

This study has been supported with grants from: DPT (State Planning Organization of Turkey), BAP (Bogazici University Research Projects), BUVAK (Boğaziçi University Foundation) Fahir İlker Scholarship and TUBITAK (The Scientific and Technological Research Council of Turkey, 104T027, 105T301 and 106T153)

ABSTRACT

DEFINING NOVEL CANCER GENES, SAGE TAGS AND CO-REGULATED REGIONS OF THE HUMAN GENOME EVOLVING FROM THE SEARCH AND IDENTIFICATION OF NOVEL WNT/TCF/B-CATENIN TARGETS

The Wnt signaling pathway is involved in many differentiation events during embryonic development and can lead to tumor formation after aberrant activation of its components. β -catenin, a cytoplasmic component, plays a major role in the transduction of canonical Wnt signaling. The first part of this study aimed at identifying novel genes that are regulated by active β -catenin/TCF signaling in hepatocellular carcinoma-derived Huh7 cells with high (transfected) and low β -catenin/TCF activities. High TCF activity Huh7 cells led to earlier and larger tumor formation when xenografted into nude mice. SAGE (Serial Analysis of Gene Expression), genome-wide microarray and *in silico* promoter analysis were performed in parallel, to compare gene expression between low and high β -catenin/TCF activity clones, and also those that had been rescued from the xenograft tumors. SAGE and genome-wide microarray data were compared and contrasted. BRI3 and HSF2 were identified as novel targets of Wnt/ β -catenin signaling after combined analysis and confirming experiments including qRT-PCR, ChIP, luciferase assay and lithium treatment. In the course of analyzing SAGE and microarray data of Huh7 cells, we felt the need to overlay this data with different data sources. Starting with this motivation in the second part of the study, we performed the first meta-analysis of SAGE and microarray data together, which we termed common gene expression patterning (COGENT). By using COGENT data, we introduced the “being multi-cancer” concept for genes being differentially expressed in several types of tumors. There are not only multi-cancer genes but also multi-cancer co-regulated regions (RIDGEs) on the genome. “Multi-canceriness” of a gene correlates well with the number publications stating that gene in a cancer context, its rank within a single study and the number of its orthologs. There are many multi-cancer genes which have not been related to tumorigenesis as yet. In addition, there are 81 tumor associated RIDGEs (TA-RIDGEs) on the human genome which change similarly in multiple types of tumors.

ÖZET

WNT/TCF/B-CATENİN YOLAĞINDA HEDEF TARAMALARINDAN BAŞLAYARAK YENİ KANSER GENLERİNİN VE İNSAN GENOMU ÜZERİNDE BİRLİKTE HAREKET EDEN BÖLGELERİN TANIMLANMASI

Wnt sinyal yolağı embryo gelişimi sırasında kilit noktalarda görev yapmakta olup, bu yolağın yersiz aktivasyonu tümör oluşumuna yol açabilmektedir. β -catenin molekülü Wnt yolağının sitoplazmik bir elemanı olup, Wnt sinyallerinin iletiminde merkezi bir öneme sahiptir. Bu tezin ilk kısmında Wnt sinyal yolağı aktif olan hücrelerle aktif olmayan hücreler birbirleriyle karşılaştırılmıştır. Bir karaciğer kanseri hücresi olan Huh7 hücrelerinde Wnt sinyal yolağı yapay olarak normalden fazla aktive edildiğinde bu hücreler, farede daha agresif ve daha büyük tümörler oluşturmaktadırlar. Wnt aktiviteleri ve tümör oluşturma yetenekleri farklı olan Huh7 hücrelerinin, gen anlatımının seri analizi (serial analysis of gene expression - SAGE) ve mikrodizin yongaları metodlarıyla karşılaştırılmaları yapılmıştır. Bu karşılaştırma sonrasında, bilgisayar destekli promotör analizinin de yardımıyla Wnt sinyal yolağının muhtemelen hedeflediği moleküller belirlenmiştir. Belirlenen aday moleküller arasından BRI3 ve HSF2 genlerinin Wnt hedefi olduğu lüsis raporlama metodu, kromatin immüno çöktürme metodu ve polimerizasyon zincir reaksiyonu gibi metodlarla doğrulanmıştır. Çalışmanın birinci kısmında yapılan SAGE ve mikrodizin yongası analizi sırasında oluşan gen listelerinin farklı listelerle karşılaştırılarak listedeki genlerden çalışılabilir bir alt küme oluşturma ihtiyacı doğmuştur. Bu ihtiyaç çerçevesinde, Wnt hedefi adaylarından hangilerinin kanser oluşum sürecinde de önemli olacağını bulmak için literatürde yayınlanmış 49 farklı tümör tipine ait SAGE ve mikrodizin yongası datası analiz edilmiştir. Bu analize kanserlerde ortak gen anlatım kalıpları (common gene expression patterning in tumors – COGENT) adı verilmiştir. COGENT analizi sayesinde bir genin çok kanser tipinde değişmesinin o gen için belirleyici bir özellik olduğu gösterilmiştir. Eğer bir gen daha fazla kanser tipinde değişiyorsa, o zaman büyük ihtimalle o genle alakalı daha fazla kanser yayını vardır. Fakat, aynı zamanda çok kanserde değişen ve hakkında hiç kanser yayını, hatta hiç yayın olmayan genler de vardır.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	iv
ABSTRACT	vi
ÖZET	vii
LIST OF FIGURES	xi
LIST OF TABLES	xiii
LIST OF ABBREVIATIONS	xv
1. INTRODUCTION	1
1.1. Wnt Pathways	1
1.2. Wnt Pathways and Cancer	3
1.3. Serial Analysis of Gene Expression	4
1.4. Meta analysis of cancer gene expression profiles and co-regulated regions on the genome (RIDGEs)	5
2. AIM	7
3. MATERIALS AND METHODS	8
3.1. Analysis of the Wnt/B-catenin/TCF4 pathway using SAGE, genome-wide microarray and promoter analysis: Identification of BRI3 and HSF2 as novel targets.	8
3.1.1. RNA isolation, Serial Analysis of Gene Expression and microarrays	8
3.1.2. Cell culture experiments	10
3.1.3. Plasmids	11
3.1.4. SAGE and microarray comparison	11
3.1.5. Promoter Analysis and proMine	12
3.1.6. Luciferase assay and promoter plasmids	12
3.1.7. Lithium treatment	13
3.1.8. ChIP Assay	13
3.2. Meta analysis of cancer gene expression profiles reveals cancer genes, SAGE tags and co-regulated regions on the genome	14
3.2.1. Data collection and processing	14
3.2.1.1. Microarray data	14

3.2.1.2. SAGE data	15
3.2.2. Being multi-cancer	17
3.2.2.1 RT-PCR and Q-RTPCR experiments	17
3.2.3. RIDGEs (Regions of increased co-regulation on the genome)	19
3.2.4. Implementation of data to MySQL database	22
4. RESULTS	23
4.1. Analysis of the Wnt/B-catenin/TCF4 pathway using SAGE, genome-wide microarray and promoter analysis: Identification of BRI3 and HSF2 as novel targets	23
4.1.1. TCF over-activation by stable over-expression of β -catenin leads to bigger and more aggressive tumors in nude mice	23
4.1.2. SAGE and genome-wide microarray analysis of differential expression between high and low TCF activity Huh7 cells	24
4.1.3. In silico promoter analysis	29
4.1.4. Novel Wnt/ β -catenin/TCF targets; BRI3 and HSF2	34
4.1.4.1 BRI3 and HSF2 promoters are regulated by the β -catenin/TCF4 complex	39
4.1.4.2 β -catenin interacts with BRI3 and HSF2 promoters	40
4.1.4.3 Inhibition of GSK3 β by lithium leads to increased BRI3 and HSF2 mRNA levels	40
4.2. Meta analysis of cancer gene expression profiles reveals cancer genes, SAGE tags and co-regulated regions on the genome	41
4.2.1. Being multi-cancer	53
4.2.2. Under-expressed ribosomal genes in cancer suggests functional roles for ribosomal genes	62
4.2.3. Genomic regions of transcriptional imbalance (RIDGEs)	63
4.2.4. Differential expression at TA-RIDGEs	69
4.2.5. GC Content is a major factor affecting presence of RIDGEs	74
4.2.6. Online re-analysis and availability of the data: www.oncoreveal.org	77
5. DISCUSSION	79

5.1. Analysis of the Wnt/ β -catenin/TCF4 pathway using SAGE, genome-wide microarray and promoter analysis: Identification of BRI3 and HSF2 as novel targets	79
5.2. Meta analysis of cancer gene expression profiles reveals cancer genes, SAGE tags and co-regulated regions on the genome	81
5.3. Conclusions and Future Perspectives	86
REFERENCES	87

LIST OF FIGURES

Figure 1.1.	Wnt/TCF/ β -catenin pathway	1
Figure 1.2.	β -catenin-independent Wnt signaling	2
Figure 3.1.	The pseudocode for score of over-representation for each tumor-normal comparison	20
Figure 3.2.	The pseudocode for calculating the significance of co-regulation between two tumor-normal comparisons	21
Figure 4.1.	TCF over-activation by stable over-expression of β -catenin leads to larger and earlier onset tumors in nude mice	23
Figure 4.2.	Verification and expansion of SAGE and Microarray results by using Q-RT-PCR and RT-PCR	25
Figure 4.3.	A particular pattern for down-regulated Wnt targets	33
Figure 4.4.	BRI3 is a novel target of Wnt/ β -catenin/TCF signaling – I	35
Figure 4.5.	BRI3 is a novel target of Wnt/ β -catenin/TCF signaling – II	36
Figure 4.6.	<i>HSF2</i> is a novel target of Wnt/ β -catenin/TCF signaling – I	37
Figure 4.7.	<i>HSF2</i> is a novel target of Wnt/ β -catenin/TCF signaling – II	38
Figure 4.8.	<i>AXIN2</i> is an immediate responder to lithium treatment of Huh7 and U373-MG cells	39
Figure 4.9.	Study design for meta analysis of cancer gene expression profiles	42
Figure 4.10.	Verification and extension of OncoReveal predictions in an independent extended brain tumor panel	47
Figure 4.11.	IQSEC1, C17ORF45 and CPLX1 but not other genes inhibit colony formation of 18k40 cells	53
Figure 4.12.	Being multi-cancer	55
Figure 4.13.	Determination of appropriate window size(s) for tumor associated RIDGEs	64
Figure 4.14.	Overview and close up of tumor associated and normal associated RIDGEs by using gene neighborhood correlation score	66
Figure 4.15.	Distributions of “family gene” percentages and peaking gene densities within RIDGEs	68
Figure 4.16.	Regulation of Keratin, SPRR and S100A gene families at skin cancer and its relation to melanoma metastasis	70

Figure 4.17.	MHC Class II RIDGE	73
Figure 4.18.	Histone 1 RIDGE	73
Figure 4.19.	Chr8_NAPRT1 an example of non-paralog RIDGE which is differentially expressed in a wide variety of tumors	73
Figure 4.20.	Online service; www.oncoreveal.org	78

LIST OF TABLES

Table 3.1.	Primer List used in the Part I	8
Table 3.2.	Primers used for RT-PCR and Q-RT-PCR at Part II	18
Table 4.1.	Comparison of SAGE and Microarray results by using Q-RT-PCR and RT-PCR	26
Table 4.2.	SAGE and microarray comparison by using tendency analysis	26
Table 4.3.	Percent common with different thresholds	27
Table 4.4.	Tendency analysis in high abundance class (tag count ≥ 15)	27
Table 4.5.	Tendency analysis in middle abundance class ($3 < \text{Tag count} < 15$) . .	27
Table 4.6.	Tendency analysis in low abundance class (Tag count ≤ 3)	27
Table 4.7.	Comparison of microarray data from S33Y expressing and control Huh7 cells (Huh7 – S33Y and Huh7 – vec) and their post tumor counterparts which were rescued cells from corresponding xenografts (PT-Huh7-S33Y and PT-Huh7- vec) with Huh7 side population . . .	29
Table 4.8.	Genes which are up-regulated in both Huh7 side population cells and post tumor cells	30
Table 4.9.	Number of TCF binding sites in the promoters of previously verified Wnt targets by using short (A) or long (B) TCF motif	31
Table 4.10.	Number of TCF binding elements (TBEs) in the promoters of up (A) and down-regulated(B) genes in Huh7-S33Y in comparison to Huh7- vec	32
Table 4.11.	Tumor types used in the meta-analysis	42
Table 4.12.	Microarray studies used in the meta-analysis.	44
Table 4.13.	SAGE datasets used in the meta-analysis.	46
Table 4.14.	SAGE Genie and SAGE map agreement statistics for short SAGE tags (A) and long SAGE tags (B)	50
Table 4.15.	COGENT-SAGE analysis by using OncoReveal. Several multi-cancer genes which are differentially expressed at CNS tumors	51
Table 4.16.	Already well established tumor-associated multi cancer genes (A) and multi-cancer genes which do not have any cancer geneRIF (B)	56
Table 4.17.	DAVID analysis to find enriched classes of genes in tumor-associated multi-cancer genes (A) and not-yet-associated to tumor multi-cancer	59

	genes (B)	
Table 4.18.	Revealed SAGE tags by using COGENT	62
Table 4.19.	Co-regulation (p value < 0.01) frequencies for over-expressed (A) and under-expressed (B) genes according to COGENT data with q value filter of 0.05 and SAGE p value filter of 0.02	75
Table 4.20.	GC rich regions are more likely to be co-regulated	77

LIST OF ABBREVIATIONS

Ca ²⁺	Calcium
cDNA	Complementary deoxyribonucleic acid
CNS	Central nervous system
CO ₂	Carbon dioxide
COGENT	Common Gene Expression Patterning
DAVID	DAVID gene classification tool of NIH (national institutes of health)
DEPC	Diethylpyrocarbonate
DGED	Digital Gene Expression Display tool under SAGE Genie SAGE data analysis utilities
DMEM	Dulbecco's Modified Eagle's Medium
DNA	Deoxyribonucleic acid
DNET	Disembryoblastic neuroepithelial tumor
dNTP	Deoxynucleosidetriphosphate
DTT	Dithiothreitol
EDTA	Ethylenediamine tetra acedic acid
ER	Endoplasmic Reticulum
EST	Expressed Sequence Tag
FBS	Fetal Bovine Serum
FITC	Fluorescein isothiocyanate
GC	Guanin or Cytosine base
GEO	Gene Expression Omnibus database at NCBI's database system
geneRIF	Gene References Into Function (a sub-database of Entrez GENE database)
Huh7-vec	Huh7 cells stably expressing the empty pCI-NEO mammalian expression vector
Huh7-S33Y	Huh7 cells stably expressing β -catenin molecule with S33Y mutation
MGC	Mammalian Gene Collection
mRNA	Messenger ribonucleic acid

NA-RIDGE	Normal tissue regulation associated Regions of Increased Co-regulation on the Genome
NCBI	National Center for Biotechnology
NMSC	Non-melanoma skin cancer
PAGE	Polyacrylamide gel electrophoresis
PCP	Planar Cell Polarity
PCR	Polymerase chain reaction
PT-Huh7-S33Y	Huh7 cells recultured from the mice xenograft which was formed by Huh7 cells stably expressing β -catenin molecule with S33Y mutation
PT-Huh7-vec	Huh7 cells recultured from the mice xenograft which was formed by Huh7 cells stably expressing the empty pCI-NEO mammalian expression vector
Q-RTPCR	Quantitative Reverse transcriptase mediated polymerase chain reaction
RIDGE	Regions of increased co-regulation on the genome
RNA	Ribonucleic acid
	Reverse transcriptase mediated polymerase chain reaction
RT-PCR	
SAGE	Serial Analysis of Gene Expression
SD	Standard Deviation
SE	Standard Error
TA-RIDGE	Tumor associated Regions of Increased Co-regulation on the Genome
TBE	TCF binding element
TCF	T cell Factor
TFBE	Transcription Factor Binding Element
TSS	Transcriptional Start Site

1. INTRODUCTION

1.1. Wnt Pathways

The Wnt/Wingless signaling pathway has been found to be important in several biological processes such as development [1], carcinogenesis [2-4] and stem cell biology [1, 5-7]. In mammals, 12 distinct Wnt protein families exist which might induce at least 4 different pathways: Wnt/ β -catenin/TCF pathway, Wnt/ Ca^{2+} pathway, Wnt/planar cell polarity (PCP) pathway and Wnt/G protein pathway [1, 8, 9]. The Wnt/ β -catenin pathway is the most established one among these pathways. β -catenin is a key regulator of the Wnt/ β -catenin signaling pathway, dysregulation of which was established in several tumors such as colorectal tumors [10], liver tumors [11], and breast tumors [12]. Considerable amount of data tells us that, these dysregulations lead to β -catenin hyperactivation that mimics active Wnt/ β -catenin pathway which might result in aberrant activation of cell cycle mechanisms, a major mark of tumor initiation.

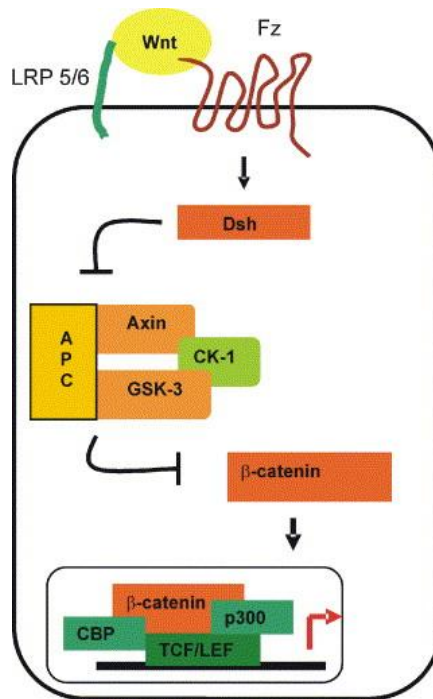


Figure 1.1. Wnt/TCF/ β -catenin pathway (Kohn et al, 2005)

Wnt/TCF/ β -catenin; the so-called ‘canonical pathway’ is known to act through β -catenin. Induction of this pathway is achieved by Wnt binding to Frizzled receptor and LRP5-6 coreceptors. This in turn activates Dishevelled which will inhibit the formation of

a GSK3 β -Axin-APC complex that phosphorylates β -catenin, thus marking it for degradation. If not thus degraded, β -catenin, travels into the nucleus; forms transcriptosomes with the TCF family of transcription factors and activates several target genes, which in turn determine cell fate (Figure 1.1). Non-canonical Wnt pathways are Wnt/calcium, Wnt/Planar Cell Polarity (PCP) and Wnt/G protein pathways. In the Wnt/calcium pathway, Wnt ligands bind to Frizzled and co-receptor Knypek (Kny) or Ror2. This binding leads to activation of calcium/calmodulin-dependent kinase II (CamKII) and protein kinase C (PKC) which will lead to multiple further changes in the cell. In the Wnt/PCP pathway, Dishevelled activation leads to activation of some GTPases such as Rho and cdc42. In addition; there is a G protein linked pathway downstream of Wnt/Frizzled binding, concordant with the fact that Frizzled is a seven trans-membrane domain containing receptor which has some homology with G protein coupled receptors, Wnt binding might lead to activation of phospholipase C (PLC) and phosphodiesterase (PDE), via G proteins. In spite of the effort to draw boundaries between non-canonical Wnt pathways, there seem to be cross-talks between Wnt/Calcium; Wnt/PCP and Wnt/G-protein pathways [9] (Figure 1.2).

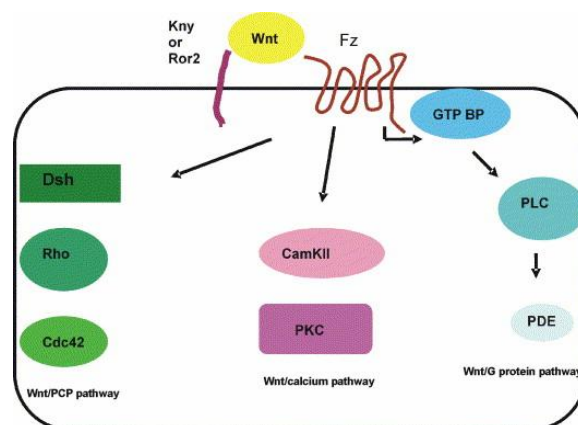


Figure 1.2. β -catenin-independent Wnt signaling. Downstream events are initiated by Wnt binding to a Frizzled (Fz) receptor and possibly the coreceptor, Knypek (Kny) or Ror2. This can lead to activation of a pathway involving calcium/calmodulin-dependent kinase II (CamKII) and protein kinase C (PKC). Frizzled may activate heterotrimeric GTP-binding proteins (GTP BP), leading to activation of phospholipase C (PLC) and phosphodiesterase (PDE). Frizzled can also recruit Dishevelled (Dsh), leading to activation of small GTP-binding proteins, such as Rho and Cdc42. Figure and Figure legend adapted from Kohn et al, 2005.

Not only tumor initiation capacity but also the role in development and stem cell biology of the Wnt pathway makes Wnt targets important. There are many targets already identified in different contexts such as colorectal, liver [13-15] and ovarian cancers [16]. Wnt pathway is also known to be the target of several different drugs in several different diseases including cancer [17-20].

1.2. Wnt Pathways and Cancer

Because Wnt targets such as cMyc and CyclinD1 have important roles in cell growth control, it is a natural fact that Wnt pathway and Wnt pathway targets are important players of tumorigenesis. In fact, lots of mutations and/or accumulations of Wnt pathway members have been reported in several cancers such as colorectal [10], breast [12], ovarian [13] liver [14] and hematologic malignancies [21].

In 90 per cent of colorectal cancers, there has been an APC and/or β -catenin mutation [14]. Besides, there is an intra-cellularly heterogeneous (cytoplasmic or nuclear) accumulation of β -catenin in colorectal cancers (CRCs). Nuclear accumulation tends to be in the invasive front of colorectal tumors [7].

Several different mutations in the APC gene, including truncation or single amino acid mutations might or might not lead to inhibition of β -catenin regulation ability. APC mutations are found in almost all cases of familial adenomatous polyposis (FAP), although this type of colorectal cancer accounts for approximately 1 per cent of all CRC cases. Besides APC mutations in colorectal cancer, β -catenin mutations are also seen in sporadic CRC cases. [10]

Wnt/ β -catenin signaling seems to act through dissimilar mechanisms in colorectal and liver cancers. In colorectal cancer; the major mechanism is APC interaction inactivation where in 90 per cent of the colorectal cancers, APC or β -catenin is mutated. However, in hepatocellular cancers, although the nuclear β -catenin accumulation rate is between 33-67 per cent, there are APC or β -catenin mutations only in 20-30 per cent of the hepatocellular carcinoma cases [14]. Thus mutation and nuclear expression levels do not

necessarily correlate. In liver cancer, β -catenin mutation status is correlated with the stage and prognosis of HCC. B-catenin is found mutated mainly in lower stage (stage I) HCCs and patients with mutant β -catenin have higher survival rates [22]. It has also been shown that, canonical Wnt targets such as MMP7 and cyclinD1 are not changed during β -catenin accumulation in hepatocarcinogenesis [23].

Thus, explaining Wnt pathway together with its targets in liver cancer will provide insight in to diagnostic and therapeutic process of liver cancers. There has already been several studies trying to identify novel Wnt targets in liver by different approaches such as; transgenic mouse models with β -catenin over-expressing liver [24] which suggested epidermal growth factor receptor (EGFR) as a target; or analysis of liver cancer cell lines which suggested novel targets such as FUT8, LYZ and STMN2 [25].

1.3. Serial Analysis of Gene Expression

Serial Analysis of Gene Expression (SAGE) [26] is a gene expression profiling technique based on high throughput sequencing, which allows fast and quantitative analysis of a wide variety of samples. Currently, it can be considered as one of the two most popular gene expression profiling methods, the second one being the microarray approach. Despite the advent and rapid adoption of microarray technology for analyzing differential gene expression, SAGE, especially with the availability of new generation sequencing technologies, continues to be used since it allows sensitivity and specificity comparable with microarrays while additionally allowing the quantization of differential expression levels for both known and unknown genes thanks to the lack of a requirement for existing probes. It can also be said that, SAGE and microarray often give different results, even with a tendency analysis the agreement is not very good [27, 28].

SAGE is based on the following three principles: (1) it extracts a short tag (14 or 19 bp for SAGE and Long-SAGE respectively) from the 3' end of each mRNA within a certain sample. These tags are used to identify the gene by aligning the tag sequence to the Unigene database. (2) In contrast to microarrays, SAGE does not need *a priori* knowledge of sequence information; it can quantify whatever is in the cell. Therefore, it allows

identification of novel genes. (3) By forming concatamers from the short tags (a typical concatamer contains 40 tags) it allows rapid and direct quantification of mRNAs.

There has been several experimental modifications [29-37] and upgrades [38-40] on SAGE. Besides experimental modification efforts, there has been plenty of research from the bioinformatics point of view. There are currently three desktop programs which are designed to analyze SAGE data: iSAGE (the original program distributed with SAGE), eSAGE [41] and SAGEparser [42]. Each has its advantages and disadvantages. USAGE [43], SAGEgenie [44] and Identitag [45] are three web based programs to better interpret SAGE results. There are some additional efforts, which try to determine the differentially expressed genes in SAGE experiments [46], which successfully cluster the SAGE data [47] and which are designed to correct some experimental artifacts [48, 49].

SAGE has been applied to compare expression profiles in a wide variety of studies [50] such as testing effects of molecules or to compare transcriptomes of several different cancers [50, 51]. One promising method, Gene Identification Signature (GIS) [52] analysis is a novel method to annotate the functional genomic regions, and it is designed on the basic principles of SAGE.

By developing high throughput sequencing systems, the cost of SAGE is being reduced and it is possible that, SAGE will remain attractive in the near future, particularly with uncharacterized genomes.

1.4. Meta analysis of cancer gene expression profiles and co-regulated regions on the genome (RIDGEs)

Several different experimental methodologies such as microarrays, SAGE, and ChIP-chip are being increasingly used globally to generate tremendous amounts of high-throughput data. SAGE, as previously introduced, is based on concatenation and cloning based counting of tags from 3' end of mRNAs. Microarrays, on the other hand, are hybridization based systems in which a prior knowledge of sequence is needed. There are lots of data available in public databases from both SAGE and microarray technologies. Regardless of the technology used, meta-analysis of biological data overcomes the

individually undetectable experimental and/or technical biases introduced within single studies.

Meta analysis of microarray based gene expression profiles is starting to be an established [53] and useful method for deciphering several concepts of oncology such as common cancer gene expression signatures [54, 55], tissue-specific expression of cancer genes [56] and common transcription factors of cancer regulation [57]. On the other hand, usage of SAGE and microarray data together in a large scale has not been reported as yet.

Co-regulated regions on the human genome were first defined by Cohen et al [58] by analyzing gene expression profiles of two mitotic cell cycles of yeast. Afterwards, Caron et al [59] defined regions of increased gene expression (RIDGEs) by using a running median over a window of 36 genes along the chromosomes while performing SAGE library analysis from different tissues. This was followed by demonstration of large domains of co-regulation on the drosophila genome [60]. And, recently Stransky et al. [61] showed that some of these RIDGEs are dependent on chromosomal amplifications/deletions while some are not. Overall, it seems important to consider the presence of wider regulatory mechanisms, either at genomic or epi-genetic levels or of any other putative mechanisms such as miRNAs which might explain higher level regulation of multiple genes.

2. AIM

First part - Analysis of the Wnt/B-catenin/TCF4 pathway using SAGE, genome-wide microarray and promoter analysis: Identification of BRI3 and HSF2 as novel targets

The primary aim was to identify novel Wnt targets by comparing gene expression profiles of β -catenin hyper-activated Huh7 cells with controls. The secondary aim was to compare SAGE and microarray technologies in terms of accuracy and sensitivity.

Second part – Meta analysis of cancer gene expression profiles reveals cancer genes, SAGE tags and co-regulated regions on the genome

In the second part, the initial motivation was to screen the gene list we found from Part I and expand our analysis. The primary aim was to perform a meta-analysis of cancer gene expression profiles in order to reveal some genes, patterns or pathways. In addition, we aimed at identifying genomic regions that are co-regulated in tumorigenic transformation.

3. MATERIALS AND METHODS

3.1. Analysis of the Wnt/B-catenin/TCF4 pathway using SAGE, genome-wide microarray and promoter analysis: Identification of BRI3 and HSF2 as novel targets

3.1.1. RNA isolation, Serial Analysis of Gene Expression and microarrays

We isolated total RNA from cell lines by using the classical guanidine isothiocyanate method. We performed SAGE experiments by using Invitrogen's I-SAGE kit (Invitrogen, San Diego, CA USA) according to the kit manual and EMBO's SAGE manual [62], with 25 µg of total RNA from both samples. Kit manual was followed until preparation of 100 bp ditags. After this step, we used EMBO's SAGE manual, which contains some minor modifications. Sequencing was done with Amersham's 96 capillary sequencer.

Microarray experiments were performed in triplicate, 10 µg of total RNA for each sample. Affymetrix kit was used to prepare the labeled oligonucleotides, and Affymetrix washing and scanning station was used to scan the images. Data analysis was done with Microsoft Excel and GenePattern software of MIT [63].

Table 3.1. Primer List used in the Part I (Wnt targets)

Primer Name	Sequence (5' to 3')	Purpose
BRI3_1F	AGGGTCTACAACATCCACAG	1F and 2R is pair for Q-RTPCR
BRI3_2R	TGTTCCCTTTAAGCGAAGGT	
HSF2_4R	AAGCTGCTTATCTGGTTTGG	4R and 5F is pair for Q-RTPCR to detect HSF2- α isoform
HSF2_5F	TAGTTCTGTGCAGATGAATC	
HSF2_6F	CTGGTTGATCTTAGAATAAAGGAT	6F and 7R is pair for Q-RTPCR to detect HSF2- β isoform
HSF2_7R	TTCTGATGATGCTGTTGTACT	
TINP_1F	ACTAAAGTGTGCTTTGTTGGAG	1F and 2R is pair for Q-RTPCR
TINP1_2R	GGATGAGGGATTCTTCTTTACAC	
CLU_1F	GAATTGGAGGCATGATGAAGAC	1F and 4R is pair for Q-RTPCR
CLU_4R	CTGATTGGACATTCCTGGAG	
SEC31L1_5F	GAAACCTATTCCAGATGAGCAC	1F and 2R is pair for Q-RTPCR
SEC31L1_6R	GTTGGTGAAAGTGTCTGTTC	

Table 3.1. (continued)

IQSEC_1F	GGCTGAAATTCCACATTGCT	1F and 2R is pair for Q-RTPCR
IQSEC_2R	TGGAGAAGCACAGGATAGAG	
APOA2_1F	ACTGTTACCAACATGAAGCTG	1F and 4R is pair for Q-RTPCR
APOA2_4R	TTTCAAAGTAAGACTTGGCCT	
NES_1F	AGACTTCCCTCAGCTTTCAG	1F and 2R is pair for Q-RTPCR
NES_2R	ATCTCTGCATCTACAGCAGG	
MGAT1_1F	CCTTGCCCGCAGTCCTA	1F and 2R is pair for RTPCR, they have not been tested in Q-RTPCR
MGAT1_2R	CGGAGCAGGCCAAGTTC	
ATP5G1_1F	GCCTGATTAGACCCCTGGTA	1F and 2R is pair for RTPCR, they have not been tested in Q-RTPCR
ATP5G1_2R	GGCTAAAGCTGGGAGACTGA	
KIAA1924_1F	GGCAGGGGTACACGACAG	1F and 2R is pair for RTPCR, they have not been tested in Q-RTPCR. KIAA1924 is a synonym of gene with current symbol WDR90
KIAA1924_2R	CTTCCTCCCAGACCCAGTC	
TPT1_1F	GATCGCGGACGGGTGT	1F and 2R is pair for RTPCR, they have not been tested in Q-RTPCR
TPT1_2R	TTCAGCGGAGGCATTTC	
BLP1_1R	CATCCCAGGAAGAAGGAGTAG	1F and 2R is pair for RTPCR, they have not been tested in Q-RTPCR
BLP1_2F	GGAATTGAGTGTGCCAGTCC	
MMP7_1F	ATGGACTTCCAAAGTGGTCACCTACAG	1F and 2R is pair for RTPCR, they have not been tested in Q-RTPCR
MMP7_2R	GGATACATCACTGCATTAGGATCAGAG	
MYC_1F	TCTGGATCACCTTCTGCTGG	1F and 2R is pair for RTPCR, they have not been tested in Q-RTPCR
MYC_2R	TGTTGCTGATCTGTCTCAGG	
HSF2_pr_1F	taagctagcAACTGGCACTGGTGGCTAAG	HSF2_pr_1F and HSF2_pr_2R were used for amplifying genomic DNA from ChIP experiment
HSF2_pr_2R	cataagcttGGAGACCGTGTTTGGATGTT	
HSF2_pr_3F	taagctagcCTGAGAATTTCAAGAGGATGAG	HSF2_pr_3F and HSF2_pr_2R are used to clone HSF2 promoter for luciferase assays
HSF2mouseF	atctgctcacaagcagctca	Used in ChIP from mouse liver for HSF2
HSF2mouseR	cggatgcactacagacaaa	Used in ChIP from mouse liver for HSF2
BRI3_pr_1F	taagctagcAACCCGTGAAGCCTTGTATG	BRI3_pr_1F and BRI3_pr_2R were used for amplifying genomic DNA in and cloning BRI3 promoter into pGL3-basic vector for luciferase experiments
BRI3_pr_2R	cataagcttGCCGCTGCTCTCTAGACCT	
BRI3_pr_3F	taagctagcCTGGAGACAGAATACAGTAACGG	BRI3_pr_3F and BRI3_pr_4R were used for amplifying genomic DNA from ChIP experiment
BRI3_pr_4R	cataagcttGGTCATGACCTTTAGGGCTC	
BRI3mouseF	gctgcagaatctgtcaacca;	Used in ChIP from mouse liver for BRI3
BRI3RmouseR	gggtgtgtcctcgttcagt	Used in ChIP from mouse liver for BRI3

3.1.2. Cell culture experiments

Huh7, which is an established hepatocellular carcinoma cell line, were cultured in DMEM containing 1 gr/liter (Biochrom), containing 1 per cent non essential amino acid (Biochrom), 1 per cent Penicillin/Streptomycin solution (Biochrom) and 10 per cent FBS (Biochrom). 293T cells will be cultured in DMEM containing 4,5 gr/liter glucose (Biochrom), containing 1 per cent non essential amino acid (Biochrom), 1 per cent Penicillin/Streptomycin solution (Biochrom) and 10 per cent FBS (Biochrom). Cells will be transfected with the appropriate plasmids by using Lipofectamine 2000 for 293T cell lines (embryonic kidney cell line) with 2.5:1 Lipofectamine to DNA ratio. Huh7 cells will be transfected with Exgen 500 transfection reagent by using 6 equivalents.

#18 cells; which are early passage (passages 10-20 starting from establishment) glioblastoma cells, were cultured in MEM containing L-glutamine (Invitrogen), 10 per cent fetal bovine serum (Sigma) and penicillin/streptomycin (Sigma). U373-MG cells were cultured in DMEM containing 4.5 gr/liter D-glucose (Biochrom), containing 1 per cent non essential amino acid (Biochrom), 1 per cent Penicillin/Streptomycin solution (Biochrom) and 10 per cent FBS (Biochrom). Cells were transfected with the appropriate plasmids by using Turbofect for #18 cells with 4:2 Exgen (ul) to DNA (ug) ratio in 12 well plates. U373-MG cells were transfected with Exgen 500 transfection reagent by using 6 equivalents (2 ug DNA, 6.6 ug Exgen in 12 well plates).

Colony-formation assays were performed by seeding 10000 #18 cells into 6-well plates and growing them for 15 days in the presence of 800 ug/ml Geneticin (G418 solution, Sigma). Afterwards, cells were fixed with 1:3 diluted Giemsa solution (Sigma). After staining, colonies were counted by using Colony counting module of GelDoc bioimaging system (Biorad), by using constant sensitivity and averaging values for all samples, which eliminated colony counting artifacts. Sensitivity defines the strength of the signal coming from a colony (the cell density of a colony), while averaging defines the radius of a colony (the spreading property of the colony).

3.1.3. Plasmids

Full length open reading frame plasmids were purchased from RZPD German Resource Center for Genome Research (<http://www.rzpd.de>) as: C17orf45 in pCMVSPORTS6 (IRAT p970G1149D), HPCAL1 in pCMVSPORTS6 (IRAT p970A1118D), NCDN in pCMVSPORTS6 (IRAT p970A0738D), IQSEC1 in pCMVSPORTS6 (IRAT p970B0136D), SOX8 in pCMVSPORTS6 (IRAT p970B0541D), CPLX1 in pOTB-7 (IRAU p969H0610D) and RIMS3 in pOTB-7 (IRAU p969C0913D). *HPCAL1* and *SOX8* full length ORFs were directionally sub-cloned into pCDNA3 by using KpnI and NotI restriction enzymes. *IQSEC1* and *NCDN* full length ORFs were directionally sub-cloned into pCDNA3 by using EcoRI and NotI restriction enzyme sites. *CPLX1* and *RIMS3* full length ORFs were directionally sub-cloned into pCDNA3 by using EcoRI and XhoI restriction enzyme sites. Clones were verified by sequencing from 5' sites for all genes and from 3' sites for all genes except *IQSEC1* and *CPLX1* due to problematic sequencing reaction. NCDN cDNA was identified to be transcriptional isoform 3 (Entrez Refseq Id: NM_014284.2, gi: 62526027), which encodes the longer protein isoform; neurochondrin isoform 1.

3.1.4. SAGE and microarray comparison

When comparing SAGE and microarray data, we considered the consensus assignments of SAGE Genie best gene mappings[44] and SAGE Map[64] tag mapping algorithms. Unigene ID values of Affymetrix probe-sets were used for microarray probe mappings. Microarray data for a certain probe-set was used, when there is at least 1 present call beyond 6 samples (triplicates of 2 samples). Robust microarray analysis macro of GenePattern software was used when merging microarray data for 6 samples.

1883 tag mappings were available among 5190 tags, when we take the consensus tag mapping and exclude tags that match to multiple genes. For tags that match to same gene, we took average. 1726 tags remained. Of these 1726 genes, 1476 were present in microarray data.

When forming the union set of genes for up-regulated genes, we took the significantly up-regulated genes in SAGE (p value < 0,05) and excluded the ones that were in down-regulation tendency according to microarray (Huh7-S33Y < Huh7-vec), upon β -catenin over-expression. Afterwards, we chose genes from microarray data for which Huh7-S33Y/Huh7-vec ratio is bigger than 2 and exclude the ones that were down-regulated according to SAGE (Huh7-S33Y < Huh7-vec). Finally, we took the union of these two sets, to find the final union set of up-regulated genes. We performed the same set of rules with opposite thresholds while forming the union set of down-regulated genes.

3.1.5. Promoter Analysis and proMine

We used a one-sided Z test in order to assess the significance of the difference between the means of the control group and the test group. As the control group, we used the promoters of the genes which are not altered upon β -catenin over-expression according to microarray data (Huh7-S33Y/Huh7-vec: 1 ± 0.03). We extracted the promoters from Homo Sapiens Promoter Database of Cold Spring Harbor Laboratory. [37] We used the best promoter option while choosing the promoter type. We extracted 5000 bases up- resp. downstream of the transcriptional start site (TSS).

To automate this promoter analysis proMine was designed as an online tool coded in the PHP web programming language. The input for the tool is two promoter lists. The tool counts infinite number of transcription factor binding sites in two promoter lists separately, calculates means and variances and automatically produces a Z value for each TF binding sequence. This Z value tells the extent of differential binding of a certain TF onto the promoters in the two lists.

3.1.6. Luciferase assay and promoter plasmids

5'-upstream fragments of *HSF2* (Heat shock transcription factor 2, Accession numbers: CR597278, BC064622) and *BRI3* (Integral Membrane Protein 2C, ITM2C, BRICD2C, Cerebral protein 14, E25, E25C, hRPC.1050_D_4, hucep-14, ITM3, NPD018, PSEC0047, Accession numbers: CR598205, BC098563) genes were amplified from human genomic DNA (isolated from peripheral blood) using the primers listed in Table 3.1

and verified by sequencing. The PCR fragments were cloned into firefly luciferase reporter vector pGL3-basic (Promega) NheI and HindIII sites which were designed into the forward and the reverse primers, respectively. pRL-TK (Promega), a Renilla luciferase internal control vector, was used for normalization of firefly luciferase signals.

For co-transfection experiments, HEK293T cells or Huh7 cells were transfected with 100ng firefly luciferase reporter plasmid and 10ng of pRL-TK Renilla luciferase internal control plasmid and promoter activities were stimulated by co-transfecting pcDNA3 or pcDNA3-S33Y- β -catenin or pcDNA3-dnTCF4 or combinations of these plasmids, 350ng each. Lipofectamine 2000 reagent (Invitrogen) was used for the transfections and the manufacturer's protocol was followed. Cells were harvested 36hrs post-transfection. For Wnt3a-conditioned medium experiments, HEK293T cells were exposed to either L-cell or L-Wnt3a cell-derived conditioned medium for 36hrs post-transfection of the reporter plasmids[6]. Cells were then harvested into 100 μ L 1X Passive Lysis Buffer (Promega). Luciferase activities were measured using Dual Glo™ Luciferase Assay System (Promega), as per procedure provided by the manufacturer.

3.1.7. Lithium treatment

Huh7 cells were treated with lithium chloride or sodium chloride (25mM). Cells were harvested at indicated time points by adding 350 μ L of RLT lysis buffer of RNeasy® Mini Kit (Qiagen) and total RNA was isolated using the RNeasy® Mini Kit (Qiagen) according to the manufacturer's protocol. For western blot analysis, the above procedure was repeated, but the cells were harvested at indicated time points by adding 700 μ L of 1X Triton-HEPES Lysis Buffer [1% Triton X-100, 25mM HEPES, pH 7.4, 15mM MgSO₄, 4mM EGTA, 1mM DTT]. The lysates were sheared through a 23G syringe 5 times and analyzed by Bradford assay for total protein levels. Equal protein amounts (100 μ g) were resolved on 10% SDS-PAGE and analyzed by western blotting using monoclonal anti- β -catenin antibody 9E10 and anti- β -actin polyclonal antibody.

3.1.8. ChIP Assay

ChIP procedure for Huh7 and #18 glioblastoma cells was performed using the Epigentek ChIP kit (catalog # P-2002). 500,000 Huh7 or #18 cells were used for

immunoprecipitation. Sonication was performed to shear genomic DNA (Sonics & Materials, Inc. Model NO: VC 50T). 4 pulses of 12 seconds with an output power of 40 produced desired fragmentation for Huh7 cells, which was verified by DNA gel electrophoresis (data not shown). Other than sonication, manufacturer's instructions were followed. 3 μ g of mouse monoclonal anti- β -catenin antibody (Santa Cruz, catalogue number SC-7963) was used with 1 μ g of anti-RNA polymerase II and mouse normal IgG antibody from the kit as controls. PCR was performed from purified DNA fragments by using Fermentas Taq polymerase, starting with an initial denaturation at 94 °C for 5 minutes, followed by 48 cycles of 30 seconds at 94 °C, 1 minute at 55 °C, 1 minute at 72 °C and finally 5 minutes of 72 °C.

ChIP by using mouse liver tissues were performed as previously described [65]. Antibodies used for immunoprecipitation were β -catenin antibody (Invitrogen, mAb # 44207M) and caspase 8 antibody (Cell Signaling Technologies: Caspase-8 (1C12) Mouse mAb #9746). PCR was performed from purified DNA fragments by using Fermentas Taq polymerase, starting with an initial denaturation at 92 °C for 5 minutes, followed by 48 cycles of 30 seconds at 92 °C, 1 minute at 55 °C, 2 minute at 70 °C and finally 5 minutes of 72 °C.

3.2. Meta analysis of cancer gene expression profiles reveals cancer genes, SAGE tags and co-regulated regions on the genome

3.2.1. Data collection and processing

170 normal, 132 cancer SAGE libraries representing 32 different types of cancers. 477 normal and 927 cancer microarray samples were analyzed from 37 different types of cancers. In total, this corresponds to 49 different types of tumors comprising all 4 major types of human cancer, which are epithelial, hematological, central nervous system and connective tissue tumors.

3.2.1.1. Microarray data. 25/37 of microarray datasets were downloaded from Oncomine. Oncomine releases gene symbols as identifiers of genes. Because Entrez gene ID is a more stable and robust identifier for Entrez genes we used Entrez gene ids after converting

Oncomine released gene symbols. Some of the Entrez gene symbols in Oncomine do not map to a gene ID (ca 1000 genes out of 19000 as of 28 October 2008). To make most of the data, we tried to match the gene ID by iterating through older versions of seq_gene.md file of Entrez Dumps. Out of 948 non-mapping gene symbols, 340 remained after this effort (oncomine_symbols_togeneids.txt). We checked for any possible duplicate gene ids in this design, there were none.

12 of 37 were downloaded from Entrez GEO. For the sets downloaded from Entrez GEO, a preprocessing was performed as in Oncomine's routine: data was log2 transformed and median was set to 0 by subtracting the median from each data point and standard deviation was set to 1 by dividing each data point with standard deviation. Q value for multiple hypothesis testing was calculated as described[66]. The Entrez gene IDs were extracted in contrast to gene symbols in oncomine data. During preprocessing any probe-sets which maps to ESTs, duplicate-matching probes or non-matching probes were ignored.

If there are multiple probes matching to a single gene, we considered the average fold, p and q values of multiple probes that point to the same direction (over or under expression). If there are probes which point to both directions we considered that gene as both over and under-expressed in the COGENT procedure. These events are probably coming from different transcriptional isoforms of the same gene.

3.2.1.2. SAGE data. The Digital Gene Expression Display (DGED: http://cgap.nci.nih.gov/SAGE/SDGED_Wizard) tool was used to find differentially expressed SAGE tags between several different tumors and corresponding normal counterparts. DGED is one of the tools under the SAGE Genie[50] web platform founded to analyze CGAP data. The two parameters of output in DGED are the F and p values; which are statistical parameters to define stringency of differential expression when comparing SAGE tag expression values in different samples. (explained in Ref[64]) Values of 1,5 for F and 0.05 for p were used as cutoff points in these analyses. For the SAGE COGENT, long SAGE tags were converted to short SAGE tags and analyzed together with short SAGE libraries. For sage and microarray COGENT, short or long SAGE tags were converted to Entrez GENE IDs by using consensus tag mapping. Consensus tag mapping

considers two mainly used SAGE tag mapping algorithms: SAGE Map (<http://www.ncbi.nlm.nih.gov/SAGE/>) and SAGE Genie. SAGE Map maps sage tags to all GENBANK entries by using only the most 3' sage restriction enzyme site (mostly NlaIII). They use a scoring system which scores each tag to Unigene ID mapping which considers (1) number of sequence types, (2) reliable source score, and (3) number of sequences as detailed in (<ftp://ftp.ncbi.nlm.nih.gov/pub/sage/mappings/README.txt>). SAGE Genie, on the other hand, performs a different approach that divides GENBANK into subsets based on the representation percentage in a confident tag list that is derived from all SAGE libraries available. In this design, a high-ranking gene set (database) contains a higher percentage of sequences whose extracted SAGE tags map to the confident tag list. Both approaches are debatable. Different studies used SAGE Map, SAGE Genie's best gene calls (the gene match in the highest ranking dataset) or their own approach for SAGE tag mapping. When the mappings of SAGE Map and SAGE Genie are compared (Table 4.14A and Table 4.14B), it is clear that, there are significant discrepancies especially in short SAGE tag mappings. In order to minimize false mappings we defined a consensus tag mapping where we considered a tag mapping only if SAGE Map points to only one gene and this mapping is the same as SAGE Genie's best gene call. By using consensus tag mapping 68 per cent and 80 per cent of all possible Entrez GENE IDs are mappable by using short and long SAGE tags.

Solexa SAGE data produces significantly more significant differences (~10000 diff. expressed genes vs. ~500 diff. expressed genes) when compared to regular SAGE libraries due to very high tag counts. In order to correct for this, we calculated the average number of differential expression events in other comparisons (other microarray comparisons for SAGE & microarray COGENT and other SAGE comparisons for SAGE COGENT) and took the same number of differentially expressed genes for microarray & SAGE COGENT (not probes or tags) and the same number of differentially expressed tags for SAGE COGENT. We called this procedure "rank selection".

3.2.2. Being multi-cancer

SPSS software was used for the statistical tests, unless otherwise stated. Entrez geneRIF (Gene References Into Function) database was used to understand if a gene is tumor-associated or not. Entrez geneRIF database is a user-curated database in which authors record gene related summary of their articles. In the release we used (November 2008), ca 33 per cent of the human genes had at least one geneRIF.

When assigning ranks to genes, we used the lowest rank (top change) in the case of multiple probe-sets/tags. For ortholog or paralog number assignment we used the Entrez Homologene database which uses an automated protein sequence alignment based algorithm to assign homologs to genes.

3.2.2.1 RT-PCR and Q-RTPCR experiments. Brain tumor samples were collected at the Brain Surgery Department of Cerrahpaşa Hospital/İstanbul/Turkey with informed consents from patients under the ethical permit from Cerrahpaşa Hospital's ethical committee. Non-tumor brain tissues were collected from epilepsy surgeries. All tumor tissues were diagnosed and non-tumor tissues verified to be normal by a professional pathologist at Cerrahpaşa Hospital's Pathology Department. Tissues were fresh-frozen within 30 minutes after removal. RNAs from brain tissues were extracted by using Qiagen Rneasy Lipid Tissue Midi Kit. First strand cDNA synthesis was performed using Improm RT cDNA synthesis System. RT-PCRs were performed using Taq Polymerase (Fermentas). qRT-PCR reactions were carried out using faststart sybGreen kit (Roche, cat# 03003230001), and Roche Lightcycler 1.5. There is a high amount of ACTB or GAPDH up-regulation in tumors when compared to normals, as suggested by Oncoreveal (the expression analysis method developed as described in the Results section) and verified in the independent CNS tumor panel. We used S18 ribosomal RNA (Genbank accession: X03205) as normalizer. The sample/normal's average ratios varied between 0.67 and 1.28 in S18. This was a reasonable interval when compared to that of ACTB's which was 0.87-2.60. The primers used for RT-PCR or QRTPCR experiments are listed in Table 3.2.

Table 3.2. Primers used for RT-PCR and Q-RT-PCR in Part II. All primer names are designated as HUGO gene symbol followed by an underscore, than the internal primer number code and primer direction (F for forward and R for reverse)

ACTB_1F	AAGATCAAGATCATTGCTCCTC
ACTB_2R	GGGTGTAACGCAACTAAGTC
BRI3_1F	AGGGTCTACAACATCCACAG
BRI3_2R	TGTTCCCTTTAAGCGAAGGT
CABP1_1F	CTTCAGAGAATTCGACAAGGAC
CABP1_2R	CATTGGTGTCAAACCTCTCGG
CLU_1F	GAATTGGAGGCATGATGAAGAC
CLU_4R	CTGATTGGACATTTCTGAG
CPLX1_1F	CTTGATGCCGTACTTGTCTC
CPLX1_4R	GGAGTTTGTGATGAAGCAGG
GAPDH_1F	TTAGCACCCCTGGCCAAGG
GAPDH_2R	CTTACTCCTTGGAGGCCATG
HPCAL_1F	AACAATGACGGCAAACCTGTC
HPCAL_2R	GACAAGCCTGTGTTTCTCTG
HSRRN18S_1F	CTGAAACTTAAAGGAATTGACGGA
HSRRN18S_2R	GTTATCGGAATTAACCAGACAAATC
IQSEC_1F	GGCTGAAATTCCACATTGCT
IQSEC_2R	TGGAGAAGCACAGGATAGAG
LOC157627_1F	GTCTGGCTTAAATGAAAGGG
LOC157627_2R	GACTCCATTCTTTGCCTCTG
C17ORF45_1F	CAAATAAAGAGCTCGAGCTTCTG
C17ORF45_4R	GGAATATAACCTTCTCTTGGGC
NCDN_1F	CACATCTCTAATGAACACCCT
NCDN_2R	GGACTGTGATAGGAAGAGGA
RIMS3_1F	AGTTCAGCGATTTCTGAT
RIMS3_2R	CTTGATATAGGTGGCTGGGAG
CNN3_1F	GCTCATGCCTGTCACCTCTT
CNN3_2R	CTCGGCCGAAGTCAAGAA
CPLX1_2F	GTCTGGGTCCTTCTCCTCGT
CPLX1_3R	ATCGCCATGGAGTTTGTGAT
AXIN2_1F	TTATGCTTTGCACTACGTCCCTCCA
AXIN2_2R	CGCAACATGGTCAACCCTCAGAC
PHYHIP_3F	TCTACCGCAACCATCACAAGG
PHYHIP_4R	CTGGCCCGTGTTGAACTCC
SOX8_3F	CAAGGGCTACGACTGGAGTCT
SOX8_4R	CATGCGGCTTGGCTTTGAG
SNAP91_3F	AGCCACAACCTGTTACGTCTCC
SNAP91_4R	TCTCCCAAAGGTCTCCCAT

3.2.3. RIDGEs (Regions of increased co-regulation on the genome)

As briefly summarized in the results section, for defining tumor associated RIDGEs (TA-RIDGEs) we assigned a score to each gene by using an expression matrix from the difference of mean values of tumor and normal samples from the microarray studies. We did not consider SAGE data at this step, in order not to deal with possible biases if merging two different types of data distributions. This score is the average Spearman rank correlation coefficient (R_s) between the genes to be scored and each neighboring gene within ± 21 window size. Genes with less than 3 data points were ignored in calculations. To define normal associated RIDGEs (NA-RIDGEs) we used the processed expression values from a wide normal tissue expression study (GSE2361) from Entrez GEO.

To assign p value thresholds to this average score (in order to see which scores are significantly high), we calculated the average R_s score for a gene by picking random N genes from different chromosomes, where N represents two times window size. We repeated this procedure for window sizes 2-48. We then picked 21 as the optimum window size, because it maximizes the number of RIDGEs when 3 different statistical thresholds are used. Window size 21 was reasonable for both tumor minus normal expression matrix and for normal expression matrix.

In order to assign member genes to each RIDGE we used a double p value approach to catch the iceberg like structure of the RIDGE as explained in the results section.

When working on differential expression of genes in RIDGEs we first assigned a score of over-representation for each tumor-normal comparison. The pseudocode for this algorithm is presented at Figure 3.1.

Note that in the seq_gene file of Entrez Gene we used (28 October 2008), there were ca 32.000 gene assignments. But, most of them are not mappable by most of the microarray platforms or by SAGE tags. Therefore, we defined a new subset of detectable genes which comprised 20433 different Entrez Gene IDs as the genes which were

detectable either by microarray (present in the expression value matrix used) or by SAGE (all Entrez GENE Ids which were mapped by any of the short or long SAGE tags by using consensus tag mapping).

The pseudocode for calculating the significance of co-regulation between two tumor-normal comparisons is stated at Figure 3.2.

3.2.4. Implementation of data to MySQL database

COGENT-SAGE, COGENT-microarray&SAGE and RIDGE data were integrated in to a relational MySQL database, which holds data from several sources as follows: Tag to unigene mappings of both SAGE Map and SAGE Genie, Entrez Gene database dump for *Homo Sapiens* linked with gene ontology and Kegg pathway information and Unigene database. Finally, a user-friendly web interface was designed to filter this MySQL database, and to integrate user input with it. This online service is available at www.oncoreveal.org after a free registration step (previously named sagExplore resp. sagereveal and found at <http://aklab.boun.edu.tr:16080/sagereveal/ustframe.php> before microarray data was integrated).

4. RESULTS

4.1. Analysis of the Wnt/B-catenin/TCF4 pathway using SAGE, genome-wide microarray and promoter analysis: Identification of BRI3 and HSF2 as novel targets

4.1.1. TCF over-activation by stable over-expression of β -catenin leads to bigger and more aggressive tumors in nude mice

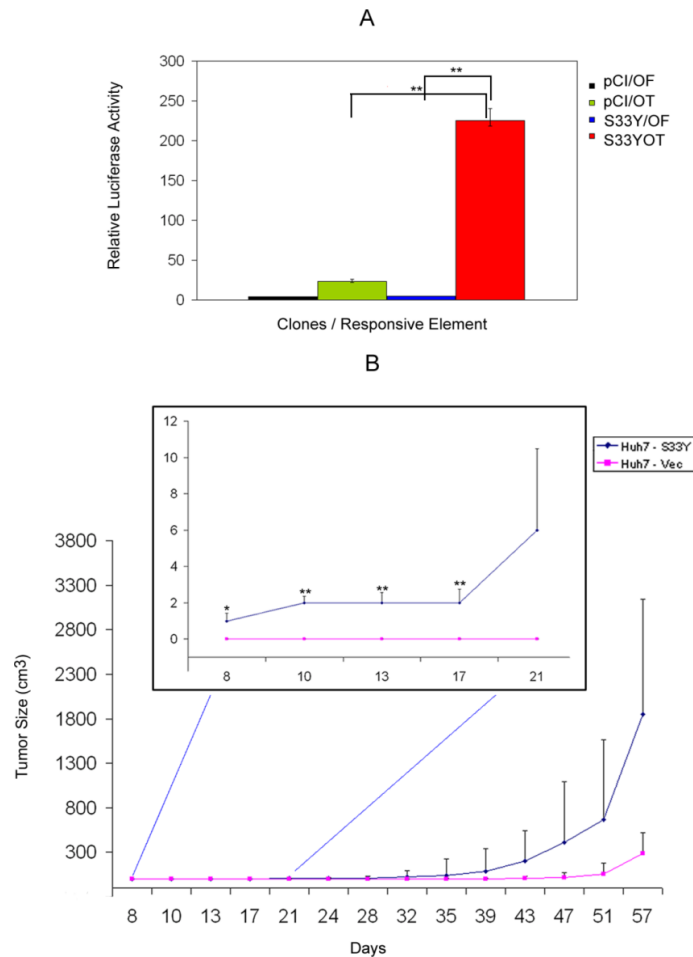


Figure 4.1. TCF over-activation by stable over-expression of β -catenin leads to larger and earlier onset tumors in nude mice. A: There is more TCF binding in hyperactive β -catenin expressing Huh7 cells (S33Y/OT), when compared to mutant TCF binding site containing controls (S33Y/OF) and empty vector controls (pCI/OF and pCI/OT). Error Bars: SD. B: Huh7-pCI-S33Y xenografts (Huh7 – S33Y) appear earlier and grow faster than Huh7- pCI – Neo xenografts (Huh7 – Vec). Error Bars: SE. For (A) and (B) ‘*’ indicates $p < 0.05$; ‘**’ indicates $p < 0.01$.

Higher activation of Wnt/ β -catenin/TCF signaling in Huh7 cells was obtained by expression of a β -catenin mutant (S33Y) that is hyperactive as it avoids degradation and accumulates sufficiently to enter the nucleus. Its activity was tested using a TCF-luciferase

reporter assay. Hyperactive β -catenin-transfected cells (Huh7-S33Y) displayed more than 10-fold higher TCF activity, when compared to control cells transfected with vector (Huh7-vec) (Figure 4.1A). Upon bilateral subcutaneous injection into immunodeficient *nude* mice, the (polyclonal) Huh7-S33Y cell line initiated tumors earlier and led to larger tumors when compared to control (Figure 4.1B). At the 13th day of tumor size measurement, Huh7-S33Y formed tumors in 80 per cent of the mice, whereas the first measurable tumors in controls appeared in only 2 of 10 mice after 28 days. After 60 days, the median tumor size was 6.4 fold more in Huh7-S33Y initiated tumors when compared to Huh7-vec initiated tumors.

4.1.2. SAGE and genome-wide microarray analysis of differential expression between high and low TCF activity Huh7 cells

In order to identify new members of the Wnt pathway and study differences between Huh7-S33Y and Huh7-vec cells possibly affecting tumorigenic capacity, we performed global gene expression analysis using both SAGE and genome-wide microarray analysis on the same RNA isolations.

In the SAGE experiment, 42,126 and 41,898 tags from hyperactive β -catenin transfected (Huh7-S33Y) and empty vector transfected control were counted respectively. 431 tags were differentially expressed (p value < 0.05) which unambiguously correspond to 126 down regulated and 33 up-regulated genes.

In the microarray experiments, 96 genes (126 probes) were up-regulated and 121 genes (152 probes) were down-regulated more than 2 fold (p value < 0.05) when median intensity values were considered among triplicate microarray analysis, each from a separate cDNA synthesis. We also performed real-time (qRT-PCR) and RT-PCR for some of the differentially expressed genes and showed that there is an acceptable level of concordance between SAGE, microarray and RT-PCR based methods (Figure 4.3 and Table 4.1).

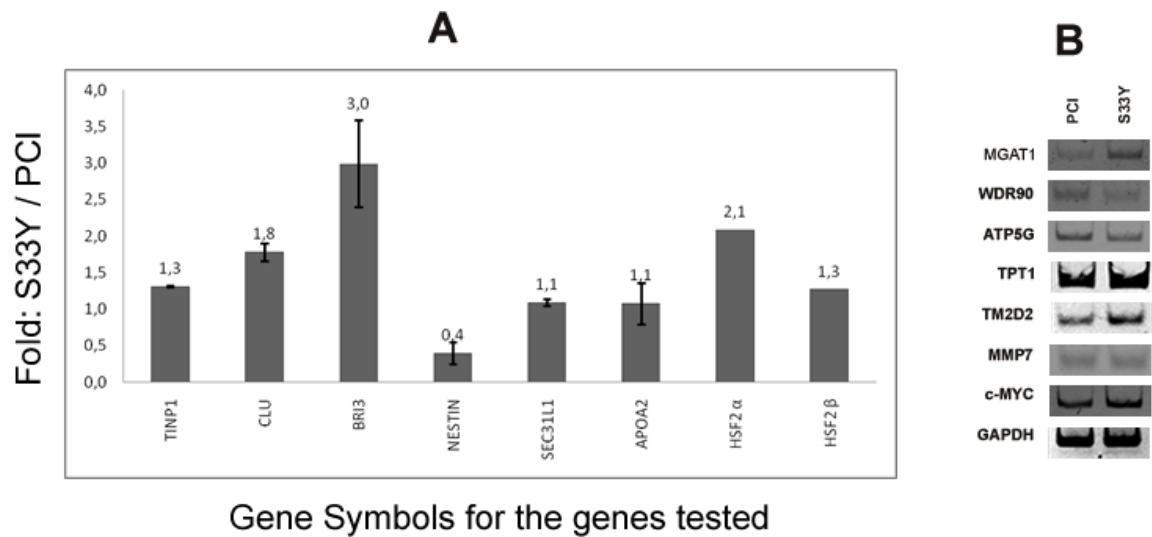


Figure 4.2. Verification and expansion of SAGE and Microarray results by using Q-RT-PCR and RT-PCR. A: Quantitative Real Time PCR data, results were normalized using the geometric average of TF and TPI expression levels because their expression level is very similar in Huh7-vec and Huh7-S33Y samples according to both SAGE and microarray data. *HSF2- α* and *HSF2- β* levels were normalized by geometric mean of *s18* and *GAPDH* expression. Error bars represent SD. Lack of error bars means SD < 0.001 B: RT-PCR data for several example genes.

We compared the results of SAGE and microarray data, by using a tendency analysis that compares the common detection of up-regulation, down-regulation or no-change tendencies. For tendency analysis, there were 1476 genes that were comparable between microarray and SAGE data (see Materials and Methods). With the default set of thresholds, the agreement in up or down regulation detection is very low (8 per cent and 16 per cent respectively) but the inverse tendency ratio is lower (3.93 per cent) (Table 4.2). For 52 per cent of the comparable genes, methods point to same tendency (either up-regulation, down-regulation or no-change tendency). For the remaining 43 per cent of the genes, one method points to negative tendency and the other method points to no-change tendency, or vice-versa. In other words, for approximately 95 per cent of the genes the methods give non-inverse results. We did the same analysis by applying different thresholds for up and down-regulation (Table 4.3). When 1st, 2nd and 3rd analysis in Table 4.2 are compared, it can be seen that no-change tendency is increasing and inverse tendency is decreasing in more stringent conditions (higher fold threshold for defining up

or down regulation). In summary, according to our data, SAGE and microarray do not give contradictory results but variably reveal differential expression for different genes, and it is best to perform them in parallel.

Table 4.1. Comparison of SAGE and Microarray results by using Q-RT-PCR and RT-PCR

Gene	Microarray Fold*	SAGE <i>p value</i>	SAGE Fold**	Q-RTPCR or RTPCR result***
TINP1	0.97	0.46	-1.02	1.31
CLU	1.29	0.00	1.97	1.78
BRI3	1.01	0.01	2.36	2.99
NES	0.81	0.00	-3.38	0.40
SEC31L1	0.89	0.00	-5.33	1.09
APOA2	0.78	0.00	-2.45	1.07
MGAT1	1.17	0.04	3.00	Up
KIAA1924 (WDR90)	1.01	0.02	-6.00	Down
ATP5G1	0.78	0.00	-5.00	Down
TPT1	1.10	0.00	1.73	Up
BLP1 (tm2d2)	0.95	0.01	6.00	Up
MMP7	ND****	ND	ND	No Change
MYC	1.13	ND	ND	Up

* Microarray Fold: Ratio of Median Huh7-S33Y / Median Huh7-vec

** SAGE Fold: Ratio of Huh7-S33Y/Huh7-vec. When Huh7-vec is 0. it is considered as 1

*** Numbers represent Q-RTPCR results. Texts represent RTPCR results.

****ND: non-detected

Table 4.2. SAGE and microarray comparison by using tendency analysis

	Up-regulation tendency*	Down-regulation Tendency**	No-change tendency***	Total	Inverse Tendency****
Microarray	219	302	955	1476	NA
SAGE	89	377	1010	1476	NA
SAGE and Microarray	23	95	665	783	58
per cent common*****	8.07	16.27	51.15	53.05	3.93

* Up-regulation tendency is defined as $p < 0.2$ and Huh7-S33Y/Huh7-vec > 1 for SAGE; Huh7-S33Y/Huh7-vec > 1.15 for microarray

** Down-regulation tendency is defined as $p < 0.2$ and Huh7-S33Y/Huh7-vec < 1 for SAGE; Huh7-S33Y/Huh7-vec < 0.87 for microarray

*** No change tendency is defined as not being in positive or negative tendency

**** Up-regulated according to SAGE and down-regulated according to microarray or vice-versa

*****The ratio of intersection over total number of genes in consideration : (genes commonly changed in microarray and SAGE) / [(genes analyzed in microarray) + (genes analyzed in SAGE) – (genes commonly changed in microarray and SAGE)]

Table 4.3. Percent common with different thresholds

Analysis #	SAGE <i>p</i> value threshold	Microarray fold threshold	Common up-regulation tendency	Common down-regulation tendency	Common no-change tendency	Inverse tendency
1	0.20	1.00	8.86	24.92	0.10	11.38
2	0.20	1.15	8.07	16.27	51.15	3.93
3	0.20	1.50	0.94	5.08	67.19	0.07
4	0.05	1.15	2.51	7.38	62.04	1.02
5	0.40	1.15	11.75	21.49	24.59	9.01

Table 4.4. Tendency analysis in high abundance class (tag count ≥ 15)

	Positive tendency	Negative Tendency	No-change tendency	Total	Inverse Tendency
Microarray	9	18	40	67	NA
SAGE	13	42	12	67	
SAGE and Microarray	4	13	8	25	6
per cent (common / SAGE count)	22.22	27.66	18.18	37.31	8.96

Table 4.5. Tendency analysis in middle abundance class ($3 < \text{Tag count} < 15$)

	Positive tendency	Negative Tendency	No-change tendency	Total	Inverse Tendency
Microarray	80	95	312	487	NA
SAGE	17	230	240	487	
SAGE and Microarray	3	59	164	226	37
per cent (common / SAGE count)	3.19	22.18	42.27	46.41	7.60

Table 4.6. Tendency analysis in low abundance class (Tag count ≤ 3)

	Positive tendency	Negative Tendency	No-change tendency	Total	Inverse Tendency
Microarray	130	189	603	922	NA
SAGE	62	117	743	922	
SAGE and Microarray	16	26	484	526	18
per cent (common / SAGE count)	9.09	9.29	56.15	57.05	1.95

We also did tendency analysis in different abundance classes of genes. The agreement ratios did not change very much, when 3 different sets of abundance subsets were compared. However, the commonality percentage is mildly decreased in the cases of low abundance and middle abundance classes when compared to the high abundance class. (Tables 4.4, 4.5 and 4.6)

Microarray is more sensitive than 40000 counts of SAGE tags, as SAGE detected 7,8 fold less genes than microarray, particularly due to the tag mapping problems. For SAGE 1727 tags unambiguously map to one gene, for microarray there are 13497 genes detected. However, in terms of sensing differential expression, the two methods perform more similarly (approximately same ratio in down-regulated genes while microarray sensed 3 fold more up-regulated genes).

Within the differentially expressed genes, several genes which seem to be novel in the context of the Wnt pathway were identified, in addition to the already known Wnt targets in liver. For instance, Glutamate-ammonia ligase (glutamine synthetase) (*GLUL*), a known target gene, is 2 fold up-regulated in Huh7-S33Y compared to Huh7-vec in accordance with previous observations [67]. On the other hand, epidermal growth factor expression did not change (Huh7-S33Y/Huh7-vec = $0,97 \pm 0,17$; from 10 different probes) although proposed as a novel liver Wnt target [24]. In agreement with earlier work, Wnt/ β -catenin targets in Huh7 cells appear to be largely different from targets in colon carcinoma, in spite of some similarities such as *c-Myc* (Figure 4.2).

As Huh7 cells have been shown to have a side population which is more tumorigenic and more stem cell-like [68, 69], we also rescued the tumor-forming cells and reanalyzed their transcriptome profile to check if any substantial changes could be identified related to in vivo subpopulation selection and compared our results to those of Haraguchi et al [69]. Intersection analysis of the genes up-regulated in Huh7 stem cells and post-tumor cells revealed some genes commonly over-expressed in both cases (Tables 4.7 and 4.8). When functional ontology distributions of these genes were analyzed, it was seen that, 8 out of 14 were plasma membrane proteins, 7 out of 14 had roles in cell communication and 4 had transmembrane transporter activity. All of these properties might be related to cancer stem cell characteristics such as invasion, metastasis and high drug export ability.

Table 4.7. Comparison of microarray data from S33Y expressing and control Huh7 cells (Huh7 – S33Y and Huh7 – vec) and their post-tumor counterparts which were rescued cells from corresponding xenografts (PT-Huh7-S33Y and PT-Huh7-vec) with Huh7 side population

Comparison	Number of genes common to Huh7-SP *	Number of genes common with 141 over-expressed or 90 under-expressed genes in Huh7 Side population from (Haraguchi et al. 2006) *
(PT-Huh7-S33Y / Huh7- S33Y) > 2	508	19
PT-Huh7- vec / Huh7- vec > 2	428	22
(PT-Huh7-S33Y / Huh7- S33Y) > 2 & (PT-Huh7- vec / Huh7- vec) > 2	202	14**
Huh7- S33Y / Huh7- vec > 2	87	1
PT-Huh7-S33Y / PT-Huh7- vec > 2	221	5
(PT-Huh7-S33Y / Huh7- S33Y) < 0,5	866	8
PT-Huh7- vec / Huh7- vec < 0,5	528	9
(PT-Huh7-S33Y / Huh7- S33Y) < 0,5 & (PT-Huh7- vec / Huh7- vec) < 0,5	327	6
Huh7- S33Y / Huh7- vec < 0,5	120	6
PT-Huh7-S33Y / PT-Huh7- vec < 0,5	371	0

* increased or decreased genes: first 5 rows are compared with Huh7-SP increased genes, and former 5 rows are compared with Huh7-SP decreased genes.

** Please see 4.8 for a list of these 14 genes.

4.1.3. *In silico* promoter analysis

To increase the stringency in our search for Wnt pathway-related genes within our microarray and SAGE data, promoter regions of the detected genes were analyzed in terms of TCF binding element (TBE) content similar to a previous approach by Schwartz *et al* [13] (Table 4.9). In summary, TBEs were counted in promoters of genes that are over or under-expressed upon stable β -catenin over-expression and compared to promoters in a no-change gene pool. We expanded the previous approach by performing a two-sided Z test and considering two different TCF binding motifs, short TCF binding motif (TCF_{short}: ‘WWCAAWG’) and long TCF binding motif (TCF_{long}: ‘RWWCAAWGG’).

Table 4.8. Genes which are up-regulated in both Huh7 side population cells and post-tumor cells ((PT-Huh7-S33Y / Huh7- S33Y) > 2 & (PT-Huh7-PCI / Huh7- PCI) > 2).

Gene Name	Gene Symbol
ATP-BINDING CASSETTE, SUB-FAMILY G (WHITE), MEMBER 2	ABCG2
ANKYRIN 3, NODE OF RANVIER (ANKYRIN G)	ANK3
ANNEXIN A3	ANXA3
CARBONIC ANHYDRASE IV	CA4
GLUTATHIONE S-TRANSFERASE A1	GSTA1
JAGGED 1 (ALAGILLE SYNDROME)	JAG1
LACTASE	LCT
PLATELET/ENDOTHELIAL CELL ADHESION MOLECULE (CD31 ANTIGEN)	PECAM1
PLACENTA-SPECIFIC 8	PLAC8
S100 CALCIUM BINDING PROTEIN A10 (ANNEXIN II LIGAND, CALPACTIN I, LIGHT POLYPEPTIDE (P11))	S100A10
SH3-DOMAIN BINDING PROTEIN 5 (BTK-ASSOCIATED)	SH3BP5
SOLUTE CARRIER FAMILY 1 (NEURONAL/EPITHELIAL HIGH AFFINITY GLUTAMATE TRANSPORTER, SYSTEM XAG), MEMBER 1	SLC1A1
SOLUTE CARRIER FAMILY 3 (CYSTINE, DIBASIC AND NEUTRAL AMINO ACID TRANSPORTERS, ACTIVATOR OF CYSTINE, DIBASIC AND NEUTRAL AMINO ACID TRANSPORT), MEMBER 1	SLC3A1
SOLUTE CARRIER FAMILY 7 (CATIONIC AMINO ACID TRANSPORTER, Y+ SYSTEM), MEMBER 7	SLC7A7

In order to evaluate the biological significance of this statistical analysis, we first used the promoters of already known targets [6]. We saw that there are more TBEs in the ‘-1500 -1000’ region (p value = 0.12), in known target’s promoters, regardless of the TCF binding motif we used (Table 4.9).

Table 4.9A. Number of TCF binding sites in the promoters of previously verified Wnt targets (short motif - TCF_{short} - 'WWCAAWG')

	Verified Wnt targets. TCF _{short}								
	Control; n = 203			Up-regulated in S33Y n = 37			Two Sided Z test		
Gene Region	T	M	V	T	M	V	SE	Z value	p value
'-5000 0'	1241	6.11	8.62	222	6.00	10.38	0.48	-0.23	0.42
'-1000 0'	177	0.87	1.06	37	1.00	1.30	0.17	0.76	0.24
'-5000 -4000'	266	1.31	1.46	38	1.03	1.27	0.20	-1.43	0.08
'-4000 -3000'	245	1.21	1.31	47	1.27	1.17	0.19	0.34	0.38
'-3000 -2000'	270	1.33	1.61	47	1.27	1.60	0.21	-0.29	0.42
'-2000 -1500'	151	0.74	0.87	24	0.65	0.61	0.15	-0.62	0.27
'-1500 -1000'	122	0.60	0.61	28	0.76	0.62	0.13	1.21	0.12
'-1000 -500'	94	0.46	0.48	20	0.54	0.52	0.11	0.68	0.27
'-500 0'	82	0.40	0.41	17	0.46	0.52	0.11	0.53	0.31
'-500 500'	147	0.72	0.99	27	0.73	0.85	0.16	0.03	0.50
'-1000 1000'	350	1.72	2.78	64	1.73	3.28	0.27	0.02	0.50
'0 500'	63	0.31	0.37	9	0.24	0.35	0.10	-0.67	0.27
'0 1000'	171	0.84	1.19	26	0.70	1.51	0.18	-0.78	0.24
'0 5000'	1241	6.11	11.32	184	4.97	10.30	0.55	-2.06	0.02
'-5000 5000'	2484	12.24	27.27	407	11.00	25.46	0.86	-1.44	0.08

Table 4.9B. Number of TCF binding sites in the promoters of previously verified Wnt targets (long motif - TCF_{long} - 'RWCAAWGG')

	Verified Wnt targets. TCF _{long}								
	Control; n = 203			Up-regulated in S33Y n = 37			Two Sided Z test		
Gene Region	T	M	V	T	M	V	SE	Z value	P value
'-5000 0'	146	0.72	0.71	23	0.62	0.88	0.14	-0.70	0.24
'-1000 0'	22	0.11	0.11	5	0.14	0.12	0.05	0.50	0.31
'-5000 -4000'	34	0.17	0.18	2	0.05	0.05	0.07	-1.63	0.05
'-4000 -3000'	32	0.16	0.15	7	0.19	0.21	0.06	0.49	0.34
'-3000 -2000'	29	0.14	0.14	4	0.11	0.10	0.06	-0.56	0.31
'-2000 -1500'	16	0.08	0.07	1	0.03	0.03	0.04	-1.17	0.14
'-1500 -1000'	12	0.06	0.06	4	0.11	0.10	0.04	1.26	0.12
'-1000 -500'	12	0.06	0.06	3	0.08	0.07	0.04	0.57	0.31
'-500 0'	10	0.05	0.05	2	0.05	0.05	0.04	0.13	0.46
'-500 500'	19	0.09	0.08	2	0.05	0.05	0.05	-0.83	0.21
'-1000 1000'	44	0.22	0.21	9	0.24	0.24	0.08	0.35	0.38
'0 500'	9	0.04	0.04	0	0.00	0.00	0.03	-1.31	0.10
'0 1000'	22	0.11	0.10	4	0.11	0.10	0.05	-0.01	0.50
'0 5000'	153	0.75	0.75	19	0.51	0.57	0.14	-1.69	0.05
'-5000 5000'	299	1.47	1.46	42	1.14	1.41	0.20	-1.70	0.04

We then defined a union set of genes by merging microarray and SAGE data on which we performed the analysis (See methods). There are different numbers and

distribution patterns of TCF binding elements (TBEs) in promoters of the genes that are affected upon hyperactive β -catenin over-expression in Huh7 cells; when compared to a no-change pool. We observed the most significant differences in ‘-1000 to Transcription Start Site (TSS)’ and ‘-1500 to -1000’ regions (Table 4.10A and 4.10B). For up-regulated targets, abundance of TBEs in promoter regions has been shown before [13]. However, our analysis hints at a particular pattern for down-regulated targets where there are no TBEs in ‘TSS to +1000’ and there are more TBEs between ‘-1500 to -1000’ (Figure 4.2). Supporting this general statistical analysis (Table 4.10B and Figure 4.3), there is no TCF_{long} in ‘TSS to +1000’ in any of 15 down-regulated genes with TCF_{long} in ‘-1500 to -1000’.

Table 4.10A. Number of TCF binding elements (TBEs) in the promoters of up-regulated genes in Huh7-S33Y in comparison to Huh7-vec when the longer motif - TCF_{long} - ‘RWWCAAWGG’ is considered

	Up-regulated in S33Y. SAGE and microarray union dataset; TCF _{long}								
	Control; n = 203			Up-regulated in S33Y n = 87			Two Sided Z test		
Gene Region	T*	M*	V*	T	M	V	SE*	Z value*	p value
'-5000 0'	146	0.72	0.71	78	0.90	0.87	0.09	1.96	0.03
'-1000 0'	22	0.11	0.11	18	0.21	0.21	0.04	2.82	0.00
'-5000 - 4000'	34	0.17	0.18	21	0.24	0.28	0.05	1.63	0.05
'-4000 - 3000'	32	0.16	0.15	20	0.23	0.20	0.04	1.73	0.04
'-3000 - 2000'	29	0.14	0.14	9	0.10	0.12	0.04	-0.97	0.18
'-2000 - 1500'	16	0.08	0.07	6	0.07	0.09	0.03	-0.34	0.38
'-1500 - 1000'	12	0.06	0.06	4	0.05	0.04	0.03	-0.52	0.31
'-1000 -500'	12	0.06	0.06	5	0.06	0.05	0.03	-0.06	0.50
'-500 0'	10	0.05	0.05	13	0.15	0.17	0.02	4.32	0.00
'-500 500'	19	0.09	0.08	20	0.23	0.22	0.03	4.36	0.00
'-1000 1000'	44	0.22	0.21	32	0.37	0.35	0.05	3.08	0.00
'0 500'	9	0.04	0.04	7	0.08	0.07	0.02	1.64	0.05
'0 1000'	22	0.11	0.10	14	0.16	0.14	0.03	1.58	0.07
'0 5000'	153	0.75	0.75	59	0.68	0.68	0.09	-0.82	0.21
'-5000 5000'	299	1.47	1.46	137	1.57	1.53	0.13	0.79	0.24

Table 4.10B. Number of TCF binding elements (TBEs) in the promoters of down-regulated genes in Huh7-S33Y in comparison to Huh7-vec when the longer motif - TCF_{long} - 'RWWCAAWGG' is considered

	Down-regulated in S33Y. SAGE and microarray union dataset. TCF _{long}								
	Control; n = 203			Down-regulated in S33Y n = 160			Two Sided Z test		
Gene Region	T	M	V	T	M	V	SE	Z value	p value
'-5000 0'	146	0.72	0.71	113	0.71	0.68	0.07	-0.19	0.46
'-1000 0'	22	0.11	0.11	14	0.09	0.08	0.03	-0.81	0.21
'-5000 -4000'	34	0.17	0.18	20	0.13	0.13	0.03	-1.27	0.12
'-4000 -3000'	32	0.16	0.15	24	0.15	0.13	0.03	-0.25	0.42
'-3000 -2000'	29	0.14	0.14	21	0.13	0.13	0.03	-0.39	0.38
'-2000 -1500'	16	0.08	0.07	14	0.09	0.10	0.02	0.41	0.34
'-1500 1000'	12	0.06	0.06	18	0.11	0.12	0.02	2.86	0.00
'-1000 -500'	12	0.06	0.06	9	0.06	0.05	0.02	-0.15	0.46
'-500 0'	10	0.05	0.05	4	0.03	0.02	0.02	-1.42	0.08
'-500 500'	19	0.09	0.08	9	0.06	0.07	0.02	-1.62	0.05
'-1000 1000'	44	0.22	0.21	23	0.14	0.15	0.04	-2.02	0.02
'0 500'	9	0.04	0.04	5	0.03	0.03	0.02	-0.80	0.21
'0 1000'	22	0.11	0.10	9	0.06	0.05	0.03	-2.12	0.02
'0 5000'	153	0.75	0.75	107	0.67	0.70	0.07	-1.24	0.12
'-5000 5000'	299	1.47	1.46	220	1.38	1.56	0.10	-1.02	0.16

*T = Total number of motifs. M = average number of motifs. V = variance of distribution of motif counts. SE = Standard Error of mean. Positive Z value indicates more, negative Z value indicates less TBE in test group.

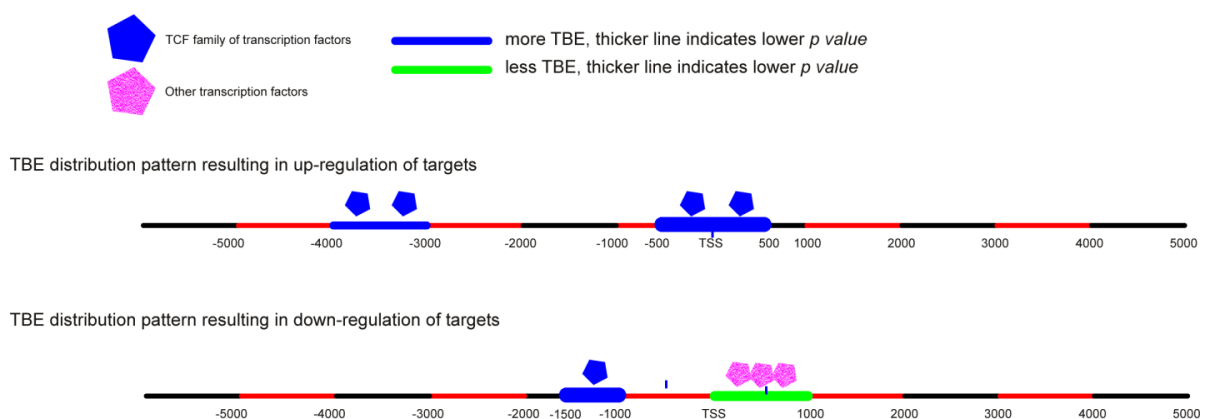


Figure 4.3. A particular pattern for down-regulated Wnt targets where there are no TBEs in 'TSS to +1000' and there are more TBEs between '-1500 to -1000'

4.1.4. Novel Wnt/ β -catenin/TCF targets; BRI3 and HSF2

After making a combined study of SAGE, microarray, real time, RT-PCR and *in silico* promoter data; some candidate genes were analyzed in further detail. Among these genes, we showed that BRI3 (Entrez Gene ID: 25798) and HSF2 (Entrez Gene ID: 3298) are novel targets of Wnt/ β -catenin/TCF signaling. BRI3 is 3.0 fold up-regulated in Huh7-S33Y when compared to Huh7-vec (*p* value = 0.002). HSF2- α and HSF2- β is up-regulated 2.1 (*p* value < 0.0001) and 1.3 (*p* value < 0.0001) fold in Huh7-S33Y when compared to Huh7-vec, respectively (Figure 4.2). We also checked BRI3 and HSF2 mRNA levels in primary cell cultures rescued from the tumors that developed in Huh7-S33Y and Huh7-vec xenografts (post-tumor cell lines PT-Huh7-S33Y and PT-Huh7-vec respectively). BRI3 mRNA increased 2.53 fold in PT-Huh7-S33Y when compared to PT-Huh7-vec (*p* value < 0.0001). HSF2- α and HSF2- β were up-regulated 1.7 (*p* value = 0.07) and 2.27 (*p* value = 0.002) fold respectively.

When a comprehensive liver cancer screen (104 liver tumor tissues versus 76 non-tumor liver tissues) with genome-wide microarray analysis by Chen *et al* was reanalyzed, *HSF2* expression was found to be slightly (1.34 fold) but significantly (*p*= 0.0003) up-regulated in liver cancers when compared to normal liver tissues whereas expression of *BRI3* was not detected with that particular array design [70, 71].

When we analyzed the -5000 to +5000 regions of the transcriptional start sites (TSS) of BRI3 and HSF2, we found that BRI3 has 5 TCF_{short} and 2 TCF_{long} motifs between -3225th and 1122nd positions relative to TSS. HSF2 has 17 TCF_{short} and 5 TCF_{long} motifs between -4504th and 4695th positions relative to TSS (Figure 4.5A and 4.7A for BRI3 and HSF2, respectively)

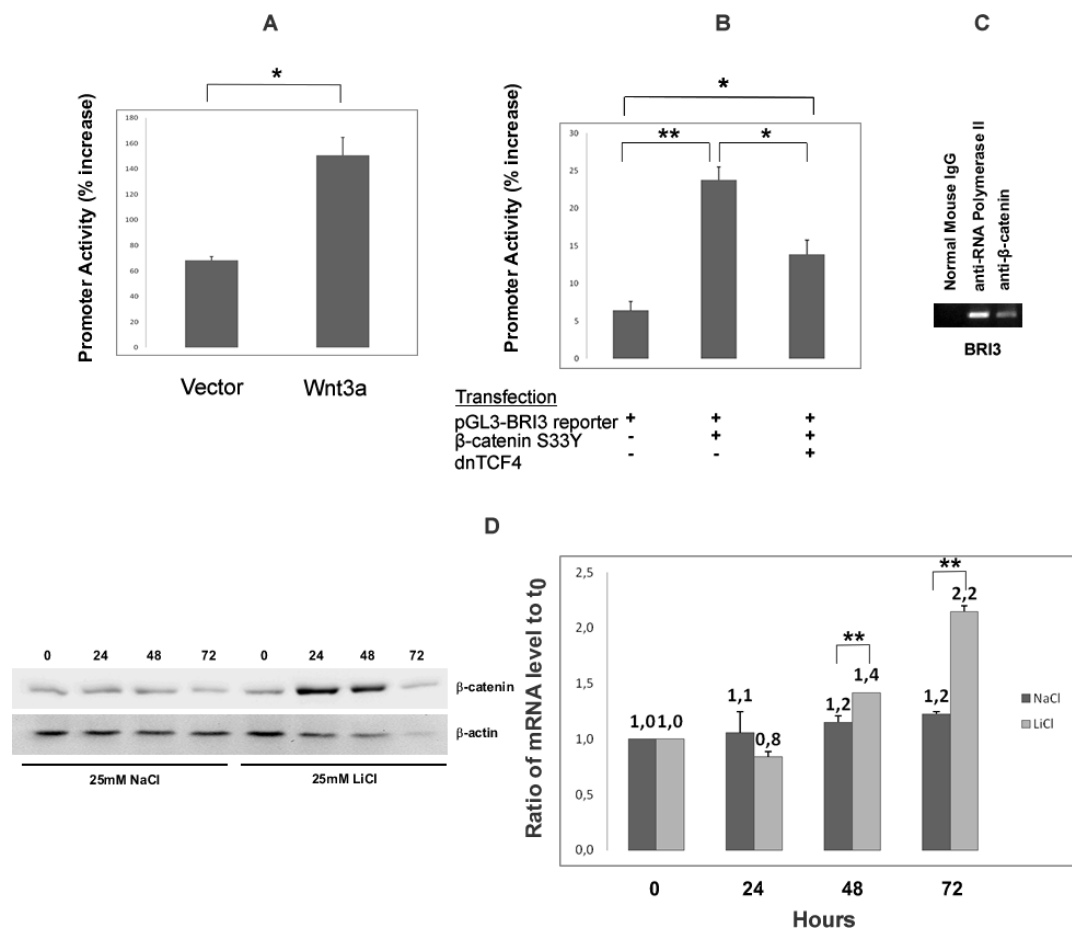


Figure 4.4. BRI3 is a novel target of Wnt/ β -catenin/TCF signaling - I. A: Luciferase reporter assays: Autocrine stimulation of Huh7 cells by Wnt3a lead to up-regulation of BRI3 promoter (Wnt3a) when compared to empty vector control (Vector). Both vector and Wnt3a plasmids were co-transfected with BRI3 promoter reporter vector (pGL3-BRI3) B: Luciferase reporter assays: BRI3 promoter activation by hyperactive β -catenin is inhibited by addition of dominant negative TCF4 (dnTCF4) in Huh7 cells. C: ChIP analysis: β -catenin binds to BRI3 promoter in Huh7 cells revealed by ChIP analysis performed with anti- β -catenin antibody, anti-RNA polymerase antibody or normal mouse IgG. The band indicates genomic DNA product of BRI3 promoter. D: Lithium treatment: BRI3 mRNA level is increased upon increased β -catenin levels (verified by Western Blot in the left panel) due to lithium treatment of Huh7 cells. RNA levels were normalized by using β -actin (ACTB) RNA levels. Error Bars: SD, (* p value < 0.05; ** p value < 0.01). Each graph is representative of at least two independent experiments.

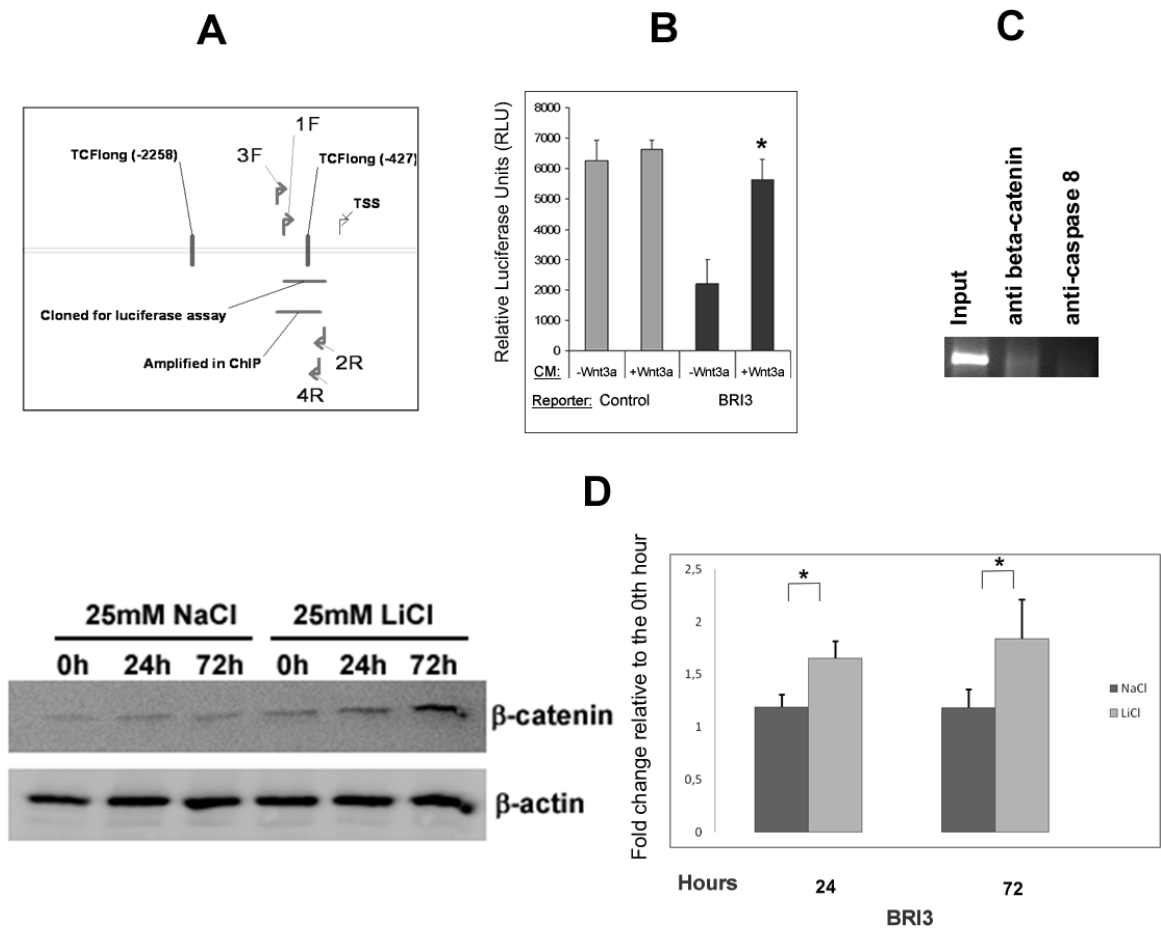


Figure 4.5. BRI3 is a novel target of Wnt/ β -catenin/TCF signaling - II A: Scheme of BRI3 promoter region. B: Luciferase reporter assays: BRI3 promoter is 2.48 fold more active in 293T cells which are exposed to conditioned medium from Wnt3a secreting cells when compared to non-secreting exposed controls (BRI3). This difference is significant, when compared to the non-significant variation in pGL3-Basic (Control). C: ChIP analysis: β -catenin binds to BRI3 promoter in mouse liver revealed by ChIP analysis performed with β -catenin antibody (Invitrogen, catalog # 44207M) or anti-caspase 8 antibody (Cell Signaling Technologies: Caspase-8 (1C12) Mouse mAb #9746). The band indicates genomic DNA product of BRI3 promoter. D: Lithium treatment: BRI3 mRNA level is increased upon increased β -catenin levels (verified by Western Blot in the left panel) due to lithium treatment of U373-MG cells. RNA levels were normalized by using β -actin (ACTB) RNA levels. Error Bars: SD, (* p value < 0.05). Each graph is representative of at least two independent experiments.

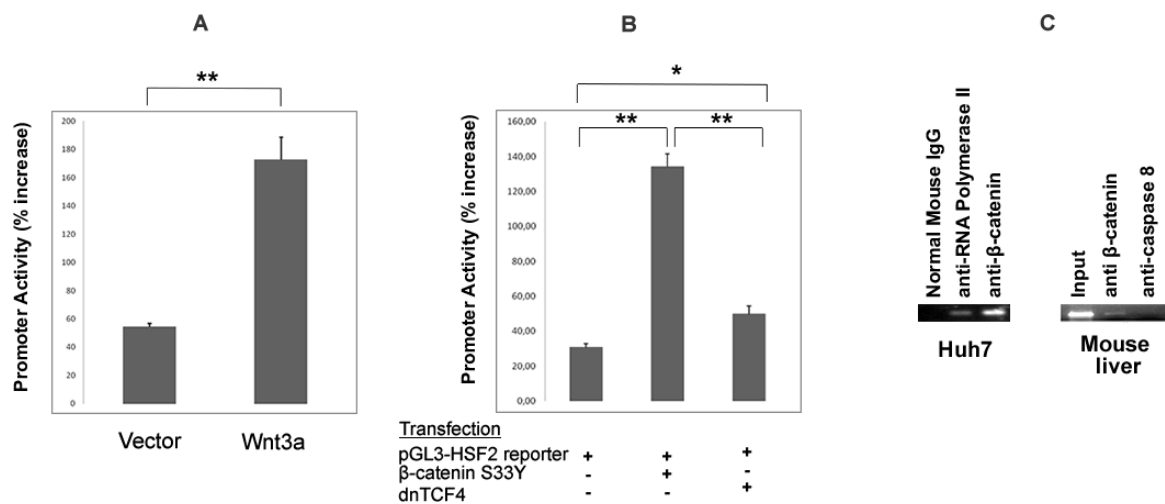


Figure 4.6. *HSF2* is a novel target of Wnt/ β -catenin/TCF signaling - I. A: Luciferase reporter assays: Autocrine stimulation of Huh7 cells by Wnt3a lead to up-regulation of *HSF2* promoter (Wnt3a) when compared to empty vector control (Vector). Both vector and Wnt3a plasmids were co-transfected with *HSF2* promoter reporter vector (pGL3-*HSF2*) B: Luciferase reporter assays: *HSF2* promoter is activated by hyperactive β -catenin in Huh7 cells which is inhibited by addition of dominant negative *TCF4* (dnTCF4). C: Left: ChIP analysis: β -catenin binds to *HSF2* promoter in Huh7 cells revealed by ChIP analysis performed with anti- β -catenin antibody, anti-RNA polymerase antibody or normal mouse IgG. The band indicates genomic DNA product of *HSF2* promoter. Right: ChIP analysis: β -catenin binds to *HSF2* promoter in mouse liver revealed by ChIP analysis performed with β -catenin antibody (Invitrogen, catalog # 44207M) or anti-caspase 8 antibody (Cell Signaling Technologies: Caspase-8 (1C12) Mouse mAb #9746). The band indicates genomic DNA product of *HSF2* promoter. Error Bars: SD, (* p value < 0.05; ** p value < 0.01). Each graph is representative of at least two independent experiments.

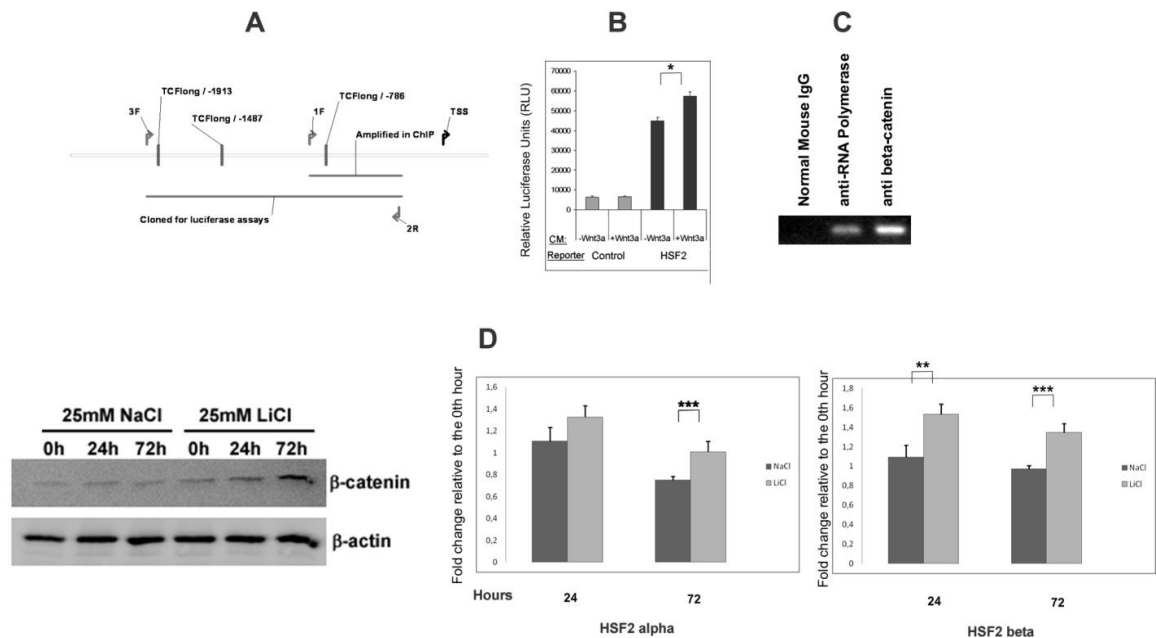


Figure 4.7. *HSF2* is a novel target of Wnt/ β -catenin/TCF signaling - II A: Scheme of *HSF2* promoter region. B: Luciferase reporter assays: *HSF2* promoter is 1.25 fold more active in Wnt3a exposed 293T cells when compared to non-exposed counterparts (*HSF2*). This difference is significant, though mild, with a p value of 0.01, when compared to the non-significant variation in pGL3-Basic (Control). C: ChIP assays: β -catenin binds to *HSF2* promoter in 18kj40 glioblastoma cells revealed by ChIP analysis performed with anti- β -catenin antibody, anti-RNA polymerase antibody or normal mouse IgG. The band corresponds to the PCR product of the genomic DNA *HSF2* promoter region captured with antibody. D: Lithium treatment: *HSF2* α and *HSF2* β mRNA level is increased upon increased β -catenin levels (verified by Western Blot in the left panel) due to lithium treatment of U373-MG glioblastoma cells. RNA levels were normalized using the geometric mean of β -actin (ACTB) and GAPDH RNA levels. Error Bars: SD, (** p value < 0.01; *** p value < 0.005).

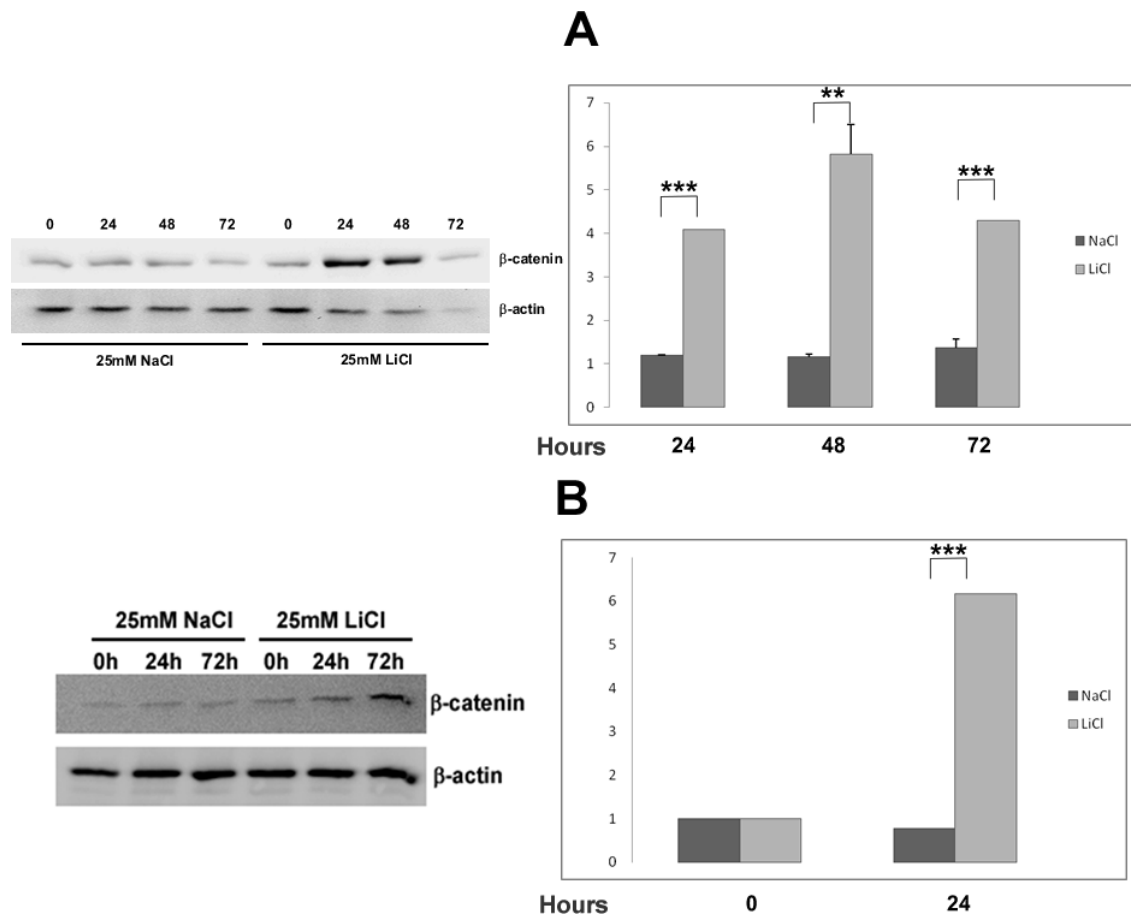


Figure 4.8. *AXIN2* is an immediate responder to lithium treatment of Huh7 and U373-MG cells. A: Lithium treatment: *AXIN2* mRNA level is increased after ≤ 24 hrs upon increased β -catenin levels (verified by Western Blot in the left panel) due to lithium treatment of Huh7. RNA levels were normalized using β -actin (*ACTB*) RNA levels. B: Lithium treatment: *AXIN2* mRNA level is increased after ≤ 24 hrs (verified by Western Blot in the left panel) due to lithium treatment of U373-MG. RNA levels were normalized by using β -actin (*ACTB*) RNA levels. Error Bars: SD, (** p value < 0.01; *** p value < 0.005).

4.1.4.1. *BRI3* and *HSF2* promoters are regulated by the β -catenin/TCF4 complex. Luciferase reporter assay was used to test the effect of over-expression of S33Y- β -catenin and dominant negative TCF4 [72] on the activity of *BRI3* and *HSF2* promoters in Huh7 cells. Over-expression of S33Y- β -catenin resulted in significant increase in promoter activities of both genes and this effect was reduced by addition of dominant negative TCF4 (Figure 4.4B and 4.6B). We then tested promoter activities of *BRI3* and *HSF2* upon autocrine Wnt3a stimulation of Huh7 cells by transient transfection with a Wnt3a plasmid;

which resulted in up-regulation of promoter activities of both genes (Figure 4.4A and 4.6A). We also tested promoter activities of *BRI3* and *HSF2* upon stimulation with Wnt3a in the 293T cell system described in [6]. This was achieved by exposing 293T cells to conditioned medium from L-cells or L-Wnt3a cells; and directly transfecting the 293T cell lines with the reporter constructs. As expected, the promoter activities were up-regulated in the L-Wnt3a exposed cells, with respect to L-cells ($p < 0.05$, student's t-test) (Figure 4.4B and Figure 4.7B for *BRI3* and *HSF2* respectively). Taken together, β -catenin and TCF4 regulate *BRI3* and *HSF2* promoters, *in vitro*.

4.1.4.2. β -catenin interacts with *BRI3* and *HSF2* promoters. We used chromatin immunoprecipitation assay (ChIP) to obtain direct evidence for the proposed interaction between the Wnt/ β -catenin pathway and the control of *BRI3* and *HSF2* gene expression. PCR analysis of the immunoprecipitated DNA showed that the anti- β -catenin and anti-RNA Polymerase II antibody, but not the normal mouse IgG, pulled-down *BRI3* and *HSF2* promoter fragments in Huh7 (Figure 4.4C and 4.6C). The promoters of *BRI3* and *HSF2* were also bound by anti- β -catenin but not by anti-caspase-8 antibody in mouse normal liver tissue (Figure 4.5C and Figure 4.6C). We also did ChIP analysis in early passage glioblastoma cells (#18 in [73]) to test *BRI3* and *HSF2* targeting by β -catenin in another cell context. *HSF2* promoter was bound by β -catenin in #18 glioblastoma cells (Figure 4.7C) whereas *BRI3* promoter was not (data not shown).

4.1.4.3. Inhibition of GSK3 β by lithium leads to increased *BRI3* and *HSF2* mRNA levels. The key kinase of the Wnt pathway GSK3 β that phosphorylates β -catenin to initiate its degradation is known to be inhibited by lithium. Cells exposed to lithium salts mimic the activated pathway and as a result of increased β -catenin accumulation and nuclear translocation, target gene expression is activated [74]. We treated Huh7 cells with lithium chloride and sodium chloride over 72 hours to follow *BRI3* and *HSF2* mRNA levels concomitantly with accumulation of β -catenin protein levels. As it is known that GSK3 β is already somewhat inhibited in Huh7 cells [75], we also included U373-MG cells (an established glioblastoma cell line) in this study. *BRI3* RNA levels were found to increase together with β -catenin protein accumulation in lithium treated but not sodium treated Huh7 and U373-MG cells (Figure 4.4D and 4.5D), later than *AXIN2* RNA which was used as a positive control for lithium treatment, being an established Wnt/TCF/ β -catenin target

[76] (Figure 4.8). HSF2 alpha and beta RNA levels increased together with β -catenin protein accumulation in U373 cells (Figure 4.6D) but did not significantly change in Huh7 cells (data not shown). Whether this might be due to different GSK3 β activities or some other lithium-related sensitivity difference between the cell lines is unknown.

4.2. Meta analysis of cancer gene expression profiles reveals cancer genes, SAGE tags and co-regulated regions on the genome

We conducted a gene and/or SAGE-tag centric meta-analysis of cancer gene expression data in order to see which transcriptional units are common to different cancers and which are specific to certain groups of tumors. We named this analysis as common gene expression patterning in tumors (COGENT – Figure 4.9). In COGENT, SAGE data and microarray data are used together or separately. Both data types were pre-processed as necessary. SAGE libraries were compared by using the DGED tool of SAGE Genie. For Solexa-SAGE libraries, a similar number of genes were selected from the top of the altered gene list by a procedure we called as rank selection. Microarray data was collected from Oncomine and Entrez GEO. Entrez GEO data was converted to Oncomine format, except that gene identifier was selected as Entrez GENE ID (for both sources), rather than gene symbol as in Oncomine. During gene mapping, consensus tag mapping was used separately to map long or short SAGE data, in order to maximize data acquisition. We have analyzed 302 SAGE libraries of 32 cancer types and 1404 microarrays of 37 different cancer types, which correspond in total to 49 different tumor types and comprise all major types which are epithelial, hematological, central nervous system and connective tissue tumors. Tumor names, classifications and used datasets are listed in Tables 4.11, 4.12 and 4.13. We have conducted different methods in order to maximally use the data and minimize the amount of false positives which might occur during several steps such as SAGE tag mapping (see Methods). We have, for the first time, used most of the available SAGE and microarray data together in a meta-analysis. As summarized in Figure 4.9, we have used SAGE alone, microarray alone or SAGE and microarray together by varying statistical thresholds for each part.

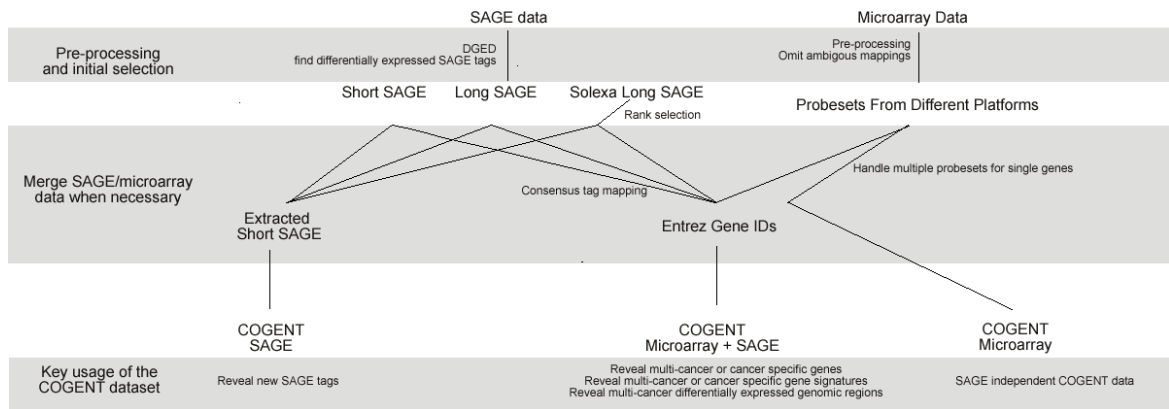


Figure 4.9. Study design

Table 4.11 Tumor types used in the meta-analysis

Cancer Type	Major Type	Tissue Source
Skin Cancer (squamous cell)	Epithelial	Squamous Skin
Ovarian Adenocarcinoma	Epithelial	Ovary
Esophageal Adenocarcinoma	Epithelial	Esophagus
Breast Carcinoma (DCIS)	Epithelial	Breast
Breast Carcinoma (IDC)	Epithelial	Breast
Renal Carcinoma (clear cell)	Epithelial	Kidney
Prostate Carcinoma	Epithelial	Prostate
Lung Adenocarcinoma	Epithelial	Lung
Colorectal Carcinoma	Epithelial	Colon
Pancreatic Adenocarcinoma	Epithelial	Pancreas
Gastric Carcinoma	Epithelial	Stomach
Thyroid Carcinoma (papillary)	Epithelial	Thyroid
Hepatocellular Carcinoma	Epithelial	Liver
Thyroid Carcinoma (follicular)	Epithelial	Thyroid
Liver cholangiocarcinoma	Epithelial	Liver Duct
Kidney Transitional Cell Carcinoma	Epithelial	Kidney Pelvis
Skin Cancer (basal cell)	Epithelial	Basal Cell Skin
Chronic Myeloid Leukemia (blast crisis)	Hematological	Myeloid
Chronic Myeloid Leukemia (chronic phase)	Hematological	Myeloid
Chronic Myeloid Leukemia (accelerated phase)	Hematological	Myeloid
Chronic Lymphocytic Leukemia	Hematological	Lymphocyte
Classical Hodgkin lymphoma	Hematological	Lymphocyte
B-Cell Acute Lymphoblastic Leukemia	Hematological	Lymphocyte

Table 4.11. (continued)

Acute Myeloid Leukemia	Hematological	Myeloid
Burkitt Lymphoma	Hematological	Lymphocyte
Hairy Cell Leukemia	Hematological	Lymphocyte
Diffuse Large B-Cell Lymphoma	Hematological	Lymphocyte
Primary Effusion Lymphoma	Hematological	Lymphocyte
Multiple Myeloma	Hematological	Plasma Cell
Mantle Cell Lymphoma	Hematological	Lymphocyte
Follicular Lymphoma	Hematological	Lymphocyte
Uterine Leiomyosarcoma	Mesenchyme	Smooth Muscle
Fibrosarcoma	Mesenchyme	Fibrous Tissue
Malignant Fibrous Histiocytoma	Mesenchyme	Fibrous Tissue
Synovial Sarcoma	Mesenchyme	Synovia
Mesothelioma	Mesothelium	Plevra
Oligodendroglioma	Nervous System	Glial Cells
Glioblastoma	Nervous System	Glial Cells
Anaplastic Oligoastrocytoma	Nervous System	Glial Cells
Astrocytoma Grade 2	Nervous System	Glial Cells
Astrocytoma Grade 3	Nervous System	Glial Cells
Astrocytoma Grade 1	Nervous System	Glial Cells
Ependymoma	Nervous System	Ependyme
Meningioma Grade 1	Nervous System	Meninges
Meningioma Grade 2	Nervous System	Meninges
Meningioma Grade 3	Nervous System	Meninges
Retinoblastoma	Nervous System	Eye
Medulloblastoma	Nervous System	Neural Progenitors
Melanoma	Neural Crest	Skin

Table 4.12. Microarray studies used in the meta-analysis

Study name	Data source*	Class I - Normal (Sample Size)	Class 2 - Tumor (Sample Size)
Nindl_nonmelanoma_skincancer	Entrez GDS2200	normal (6)	squamous cell carcinoma of the skin (5)
Hendrix_Ovarian	Oncomine	Normal Ovary (4)	Ovarian Serous Adenocarcinoma (41)
Hao_Esophagus	Oncomine	Normal Esophagus (15)	Esophageal Adenocarcinoma (5)
Boer_renal	Oncomine	Normal Kidney (162)	Clear Cell Renal Cell Carcinoma (160)
Lapointe_prostate	Oncomine	Normal Prostate (41)	Prostate Carcinoma (62)
Su_Lung	Oncomine	Normal Lung (31)	Lung Adenocarcinoma (31)
Graudens_colon	Oncomine	Normal Colon (12)	Colorectal Carcinoma (18)
Buchholz_Pancreas	Oncomine	Normal Pancreas Ducts (6)	Pancreatic Ductal Adenocarcinoma (8)
Hippo_gastric	Oncomine	Normal Gastric tissues (8)	Gastric Carcinoma (22)
He_thyroid	Entrez GDS1665	Normal thyroid (9)	Papillary thyroid carcinoma (9) -- matched samples
Chen_Liver	Oncomine	Non-tumor Liver (76)	Hepatocellular Carcinoma (104)
Hoek_Melanoma	Oncomine	Normal Melanocyte (3)	Cutaneous Melanoma (18)
Radich_CML_blastcrisis	Entrez GEO GSE4170	Normal CD34+ cells (6)	CML in blast crisis (12)
Radich_CML_chronicphase	Entrez GEO GSE4170	Normal CD34+ cells (6)	CML in chronic phase (12)
Radich_CML_accphase	Entrez GEO GSE4170	Normal CD34+ cells (6)	CML in accelerated phase (8)
Rosenwald_Multi-cancer_CLL	Oncomine	Normal B-Cells (8)	Chronic Lymphocytic Leukemia (39)
Brune_hodgkins	Entrez GEO GSE12453	Naïve B cells (5) and Memory Bcells(5) from the peripheral blood of patients	classical Hodgkin lymphoma (12)
Andersson_Leukemia BALL	Oncomine	Normal Bone Marrow (6)	B-Cell Acute Lymphoblastic Leukemia (87)
Andersson_Leukemia_AML	Oncomine	Normal Bone Marrow (6)	Acute Myeloid Leukemia (23)

Table 4.12. (continued)

Basso_lymphoma_Burkitt	Oncomine	Normal (25)	B-Cells	Burkitt Lymphoma (31)
Basso_lymphoma_hairy	Oncomine	Normal (25)	B-Cells	Hairy Cell Leukemia (16)
Basso_lymphoma_DLBCL	Oncomine	Normal (25)	B-Cells	Diffuse Large B-Cell Lymphoma (69)
Basso_lymphoma_PEL	Oncomine	Normal (25)	B-Cells	Primary Effusion Lymphoma (9)
Basso_lymphoma_myeloma	Oncomine	Normal (25)	B-Cells	Multiple Myeloma (4)
Basso_lymphoma_mantle	Oncomine	Normal (25)	B-Cells	Mantle Cell Lymphoma (8)
Basso_lymphoma_follicular_lymphoma	Oncomine	Normal (25)	B-Cells	Follicular Lymphoma (6)
Quade_leiomyosarcoma	Entrez GEO GSE764	myometrium(4)		uterine leiomyosarcoma (9)
Yoon_sarcoma_fibrosarcoma	Entrez GEO GDS1209	Normal (15)	tissues	fibrosarcoma (7)
Yoon_sarcoma_MFH	Entrez GEO GDS1209	Normal (15)	tissues	malignant fibrous histiocytoma (9)
Yoon_sarcoma_Synovial	Entrez GEO GDS1209	Normal (15)	tissues	synovial sarcoma (4)
Bredel_brain_2_oligodendroglioma	Oncomine	Normal Brain (4)		Oligodendroglioma (8)
Bredel_brain_2_GBM	Oncomine	Normal Brain (4)		Glioblastoma (31)
Bredel_brain_2_Anap_Oligoastro	Oncomine	Normal Brain (4)		Anaplastic Oligoastrocytoma (6)
Sun_brain_astrocytoma_Grade_2	Entrez GEO: GSE4290	Non-tumor (23)	Brain	Astrocytoma Grade 2 (7)
Sun_brain_astrocytoma_Grade_3	Entrez GEO: GSE4290	Non-tumor (23)	Brain	Astrocytoma Grade 3 (19)
Gutmann_Brain_astrocytoma_Grade_1	Oncomine	Normal Matter (3)	White	Pilocytic Astrocytoma (8)

*For none of the Entrez records, CEL files were available, thus processed Series matrix files were used

Table 4.13. SAGE datasets used in the meta-analysis

Study name	SAGE type	Class I size	Class 2 size
SAGE_astrocytoma_grade_1	Short	5	1
SAGE_astrocytoma_grade_2	Short	5	8
SAGE_astrocytoma_grade_3	Short	5	11
SAGE_thyroid_follicular	Short	1	1
SAGE_colon_adenocarcinoma	Short	2	2
SAGE_liver	Short	1	2
SAGE_kidney_trcell	Short	1	1
SAGE_prostate	Short	2	3
SAGE_lung_adenocarcinoma	Short	1	3
SAGE_Breast_DCIS	Short	19	8
SAGE_Breast_IDC	Short	19	11
SAGE_pancreas_adenocarcinoma	Short	1	2
SAGE_mesothelioma	Short	1	1
SAGE_stomach	Short	4	5
SAGE_medulloblastoma	Short	2	20
SAGE_glioblastoma	Short	7	10
SAGE_ependymoma	Short	7	11
SAGE_oligodendroglioma	Short	7	2
SAGE_meningioma_grade_1	Short	1	2
SAGE_meningioma_grade_2	Short	1	2
SAGE_meningioma_grade_3	Short	1	1
SAGE_retinoblastoma	Short	10	1
LSAGE_retinoblastoma	Long	1	1
LSAGE_Sol_melanoma	Long-Solexa	7	4
LSAGE_Sol_basalcell_skincancer	Long-Solexa	7	2
LSAGE_Sol_squamouscell_skincancer	Long-Solexa	7	2
LSAGE_Breast_IDC	Long	12	6
LSAGE_Breast_DCIS	Long	12	1
SAGE_Sol_melanoma	Short - Solexa	7	4
SAGE_Sol_basalcell_skincancer	Short - Solexa	7	2
SAGE_Sol_squamouscell_skincancer	Short - Solexa	7	2

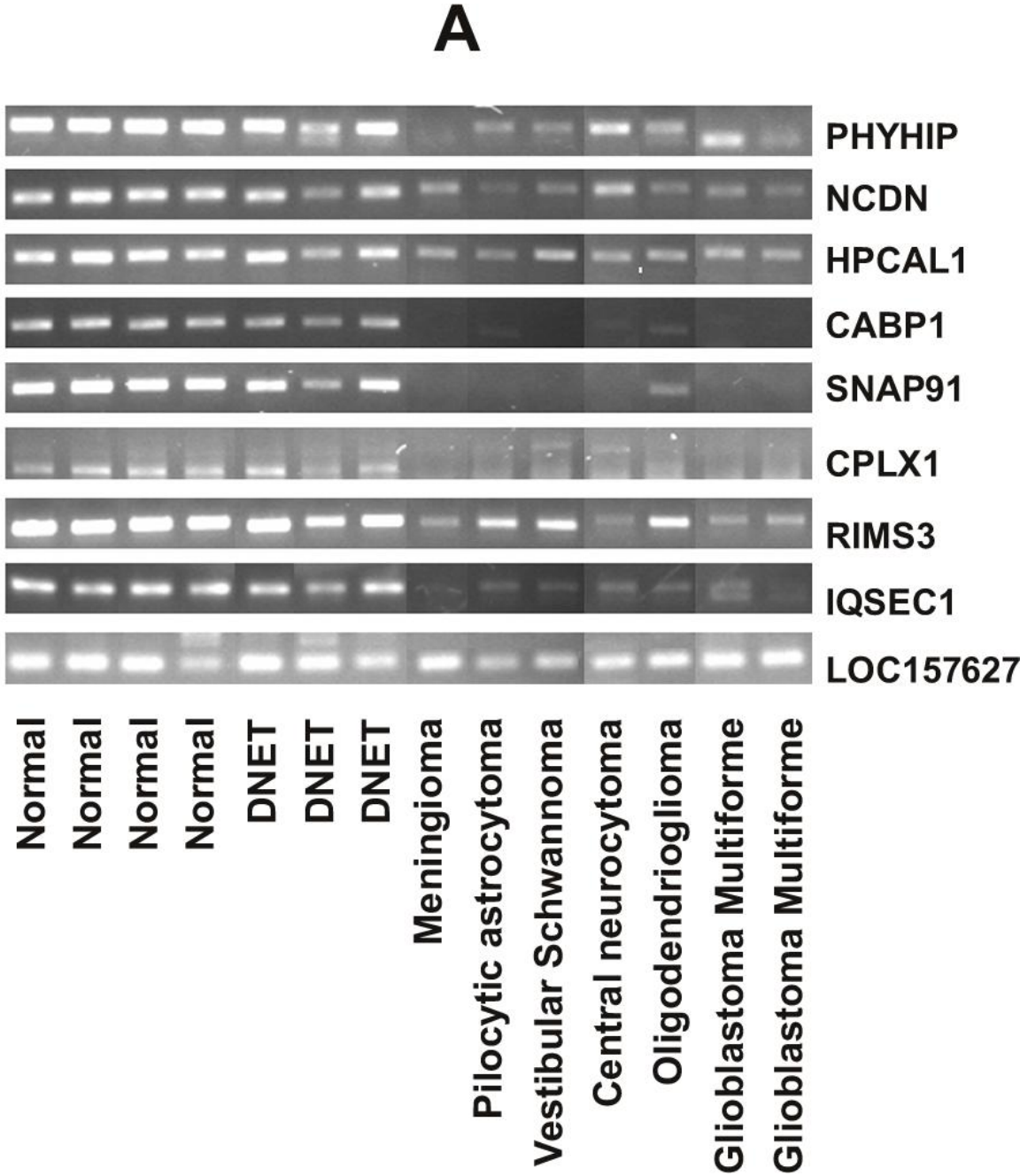


Figure 4.10. Verification and extension of Oncoreveal predictions in an independent extended brain tumor panel. A: Verification of 9 selected down-regulated genes not-yet-associated to tumors

B

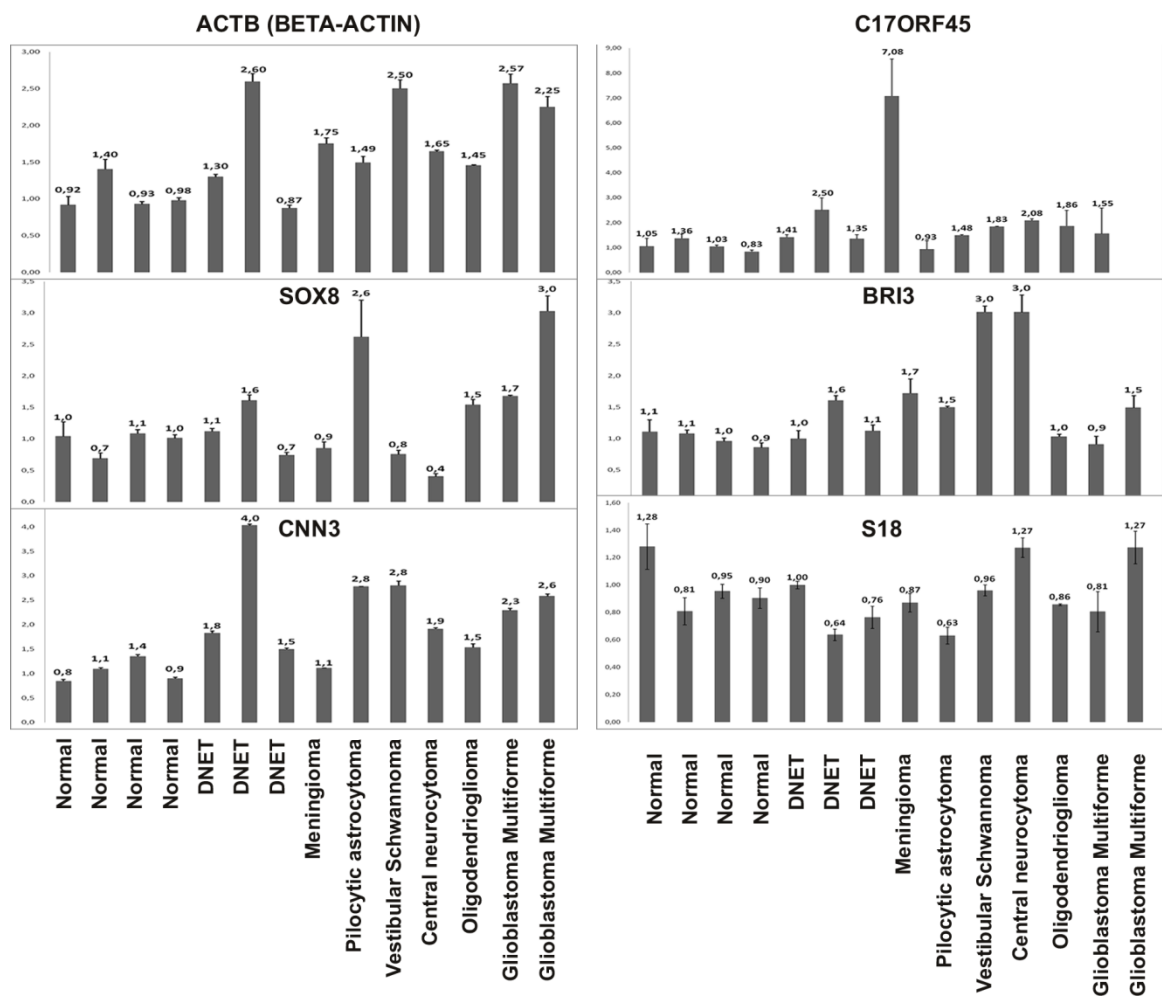


Figure 4.10. (continued) B: Verification of 3 selected up-regulated genes not-yet-associated to tumors. Real time PCR data was normalized using s18 levels. All PCR were done on the same cDNA panel.

Assigning an Entrez GENE id to a SAGE tag is problematic due to high level of multiple matches. Two mainly used approaches are SAGE Map and SAGE Genie. 43 per cent of the short tag mappings and 7 per cent of the long tag mappings do not agree when only single mappings of SAGE Map are compared with best gene calls of SAGE Genie (Table 4.14). In order to minimize false positives, we used the SAGE tags for which SAGE Map maps to a single gene and SAGE Genie's best gene is same with that mapping. We termed this consensus tag mapping. In order to merge SAGE data with microarray data using a true tag mapping approach was of great importance. In order to see both how consensus tag mapping performs and to verify some of the multi-cancer genes as yet not

associated to tumors (See below for explanation of multi-cancer genes not yet associated to tumors), we have verified the expression levels of 13 genes in an independent brain tumor – non-tumor brain panel (Figure 4.10) all of which were assigned by consensus tag mapping. RT-PCRs and Q-RTPCRs were all consistent with what Oncoreveal suggested, although not perfectly overlapping (compare Figure 4.10 & Table 4.15). Overall it is clear, also as verified by RT-PCR, that consensus tag mapping is the best available option that can be used to assign trusted gene mappings to SAGE tags. To be functionally related to tumorigenesis, more data than differential expression is needed. The 13 genes in the PCR panel were selected for their virtue of being differentially expressed in CNS tumors but not yet recognized as tumor- associated.

After verifying the expression pattern of genes in an independent CNS tumor panel, we intended to see if some of these genes have an effect on tumorigenesis. Due to the fact that most of the candidate genes were shown to be under or over-expressed in glioblastomas by Oncoreveal; several of the genes from the verification panel were cloned into expression vector pCDNA3 and transiently transfected into early passage fast growing, Imatinib insensitive glioblastoma cells (#18 in ref[73]) and an established glioblastoma cell line U373-MG.

Table 4.14A SAGE Genie and SAGE map agreement statistics for short SAGE tags

			SAGE Map no mapping		SAGE Map single mapping	SAGE Map multiple mapping
SAGE	Genie's	best	0		51810 tags	36458 tags
gene call agrees*					22329 Unigenes (16260 Entrez Gene ID mappings ; 16063 unique Entrez Gene IDs**)	
SAGE	Genie's	best	112416	tags***	39071 tags	20010 tags
gene call disagrees						

*at least with 1 of the SAGE maps highest scored 2 mappings

** In the current Unigene dump; 122727 Unigene clusters map to 23739 Entrez Gene IDs out of 32736 total genes in 28 October 2008 seq_gene.md release of NCBI. So, out of 23739 possible 16063 (68 per cent) can be captured by using consensus tag mapping.

*** SAGE genie calls for a unigene whereas SAGE map does not call

Table 4.14B SAGE Genie and SAGE map agreement statistics for long SAGE tags

	SAGE Map no mapping	SAGE Map single mapping	SAGE Map multiple mapping
SAGE Genie's best gene call agrees	0	98490 tags 30562 Unigenes 19098 Gene IDs**	8035 tags
SAGE Genie's best gene call disagrees	246219 tags	7939 tags	1049 tags

**In the current Unigene dump; 122727 Unigene clusters map to 23739 Entrez Gene IDs out of 32736 total genes in 28 October 2008 seq_gene.md release of NCBI. So, out of 23739 possible 19098 (80 per cent) can be captured by using consensus tag mapping.

Five of the genes screened in this assay, *HPCAL1*, *NCDN*, *RIMS3*, *CPLX1* and *IQSEC1* were under-expressed in glioblastomas. *C17ORF45* was similarly expressed in glioblastomas when compared to non-tumor CNS samples, whereas it was over-expressed in some other CNS tumors. *SOX8* was over-expressed in one glioblastoma sample and pilocytic astrocytoma. It was shown that *C17ORF45*, *IQSEC1* and *CPLX1* inhibited colony formation ability of #18 cells, revealed with an assay that monitors cell-cell contact independent growth. In addition, *IQSEC1* and *CPLX1* also inhibited colony formation ability of U373-MG cells (Figure 4.11).

Table 4.15. COGENT-SAGE analysis by using Oncoreveal. Several multi-cancer genes which are differentially expressed in CNS tumors (See Figure 4.10 for RT-PCR or Q-RTPCR analysis of these genes in an independent brain tumor, non-tumor panel)

Gene Symbol	SAGE tag	Increased in	# Inc.	Decreased in	# Dec.
HPCAL1	TGTGCCCTGA			Astrocytoma Grade 1 Astrocytoma Grade 2 Astrocytoma Grade 3 Medulloblastoma Glioblastoma Ependymoma Oligodendroglioma Meningioma Grade 1 Meningioma Grade 3	9
CABP1	TCGGGGCCCC	Breast Carcinoma (IDC)	1	Astrocytoma Grade 1 Astrocytoma Grade 2 Astrocytoma Grade 3 Medulloblastoma Glioblastoma Ependymoma Oligodendroglioma Retinoblastoma	8
SNAP91	TACGCCCTCT			Astrocytoma Grade 1 Astrocytoma Grade 2 Astrocytoma Grade 3 Medulloblastoma Glioblastoma Ependymoma Oligodendroglioma	7
CPLX1	TCGGGGCCCC	Breast Carcinoma (IDC)	1	Astrocytoma Grade 1 Astrocytoma Grade 2 Astrocytoma Grade 3 Medulloblastoma Glioblastoma Ependymoma Oligodendroglioma Retinoblastoma	8
IQSEC1	TGGGGAAATC	Breast Carcinoma (DCIS) Breast Carcinoma (IDC)	2	Astrocytoma Grade 1 Astrocytoma Grade 2 Astrocytoma Grade 3 Medulloblastoma Glioblastoma Ependymoma Oligodendroglioma Skin Cancer (squamous cell)	8
LOC157627	AGGAGTTGTG			Astrocytoma Grade 1 Astrocytoma Grade 2 Astrocytoma Grade 3 Medulloblastoma Glioblastoma Ependymoma Oligodendroglioma	7

Table 4.15. (continued)

ACTB	GCTTTATTTG	Astrocytoma Grade 2 Pancreatic Adenocarcinoma Gastric Carcinoma Glioblastoma Retinoblastoma	5	Prostate Carcinoma Lung Adenocarcinoma Meningioma Grade 1 Meningioma Grade 3	4
SOX8	CAGCGTCTCC	Astrocytoma Grade 1 Astrocytoma Grade 2 Astrocytoma Grade 3 Oligodendroglioma	4	Ependymoma	1
CNN3	TACAAAAGTG	Astrocytoma Grade 2 Astrocytoma Grade 3 Medulloblastoma Glioblastoma Ependymoma	5	Breast Carcinoma (DCIS)	1
BRI3	TTTTTACTGA	Astrocytoma Grade 2 Astrocytoma Grade 3 Ependymoma	3	Colorectal Carcinoma	1
C17ORF45	AACTAATACT	Astrocytoma Grade 2 Astrocytoma Grade 3 Thyroid Carcinoma (follicular) Medulloblastoma Ependymoma Oligodendroglioma Retinoblastoma	7	Lung Adenocarcinoma Breast Carcinoma (DCIS)	2
PHYHIP	GCCTGAGGGC			Astrocytoma Grade 1; Astrocytoma Grade 2; Astrocytoma Grade 3 Medulloblastoma Glioblastoma Ependymoma Oligodendroglioma Meningioma Grade 1 Meningioma Grade 2 Meningioma Grade 3	10
NCDN	TGGACACTCA	Breast Carcinoma (DCIS)	1	Astrocytoma Grade 1 Astrocytoma Grade 2 Astrocytoma Grade 3 Medulloblastoma Glioblastoma Ependymoma Oligodendroglioma Meningioma Grade 1 Meningioma Grade 2 Meningioma Grade 3	10
RIMS3	ATTAAAGTCA			Astrocytoma Grade 1 Astrocytoma Grade 2 Astrocytoma Grade 3 Glioblastoma Ependymoma Oligodendroglioma Meningioma Grade 1 Meningioma Grade 2 Meningioma Grade 3	9

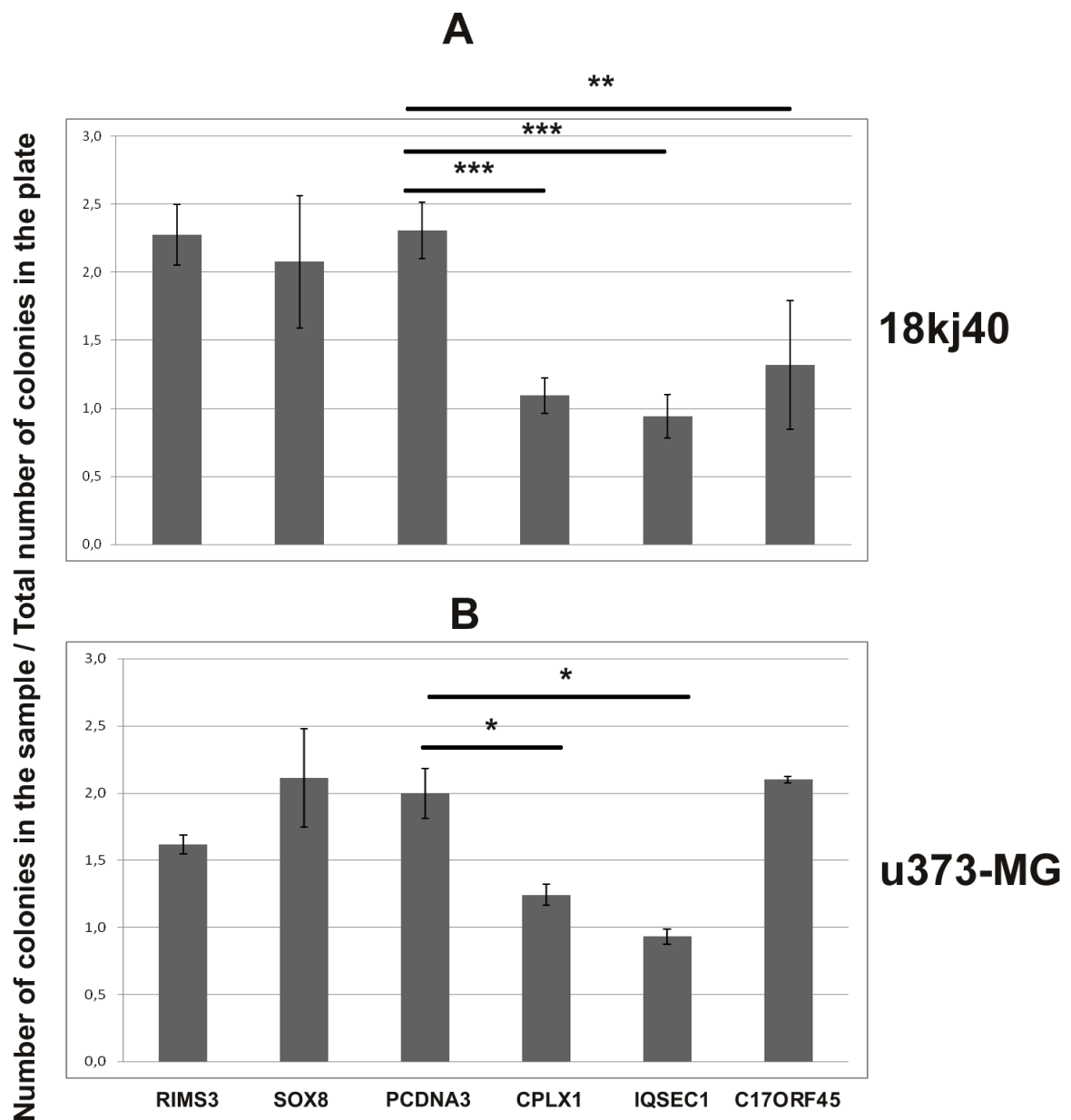


Figure 4.11 A: IQSEC1, C17ORF45 and CPLX1 but not other genes inhibit colony formation of 18kj40 cells. B: IQSEC1 and CPLX1 but not other genes inhibit colony formation of u373-MG cells. All values are representative of at least 3 independent day experiments. Y axis is normalized by using the total number of colonies per plate due to the fact that there is a plate effect on colony outgrowth. Error bars represent SE. * : p value < 0.05; ** : p value < 0.01; ***: p value < 0.005

4.2.1. Being multi-cancer

We aimed to see if being differentially expressed in more types of tumors (being multi-cancer) means something. We analyzed several putative relationships for a certain

gene, between being multi-cancer and several properties of a gene such as number of publications rank of the gene expression change, number of orthologs and number of paralogs. We have found that, the more the number of cancers a gene is differentially expressed (over- or under-expressed) the more publications (having at least 1 related publication - cancer publications or all publications) it has (Figure 4.12A). To assign publications to genes, we used Entrez geneRIF database which consists of user curated summary of publications. Number of cancer generifs were defined as the sentences recorded (by article authors) in Entrez Generif database containing “tumor”, “carcinoma”, “cancer” or “neopla” keywords. Currently, for 10000 of the genes there is at least one geneRIF (Half of the genes covered in this study). This correlation was strikingly significant (Kendall Tau < 0.0000). In addition, the rank of a gene is inversely correlated with being multi-cancer (Figure 4.12B). So, if a gene’s expression change is high in a single type of cancer, this means that it will probably be a multi-cancer gene (and vice versa). Although there is a trend of multi-cancer genes being associated with tumorigenesis, there are lots of multi-cancer genes which have not been related to tumorigenesis. For instance; TP53, EGFR, ERBB2, VEGFA, PTGS2, BRCA1, CDKN2A, ESR1, PTEN, CCND1 and HIF1A are some well-established tumor genes among the most multi-cancer genes (altered > 10 types; see Table 4.16A for a list), whereas SYNCRIP, P4HA1, GART and ASCC3 and a lot more were over-expressed in similar numbers of tumors and have not been related to tumorigenesis at all (see Table 4.16B for a representative list and for a more comprehensive analysis of multi-cancer genes see www.oncoreveal.org). There are 438 and 110 genes which do not have any cancer geneRIFs but are found to be over- or under-expressed in more than 10 resp. 13 types of tumor-normal comparisons. Functional classifications of multi-cancer genes were also analyzed. As expected, overall multi-cancer genes are significantly enriched (DAVID: Benjamini < 0.05) for cancer related functions such as cell cycle and apoptosis (see Table 4.17A for the top 50 enriched classes). In order to find if there is a new functional class identified by our approach we did the same analysis by using multi-cancer genes not yet associated to tumors with a background of tumor-associated multi-cancer genes (rather than the whole genome as we used for tumor-associated multi-cancer genes). We find 4 such classes (Benjamini < 0.05) which are organelle membrane, endomembrane system, protein transport and endoplasmic reticulum (Table 4.17B). 4 classes mostly overlap with each other (e.g. endomembrane system genes are a subset of organelle membrane).

Therefore, we concluded that the endoplasmic reticulum related genes may be important in tumorigenesis but have not yet been identified as such until now (see Discussion).

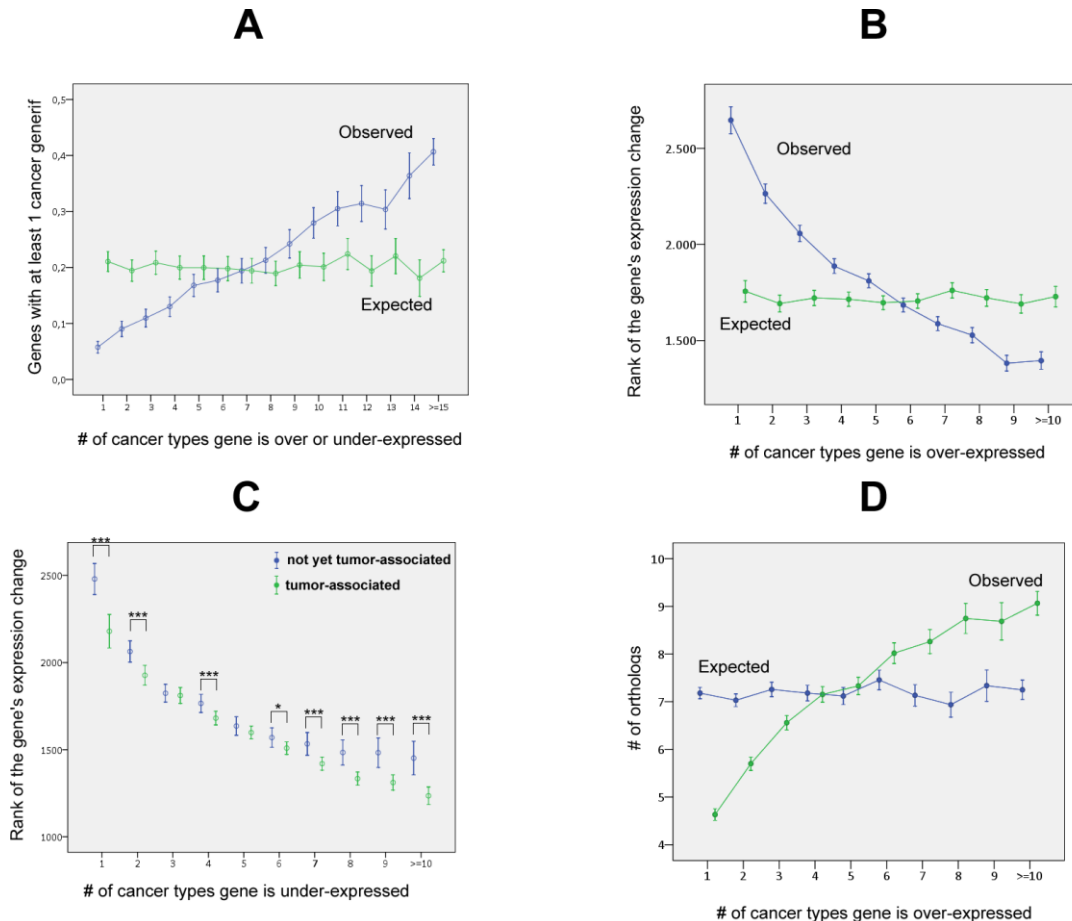


Figure 4.12. Being multi-cancer: Meta-analysis of SAGE-microarray merged tumor expression data points to several patterns dependent on the number of tumor-normal comparisons in which a gene is differentially expressed

A: The multi-cancer genes are more likely to be associated with tumorigenesis with an uninterrupted trend. The sum of # of tumor-normal comparisons where a gene is up-regulated or down-regulated is used for the analysis. Real number of generifs are recoded to 1 to denote being tumor-associated, and the rest of the genes are 0 to denote not being tumor associated)

B: Rank of a gene's expression change (when sorted by p value) decreases with multi-canceriness.

C: Rank of a gene's expression change in a certain observation explains being associated to tumorigenesis.

Ranks of the tumor-associated and non-tumor-associated genes are significantly ($p < 0.000000$, Mann Whitney U test) different when the data is not sub-divided by the number of tumor type differential expression was observed, rank is significantly lower.

D: Multi-cancer genes have more orthologs (homologs in different species) but not paralogs (homologs in the same species, data not shown).

***: p value < 0.005, **: p value < 0.01, *: p value < 0.05. In all of the graphs, the point markers are means of the data, and the error bars indicate 95% confidence interval.

Table 4.16A. Already well established tumor-associated multi cancer genes. All data was filtered for the genes which are altered (Up + Down) in more than 10 types of tumors when compared to corresponding normals. And genes are sorted by number of cancer generifs in decending order.

Symbol	Description	# Up	# Down	# Total geneRIFs	# Cancer geneRIFs
TP53	tumor protein p53	7	8	2371	782
EGFR	epidermal growth factor receptor (erythroblastic leukemia viral oncogene homolog, avian)	10	4	887	326
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	10	5	530	321
VEGFA	vascular endothelial growth factor A	14	5	1058	309
PTGS2	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	4	8	779	291
BRCA1	breast cancer 1, early onset	7	5	818	254
CDKN2A	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	16	3	609	235
ESR1	estrogen receptor 1	6	12	875	161
PTEN	phosphatase and tensin homolog	3	16	366	161
CCND1	cyclin D1	16	7	380	153
HIF1A	hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)	11	10	502	149
MMP9	matrix metalloproteinase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)	13	5	638	142
TGFB1	transforming growth factor, beta 1	9	9	1118	129
BCL2	B-cell CLL/lymphoma 2	8	14	446	126
MMP2	matrix metalloproteinase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)	7	4	423	126
BIRC5	baculoviral IAP repeat-containing 5	13	4	269	123
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	6	7	244	113
STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)	9	10	349	96
NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	5	14	661	95
MSH2	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	13	5	237	87
GSTP1	glutathione S-transferase pi 1	13	3	617	87
MYC	v-myc myelocytomatosis viral oncogene homolog (avian)	13	9	327	84
IL6	interleukin 6 (interferon, beta 2)	9	6	1224	84
CD44	CD44 molecule (Indian blood group)	12	10	193	82

Table 4.16B. Multi-cancer genes which do not have any cancer geneRIF. Genes were sorted by the # of tumor-normal comparisons they are altered (Up + Down)

Symbol	Description	# Up	# Down	# Total geneRIFs
ITPK1	inositol 1,3,4-triphosphate 5/6 kinase	15	16	2
H2AFV	H2A histone family, member V	18	12	0
ACVR1B	activin A receptor, type IB	10	18	8
QDPR	quinoid dihydropteridine reductase	11	16	2
ALDOA	aldolase A, fructose-bisphosphate	13	13	4
DCTN2	dynactin 2 (p50)	10	16	5
SEPT11	septin 11	14	11	3
ATP1B3	ATPase, Na ⁺ /K ⁺ transporting, beta 3 polypeptide	14	11	3
RAB6A	RAB6A, member RAS oncogene family	11	14	12
CAPN3	calpain 3, (p94)	9	16	28
RBBP6	retinoblastoma binding protein 6	8	17	2
MBP	myelin basic protein	7	18	26
CAMK2G	calcium/calmodulin-dependent protein kinase II gamma	7	18	11
SYNCRIP	synaptotagmin binding, cytoplasmic RNA interacting protein	19	5	4
ZNF207	zinc finger protein 207	18	6	0
ILF3	interleukin enhancer binding factor 3, 90kDa	17	7	14
MAP4	microtubule-associated protein 4	15	9	8
TBL1X	transducin (beta)-like 1X-linked	11	13	3
HEXIM1	hexamethylene bis-acetamide inducible 1	10	14	20
GART	phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase, phosphoribosylaminoimidazole synthetase	20	3	5
PCDHGC3	protocadherin gamma subfamily C, 3	17	6	1
IFI6	interferon, alpha-inducible protein 6	17	6	1
PDXK	pyridoxal (pyridoxine, vitamin B6) kinase	14	9	3
RAB2A	RAB2A, member RAS oncogene family	13	10	3
PFKP	phosphofructokinase, platelet	11	12	5
SORL1	sortilin-related receptor, L(DLR class) A repeats-containing	9	14	23
ALDH3A2	aldehyde dehydrogenase 3 family, member A2	9	14	11
SEPT6	septin 6	8	15	7
MTMR9	myotubularin related protein 9	4	19	1

Table 4.16B. (continued)

MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase	18	4	1
TMCO1	transmembrane and coiled-coil domains 1	16	6	0
RCN1	reticulocalbin 1, EF-hand calcium binding domain	16	6	0
SMC4	structural maintenance of chromosomes 4	15	7	1
YARS	tyrosyl-tRNA synthetase	14	8	10
COL6A1	collagen, type VI, alpha 1	14	8	19
CTPS	CTP synthase	14	8	3
MAP1B	microtubule-associated protein 1B	14	8	5
FCGRT	Fc fragment of IgG, receptor, transporter, alpha	14	8	25
AGPAT7	lysophosphatidylcholine acyltransferase 4	14	8	2
CPNE1	copine I	12	10	2
METTL7A	methyltransferase like 7A	10	12	0
TGOLN2	trans-golgi network protein 2	9	13	0
KLC1	kinesin light chain 1	9	13	8
TRAF5	TNF receptor-associated factor 5	9	13	4
IDS	iduronate 2-sulfatase	8	14	14
PHF15	PHD finger protein 15	8	14	0
PPP1R12B	protein phosphatase 1, regulatory (inhibitor) subunit 12B	7	15	1
JMJD2B	jumonji domain containing 2B	7	15	1
ADARB1	adenosine deaminase, RNA-specific, B1 (RED1 homolog rat)	6	16	12
CYFIP2	cytoplasmic FMR1 interacting protein 2	5	17	1

Table 4.17A. DAVID analysis to find enriched classes of genes in multi-cancer tumor-associated genes. The analysis were performed by using 317 genes increased (out of 795) > 10 types of tumor-normal comparisons which have at least 1 CANCER GENE RIF.

Category	Term	Count	%	PValue	Fold Enrichment	Benjamini
SP_PIR_KEYWORDS	Direct protein sequencing	160	44,82 %	4,71E-46	3,207651	5,02E-43
GOTERM_MF_ALL	protein binding	265	74,23 %	4,68E-39	1,807181	1,35E-35
GOTERM_BP_ALL	cell cycle	83	23,25 %	2,26E-29	4,169708	1,18E-25
GOTERM_BP_ALL	cell cycle process	74	20,73 %	1,11E-27	4,437261	2,92E-24
SP_PIR_KEYWORDS	phosphoprotein	186	52,10 %	8,45E-26	1,982139	4,50E-23
GOTERM_BP_ALL	regulation of progression through cell cycle	59	16,53 %	1,19E-24	5,037689	2,09E-21
GOTERM_BP_ALL	regulation of cell cycle	59	16,53 %	1,59E-24	5,00912	2,09E-21
SP_PIR_KEYWORDS	cell cycle	49	13,73 %	2,01E-23	5,963205	7,14E-21
GOTERM_BP_ALL	mitotic cell cycle	45	12,61 %	9,61E-23	6,355511	1,01E-19
GOTERM_BP_ALL	cell cycle phase	47	13,17 %	1,74E-22	5,92943	1,53E-19
GOTERM_BP_ALL	M phase	40	11,20 %	5,74E-20	6,259551	4,30E-17
GOTERM_BP_ALL	mitosis	35	9,80 %	7,09E-19	6,924801	4,66E-16
KEGG_PATHWAY	hsa04110:Cell cycle	32	8,96 %	2,40E-18	6,83744	4,81E-16
GOTERM_BP_ALL	M phase of mitotic cell cycle	35	9,80 %	9,39E-19	6,864322	5,48E-16
SP_PIR_KEYWORDS	cell division	31	8,68 %	4,58E-18	7,771638	9,77E-16
SP_PIR_KEYWORDS	mitosis	26	7,28 %	4,50E-18	10,18461	1,20E-15
GOTERM_BP_ALL	cell proliferation	61	17,09 %	3,19E-17	3,44177	1,68E-14
SP_PIR_KEYWORDS	cell cycle control	16	4,48 %	1,37E-16	21,68199	1,98E-14
GOTERM_BP_ALL	cell division	33	9,24 %	5,70E-16	6,049409	2,65E-13
GOTERM_BP_ALL	response to stress	69	19,33 %	2,65E-15	2,866741	1,17E-12
SP_PIR_KEYWORDS	atp-binding	66	18,49 %	1,64E-14	2,862642	2,49E-12
SP_PIR_KEYWORDS	nucleotide-binding	76	21,29 %	3,20E-14	2,562616	4,26E-12
GOTERM_MF_ALL	binding	318	89,08 %	4,30E-14	1,198106	6,20E-11
GOTERM_BP_ALL	programmed cell death	54	15,13 %	4,38E-13	3,066072	1,77E-10
GOTERM_BP_ALL	developmental process	130	36,41 %	1,13E-12	1,789882	4,25E-10
GOTERM_BP_ALL	biological regulation	180	50,42 %	1,39E-12	1,543377	4,85E-10
GOTERM_BP_ALL	regulation of biological process	168	47,06 %	1,84E-12	1,585472	6,03E-10
GOTERM_BP_ALL	cell death	54	15,13 %	3,44E-12	2,907989	1,06E-09

Table 4.17A (continued)

GOTERM_BP_ALL	death	54	15,13 %	3,44E-12	2,907989	1,06E-09
GOTERM_BP_ALL	apoptosis	52	14,57 %	3,91E-12	2,978876	1,08E-09
SP_PIR_KEYWORDS	acetylation	41	11,48 %	1,96E-11	3,431926	2,33E-09
GOTERM_MF_ALL	ATP binding	74	20,73 %	4,95E-12	2,351107	4,75E-09

GOTERM_BP_ALL	regulation of apoptosis	41	11,48 %	1,90E-11	3,422683	4,98E-09
GOTERM_MF_ALL	adenyl ribonucleotide binding	74	20,73 %	8,86E-12	2,322599	6,37E-09
GOTERM_BP_ALL	regulation of mitosis	16	4,48 %	2,76E-11	10,26566	6,59E-09
GOTERM_BP_ALL	regulation of programmed cell death	41	11,48 %	2,66E-11	3,384933	6,67E-09
UP_SEQ_FEATURE	mutagenesis site	74	20,73 %	3,61E-13	2,455181	6,82E-09
GOTERM_BP_ALL	cellular component organization and biogenesis	111	31,09 %	3,73E-11	1,830798	8,53E-09
GOTERM_BP_ALL	interphase	18	5,04 %	4,65E-11	8,249194	1,02E-08
GOTERM_BP_ALL	cell cycle checkpoint	15	4,20 %	6,02E-11	10,86587	1,26E-08
GOTERM_BP_ALL	regulation of cellular process	155	43,42 %	6,40E-11	1,574266	1,29E-08
SP_PIR_KEYWORDS	heterodimer	17	4,76 %	1,71E-10	8,3566	1,83E-08
GOTERM_MF_ALL	purine ribonucleotide binding	83	23,25 %	4,16E-11	2,111824	2,00E-08
GOTERM_MF_ALL	ribonucleotide binding	83	23,25 %	4,16E-11	2,111824	2,00E-08
SP_PIR_KEYWORDS	cytoplasm	95	26,61 %	2,12E-10	1,916806	2,06E-08
GOTERM_BP_ALL	DNA metabolic process	52	14,57 %	1,15E-10	2,715626	2,24E-08
GOTERM_BP_ALL	cell development	65	18,21 %	1,54E-10	2,344818	2,89E-08
GOTERM_BP_ALL	interphase of mitotic cell cycle	17	4,76 %	1,68E-10	8,299008	3,04E-08
GOTERM_MF_ALL	adenyl nucleotide binding	74	20,73 %	9,06E-11	2,209783	3,72E-08

After observing the correlation between the rank of a gene and multi-canceriness, we aimed to find what makes some genes tumor-associated and some not-yet-associated when they are differentially expressed in the same number of tumors. The minimum rank of the gene within a single tumor-normal comparison (the minimum rank of a gene was considered when multiple probesets are mapping to a single gene, when the dataset is sorted by p value for microarray data, and first by p value and then fold value for SAGE data), but not p-value or number of orthologs explains being tumor-associated or not (Figure 4.12C). In other words, if a gene ranks high (when sorted by q value for microarray data and first by p value then by fold for SAGE) in a single type of tumor-normal comparison, this means that it will more probably be multi-cancer and tumor-associated (has been studied by some group)

Table 4.17B. DAVID analysis to find enriched classes in not-yet-associated to tumor multi-cancer genes when compared to tumor-associated multi-cancer genes

Category	Term	Count	PValue	Fold Enrichment	Benjamini	FDR
GOTERM_CC_ALL	organelle membrane	80	2,83E-07	1,441169	1,23E-04	3,99E-04
GOTERM_CC_ALL	endomembrane system	64	7,55E-06	1,441169	1,65E-03	1,07E-02
SP_PIR_KEYWORDS	protein transport	33	2,88E-05	1,650882	1,53E-02	0,041802
SP_PIR_KEYWORDS	endoplasmic reticulum	50	3,16E-05	1,485168	0,008428571	0,045915

In addition, we found that multi-cancer genes have more orthologs when compared to non-multi-cancer genes which can explain or be explained by several different points of view (Figure 4.12C). One example hypothesis is that the genes which are responsible for core processes in the cell are conserved more and they are altered in all tumorigenesis transitions independent of the tissue source.

Overall, we present data that indicates that being multi-cancer is a discriminating property for a gene. We also present an online tool for user data overlay with COGENT data (www.oncoreveal.org) which makes the newly revealed multi-cancer genes workable.

In addition to identifying multi-cancer genes, we worked on multi-cancer SAGE tags (SAGE COGENT). This approach reveals important SAGE tags which have been ignored in single SAGE studies (Table 4.18). In the routine SAGE library analysis, users tend to ignore multiple mapping or ambiguously mapping SAGE tags, due to extra requirements of money and time for clarification of their identity. If a SAGE tag is differentially expressed in many types of tumors, identifying that tag's corresponding gene will probably be worth the effort. For instance, there are 7 over-expressed and 58 under-expressed SAGE tags which are differentially expressed in more than 3 types of tumors and does not map reliably to a gene according neither to SAGE Map or SAGE Genie.

Table 4.18. SAGE tags revealed using COGENT

# of cancer types*	Over-expressed tags			Under-expressed tags		
	Single	2 & 3	> 3	Single	2 & 3	> 3
Tags which don't map stringently to a single gene	3558	1522	639	2974	1419	707
Tags which dont map to reliable 3'ends**	508	92	7	496	159	58
Tags which don't map to any existing cDNA sequence	19	2	0	23	8	2

*Being differentially expressed in this many number of tumor types

** SAGE Maps polyadenylated MGC, RefSEQ, mRNA or EST called tags and SAGE Genie's reliable 3'end

4.2.2. Under-expressed ribosomal genes in cancer suggests funtctional roles for ribosomal genes

In addition to the ER genes, Oncoreveal points to several other classes of genes which were not known to behave likewise. A noteworthy example is down regulation of several different ribosomal proteins in a wide variety of tumor types. For instance according to middle stringency COGENT microarray & SAGE; 121 ribosomal protein genes are over-expressed at least in 2 types of tumors and 92 per cent of them are also under-expressed in at least 1 type of tumor. RPL18A, for instance, is under-expressed in 6 types (3 blood tumors and 3 epithelial tumors) and over-expressed in 2 types of tumors (astrocytoma grade 2 and 3). In addition RPS16, RPL15, RPL31, RPS15A, RPL22 and RPS6 are under-expressed in more than 7 types of tumors. Ribosomal genes are tightly regulated in case of cell proliferation and cell growth by several cancer genes such as p53, RB, MYC and PTEN. p53 and RB regulate polymerase I and polymerase III activity, which results in rDNA transcription regulation. MYC and PTEN are known to act on regulation of ribosome biogenesis and translational control[77]. Besides, several ribosomal genes are shown to be important in tumor development. RPS9 ribosomal protein was found to be over-expressed in colorectal cancer[78]. RPL7 and RPL37 ribosomal proteins are

found to be over-expressed in prostate cancer cell lines[79]. RPL15 is over-expressed in esophageal cancer[80]. Many of the ribosomal genes are found to be important in cancer in a zebra fish model[37]. RPL19 was suggested as a prognostic marker in prostate cancer[81]. In addition, it has been shown that the epigenetic status of 18S and 28S ribosomal genes predicts survival in ovarian cancer[82]. Ribosomal proteins are generally believed to be related to excessive proliferation in tumor cells and commonly regarded as being functionally irrelevant. All these data suggests that, ribosomes might play a role in cancer, other than controlling the overall RNA level in the cell. Accordingly, Oncoreveal points not only to over-expression but also to under-expression of several ribosomal proteins (see www.oncoreveal.org).

4.2.3. Genomic regions of transcriptional imbalance (RIDGES)

Considering previous studies that indicate co-regulated regions[61] in different cancers, we aimed to define multi-cancer RIDGES (tumor associated – TA-RIDGES) on the chromosomes in comparison with RIDGES regulated among normal tissues (normal associated - NA-RIDGES). To define TA-RIDGES we used normalized difference of tumor and normal mean values from microarray studies, whereas to define NA-RIDGES we used a comprehensive microarray study which comprises 36 different normal tissues[83]. Afterwards, as previously proved to be useful[61] we adapted a gene neighborhood correlation method to identify RIDGES. Briefly, this method assigns a neighborhood correlation score to each gene by scanning the gene's neighbors in a certain window size and calculating the average spearman correlation coefficient. This window size was optimized by maximizing the number of genes above the statistical threshold for a certain window size in Stransky et al. In our study, we optimized the window size in order to maximize the number of RIDGES rather than the number of genes (Figure 4.13). This determined the appropriate window-size as 21. In order to identify the RIDGE members, we used a double p value approach, one primary (smaller p value) and one secondary (bigger p value). When a gene with a score higher than the primary threshold is detected, the algorithm scans -1 window size and +2 window size of that gene for genes with a score higher than the secondary threshold. We used this method after analyzing all of the regions manually which results in the observation that, using a single threshold will either miss lots of RIDGE members if the p value is low or will identify many false positives for high p

values (e.g. 0.05). Due to the mathematical nature of the algorithm, genes towards the middle of the RIDGEs get a higher score and the genes on the edges get a lower score. Because of the iceberg like appearance of these RIDGEs, we called this double p value approach the “ICEBERG” algorithm.

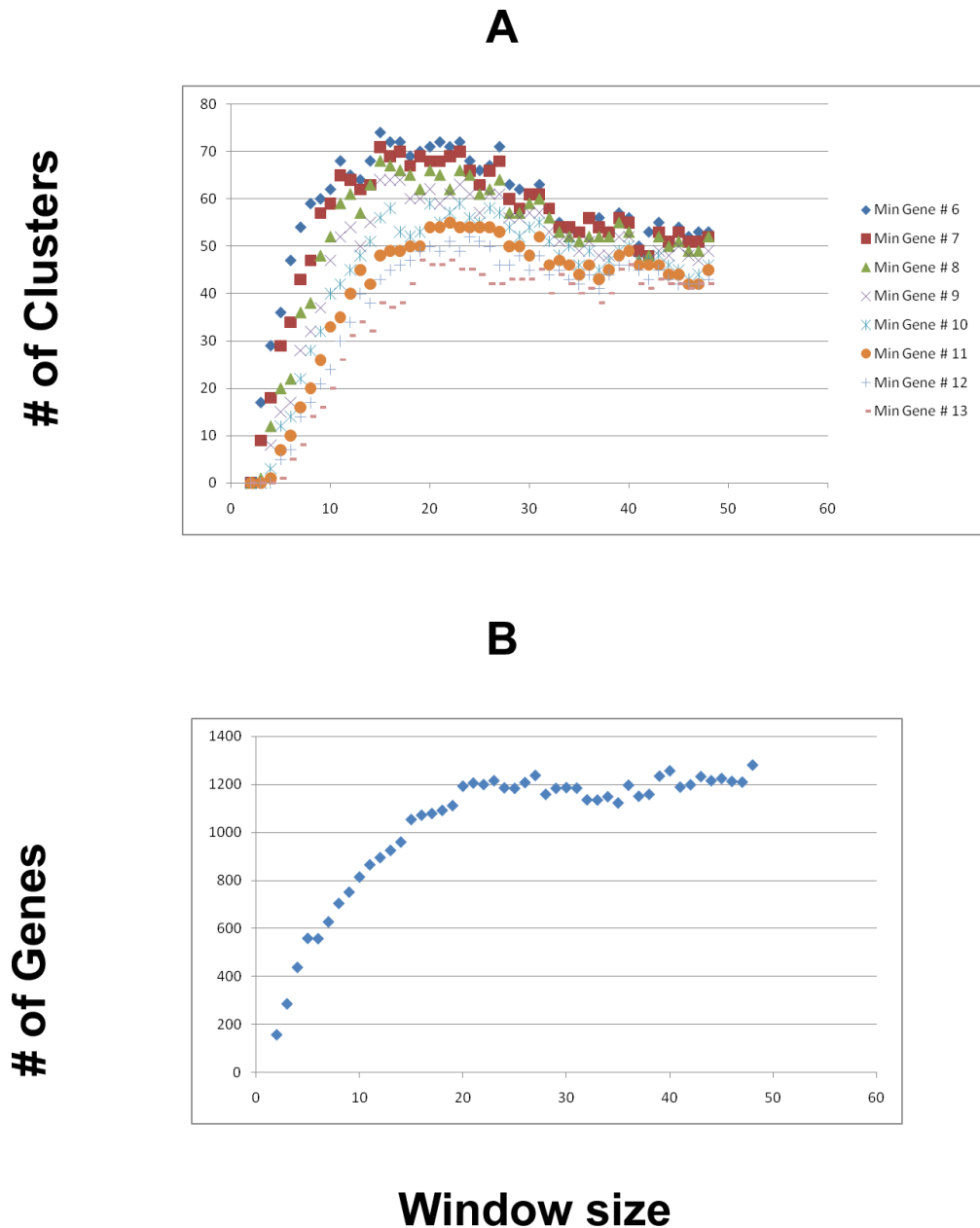


Figure 4.13. Determination of appropriate window size(s) for tumor associated RIDGEs. A: Number of clusters detected by using ICEBERG algorithm by using a primary p value cutoff of 0.002 and secondary p value of 0.05. B: Number of genes reaches a plateau around 20-25. We selected 21 as the minimum window size which gives as much information as bigger window sizes.

When using a primary p value of 0.002 and a secondary p value of 0.05; the ICEBERG algorithm identifies 82 TA-RIDGES and 81 NA-RIDGES 45 of which overlap with each other. In this study, our focus will be TA-RIDGES unless otherwise stated. Figure 4.14A presents an overall view of 2 representative chromosomes. An important finding was that 41 per cent (21 of 51) and 83 per cent (25 of 30) of the genes on the Y chromosome were identified as TA-RIDGE members and NA-RIDGE members, respectively (Figure 4.14B). Paralog clusters which are results of gene duplication, gene conversion and transposon mediated mechanisms are common in the human genome. We observed several paralog clusters as being TA-RIDGES and NA-RIDGES, likely to be derived from gene duplication events. In order to summarize the paralog cluster concept in relation to RIDGES, we defined a “family gene” concept based on the similarity of gene symbols, as a gene which has at least one family member within the same RIDGE. According to the “family gene” index distribution (Figure 4.15), 20 per cent of the TA-RIDGES have more than 50 per cent “family gene”s. 40 per cent of the RIDGES did not have any “family gene”. 3 “family gene” dense examples and 3 non-family TA-RIDGES are presented in Figure 4.14B.

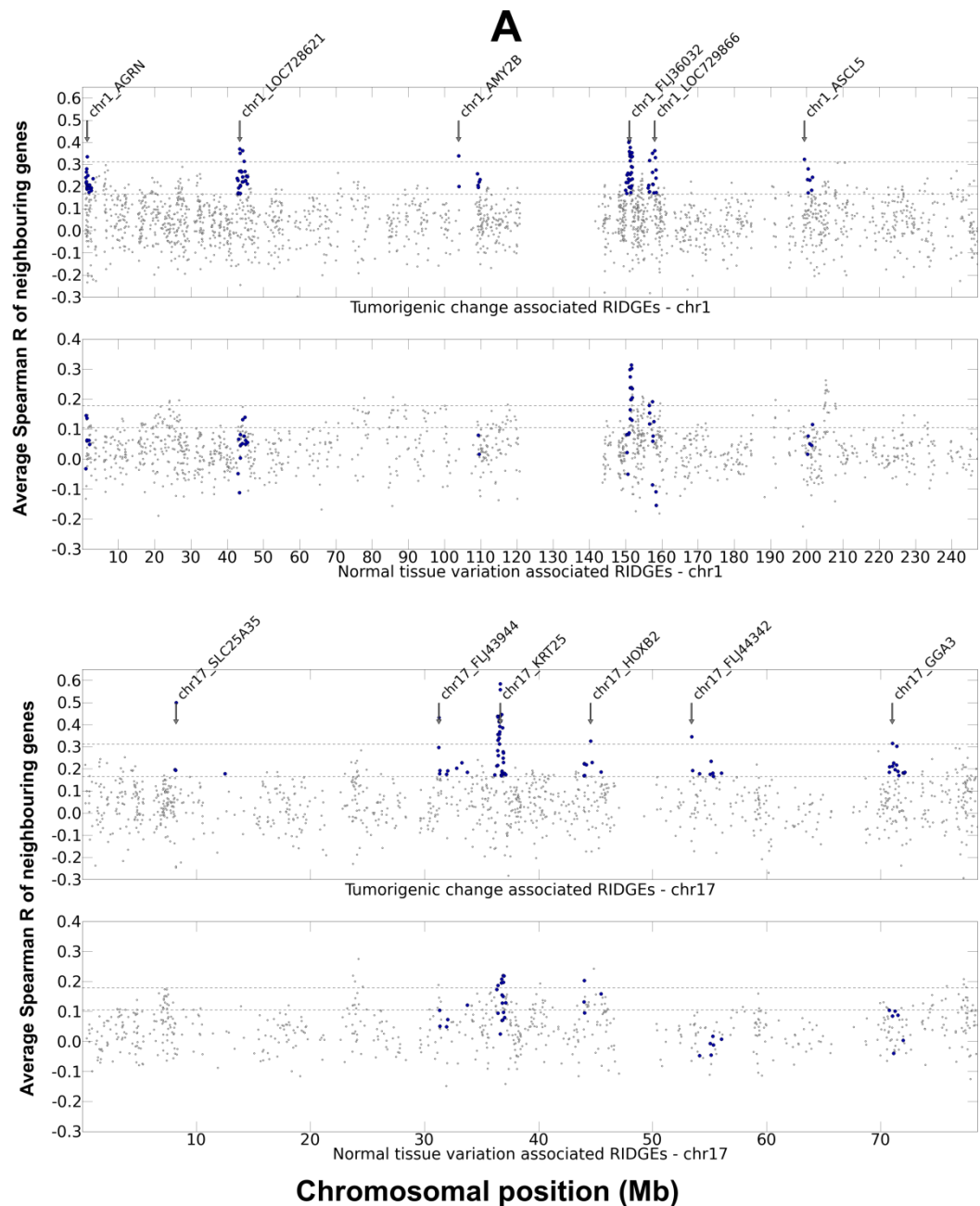


Figure 4.14. Overview and close up of tumor associated and normal associated RIDGEs by using gene neighborhood correlation score. A: 2 representative chromosomes that summarize the overall appearance of RIDGEs. All dots represent gene neighborhood correlation scores (average of R s with ± 21 neighboring genes). Blue dots were identified as being member of a tumor-associated RIDGE by the ICEBERG algorithm. The RIDGEs were named by using the chromosome they reside in and the most 5' member of the RIDGE. RIDGEs are annotated with this naming convention in the figure. Dashed lines represent the secondary p value cutoff, primary p value cutoff and promoter similarity cutoff respectively from low to high in the y-axis. Y-axis represents position in megabases (MB).

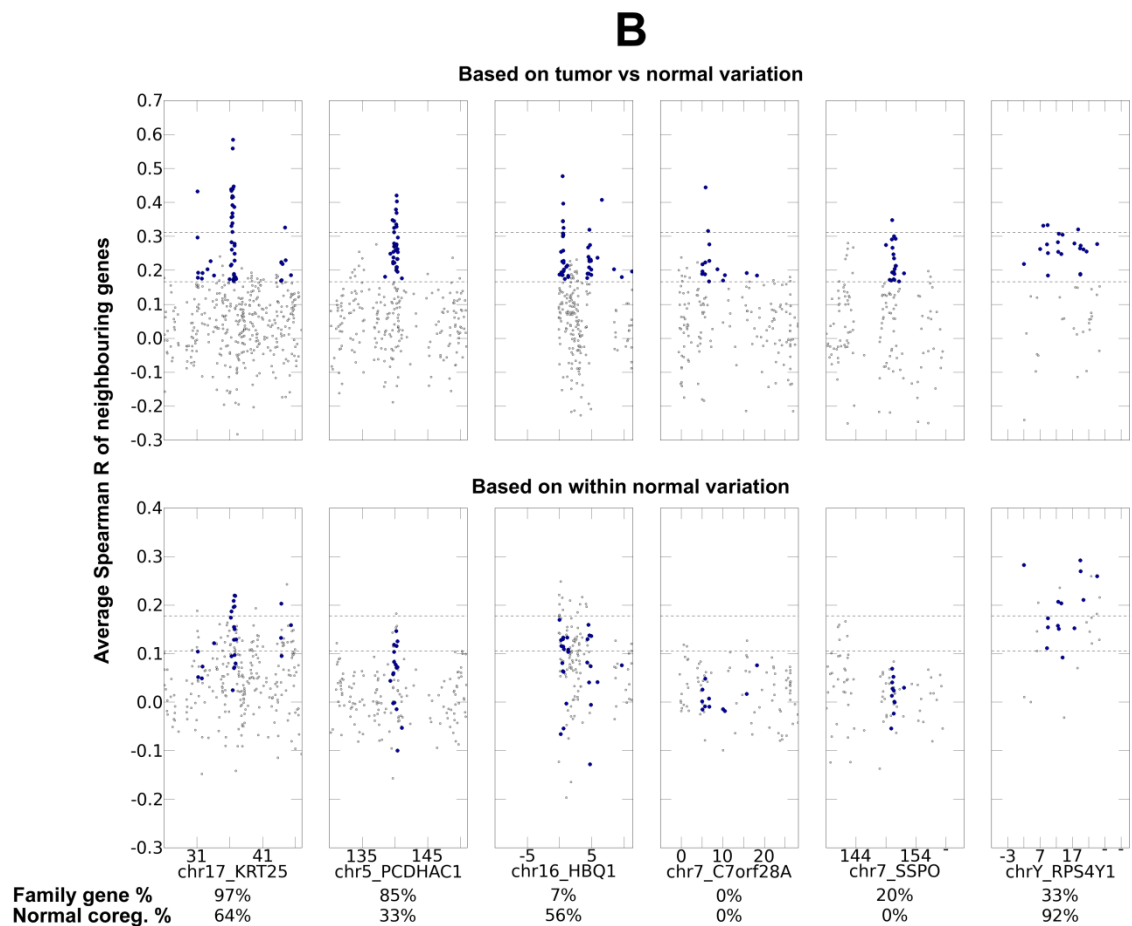


Figure 4.14. (continued) B: 6 representative TA-RIDGES of varying “family gene” per cent and being normal-associated per cent (Normal coreg. per cent). Axese are the same as in A. Chr5_PCDHAC1, Chr17_Krt25 are examples of “family gene” dense RIDGES, which comprise 28 per cent of all TA-RIDGES. Chr16_HBQ1 is a non-family RIDGE which is common to tumor and normal context. chr7_C7orf28A and chr7_SSPO are 2 examples of tumor context specific RIDGES. Almost the entire chromosome Y is co-regulated among normal tissues and also among cancer/normal transitions and identified as one RIDGE, chrY-RPS4Y1.

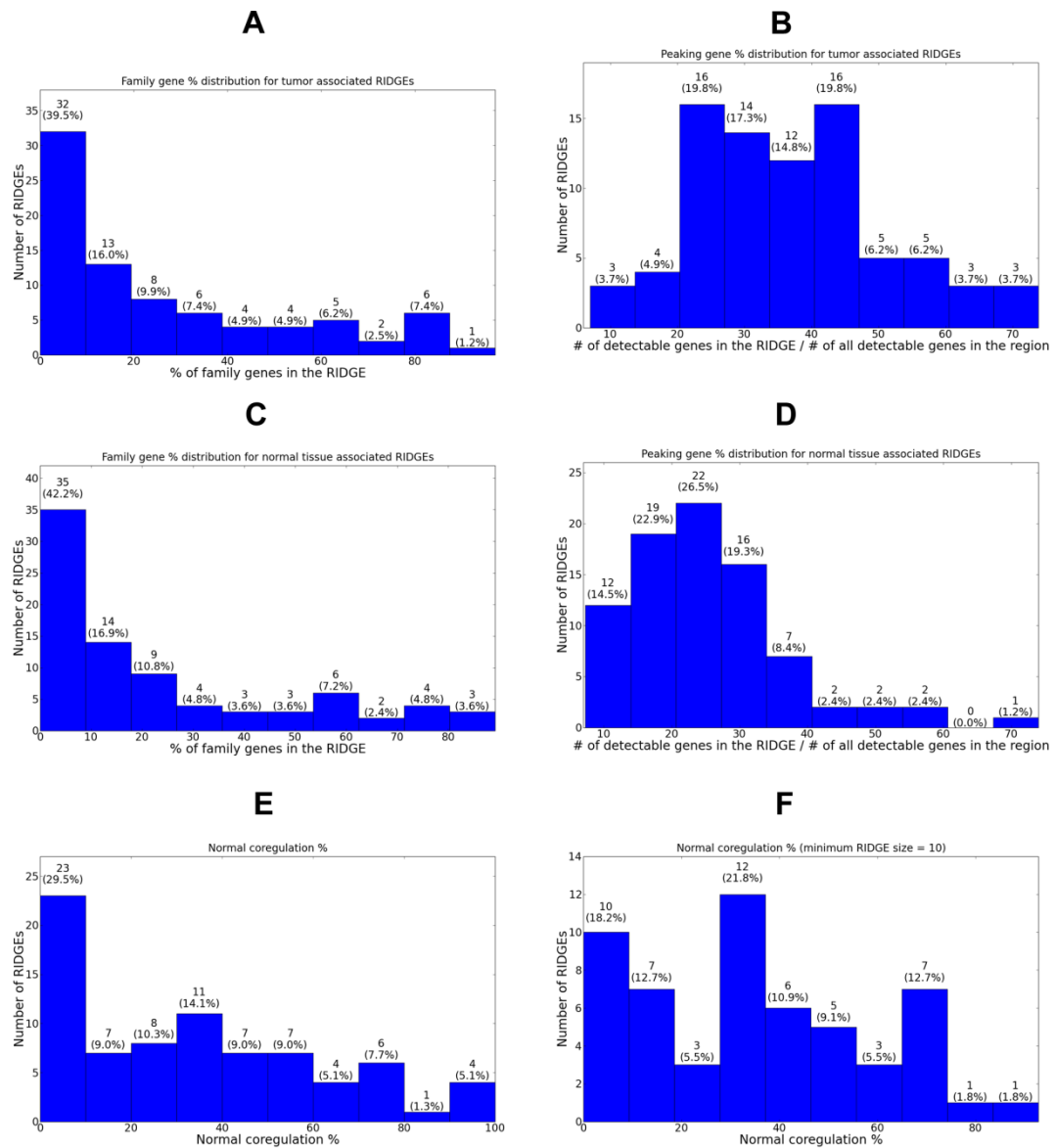


Figure 4.15. Distributions; A: “family gene” (a gene which has at least one family member within the same RIDGE) density over all tumor associated RIDGEs. B: Peaking gene (a gene identified to be as significantly co-regulated with its neighbors) density calculated as the ratio of peaking genes / all genes in the same genomic region. C and D are “family gene” and peaking gene distributions for normal tissue associated RIDGEs. E: Normal co-regulation per cent distribution among all TA-RIDGEs. Normal co-regulation per cent is the percentage of genes at a TA-RIDGE which scores above the $p=0.05$ threshold at the corresponding region from normal tissue matrix. F: Normal co-regulation per cent distribution among TA-RIDGEs which have 10 or more genes (detected as the member of the RIDGE not in the whole RIDGE region).

4.2.4. Differential expression at TA-RIDGES

In order to make TA-RIDGES workable, we used several methods to delineate differential expression in a single tumor-normal comparison and common differential expression among multiple tumor-normal comparisons and we made this data analyzable by using the online tool: www.oncoreveal.org.

By using randomizations we developed two different scores (See Methods for details). The first score assigns each tumor type a p value which is the random chance of having the same number of genes differentially expressed in a region of the same size when the same number of differentially expressed genes is randomly distributed over the genome. The second p value measures the significance of the similarity score between two different tumor-normal comparisons, based on the same randomization principle. Similarity score is defined as the binary similarity between 2 different tumor-normal comparisons, which is # of common differential expression events / # of total differential expression events. Although co-regulation does not necessarily mean differential expression, the vast majority (80 of 82) TA-RIDGES are over or under-expressed in one type of tumor when compared to the corresponding normal.

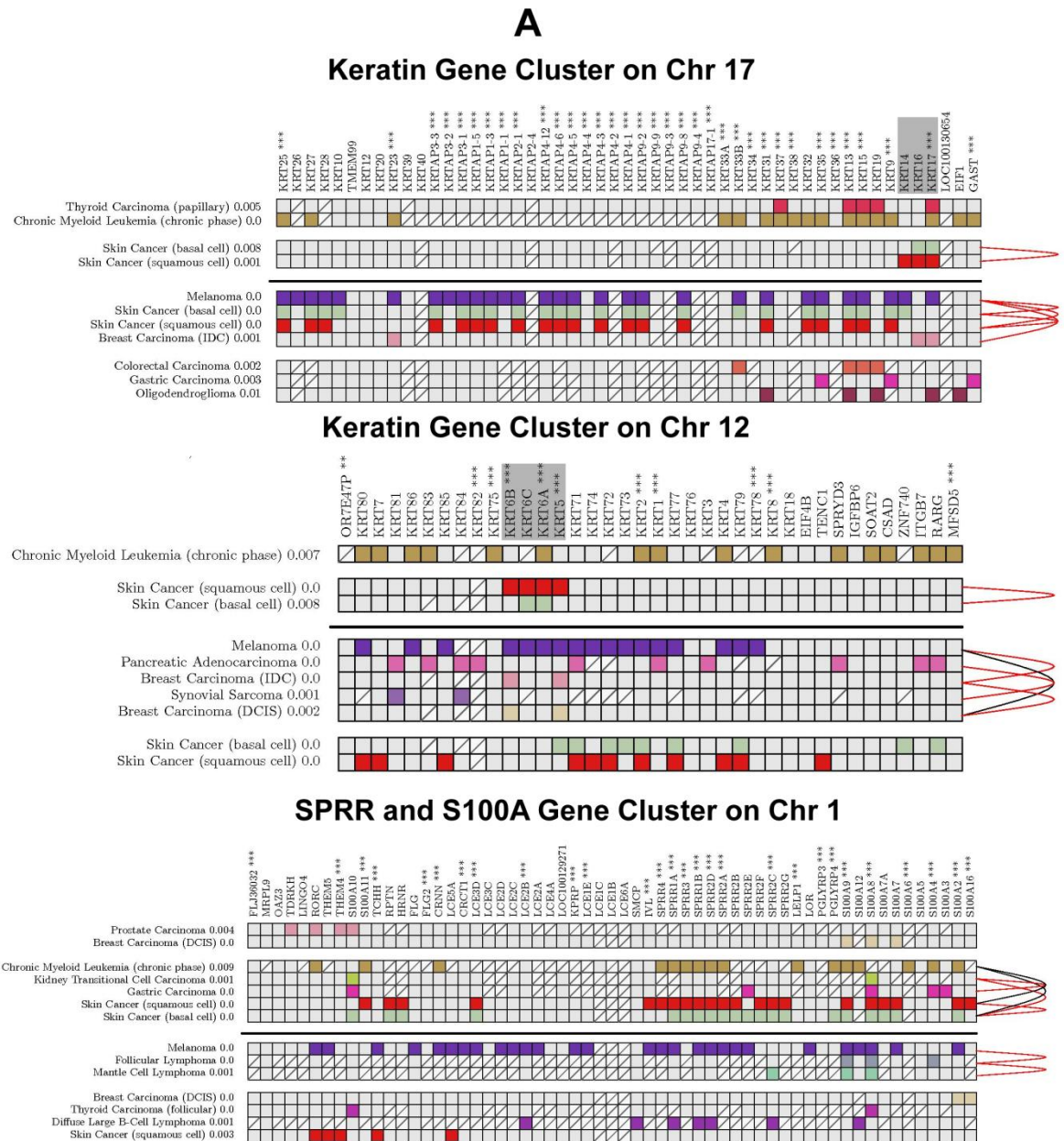


Figure 4.16. Regulation of Keratin, SPRR and S100A gene families in skin cancer and its relation to melanoma metastasis. A: KRT5, KRT6A, KRT6B and KRT6C (on chr12, shaded gray) and KRT14, KRT16 and KRT17 (on chr17, shaded gray) are up-regulated in squamous cell skin cancers and 4 of them are also up-regulated in basal cell skin carcinoma. Each colored box represents differential expression with the moderate COGENT merged dataset (q value < 0.05 for microarray, p value < 0.02 for SAGE). Cross-dashed boxes represent missing data. The p values for over-representation of differentially expressed genes in a RIDGE, determined by randomization, are stated beside the tumor name legends. Only cancers which are significantly ($p < 0.01$) represented are shown in the figures. Close up of RIDGEs with different filters can be visualized at www.oncoreveal.org. The arcs between rows represent significant ($p < 0.002$) co-differential expression events.

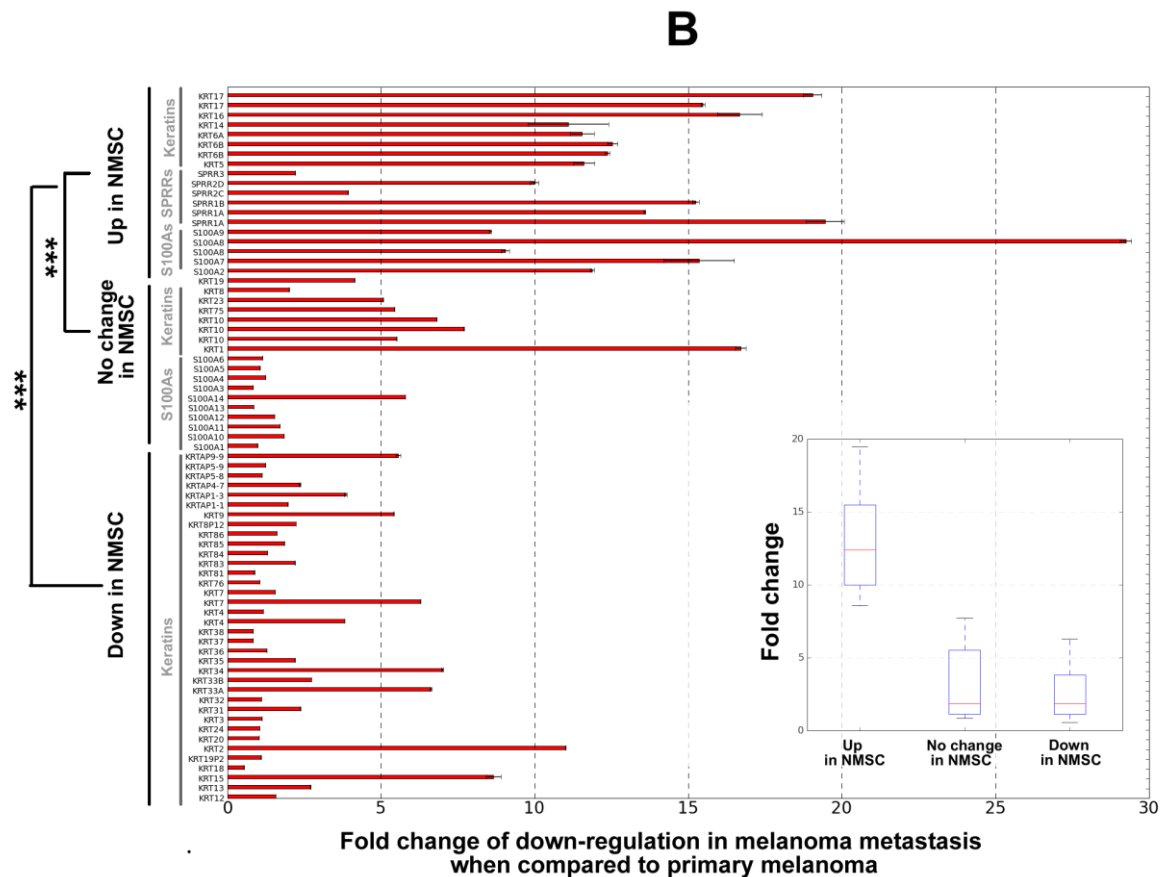


Figure 4.16. (continued) B: SPRR genes, S100A genes and Keratin genes which are over-expressed at squamous cell skin cancer are among the top down-regulated genes in melanoma metastasis when compared to primary melanoma. Inset: Distribution of fold change in different classes *** indicates $p < 0.001$ by Mann Whitney U test.

Analyzing genes in the genomic context brings in a lot of new insight. We present one example finding which points to a subgroup of keratin genes (KRT5, KRT6 (A, B and C), KRT14, KRT16 and KRT17). There are 2 big keratin clusters on the genome on chr12 and chr17 almost all of which are down regulated in squamous cell, basal cell skin cancer and melanoma when compared to corresponding normal tissues. The new subclass of keratin genes which is present on both chr12 (KRT5, KRT6) and chr17 (KRT14, KRT16 and KRT17) are over-expressed in non melanoma skin cancers (NMSCs) and among top down-regulated genes in melanoma metastasis. The expression pattern of this subclass of keratins is very similar to SPRR and S100A gene families which reside together in the epidermal differentiation complex (EDC) at 1q21 (Figure 5A and 5B). The TA-RIDGE which corresponds to the epidermal differentiation complex that contains SPRRs and

S100A “family gene”s are mostly down-regulated in melanoma and up-regulated in the other 2 types of skin cancers. All 3 types of gene families (keratins, SPRRs and S100As) are over-expressed in CML (Chronic Myeloid Leukemia), as well. Therefore, the genes which are over-expressed at NMSCs in comparison to normal skin are significantly more down-regulated in melanoma metastasis in reference to primary melanoma when compared to the other keratin, SPRR or S100A genes which are down-regulated or not altered at NMSCs (Figure 5B). The newly identified subclass of 5 keratin genes does not fall into the same functional category. KRT5 and KRT6 are type II epithelial keratins whereas KRT14, KRT16 and KRT17 are type I epithelial keratins [84]. In addition to identifying this subclass of keratin family, RIDGE analysis shows that over-expression at NMSCs is confined to S100A and SPRR families in the whole EDC whereas down-regulation in melanoma is homogenously distributed among the EDC (Figure 5A, lower panel). Overall, this suggests a functional role for these newly out-pointed 5 keratin genes, SPRRs and S100A “family gene”s in NMSC and melanoma metastasis.

MHC Class II clusters (Figure 4.17) are over-expressed in a wide-variety of malignancies including hematological, CNS and epithelial tumors. On the other hand, a clear under-expression is observed in 3 different types of lymphomas. Histone 1 cluster (chr6_C6orf62 - Figure 4.18) is over-expressed in ovarian adenocarcinoma, hepatocellular carcinoma, lung adenocarcinoma and under-expressed in esophageal adenocarcinoma, chronic myeloid leukemia, diffuse large B-Cell lymphoma and Burkitt lymphoma. S100A RIDGE and MHC Class II RIDGE have been functionally linked to tumorigenesis whereas Histone 1 RIDGE is not.

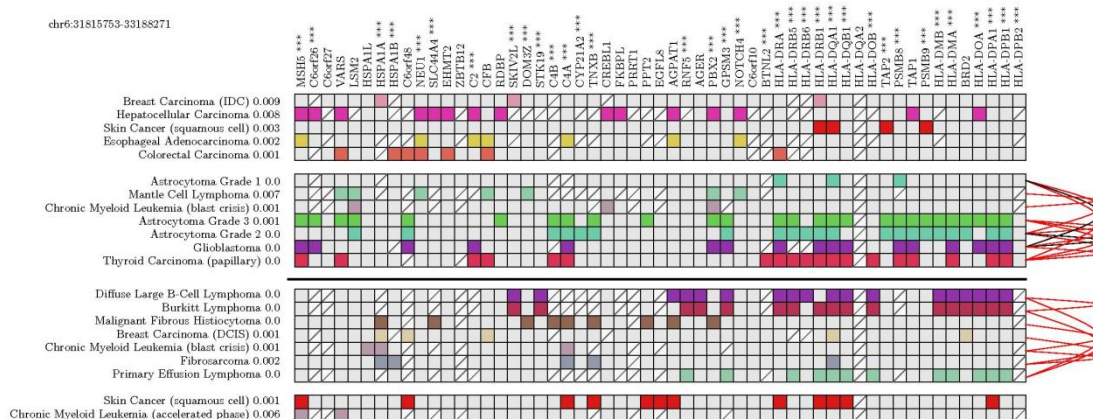


Figure 4.17. MHC Class II RIDGE. Only significantly differentially expressed tumor types are displayed. Full version of all RIDGEs are browsable at www.oncoreveal.org

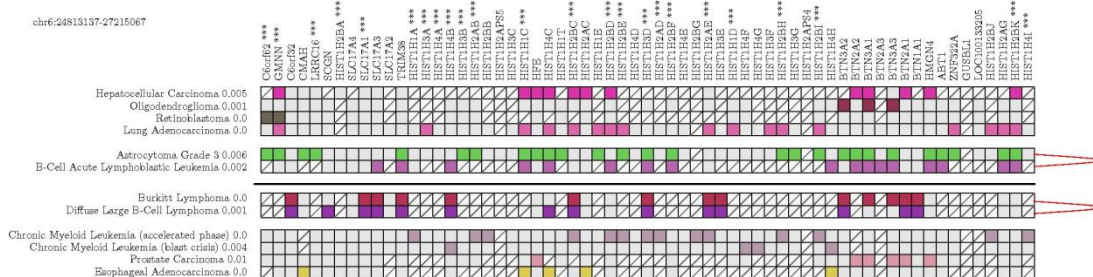


Figure 4.18. Histone 1 RIDGE. Only significantly differentially expressed tumor types are displayed. Full version of all RIDGEs is analyzable at www.oncoreveal.org

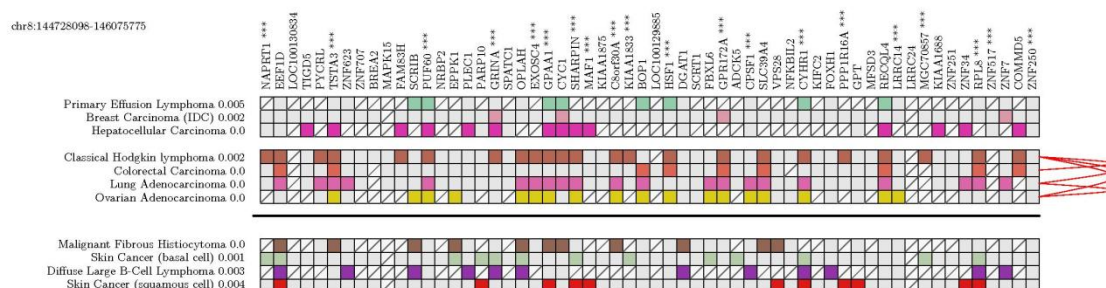


Figure 4.19 Chr8_NAPRT1, an example of a non-paralog RIDGE which is differentially expressed in a wide variety of tumors. Only significantly differentially expressed tumor types are displayed. Full version of the RIDGEs is analyzable at www.oncoreveal.org

Due to the fact that there are not many studies on co-regulated regions on the genome, it is almost impossible to summarize the non-paralog clusters. However, we also observed significant differential expression of tumors at these non-paralog TA-RIDGES, as well. For instance, Chr8_NAPRT1 at the 3' end of the 8th chromosome is mainly over-expressed in a variety of tumors including epithelial, CNS and blood tumors (Figure 4.19).

When we analyze the frequency of significant ($p < 0.01$) co-regulations among cancers in all 81 RIDGES; as an expected result, different grades or states of same cancers (e.g. astrocytoma grade 2 vs. astrocytoma grade 3 or chronic phase and accelerated phase of CML) were among top co-regulated cancer types. In addition, different types of blood cancers were also among the top co-regulated cancer types which suggest a common mechanism of action at the genomic level for hematological malignancies (Table 4.19).

4.2.5. GC Content is a major factor affecting presence of RIDGES

43 per cent of the TA-RIDGES correspond to high GC content genomic regions (H2 and H3 isochors; GC per cent > 46 per cent) which corresponds to a significant 1.72 fold enrichment when compared to non-RIDGE regions of varying window sizes (See Table 4.20; chi-square p value = 0.019 when compared to non-RIDGES with window size of 19 and $p = 0.0104$ when compared to non-RIDGES with window size of 27). The enrichment was 3.8 and 7 fold for TA-RIDGES and non-paralog TA-RIDGES when highest GC content class is considered (chi-square p values 0.0129 and 0.0001). The same enrichment was seen for NA-RIDGES as well (Table 4.20).

Table 4.19A. Co-regulation (p value < 0.01) frequencies for over-expressed genes according to COGENT data with q value filter of 0.05 and SAGE p value filter of 0.02

Co-regulated tumors	#
Astrocytoma Grade 2//Astrocytoma Grade 3	3
	7
Acute Myeloid Leukemia//B-Cell Acute Lymphoblastic Leukemia	2
	7
Chronic Myeloid Leukemia (chronic phase)//Chronic Myeloid Leukemia (accelerated phase)	1
	5
Astrocytoma Grade 3//Glioblastoma	1
	2
Oligodendroglioma//Glioblastoma	1
	1
Burkitt Lymphoma//Diffuse Large B-Cell Lymphoma	1
	0
Astrocytoma Grade 2//Glioblastoma	9
Diffuse Large B-Cell Lymphoma//Multiple Myeloma	9
Skin Cancer (basal cell)//Skin Cancer (squamous cell)	8
Fibrosarcoma//Malignant Fibrous Histiocytoma	8
Breast Carcinoma (IDC)//Breast Carcinoma (DCIS)	8
Malignant Fibrous Histiocytoma//Lung Adenocarcinoma	7
Malignant Fibrous Histiocytoma//Diffuse Large B-Cell Lymphoma	7
Burkitt Lymphoma//Multiple Myeloma	7
Lung Adenocarcinoma//Breast Carcinoma (IDC)	6
Burkitt Lymphoma//Lung Adenocarcinoma	6
Astrocytoma Grade 3//Astrocytoma Grade 2	6
Ovarian Adenocarcinoma//Lung Adenocarcinoma	6
Diffuse Large B-Cell Lymphoma//Burkitt Lymphoma	6
Astrocytoma Grade 3//Colorectal Carcinoma	5
Classical Hodgkin lymphoma//Breast Carcinoma (IDC)	5
Melanoma//Skin Cancer (basal cell)	5
Astrocytoma Grade 3//Classical Hodgkin lymphoma	5
Astrocytoma Grade 2//Ependymoma	5
Diffuse Large B-Cell Lymphoma//Primary Effusion Lymphoma	5
Esophageal Adenocarcinoma//Mantle Cell Lymphoma	5
Burkitt Lymphoma//Malignant Fibrous Histiocytoma	5
Ovarian Adenocarcinoma//Hepatocellular Carcinoma	5

Table 4.19B. Co-regulation (p value < 0.01) frequencies for under-expressed genes according to COGENT data with q value filter of 0.05 and SAGE p value filter of 0.02

Co-regulated tumors	#
Astrocytoma Grade 2//Astrocytoma Grade 3	52
Oligodendroglioma//Glioblastoma	34
Burkitt Lymphoma//Diffuse Large B-Cell Lymphoma	26
Acute Myeloid Leukemia//B-Cell Acute Lymphoblastic Leukemia	23
Astrocytoma Grade 3//Glioblastoma	22
Astrocytoma Grade 2//Glioblastoma	20
Skin Cancer (basal cell)//Skin Cancer (squamous cell)	18
Burkitt Lymphoma//Primary Effusion Lymphoma	16
Chronic Myeloid Leukemia (chronic phase)//Chronic Myeloid Leukemia (accelerated phase)	14
Astrocytoma Grade 2//Ependymoma	12
Glioblastoma//Ependymoma	12
Burkitt Lymphoma//Multiple Myeloma	12
Astrocytoma Grade 3//Ependymoma	10
Fibrosarcoma//Malignant Fibrous Histiocytoma	9
Diffuse Large B-Cell Lymphoma//Primary Effusion Lymphoma	9
Oligodendroglioma//Astrocytoma Grade 2	9
Oligodendroglioma//Anaplastic Oligoastrocytoma	8
Melanoma//Skin Cancer (squamous cell)	7
Oligodendroglioma//Astrocytoma Grade 3	7
Ovarian Adenocarcinoma//Lung Adenocarcinoma	7
Multiple Myeloma//Primary Effusion Lymphoma	7
Hairy Cell Leukemia//Primary Effusion Lymphoma	6
Oligodendroglioma//Ependymoma	6
Anaplastic Oligoastrocytoma//Glioblastoma	5
Melanoma//Skin Cancer (basal cell)	5
Mantle Cell Lymphoma//Primary Effusion Lymphoma	5
Breast Carcinoma (IDC)//Breast Carcinoma (DCIS)	5
Burkitt Lymphoma//Classical Hodgkin lymphoma	5

Table 4.20. GC rich regions are more likely to be co-regulated. Values show frequency of RIDGE type among all RIDGEs. Letter designation represents isochor classification by GC content. L1 < 37 per cent, 37 per cent <= L2 < 41 per cent, 41 per cent <= H1 < 46 per cent, 46 per cent <= H2 < 53 per cent, H3 => 53 per cent

	L1	L2	H1	H2	H3
TA-RIDGEs	6,2	21,0	29,6	37,0	6,2
Non-”family gene” TA-RIDGEs (“family gene” per cent < 20)	8,9	22,2	26,7	31,1	11,1
NA-RIDGEs	3,6	27,7	28,9	30,1	9,6
NonRIDGEs (window size 19)	5,9	29,8	37,0	25,5	1,8
NonRIDGEs (window size 27)	5,2	33,7	35,7	24,1	1,4

4.2.6. Online re-analysis and availability of the data: www.oncoreveal.org

We prepared a PHP/MySQL based online service in order to make most of the data presented in the second part of this thesis analyzable and filterable. Via www.oncoreveal.org users can analyze COGENT-SAGE, COGENT-microarray&SAGE and RIDGEs interconnected with the COGENT browser module (Figure 4.20). Several filtering, sorting and exporting options are available at the web service, as well as user data overlay. A unique property of Oncoreveal is that users can create complex gene signatures by asking complex questions. For instance, one can get the gene list which is over-expressed specifically (only) at least in 5 glial cell tumors and under-expressed in none of the tumors. Or, one can easily find cell cycle genes which are over-expressed in melanoma but not in other types of skin tumors. In addition to work on cancer types and genes/tags, users can visualize RIDGEs under broad chromosome view (Figure 4.14B) or differential expression view (Figures 4.17-19).

www.oncoreveal.org

	Cogent-SAGE analysis	COGENT microarray + SAGE analysis	RIDGE analysis
User Data Overlay	Short SAGE tags DGED result	Entrez GENE IDs Entrez GENE Symbols	NA
Cancer Type Filtering	26 tumor types Organ/Tissue source based (N = 18) Major tumor types Different thresholds	49 tumor types Organ/Tissue source based (N = 27) Major tumor types Different thresholds	49 tumor types
Tag mapping	No mapping SAGE Genie or SAGE Map Consensus tag mapping	NA	NA
Filtering	Gene Ontology, KEGG Folds of differential expression Genomic location		Chromosome based
Other		Selection of fields to be visualized Sorting (with different options) Data export	Broad chromosome view Focused view

Figure 4.20. Online service; www.oncoreveal.org. Users can analyze COGENT SAGE, COGENT microarray + SAGE and RIDGEs by using Oncoreveal. Short SAGE tags or DGED results for COGENT-SAGE and Entrez Gene IDs or symbols can be used as user input. Users can filter genes by the tumor-normal comparisons genes altered, or the number of total changes. Finding common changes as well as tumor-type specific changes is possible. Several different filtering options from commonly used data sources such as gene ontology and different visualization, sorting and export options are possible.

5. DISCUSSION

5.1. Analysis of the Wnt/ β -catenin/TCF4 pathway using SAGE, genome-wide microarray and promoter analysis: Identification of BRI3 and HSF2 as novel targets

In this part of the study it was shown that mutant β -catenin-transfected TCF4-hyperactive Huh7 cells lead to earlier and larger tumors in nude mice. SAGE, genome-wide microarray analysis and in silico promoter analysis were used in parallel in order to reveal transcriptomic changes and putative Wnt targets. BRI3 and HSF2 were shown to be novel targets of Wnt/TCF/ β -catenin signaling.

We compared SAGE and microarray data from identical total RNA by using a tendency analysis approach. Other comparison methods are available such as scaling two data structures into the same dimension and comparing the values, or comparing the folds by using direct values or logs [85]. However, in our opinion, the best method to understand if these methods are giving similar results is to check the tendency of the resulting differences, because neither of the methods gives well correlated fold ratios in comparison to more exact methods such as quantitative real-time PCR. On the other hand, SAGE and microarray sense differential expression of different genes. With new mass-sequencing technologies, it will be feasible to routinely use them in parallel in the near future. Indeed, there is a recent and good example of using sequencing and hybridization based methods in parallel for identifying nucleosome distribution on the genome of T cells [86]. In terms of sensing presence of RNAs, microarray sensed more genes when compared to SAGE, mainly originating from the tag mapping problem, definition of thresholds and the limiting total tag counts in SAGE. With the rapidly evolving new sequencing technologies, counting much higher tag numbers with SAGE could reach or even surpass microarray sensitivity.

We also integrated in silico promoter analysis to our SAGE-microarray merged expression data. This enabled us to narrow down the target list rapidly. We did the same promoter analysis in several different datasets other than the merged dataset stated in the Results section. The results somewhat varied when using SAGE data or microarray data alone, as expected from the overall differences observed when comparing the two

techniques. When the most consistent results as represented by the union dataset are considered; regardless of used motif (TCF_{short} or TCF_{long}) genes that are differentially expressed have significantly different TBE patterns when compared to promoters of non-differentially expressed genes. Identifying a putative pattern which is different for down-regulated genes when compared to up-regulated genes was promising for elucidating how a single stimulus (hyperactive β -catenin) can lead to up and down regulation simultaneously.

Combined analysis pointed out potential novel Wnt-targets among which BRI3 and HSF2 were confirmed through further experiments described in this study. Identification of BRI3, previously implicated as a marker of colon cancer progression, as a Wnt target reemphasizes the role of the Wnt pathway in cancer. Although there might be many different explanations for the role of HSF2 as a Wnt target; one possible path is through PP2A. HSF2 regulates PP2A function, a multiunit phosphatase; by binding on the PR65 unit of the enzyme complex [87]. Whether PP2A inhibits or induces Wnt pathway remains controversial, it is well known that it regulates Wnt pathway through GSK3 β and AXIN [4]. This might be a starting point for further analysis of HSF2 involvement in the Wnt pathway. On the other hand, to our knowledge, this is the first report which proposes a possible connection between stress-response regulation and Wnt/TCF/ β -catenin signaling through HSF2.

We showed that high TCF activity Huh7 cells lead to earlier setup and larger tumors in nude mice and at least two genes were identified as novel downstream targets of the Wnt pathway. We hypothesized that post-tumor cells might resemble Huh7 stem cells more closely than pre-tumor cells. The overall pattern suggested that the *in vivo* passage somewhat increased stem-cell-like properties in the transcriptome profile of both Huh7-S33Y and Huh7-vec cells rescued post-tumor, but as the side-population has been shown to maintain a steady-state ratio with the differentiated population *in vitro*, no further speculation is warranted on the origin cells of the enhanced profiles. Concordant with the stemness characteristics which will facilitate invasion and metastasis of the tumor, however, several of the post-tumor-enhanced genes were plasma membrane proteins which have roles in cell-cell communication. Both new targets appeared to be similarly active in the post-tumor profiles, consistent with Wnt pathway involvement in stem cell-like behavior [5]

5.2. Meta analysis of cancer gene expression profiles reveals cancer genes, SAGE tags and co-regulated regions on the genome

In this part of the study, we present the widest meta-analysis using SAGE and microarray data in conjunction. There are 2 different procedures we applied to make this possible, one of them is the consensus tag mapping and the other one is the rank selection which we used for filtering Solexa-SAGE data. Due to the low level of agreement between SAGE tag mapping algorithms (see Methods), especially for short SAGE tags, using consensus tag mapping rather than using any of the algorithms alone eliminated significant amount of error. By using consensus tag mapping 68 per cent and 80 per cent of all possible Entrez gene IDs are mappable by using short and long SAGE tags. Overall, consensus tag mapping provides acceptable coverage by minimizing false positive tag mappings. Rank selection on the other hand appeared as a necessity due to high amount of data produced by new generation sequencing techniques.

We found that multi-cancer genes, genes which are differentially expressed in more types of tumors, ranks higher in single types of tumor-normal comparisons and are studied/published more (tumor-associated). Besides, being published or not is dependent on the rank of a gene's expression change. In addition to pointing to the importance of multi-cancer genes, we identified several not-yet-tumor-associated multi-cancer genes and made them analyzable. Multi-cancer genes are often used as drug targets (e.g. EGFR) due to the fact that the real challenge comes during validation of the drug in the clinic. Therefore, once a drug is clinically feasible, it can then be tried also in other types of diseases. This makes over-expressed multi-cancer genes specifically important drug targets and under-expressed multi-cancer genes important gene therapy targets. By showing that there are still many multi-cancer genes not yet recognized to be tumor-associated, we point to a tendency towards working with multi-cancer genes. In other words, working and publishing multi-cancer gene behavior is not dependent on the knowledge of gene expression data; otherwise we would have expected to see at least some publications about the non-tumor-associated multi-cancer genes.

It should be acknowledged that the best current example of analyzing different cancer related microarray data is presented by Oncomine software (www.oncomine.org).

They allow analysis of single gene expression levels or analysis of multiple microarray based tumor-normal comparisons. Oncoreveal online service, on the other hand, allows unlimited size of gene list overlay. In addition, it allows SAGE data analysis in addition to the SAGE-microarray merged data analysis. Besides, Oncoreveal makes the RIDGEs workable by basic filtering options as a starting point and detailed visualization options. In our opinion, having most of the data presented in this study available, analyzable and filterable online makes this data as open and transparent as possible.

While defining being associated with tumorigenesis we used Entrez geneRIF database which holds author curated summary sentences about articles in relation to Entrez Gene entries. This database will have a degree of false negatives. However, in the release we used (November 2008), 10000 of the human genes had at least one geneRIF (ca half of the detectable gene set used in this study). Besides, using geneRIFs instead of scanning all titles/abstracts of articles minimizes false associations of genes to tumorigenesis which might occur due to common non-specific usage of tumor/cancer keywords in abstracts. However, one should still expect to find some genes which have been related to tumorigenesis with few articles but not supported by any geneRIF. But, it is also likely that, if a gene is over-expressed in 15 types of tumors and it has no cancer geneRIFs, it is probably worth going further with that gene.

Middle stringency ($q < 0.05$, sage $p < 0.02$) COGENT points to 64 different genes which are associated with endoplasmic reticulum. Although endoplasmic reticulum is not well known to be involved in tumorigenesis, it is unlikely that any single organelle will not be related with tumorigenesis. Indeed, endoplasmic reticulum and related genes have been directly or indirectly related to multiple cancer related pathways such as apoptosis or angiogenesis [88].

Co-regulated regions have been studied in the literature considering various contexts. However, this is the first study which aims to study TA-RIDGEs in a multi-cancer context by analyzing differential expression at these RIDGEs. One concern about these RIDGEs is their identification and gene assignment. Using different expression matrixes will lead to identification of different RIDGEs. For instance, 27 of these 81 TA-RIDGEs were also identified to be RIDGEs in the Stransky et al. study (at least 1 common

gene was commonly identified as being significantly co-regulated). The common RIDGEs included HLA-DR (MHC Class II), S100A family cluster and histone 1 cluster. There are several hypotheses which might explain the occurrence of RIDGEs, such as genomic amplifications/deletions, epi-genetic regulations and regional lacks of DNA repair. In the bladder cancer context[61] of Stransky et al, genomic amplifications/deletions detected by comparative genomic hybridization (CGH) method turned out to explain some fraction of the RIDGEs, but not all. In this study, we point to a significant association between RIDGEs and GC rich regions. This finding is consistent with previous observations which showed that GC rich regions are expressed more in normal tissues according to SAGE data [89].

Closer analysis of keratin, SPRR and S100A “family gene” RIDGEs pointed to two findings: First, 5 keratin genes (K17, K5, K6, K14 and K16) are regulated differently from other keratin genes in the contexts of skin cancers and melanoma metastasis. Second, all of the 3 gene families are down-regulated in melanoma but the genes which are up-regulated at NMSC are significantly more silenced in melanoma metastasis than the genes which are not up-regulated in NMSC. Keratins are epithelial cell markers and EDC is named after its raised activity at terminal differentiation of human epidermis. Therefore, it is expectable to observe down-regulation of these genes at tumorigenesis and tumor metastasis. On the other hand, up-regulation of the SPRR and S100A genes within EDC in NMSC suggests a functional role for them rather than reflecting the differentiation state of the tumor bulk. NMSCs are the most common skin cancers, although melanoma is more lethal not only among skin but also among all other cancer types. Melonoma is derived from melanocytes whereas squamous and basal cell skin cancers arise from keratonicytes. Keratins are intermediate filament proteins which are well established epithelial cell markers. In addition, there are lots of studies which linked keratins to tumorigenesis. Among the keratin gene family keratin 10 was shown to inhibit cell cycle progression and contribute to terminal skin cell differentiation [90]. Furthermore, keratin 17 was identified as a regulator of cell growth and protein synthesis [91]. Keratins 8 and 18 are necessary for invasion of melanoma cells [92]. In addition, keratins are starting to be associated to new diseases which suggests new functions [84]. SPRRs are keratinocyte differentiation markers [93]. SPRRs contribute to the formation of the cornified envelope of skin cells [94]. In addition, SPRRs are over-expressed under stress conditions by regulation via at least one enhancer

element,[94]. SPRR3 acts like a tumor suppressor on esophageal squamous cell carcinoma cells [95]. The underlying reason of up-regulation of SPRRs in non-melanoma skin cancer might be linked to their role in coping with stress. S100 family has been shown as over-expressed in gastric carcinoma [96]. In addition there are lots of studies which points to the importance of different S100 proteins in tumorigenesis such as S100A2 as a distant metastasis and survival predictor in non small-cell lung cancer [97], S100A4 is a well-established protein which contributes to tumor cell motility and metastasis [98]. In conclusion, although as yet not directly associated with NMSCs, the new class of keratins, SPRRs and S100A gene family are likely to be involved in NMSC tumorigenesis.

MHC Class II TA-RIDGE resides on the 6th chromosome. Although the general view about MHC Class II proteins is that they are generally not expressed in tumor cells, they are known to be over-expressed in several tumors [99] such as gliomas, and tumor cells can be recognized by CD4⁺ T cells due to the presence of MHC Class II based antigen presentation on the tumor cell [100]. To the best of our knowledge, the under-expression of MHC Class II proteins in lymphocytes is not an established fact, in spite of some minor findings such as lack of detection of HLA-DR in diffuse large cell lymphoma patients is a sign of shorter survival [101].

The histone RIDGE resides on the 6th chromosome. Histone proteins are being expressed from 4 different clusters on the human genome (Histone clusters 1-4). 66 histone genes reside at histone cluster 1. Every class of 5 different histone proteins (H1-H2A,H2B,H3 and H4) are being expressed from histone cluster 1. And according to Chr6_C6orf62 cluster mainly histone 1 (e.g.HIST1H1T) and histone 2 (e.g. HIST1H2BC, HIST1H2AC and HIST1H2BD) genes are over-expressed in cancer tissues in reference to corresponding normals. There is a considerable amount of data about the effect of post-translational modifications of histone proteins such as deacetylation or methylation which in turn epi-genetically regulate gene transcription. However, there are only few examples of histone proteins which are related to tumorigenesis directly [102]. Indeed, when the Entrez Gene RIFs are considered, none of the histone genes have a cancer geneRIF and the majority of the histone genes did not have any geneRIFs at all (38 of 47 histone genes). Therefore, although post-translational modification of histone proteins is a very important

concept, the expression of histones themselves seems varying in different tumors and might functionally be relevant.

Therefore, in all of the paralog RIDGEs exemplified, some aspects of these RIDGEs have already been verified by different studies. However a general knowledge of multi-cancer property of these RIDGEs is lacking. In the differential expression drawings and scorings of RIDGEs, we present a complete picture about these RIDGEs by providing 4 different dimensions together: genomic position, differential expression, tumor-normal comparisons and common differential expression in multiple tumor-normal comparisons.

In addition to the paralog RIDGEs, approximately half of the distribution was non-paralog RIDGEs. One unlikely possibility is that, the differential expression events at these RIDGEs are just side-effects of a previously important key tumorigenic event. Another possibility is that, only some of the genes within these RIDGEs contribute to the key processes during tumorigenic transformation. For instance, TNF is co-regulated with its neighbors (chr6_CDSN). Some other tumor-associated genes within the non-family RIDGEs (has the most cancer-geneRIFs) are RHOA, MAPK3, CYP2E1, RELA, GJA1, DAPK1, CCKBR, FASN, GPX1 and HSPB1.

5.3. Conclusions and Future Perspectives

This study started with the aim of identification of novel Wnt pathway targets in liver cancer. After the first part of the study, by combining SAGE and microarray data, promoter analysis and functional confirmation experiments we identified two targets; BRI3 and HSF2. Within the context of this study other candidate targets are also being explored by our laboratory members. Therefore, this initial study served as a core analysis of high throughput data which enabled further routes.

In the second part of the study; the initial motivation was to screen the Wnt target list with a tumorigenesis point of view. Indeed, this integration helped us to choose several molecules such as BRI3 and several others (e.g. C17orf45) which are candidate Wnt targets and candidate cancer genes. In addition to this integration this meta-analysis had very unique properties such as being the first SAGE and microarray merged meta-analysis, first to relate multi-cancer with being tumor-associated and first to work on multi-cancer RIDGEs. The cancer gene expression data meta-analysis reveals many new multi-cancer genes which are likely to be functionally important in tumorigenesis, independent from the Wnt target identification process. As presented in the Results section, there are lots of genes which are over/under-expressed in many types of tumors when compared to corresponding normals, and have not been studied before. In addition, we present the first multi-cancer RIDGE approach. Although, studying on these regions is hard due to high number of genes and putative regulatory processes on these regions, we zoomed on these regions in terms of differential expression in tumor-normal comparisons. This enabled identification of a new class within the keratin family which probably functions in NMSC tumorigenesis and melanoma metastasis. In addition, with the developing high throughput approaches in every experimental field such as PCR, sequencing, or immunohistochemistry it will also be possible to work on the multi-cancer RIDGEs.

Finally, we make all of our data available and workable at www.oncoreveal.org. Sharing our findings via this service with the scientific community will help expansion of our findings by other researchers.

REFERENCES

1. Clevers, H., Wnt/beta-catenin signaling in development and disease. *Cell*, 2006. Vol. **127**(3), pp. 469-80.
2. van Noort, M. and H. Clevers, TCF transcription factors, mediators of Wnt-signaling in development and cancer. *Dev Biol*, 2002. Vol. **244**(1), pp. 1-8.
3. Brabletz, T., A. Jung, and T. Kirchner, Beta-catenin and the morphogenesis of colorectal cancer. *Virchows Arch*, 2002. Vol. **441**(1), pp. 1-11.
4. Polakis, P., The many ways of Wnt in cancer. *Curr Opin Genet Dev*, 2007. Vol. **17**(1), pp. 45-51.
5. Reya, T. and H. Clevers, Wnt signalling in stem cells and cancer. *Nature*, 2005. Vol. **434**(7035), pp. 843-50.
6. Willert, K., et al., Wnt proteins are lipid-modified and can act as stem cell growth factors. *Nature*, 2003. Vol. **423**(6938), pp. 448-52.
7. Fodde, R. and T. Brabletz, Wnt/beta-catenin signaling in cancer stemness and malignant behavior. *Curr Opin Cell Biol*, 2007. Vol. **19**(2), pp. 150-8.
8. Katoh, M., WNT/PCP signaling pathway and human cancer (review). *Oncol Rep*, 2005. Vol. **14**(6), pp. 1583-8.
9. Kohn, A.D. and R.T. Moon, Wnt and calcium signaling: beta-catenin-independent pathways. *Cell Calcium*, 2005. Vol. **38**(3-4), pp. 439-46.
10. Segditsas, S. and I. Tomlinson, Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene*, 2006. Vol. **25**(57), pp. 7531-7.

11. Laurent-Puig, P. and J. Zucman-Rossi, Genetics of hepatocellular tumors. *Oncogene*, 2006. Vol. **25**(27), pp. 3778-86.
12. Howe, L.R. and A.M. Brown, Wnt signaling and breast cancer. *Cancer Biol Ther*, 2004. Vol. **3**(1), pp. 36-41.
13. Schwartz, D.R., et al., Novel candidate targets of beta-catenin/T-cell factor signaling identified by gene expression profiling of ovarian endometrioid adenocarcinomas. *Cancer Res*, 2003. Vol. **63**(11), pp. 2913-22.
14. Jheng, F.F., et al., Functional contribution of Ca²⁺ and Mg²⁺ to the intermolecular interaction of visinin-like proteins. *Protein J*, 2006. Vol. **25**(4), pp. 250-6.
15. Levy, L., et al., Transcriptional activation of interleukin-8 by beta-catenin-Tcf4. *J Biol Chem*, 2002. Vol. **277**(44), pp. 42386-93.
16. Hendrix, N.D., et al., Fibroblast growth factor 9 has oncogenic activity and is a downstream target of Wnt signaling in ovarian endometrioid adenocarcinomas. *Cancer Res*, 2006. Vol. **66**(3), pp. 1354-62.
17. Luu, H.H., et al., Wnt/beta-catenin signaling pathway as a novel cancer drug target. *Curr Cancer Drug Targets*, 2004. Vol. **4**(8), pp. 653-71.
18. Moon, R.T., et al., WNT and beta-catenin signalling: diseases and therapies. *Nat Rev Genet*, 2004. Vol. **5**(9), pp. 691-701.
19. Dihlmann, S. and M. von Knebel Doeberitz, Wnt/beta-catenin-pathway as a molecular target for future anti-cancer therapeutics. *Int J Cancer*, 2005. Vol. **113**(4), pp. 515-24.
20. Roose, J. and H. Clevers, TCF transcription factors: molecular switches in carcinogenesis. *Biochim Biophys Acta*, 1999. Vol. **1424**(2-3), pp. M23-37.

21. Chung, E.J., et al., Regulation of leukemic cell adhesion, proliferation, and survival by beta-catenin. *Blood*, 2002. Vol. **100**(3), pp. 982-90.
22. Hsu, H.C., et al., Beta-catenin mutations are associated with a subset of low-stage hepatocellular carcinoma negative for hepatitis B virus and with favorable prognosis. *Am J Pathol*, 2000. Vol. **157**(3), pp. 763-70.
23. Prange, W., et al., Beta-catenin accumulation in the progression of human hepatocarcinogenesis correlates with loss of E-cadherin and accumulation of p53, but not with expression of conventional WNT-1 target genes. *J Pathol*, 2003. Vol. **201**(2), pp. 250-9.
24. Tan, X., et al., Epidermal growth factor receptor: a novel target of the Wnt/beta-catenin pathway in liver. *Gastroenterology*, 2005. Vol. **129**(1), pp. 285-302.
25. Lee, H.S., et al., Novel candidate targets of Wnt/beta-catenin signaling in hepatoma cells. *Life Sci*, 2007. Vol. **80**(7), pp. 690-8.
26. Velculescu, V.E., et al., Serial analysis of gene expression. *Science*, 1995. Vol. **270**(5235), pp. 484-7.
27. Kim, H.L., Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from CD34+ cells. *Exp Mol Med*, 2003. Vol. **35**(5), pp. 460-6.
28. Koon, N., et al., Clustering of molecular alterations in gastroesophageal carcinomas. *Neoplasia*, 2004. Vol. **6**(2), pp. 143-9.
29. Vilain, C. and G. Vassart, Small amplified RNA-SAGE. *Methods Mol Biol*, 2004. Vol. **258**, pp. 135-52.

30. Heidenblut, A.M., et al., aRNA-longSAGE: a new approach to generate SAGE libraries from microdissected cells. *Nucleic Acids Res*, 2004. Vol. **32**(16), pp. e131.
31. Hashimoto, S., et al., 5'-end SAGE for the analysis of transcriptional start sites. *Nat Biotechnol*, 2004. Vol. **22**(9), pp. 1146-9.
32. Gowda, M., et al., Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. *Plant Physiol*, 2004. Vol. **134**(3), pp. 890-7.
33. Vilain, C., et al., Small amplified RNA-SAGE: an alternative approach to study transcriptome from limiting amount of mRNA. *Nucleic Acids Res*, 2003. Vol. **31**(6), pp. e24.
34. Luo, J., et al., Wnt signaling and human diseases: what are the therapeutic implications? *Lab Invest*, 2007. Vol. **87**(2), pp. 97-103.
35. Datson, N.A., et al., MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. *Nucleic Acids Res*, 1999. Vol. **27**(5), pp. 1300-7.
36. Saha, S., et al., Using the transcriptome to annotate the genome. *Nat Biotechnol*, 2002. Vol. **20**(5), pp. 508-12.
37. Amsterdam, A., et al., Many ribosomal protein genes are cancer genes in zebrafish. *PLoS Biol*, 2004. Vol. **2**(5), pp. E139.
38. Mathupala, S.P. and A.E. Sloan, "In-gel" purified ditags direct synthesis of highly efficient SAGE Libraries. *BMC Genomics*, 2002. Vol. **3**(1), pp. 20.

39. Angelastro, J.M., L.P. Klimaschewski, and O.V. Vitolo, Improved NlaIII digestion of PAGE-purified 102 bp ditags by addition of a single purification step in both the SAGE and microSAGE protocols. *Nucleic Acids Res*, 2000. Vol. **28**(12), pp. E62.
40. van den Berg, A., J. van der Leij, and S. Poppema, Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags. *Nucleic Acids Res*, 1999. Vol. **27**(17), pp. e17.
41. Margulies, E.H. and J.W. Innis, eSAGE: managing and analysing data generated with serial analysis of gene expression (SAGE). *Bioinformatics*, 2000. Vol. **16**(7), pp. 650-1.
42. Dinel, S., et al., Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome. *Nucleic Acids Res*, 2005. Vol. **33**(3), pp. e26.
43. van Kampen, A.H., et al., USAGE: a web-based approach towards the analysis of SAGE data. Serial Analysis of Gene Expression. *Bioinformatics*, 2000. Vol. **16**(10), pp. 899-905.
44. Liang, P., SAGE Genie: a suite with panoramic view of gene expression. *Proc Natl Acad Sci U S A*, 2002. Vol. **99**(18), pp. 11547-8.
45. Keime, C., et al., Identitag, a relational database for SAGE tag identification and interspecies comparison of SAGE libraries. *BMC Bioinformatics*, 2004. Vol. **5**(1), pp. 143.
46. Spilker, C., et al., The neuronal EF-hand calcium-binding protein visinin-like protein-3 is expressed in cerebellar Purkinje cells and shows a calcium-dependent membrane association. *Neuroscience*, 2000. Vol. **96**(1), pp. 121-9.

47. Renard, C.A., et al., Tbx3 is a downstream target of the Wnt/beta-catenin pathway and a critical mediator of beta-catenin survival functions in liver cancer. *Cancer Res*, 2007. Vol. **67**(3), pp. 901-10.
48. Colinge, J. and G. Feger, Detecting the impact of sequencing errors on SAGE data. *Bioinformatics*, 2001. Vol. **17**(9), pp. 840-2.
49. Margulies, E.H., S.L. Kardia, and J.W. Innis, Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res*, 2001. Vol. **29**(12), pp. E60-0.
50. Boon, K., et al., An anatomy of normal and malignant gene expression. *Proc Natl Acad Sci U S A*, 2002. Vol. **99**(17), pp. 11287-92.
51. Tuteja, R. and N. Tuteja, Serial analysis of gene expression (SAGE): application in cancer research. *Med Sci Monit*, 2004. Vol. **10**(6), pp. RA132-40.
52. Ng, P., et al., Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods*, 2005. Vol. **2**(2), pp. 105-111.
53. Ramasamy, A., et al., Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets. *PLoS Med*, 2008. Vol. **5**(9), pp. e184.
54. Xu, L., D. Geman, and R.L. Winslow, Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics*, 2007. Vol. **8**, pp. 275.
55. Rhodes, D.R., et al., Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A*, 2004. Vol. **101**(25), pp. 9309-14.

56. Axelsen, J.B., et al., Genes overexpressed in different human solid cancers exhibit different tissue-specific expression profiles. *Proc Natl Acad Sci U S A*, 2007. Vol. **104**(32), pp. 13122-7.
57. Rhodes, D.R., et al., Mining for regulatory programs in the cancer transcriptome. *Nat Genet*, 2005. Vol. **37**(6), pp. 579-83.
58. Cohen, B.A., et al., A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet*, 2000. Vol. **26**(2), pp. 183-6.
59. Caron, H., et al., The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, 2001. Vol. **291**(5507), pp. 1289-92.
60. Spellman, P.T. and G.M. Rubin, Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol*, 2002. Vol. **1**(1), pp. 5.
61. Stransky, N., et al., Regional copy number-independent deregulation of transcription in cancer. *Nat Genet*, 2006. Vol. **38**(12), pp. 1386-96.
62. *EMBO SAGE Congress SAGE Manual*. 2005.
63. Reich, M., et al., GenePattern 2.0. *Nat Genet*, 2006. Vol. **38**(5), pp. 500-1.
64. Lal, A., et al., A public database for gene expression in human cancers. *Cancer Res*, 1999. Vol. **59**(21), pp. 5403-7.
65. Chen-Plotkin, A.S., et al., Decreased association of the transcription factor Sp1 with genes downregulated in Huntington's disease. *Neurobiol Dis*, 2006. Vol. **22**(2), pp. 233-41.
66. Hochberg, Y. and Y. Benjamini, More powerful procedures for multiple significance testing. *Stat Med*, 1990. Vol. **9**(7), pp. 811-8.

67. Cadoret, A., et al., New targets of beta-catenin signaling in the liver are involved in the glutamine metabolism. *Oncogene*, 2002. Vol. **21**(54), pp. 8293-301.
68. Chiba, T., et al., Side population purified from hepatocellular carcinoma cells harbors cancer stem cell-like properties. *Hepatology*, 2006. Vol. **44**(1), pp. 240-51.
69. Haraguchi, N., et al., Characterization of a side population of cancer cells from human gastrointestinal system. *Stem Cells*, 2006. Vol. **24**(3), pp. 506-13.
70. Chen, X., et al., Gene expression patterns in human liver cancers. *Mol Biol Cell*, 2002. Vol. **13**(6), pp. 1929-39.
71. Rhodes, D.R., et al., Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, 2007. Vol. **9**(2), pp. 166-80.
72. Korinek, V., et al., Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC^{-/-} colon carcinoma. *Science*, 1997. Vol. **275**(5307), pp. 1784-7.
73. Hagerstrand, D., et al., Characterization of an imatinib-sensitive subset of high-grade human glioma cultures. *Oncogene*, 2006. Vol. **25**(35), pp. 4913-22.
74. Klein, P.S. and D.A. Melton, A molecular mechanism for the effect of lithium on development. *Proc Natl Acad Sci U S A*, 1996. Vol. **93**(16), pp. 8455-9.
75. Desbois-Mouthon, C., et al., Dysregulation of glycogen synthase kinase-3beta signaling in hepatocellular carcinoma cells. *Hepatology*, 2002. Vol. **36**(6), pp. 1528-36.
76. Jho, E.H., et al., Wnt/beta-catenin/Tcf signaling induces the transcription of Axin2, a negative regulator of the signaling pathway. *Mol Cell Biol*, 2002. Vol. **22**(4), pp. 1172-83.

77. Ruggero, D. and P.P. Pandolfi, Does the ribosome translate cancer? *Nat Rev Cancer*, 2003. Vol. **3**(3), pp. 179-92.
78. Pogue-Geile, K., et al., Ribosomal protein genes are overexpressed in colorectal cancer: isolation of a cDNA clone encoding the human S3 ribosomal protein. *Mol Cell Biol*, 1991. Vol. **11**(8), pp. 3842-9.
79. Vaarala, M.H., et al., Several genes encoding ribosomal proteins are over-expressed in prostate-cancer cell lines: confirmation of L7a and L37 over-expression in prostate-cancer tissue samples. *Int J Cancer*, 1998. Vol. **78**(1), pp. 27-32.
80. Wang, Q., et al., Cloning and characterization of full-length human ribosomal protein L15 cDNA which was overexpressed in esophageal cancer. *Gene*, 2001. Vol. **263**(1-2), pp. 205-9.
81. Bee, A., et al., Ribosomal protein l19 is a prognostic marker for human prostate cancer. *Clin Cancer Res*, 2006. Vol. **12**(7 Pt 1), pp. 2061-5.
82. Chan, M.W., et al., Hypermethylation of 18S and 28S ribosomal DNAs predicts progression-free survival in patients with ovarian cancer. *Clin Cancer Res*, 2005. Vol. **11**(20), pp. 7376-83.
83. Ge, X., et al., Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, 2005. Vol. **86**(2), pp. 127-41.
84. Gu, L.H. and P.A. Coulombe, Keratin function in skin epithelia: a broadening palette with surprising shades. *Curr Opin Cell Biol*, 2007. Vol. **19**(1), pp. 13-23.
85. van Ruissen, F., et al., Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. *BMC Genomics*, 2005. Vol. **6**, pp. 91.

86. Boyle, A.P., et al., High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 2008. Vol. **132**(2), pp. 311-22.
87. Cao, X., et al., Singapore familial adenomatous polyposis (FAP) patients with classical adenomatous polyposis but undetectable APC mutations have accelerated cancer progression. *Am J Gastroenterol*, 2006. Vol. **101**(12), pp. 2810-7.
88. Moenner, M., et al., Integrated endoplasmic reticulum stress responses in cancer. *Cancer Res*, 2007. Vol. **67**(22), pp. 10631-4.
89. Versteeg, R., et al., The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res*, 2003. Vol. **13**(9), pp. 1998-2004.
90. Chen, J., et al., An unexpected role for keratin 10 end domains in susceptibility to skin cancer. *J Cell Sci*, 2006. Vol. **119**(Pt 24), pp. 5067-76.
91. Kim, S., P. Wong, and P.A. Coulombe, A keratin cytoskeletal protein regulates protein synthesis and epithelial cell growth. *Nature*, 2006. Vol. **441**(7091), pp. 362-5.
92. Hendrix, M.J., et al., Biologic determinants of uveal melanoma metastatic phenotype: role of intermediate filaments as predictive markers. *Lab Invest*, 1998. Vol. **78**(2), pp. 153-63.
93. Gibbs, S., et al., Molecular characterization and evolution of the SPRR family of keratinocyte differentiation markers encoding small proline-rich proteins. *Genomics*, 1993. Vol. **16**(3), pp. 630-7.
94. Martin, N., S. Patel, and J.A. Segre, Long-range comparison of human and mouse Sprr loci to identify conserved noncoding sequences involved in coordinate regulation. *Genome Res*, 2004. Vol. **14**(12), pp. 2430-8.

95. Zhang, Y., et al., Exogenous expression of Esophagin/SPRR3 attenuates the tumorigenicity of esophageal squamous cell carcinoma cells via promoting apoptosis. *Int J Cancer*, 2008. Vol. **122**(2), pp. 260-6.
96. Liu, J., et al., In silico analysis and verification of S100 gene expression in gastric cancer. *BMC Cancer*, 2008. Vol. **8**, pp. 261.
97. Diederichs, S., et al., S100 family members and trypsinogens are predictors of distant metastasis and survival in early-stage non-small cell lung cancer. *Cancer Res*, 2004. Vol. **64**(16), pp. 5564-9.
98. Tarabykina, S., et al., Metastasis-associated protein S100A4: spotlight on its role in cell migration. *Curr Cancer Drug Targets*, 2007. Vol. **7**(3), pp. 217-28.
99. Marsman, M., et al., *Chaperoning Antigen Presentation by MHC Class II Molecules and their Role in Oncogenesis*, in *Advances in Cancer Research*. 2005, Academic Press. p. 129-158.
100. King, J., J. Waxman, and H. Stauss, Advances in tumour immunotherapy. *QJM*, 2008. Vol. **101**(9), pp. 675-683.
101. Miller, T.P., et al., HLA-DR (Ia) immune phenotype predicts outcome for patients with diffuse large cell lymphoma. *J Clin Invest*, 1988. Vol. **82**(1), pp. 370-2.
102. Ho, C.C., et al., Possible relation between histone 3 and cytokeratin 18 in human hepatocellular carcinoma. *In Vivo*, 2008. Vol. **22**(4), pp. 457-62.