Global Patterns Of Changes In The Gene Expression Associated With Genesis Of Cancer

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Ganiraju Manyam
Master of Science
IIIT-Hyderabad, 2004
Bachelor of Engineering
Bharatiar University, 2002

Director: Dr. Ancha Baranova, Associate Professor
Department of Molecular & Microbiology

Fall Semester 2009
George Mason University
Fairfax, VA

DEDICATION

To my parents
Pattabhi Ramanna and *Veera Venkata Satyavathi*
who introduced me to the joy of learning.

To friends, family and colleagues
who have contributed in work, thought, and support to this project.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Ancha Baranova, whose tolerance, patience, guidance and encouragement helped me throughout the study. This dissertation would not have been possible without her ever ending support. She is very sincere and generous with her knowledge, availability, compassion, wisdom and feedback.

I would also like to thank Dr. Vikas Chandhoke for funding my research generously during my doctoral study at George Mason University.

Special thanks go to Dr. Patrick Gillevet, Dr. Alessandro Giuliani, Dr. Maria Stepanova who devoted their time to provide me with their valuable contributions and guidance to formulate this project.

Thanks to the faculty of Molecular and Micro Biology (MMB) department, Dr. Jim Willett and Dr. Monique Vanhoek in embedding valuable thoughts to this dissertation by being in my dissertation committee. I would also like to thank the present and previous doctoral program directors, Dr. Daniel Cox and Dr. Geraldine Grant, for facilitating, allowing, and encouraging me to work in this project.

My heartful thanks goes to the formal and current graduate students in the MMB department Dr. Mohammed Jarrar, Dr. Michael Estep, Dr. Manpreet Randhawa, Dr. Aybike Birerdinc, Subashini Iyer, Beatrix Meltzer and Eshwar Iyer who contributed their time and effort in facilitating and providing the scientific environment for this project.

A very special thanks to my friends in *galigang* for their continual moral support to enhance my spirit in progressing this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

GLOBAL PATTERNS OF CHANGES IN THE GENE EXPRESSION ASSOCIATED WITH GENESIS OF CANCER

Ganiraju Manyam, PhD

George Mason University, 2009

Dissertation Director: Dr. Ancha Baranova

Cancer arises from a stepwise accumulation of genetic changes through expansion of the malignant cell clones in the population of pre-malignant cells undergoing the Darwinian selection process. In other words, cancer is an outcome of continuous and random acquisition of the changes in the genomes of individual cells. These modifications gradually and progressively change the phenotype of the normal cell making it more malignant through a loss of an overall stability of genome. To gain the comprehension of the mechanisms underlying tumor development, a number of high-throughput expression studies have been performed. The objective of the current study is to use publicly available datasets in order to analyze the most general features of the malignant cell, thus, investigating molecular phenomena common for all tumor cells, with no regard to the characteristics related to tumor's tissue of origin. Thus, we analyzed and compared the transcript diversity patterns in tumor and normal cells, studied an expression of the genes located adjacent to the telomeres and provided an evidence for the hypothesis that tumor

state behaves as stable “attractor” state. An intermediate regulatory framework hypothesis implying a set of local ‘vantage points’ genes that control the transcription of all other genes in a semi-democratic fashion has been endorsed.

A SUMMARY

The central idea of this dissertation is to explore gene expression patterns in the cancer cells having in mind a systems biology perspective on this disease. In this study we attempted to uncover the most general features of the malignant cell, thus, investigating molecular phenomena common for all tumor cells, with no regard to the tissue-specific characteristics of individual tumors. To perform large-scale data analysis, we have developed a number of novel bioinformatics techniques and as well employed some algorithms previously developed for other purposes and published elsewhere. The results obtained in this study enhance a general understanding of cancer as an expression system comprising of the components dynamically interacting with each other. The results reported in this dissertation and novel tools for *in silico* analysis of the cancer cell provide novel avenues for the functional genomics and systems biology of cancer and may be of help for large-scale computational modeling of cancer.

The first chapter provides general information on cancer, the molecular events causing this pathology as well as an introduction to human gene expression and its diversity. This summarization helps to place cancer in the perspective of gene expression both for the biologists and bioinformatic researchers. A review of bioinformatics studies of cancer is aimed to introduce the readers to the *in silico* analysis of this disorder. An overview of

the high-throughput experimental platforms, particularly, microarrays, associated methods of the gene expression analysis, an utility of these methods for the cancer transcriptome studies is accompanied by a special emphasis on the meta-analysis of expression data. Finally, the systems biology perspective of carcinogenesis is outlined, stressing the importance of inter-disciplinar nature of these studies allowing comprehension of the malignant cell at the systems biology level.

The second chapter describes the analysis of the cancer expression system in the perspective of attractor states. First, the concept of the cell as a dynamic system and the concept of attractor states are reviewed. The modeling strategy for the distance based statistical approach is described in details. The data produced as a result of the distance analysis of the two-point and the multi-point datasets are presented in two different sections. Finally, the results of the principal component analysis (PCA) are described. These results are shedding more light on the attractor behavior of cancer.

The third chapter explores the bioinformatics methods developed to analyze transcriptional abundances. We reviewed the UniGene database system that provides information on the clusters of ESTs obtained by sequencing of cDNA libraries prepared form various human tumors and tissues. The adaptation of the Shannon' statistics for estimation of the diversity to the gene expression systems is detailed in the Methods section. The classification of human tissues and tumor types as well as quantitative estimation of unique and common gene clusters in the normal and cancer tissues are described in two separate sections. Diversity estimations for the various human tissues and the lists of the potential tumor biomarkers and biomarkers of normal tissue functioning (anti-cancer biomarkers) are given. Finally, we present the results of the functional analysis of the protein-coding subset of the tumor biomarkers and anti-cancer biomarkers.

The last chapter reviews the role of telomeres in cancer development and the previous studies on the telomere position effect as well as the methodology we employed to define subtelomeric and non-telomeric region of the human chromosomes. The statistical techniques allowing comparison of the expression levels for the genes located in sub-telomeric region to that in other regions of the human genome are discussed and results presented. An exploratory approach allowing estimation of the optimal length of the sub-telomeric regions is presented. Finally, we have used the distance analysis technique previously discussed in the second chapter to compare functional properties of the expression patterns of human genes located in these two regions.

Finally, an appendix with Supplementary information containing the data associated with the chapters two, three and four was put together. Of particular interest, this appendix contains a description of a novel publicly available tool, KEGG Pathway Painter. This tool was specifically developed to provide automated summarization of the functional information pertinent to the genes comprising “top hit” lists routinely obtained in large-scale transcriptome studies.

1. Introduction

Cancer and its genesis

Cancer can be attributed to as many as 25% of deaths in the United States, making it the second most common cause of death (ACS 2009). Cancer arises from a stepwise accumulation of genetic changes through clonal expansion events in the population of pre-malignant cells undergoing the Darwinian selection process (Weinberg 2007). In a genetics perspective, it is a micro-evolutionary phenomenon resulting in the cooperative malfunction of a number of human genes enhancing the selfish survival and metastasis of cells. Each individual tumor is an outcome of continuously acquisition of the random mutations in the genomes of individual cells, a process supported by a natural selection.

Carcinogenesis is complex process initiated in a single cell or a group of cells, progressively leading to the disruption of the normal tissue architecture. The subset of cells tolerates deleterious mutations, accelerates their proliferation and invades the surrounding tissue followed by metastasis to the distant organs. This uncontrolled growth of cellular population is often due to the damage of the molecular circuitry responsible for programmed cell death or apoptosis. The cause of carcinogenesis is unknown. There are two common theories that explain the process. The somatic mutation theory states that DNA mutations that occur in the genes regulating cell cycle and proliferation of a single somatic cell cause its transformation into the malignant one through the disruption of the otherwise quiescence state (Varmus 2006). The tissue organization field theory suggests that proliferation is the default state of the cell and that carcinogenesis is a problem of tissue restructuring and organization (Sonnenschein and Soto 2008). To the most part, a combined understanding provided by these theories enables one to comprehend the causation for the genesis of cancer.

In a combined view, cancer is produced as the result of a Darwinian evolution of cells located in the microenvironment within the particular tissue of multicellular organism (Stratton, Campbell et al. 2009). The progression of cancer leads to an increase of the overall mutation rates due to the changes in certain genes that serve as caretakers of the genome. Cumulative load of the mutations pushes cell to proliferate faster than normal by whatever mean possible, direct or indirect (Strauss 1998). Usually, the hypermutability is observed at early stage of the tumor development, often in the premalignant cell before it successful transforms into the malignant one. Gene expression level changes observed in tumors are caused by a combination of genetic and epigenetic events (Gronbaek, Hother et al. 2007). Mutated genes tend to express ectopically, in the tissues where they are normally silenced.

Genetic and epigenetic events causing cancer

The genetic events causing cancerogenesis include point mutations and deletions or insertions in small as well as large DNA segments. The rearrangements of the genome and increase/decrease of the gene/chromosome copy number are also quite common events in this process. Additionally, somatic mutations in mitochondrial genomes were also reported in many tumor types, although the precise role of these alterations in cancer progression is not well comprehended (Chatterjee, Mambo et al. 2006). Mutations can also happen due to disruption of the gene by insertion of the completely exogenous DNA, for example, some tumorigenic viruses, including HPV, EBV, HBV, HHV8 which can also possess viral oncogenes (Talbot and Crawford 2004). Of all these mutations described above, a small portion can be fixed in the cellular lineage by the process of natural selection. The mutational rates for these different kinds of genetic alterations vary. Generally, these rates tend to increase in response to the exposure to exogenous mutagens or clastogens, for example, tobacco smoke or X-ray. Mutation rates are also increased in subjects with certain inherited diseases associated with increased cancer risk, for example, xeroderma pigmentosum, Fanconi anemia or ataxia-telangiectasia (Kennedy and D'Andrea 2006; Stratton, Campbell et al. 2009). The increased rate of mutation would yield increased DNA sequence diversity, providing the selection with the raw material to choose from and support the drive for increased proliferation, invasion and cancer.

Mutations found in tumors often damage the key regulatory genes controlling cell cycle, proliferation, apoptosis, genomic stability and other biological processes or functions, which normally prevent the cancer. The key genes in this context can be subdivided into two categories: cancer causing genes (oncogenes) and the tumor suppressor genes (often abbreviated as TSGs).

Oncogenes code for proteins that positively regulate cell proliferation and/or negatively affect apoptosis. Usually, these genes are activated by point mutation or fusion with other gene or juxtaposition to enhancer element that drive their expression to higher level (Konopka, Watanabe et al. 1985; Tsujimoto, Gorham et al. 1985). The cellular proto-oncogenes may acquire genetic mutation or increase their copy number (become amplified) and, thus, transform into oncogenes (Croce 2008). Oncogenes are identified by their tendency to increase the tumor growth rate and their ability to transform the cell morphologically. The protein products of oncogenes include transcription factors, chromatin remodelers, signal transducers, apoptosis suppressors, growth factors and their receptors.

As opposed to oncogenes, tumor supressors are capable of the suppression of oncological transformation by negative regulation of the genes in the pathways related to oncogenesis, for example, cell cycle (Kopnin 2000) or positive regulation of apoptosis. Unlike oncogenes, tumor suppressor genes require mutations in both of their alleles in the diploid genome as they act recessively. The mutation patterns of TSGs range from single base substitutions to whole gene deletions, generally tending to abolish the functioning of the protein product of tumor suppressor gene.

In the biological knowledge bases summarizing information of the genes involved in malignization of human cells, proven tumor suppressors are relatively rare (10%) as compared to the proven oncogenes (90%) (Stratton, Campbell et al. 2009). The mutational events promoting the tumor growth may either upregulate oncogenes or downregulate the tumor suppressor genes. The overall increase in the transcription rates were noted during cancer progression which might bias in favor of the outnumbered oncogenes and push the cell through malignization threshold.

In their natural evolution, the cancer cells gradually and progressively change their phenotype toward most malignant as result of the accumulation of the mutations in the oncogenes and the tumor suppressor genes (Barrett, Oshimura et al. 1986), the rearrangements of chromosomes (Radman, Jeggo et al. 1982) and the perturbation of the gene expression levels (Nicolson 1991). Many genes and molecular pathways involved in the underlying processes are well studied (Vogelstein and Kinzler 2004), but the complex spatiotemporal patterns of interactions between the involved molecules contribute to the difficulties of the comprehension of the malignant cell system (Hornberg, Bruggeman et al. 2006). On top of the genetic modifications, epigenetic events also play a vital role in the initiation and progression of a tumor.

Epigenetic alterations comprise mitotically and meiotically heritable changes in gene expression that are not caused by changes in the primary DNA sequence. These changes are increasingly being recognized for their roles in carcinogenesis (Gronbaek, Hother et al. 2007). Darwinian selection can act upon the phenotypic effects that are generated by epigenetic changes for cancer evolution, analogous to selection of mutations in the case of genetic alterations. Methylation is the most common epigenetic event, which often leads to gene silencing through subsequent histone deacetylation and chromatin condensation (Worm and Guldberg 2002). Hypermethylation at the promoters of tumor suppressors is a distinctive feature seen in many tumors, which can happen very early in the tumor progression. Hypermethylation is known to mediate an imbalance in many important signaling pathways (Baylin, Esteller et al. 2001; Gronbaek, Hother et al. 2007). CpG islands are the hot spots for methylation, with 50-70% of the cytosines in these sites are being methylated in human tissues (Ehrlich, Gama-Sosa et al. 1982; Esteller 2006). These cytosines often undergo spontaneous transition to thymines by deamination (Rideout, Coetzee et al. 1990). If these transitions are not corrected, they become either somatic or germline point mutations.

The change in the DNA methylation landscape of a cancer cell usually occurs in the context of other epigenetic changes. DNA methylation attracts methyl-CpG binding proteins and DNA methyltransferases. In turn, these proteins associate with histone deacetylases and histone methyltransferases, two types of enzymes playing a key role in chromatin remodeling (An 2007). Some well known oncogenes, for example, PML/RARα fusion, an archetypal chimeric oncoprotein, were shown to bring complexes of histone deacetylases (HDACs), histone methyltransferases (HMTs), and DNA methyl transferases (DNMTs) to target genes (Hormaeche and Licht 2007).

Epigenetic gene silencing has always been envisaged as a local event, silencing genes one by one. However, recent data indicate that large regions of chromosomes can be coordinately suppressed by a process termed as long range epigenetic silencing (LRES) (Frigola, Song et al. 2006; Stransky, Vallot et al. 2006; Hitchins, Lin et al. 2007). LRES can span megabases of DNA. It involves formation of a broad heterochromatin regions accompanied by hypermethylation in the contiguous clusters of CpG islands. It is not clear if LRES is initiated by one critical gene target, then spreads to cloak innocent bystanders, analogous to large chromosome deletions, or if coordinated silencing of multiple genes occur (Clark 2007).

It is important to note that unlike the genetic alterations, gene silencing by epigenetic modifications is potentially reversible. Treatment by agents that inhibit cytosine methylation and histone deacetylation can initiate chromatin decondensation, demethylation and re-establishment of gene transcription of the silenced tumor suppressor genes that, in turn, might help to restore normal phenotype. On the other hand, it is likely that application of said therapeutics will provoke further deregulation of cancer cell transcriptome, and, possibly, further malignization.

Interestingly, both genetic and epigenetic changes can be seen as either the cause for the malignization or as consequences of the tumor progression. Despite tremendous efforts to connect gene expression profiles in tumor cells to particular milestones of the tumor progression (Ellis 2003; Lee and Thorgeirsson 2004; Liu 2004; Wang 2005; Driouch, Landemaine et al. 2007; Henrickson, Hartmann et al. 2007), the exact mechanism of the tumorigenesis remains unclear.

Diversity of human gene expression

Tumors arising from different organ/tissue systems of the tumor body are often considered as different diseases due to observed differences in their cell phenotypes and particular prognoses. However, the difference of one type of the tumor from another type of the tumor might be explained by the underlying differences in the tissues of their respective origins. Each of the normal human tissues has a distinct gene expression pattern, with an exception of the constitutively expressing housekeeping genes shared between all the tissues (Butte, Dzau et al. 2001). Recent studies shown that even the housekeeping genes are not necessarily expressed at the same level across all tissues; rather, each tissue seems to have a specific expression profile of housekeeping genes. Interestingly, housekeeping genes are less compact and are evolutionary older than tissue-specific genes, and they evolve more slowly in terms of both coding and core promoter sequences (Zhang and Li 2004; Zhu, He et al. 2008). Housekeeping genes primarily use CpG-dependent core promoters, whereas the majority of tissue-specific genes possess neither CpG-islands nor TATA-boxes in their core promoters (Zhu, He et al. 2008).

Generally, the distinctive patterns of gene expression in various tissue types are explained by the differences in the functions of the resulting proteins in that particular tissue.

Typically, this phenomenon is illustrated by the tissue-specific expression of mRNA encoding for secreted hormone insulin, that is present only in Langerhans' islets of the pancreatic gland (Bliss 1982). Another important note about the expression of the human genes is that many of them are alternatively spliced. Alternative splicing leads to the expression of multiple mRNA transcripts with different sets of exons joined together. It is a prevalent phenomenon that is observed in around half of the human genes (Modrek and Lee 2002). The use of alternative promoters and the alternative exons endings (splice donor sites) may also result in alternative mRNA transcripts. There is lot of variation in the alternative transcript expression patterns across the human tissues (Landry, Mager et al. 2003; Yeo, Holste et al. 2004). The alternative splicing code that controls and coordinates the transcriptome in complex multicellular organisms remains poorly understood. It has long been argued that regulation of alternative splicing relies on combinatorial interactions between multiple proteins, and that tissue-specific splicing decisions most likely result from differences in the concentration and/or activity of these proteins (Matlin, Clark et al. 2005; Singh and Valcarcel 2005).

Among human organs, the testis shows unusually diverse gene expression pattern. In part, this pecularity of the testicular expression signature is explained by presence of the specific gene isoforms that are transcribed as a result of the chromatin remodeling and the activation of specialized transcription complexes activated during the differentiation program of spermatogenesis (Kimmins, Kotaja et al. 2004). Along with testis, brain also shows a complex and diversified transcription. On the other extreme, tissues with a secretory function such as pancreas, salivary gland, and stomach show more specialized and narrow gene expression patterns (Jongeneel, Delorenzi et al. 2005; Shyamsundar, Kim et al. 2005). The expression signatures of the splicing factor encoding genes correlate with the degree of the expression variation seen among the human tissues (Grosso, Gomes et al. 2008). Respectively, brain and testis, the two tissues with highest levels of alternative splicing events, have the largest number of expressed genes encoding for splicing factors. Additionally, SR protein kinases and small nuclear ribonucleoprotein particle (snRNP) proteins that modulate the association of core components of the spliceosome with the pre-mRNA were identified as most highly differentially expressed in the particular tissues (Grosso, Gomes et al. 2008). Concerning the brain-specific splicing factor gene expression signature, the gene list includes the brain-splicing regulators PTB1, NOVA1, A2bp1/FOX1, and members of the CELF/BRUNOL and ELAVL families, the non-SR splicing regulator Y-box protein 1 and the core snRNP protein SmN. The testis-specific signature included the splicing factor 3a subunit 2 (SF3A2) and the SR protein kinases 1 and 2 (SRPK1 and SRPK2) (Grosso, Gomes et al. 2008).

Nearly all of the cancer-upregulated genes with tissue-selective expression tend to show their selective expression in tissues, which are different from the tissue of origin of cancer (Axelsen, Lotem et al. 2007). Interestingly, different types of cancers, including different brain cancers arising from the same lineage, showed differences in the tissue-selective genes they overexpressed. Cancer cells ectopically expressing such genes may acquire phenotypic modifications that contribute to cancer cell growth and metastasis (Axelsen, Lotem et al. 2007). Of all of the genes with tissue-selective expression, those selectively expressed in testis showed the highest frequency of genes that are overexpressed in at least two types of cancer (Axelsen, Lotem et al. 2007). Respectively, in many types of human cancers the phenomenon of coordinated up-regulation of so called cancer/testis antigens is observed; these genes are particularly expressed only in testis and in tumor tissues. The specific expression of these genes in tumors and their restriction to testis tissue, make them good candidates for the cancer vaccines (Scanlan, Simpson et al. 2004).

Bioinformatics & Cancer

An increase of the amount of biotechnological datasets and the demand for the maintenance and analysis of such data lead to the development of specialized scientific discipline known as bioinformatics. It can be broadly described as the application of computer technology to explore biological processes through an analysis of the large-scale and multi-dimensional data generated from numerous sources. It encompasses organized storage of the data, development of tools to analyze the data and the actual analysis of data. Bioinformatics uses both informatic and statistical tools implemented to extract and analyze information (Wu 2001). The tools utilized to analyze the data depend on the type of data and nature of biological question addressed through the use of the particular dataset. In a broader context, bioinformatics can be viewed as the management information system (MIS) with the scope on the biotechnology and the molecular biology (Umar 2004).

Bioinformatics became an inevitable research component of the molecular biology. Due to the accelerated generation of large-scale datasets by high throughput research platforms, cancer biology also got tightly bonded to informatics. Informatics occupies a special role in the translation cancer research to aggregate and perform integrative analysis on cancer biomarkers with ultimate goal of the improvement of both prevention and therapy. Bioinformatics is being used primarily to identify cancer biomarkers, their function and molecular mechanisms underlying cancer progression. The feasibility of the targeting of the biomolecule with the drug can also be assessed by structural bioinformatics, thus, reducing the drug development timeline (Wishart 2005). The addition of the biological function and/or the associated molecular cascade of the marker gene along with the quantified transcript or protein data significantly improved the relevance of biomarkers discovered. A number of novel bioinformatics methods, for example, gene set enrichment and pathway-based analysis, already reached the level of sophistication allowing to perform the functional analysis in semi-automated manner (Subramanian, Tamayo et al. 2005; Yi, Horton et al. 2006).

So far, expression analysis spotted a number of target genes, which laid the path for the development of novel cancer diagnostics and drugs. For example, AMACR (alpha-methylacyl CoA racemase) was at first identified as prostate cancer specific protein.

Currently, it is one of the prominent biomarkers for prostate cancer, with excellent specificity and sensitivity (Jiang, Woda et al. 2004). Other studies identified specific markers for progressing of the early stage cancers that permits individualized application of the treatments to patients through precise assessment of their prognoses. An example of such biomarker is EZH2 (enhancer of zeste homolog 2), a polycomb group protein that seems to be overexpressed in metastatic prostate cancer and specifically in the progressive tumors (Varambally, Dhanasekaran et al. 2002). Using multivariate model, this protein was later identified along with E-cadherin as powerful predictor of the disease recurrence following surgery (Rhodes, Sanda et al. 2003).

Functional analysis of biomarkers identified in large-scale datasets can be performed using a wide variety of bioinformatics tools. Gene annotation is the first and foremost requirement for the initiation of functional analysis. The annotation information is provided both by various biotechnological database management organizations like NCBI (Gene), EBI (Ensembl) and by independent groups like GeneCards (Safran, Solomon et al. 2002; Curwen, Eyras et al. 2004; Maglott, Ostell et al. 2007). Pathway information systems, for example, Kyoto encyclopedia of genes and genomes (KEGG) and Biocarta provide information concerning the molecular interactions of gene products in an organized fashion (Biocarta ; Aoki-Kinoshita and Kanehisa 2007). Association of genes with their corresponding ontological terms is another way of enhancing functional analysis. The gene ontology (GO) database was developed to standardize the gene and gene product attributes across the unending list of biological databases (Ashburner, Ball et al. 2000). Database for annotation, visualization, and integrated discovery (DAVID) is a hybrid functional analysis tool combining all these systems in a single web interface (Dennis, Sherman et al. 2003; Huang, Sherman et al. 2009). Such integrative tools enhanced the functional analysis of the genes and/or proteins while providing a holistic perspective of large-scale datasets.

Data collection and maintenance remains an important aspect of clinical research as the aggregated data improves the statistical power to ascertain the results (Mathew, Taylor et al. 2007). The cancer biomedical informatics grid (caBIG^{TM}) is an ambitious initiative to connect cancer research centers in a network that allows to share biomedical data (Kakazu, Cheung et al. 2004). This co-operative bioinformatics system brings translational cancer research into the next level, changing the whole cancer research paradigm. Its seamless architecture defines compatibility standards based on ontologies, interfaces, data elements and information models for tool development enhancing the semantic interoperability between the research groups (Cimino, Hayamizu et al. 2009). An integrating informatics system for the annotation and exchange of array based data, caArray, was recently developed for acquisition, dissemination and aggregation of interoperable array data as a part of caBIG^{TM}. caArray enhances the translational cancer research by supporting analysis of array data by tools and services on and off the grid. The resources developed on the grid are generally applicable beyond cancer research, promoting the comprehension of other complex diseases.

The analysis of microarrays and the cancer transcriptome

From its inception in the mid-90s, microarray analysis was instrumental in the discovery of clinically relevant knowledge by associating changes in gene expression (GE) patterns with particular pathological conditions. In a typical experiment, mRNA expression profiles are generated for thousands of genes across a collection of samples that belong to either one of two classes, for example, pathological specimen vs. healthy tissue controls. The registered changes in expression of individual genes can be ordered in a ranked list. However, extraction of a biological insight from the long lists of individual genes generated in typical microarray experiment remains a significant challenge.

A common approach to the analysis of GE data involves focusing on a handful of genes at the top and bottom of the ordered list (i.e., those showing the largest difference in the expression level in up and down regulatory fashion) and attempting an interconnection of these genes into the plausible biological network or pathway. This approach, however, has major limitations that may potentially nullify the experimental results. In particular, the required correction for multiple hypotheses testing (e.g. by Bonferroni or by Benjamini-Hochberg methods) may lead to the situation when no individual gene meets the threshold for statistical significance indicating that the relevant biological differences between two sample sets are modest relative to the technical noise. This outcome is not surprising since severe over-fitting is regarded as inevitable plague of the microarray studies where the typical number of observations (dozens or, rarely, hundreds of patients' samples) is dramatically smaller than the number of measured end-points (typically 5 to

30 thousands of genes). Alternative approaches not controlled by multiple hypothesis tests often leave scientists with extensive lists of statistically significant genes that cannot be bound together by any common biological theme. Subsequently, an interpretation of these gene lists is left to subjective opinions of the expert biologists.

One of the current standard technologies of GE analysis that partially resolves the problem of over-fitting in microarray study design is Significance Analysis of Microarrays (SAM) (Tusher, Tibshirani et al. 2001). Based on stochastic procedure of iterative cross-validation of false discovery rate, it allows eliminating a majority of spurious and potentially non-reproducible findings. This technology has already successfully replaced conventional t-tests. However, it remains a subject of the commonest drawback in all single-gene methods: statistically significant genes often lack biological meaning. Additionally, SAM, on par with other single-gene based methods of GE analysis, may miss pathway wide effects. For example, effects of the coordinated 20% increases in the expression levels of all genes that belong to the same metabolic pathway may be masked by spurious 500% decrease in an expression of a single gene with redundant function.

Identification of the pathological mechanisms underlying differentially expressed gene lists may be facilitated by pre-grouping these genes into the relevant biological pathways or networks. Thus, instead of the traditional “bottom-up” approach that relies on post factum integration of GE lists with existing literature describing the potential biological roles of individual genes, enrichment-based analysis of gene sets may be executed. In this type of knowledge-based analysis (KBA), previously collected gene lists are used for the ranking of the functional categories or pathways and subsequent evaluation of the ranked pathways according to their gene enrichment levels. The significantly enriched pathways are further explored as primary biological themes. For this purpose, the gene sets known as signature databases were generated in accordance with the prior biological knowledge accumulated in wet-lab experiments or using computational models that identify functionally similar genes by their sequence or structural similarities.

The pioneering tool for enrichment-based analysis of microarray datasets named GSEA (Gene Set Enrichment Analysis) has been developed by a group of bioinformaticians at MIT (Subramanian, Tamayo et al. 2005). Though it's been more than three years since the method was published, GSEA have just started to gain popularity among microarray researchers, mostly due to the time required for generation of extensive knowledge basis justifying the grouping of genes into gene sets. Today, an innovative idea of GSEA has reached its maturity. Particularly, several gene signature databases became available, including those with gene sets grouped according to their chromosomal locations, an a priori placement within certain molecular pathways, a commonality of regulatory mechanisms, and an annotation under the same gene ontology terms. These gene signature databases could be used for extracting novel and often unexpected knowledge about pathologic processes and disease mechanisms. Some of the examples of successful GSEA applications include studying differential gene expression associated with adenocarcinoma of esophagus (Lagarde, Ver Loren van Themaat et al. 2008), advanced pancreatic cancer (Campagna, Cope et al. 2008), breast cancer (Anders, Hsu et al. 2008), nasopharyngeal carcinoma (Pegtel, Subramanian et al. 2005).

To gain the comprehension of the underlying mechanism of tumor development, particularly the regulation of cell growth, apoptosis and genomic stability (Hanahan and Weinberg 2000), a number of high-throughput microarray initiatives have been started. Cancer cell lines have served as the primary experimental system for exploring cancer molecular biology and pharmacology. For instance, the Nation Cancer Institute developed an NCI-60 panel consisting of 60 diverse human cancer cell lines. These cell lines underwent both selected gene sequencing and gene expression profiling in order to facilitate screenings for drugs aimed at cancer therapy (Ikediobi, Davies et al. 2006; Shoemaker 2006). DNA microarrays expression analysis of these heterogeneous cell lines provided initial snapshots of the genes and related molecular pathways pertaining to the malignization in a broad sense and to the response to specific chemotherapeutic drug treatments (Efferth 2005; Shankavaram, Reinhold et al. 2007).

The contribution of microarray based studies to cancer research is not limited to the study of NCI-60 cell lines. For instance, microarray profiling is widely used for an identification of aberrant chromosomal regions and expression signatures of various primary tumors (Buness, Kuner et al. 2007; Xu, Geman et al. 2007). Numerous high-throughput experiments generated quantitative profiles of global gene expression for most of the common forms of cancer. Based on these profiles, the diagnostic signatures were generated for specific cancer types, in many cases producing substantial insights into the tumor biology. Meta-analysis of these datasets has been attempted to extract common or generic cancer signature pattern (Xu, Geman et al. 2007). The utility of gene expression profiling for clinical settings was demonstrated by bringing the breast and lymphoma tumor signatures into the process of the clinical decision making (van de Vijver, He et al. 2002; Dave, Wright et al. 2004). Proven to be efficient means for the cancer researchers, DNA microarrays became a staple of cancer transcriptome studying labs. For example, study of Golub et al demonstrated the feasibility of cancer classification based solely on gene expression monitoring and suggested a general strategy for discovering and predicting cancer classes, independent of previous biological knowledge (Golub, Slonim et al. 1999).

The increasing number of the microarray datasets in the research of cancer and other diseases, demanded a central system to congregate this valuable profile data. The National Central for Biotechnology Information (NCBI) took an initiative to develop the database called Gene Expression Omnibus (GEO), which archives and freely distributes raw microarray files as well as other forms of large-scale data generated by the scientific community (Barrett and Edgar 2006; Barrett, Troup et al. 2007). Until recently, ArrayExpress was an analogous system to GEO developed by the European Bioinformatics Institute (EBI) (Parkinson, Kapushesky et al. 2007). Nowadays, GEO has emerged into the principal repository for microarray data storage and retrieval that covers an assortment of microarray platforms. Standard analyses of the cancer array dataset stored in these repositories are maintained in a database known as Oncomine. This database enables one to query for the specific up and down regulated genes across microarray datasets according to the tumor or tissue types of interest (Rhodes, Kalyana-Sundaram et al. 2007).

Public availability of cancer microarray datasets lead to the development of the methods to derive common cancer signatures across datasets (Xu, Geman et al. 2007). Integrative analyses are also performed combining sub-datasets in order to borrow information and using statistical and clustering analysis techniques to derive relevant markers (Golub, Slonim et al. 1999; Tusher, Tibshirani et al. 2001). The process of the combination of expression profiles resulted from independent microarray experiments often called the meta-analysis, analogous to meta analysis in clinical research that is directed to test single hypothesis using material from multiple studies (Rhodes and Chinnaiyan 2005). An example of meta-analysis of microarray datasets is the study of Wren, 1999 that included all publicly available GEO two-channel human microarray datasets (a total of 3551 individual profiling experiments). This study was conducted to identify genes with recurrent, reproducible patterns of co-regulation across different conditions. Patterns of co-expression were divided into parallel (i.e. genes are up and down-regulated together) and anti-parallel. Several ranking methods to predict a gene's function based on its top 20 co-expressed gene pairs were compared. The data matrix describing differential expression of the human genes with unknown function was made available to the scientific community (Wren 2009).

Pertinent to cancer, a handful of meta-analysis studies were performed. In the study of Alles et al., 2009 gene expression values from five large microarray datasets describing breast carcinoma expression patterns relative to ER status of the tumors were subjected to Gene Set Enrichment Analysis (GSEA). As expected, the expression of the direct transcriptional targets of the ER was muted in ER- tumors, but the expression of genes indirectly regulated by estrogen was enhanced. An enrichment of independent MYC- and E2F-driven transcriptional programs was also observed. A conclusion concerning increased transcriptional activity of MYC as a characteristic of basal breast cancers capable of mimicking a large part of an estrogen response in the absence of the ER was made, thus, suggesting a mechanism by which these cancers achieve estrogen-independence and providing a potential therapeutic target for this poor prognosis sub group of breast cancer (Alles, Gardiner-Garden et al. 2009). Another study compared expression data from a diverse collection of 9 breast tumor array datasets generated on either cDNA or oligonucleotide arrays from the Oncomine database and identified genes that were universally up or down regulated with respect to ER+ versus ER- tumor status (Smith, Saetrom et al. 2008). Liang et al. analyzed miRNA target genes across Oncomine datasets profiling gene expression in the patients with lung adenocarcinoma (AD) and squamous cell carcinoma (SCC), two major histologic subtypes of lung cancer. Expression of a minimal set of 17 predicted miR-34b/34c/449 target genes was identified from a training set to classify 41 AD and 17 SCC, and correctly predicted in average 87% of 354 AD and 82% of 282 SCC specimens from total 9 independent published datasets. Expression of this signature in two published datasets of epithelial cells obtained at bronchoscopy from cigarette smokers, if combined with cytopathology of the cells, yielded 89-90% sensitivity of lung cancer detection and 87-90% negative predictive value to non-cancer patients (Lagarde, Ver Loren van Themaat et al. 2008). Romualdi et al. used meta-analysis to define the gene expression signature of rhabdomyosarcoma, a highly malignant soft tissue sarcoma (Romualdi, De Pitta et al. 2006), and demonstrated a general downregulation of the energy production pathways, suggesting a hypoxic physiology for rhabdomyosarcoma cells.

Meta analysis of cancer microarrays was also useful in identifying the aberrant genomic loci in the tumor cells. Eight datasets including more than 1200 breast tumors were investigated to identify chromosomal regions and candidate genes possibly causal for breast cancer metastasis. By utilizing Gene Set Enrichment Analysis chromosomal regions were ranked according to their relation to metastasis. Over-representation analysis identified regions with increased expression for chromosome 1q41-42, 8q24, 12q14, 16q22, 16q24, 17q12-21.2, 17q21-23, 17q25, 20q11, and 20q13 among metastasizing tumors and reduced gene expression at 1p31-21, 8p22-21, and 14q24. Analysis of genes with extremely imbalanced expression in these regions, DIRAS3 at 1p31, PSD3, LPL, EPHX2 at 8p21-22 and FOS at 14q24 pinpointed them as candidate metastasis suppressor genes. Potential metastasis promoting genes list included RECQL4 at 8q24, PRMT7 at 16q22, GINS2 at 16q24, and AURKA at 20q13 (Thomassen, Tan et al. 2009). An association was also established between the tumor stage of the breast carcinoma and the recognized genomic regions identified by the meta analysis applied to

12 independent human breast cancer microarray studies comprising 1422 tumor samples (Buness, Kuner et al. 2007).

Cancer – A systems biology perspective

Understanding the biology at the system level should improve the way a medical condition is perceived and thus affects the methods to prevent, diagnose and treat it (Ahn, Tewari et al. 2006). The capability of current biotechnological methods to extract global gene expression patterns along with levels of proteins and metabolites from the same sample opened avenues for the conception of network level data. Sequencing of genomes, high-throughput data generation technologies combined with organized informatic systems further enabled the collection of comprehensive datasets providing the system-wide overview of the complex biological objects (Kitano 2002). The elucidation of the network of biological networks with their dynamic interactions forms the basic premise for systems biology in general. System-level models require quantification of the network elements as they change over time in response to various perturbations, therefore, implying the computational modeling of the dynamic interactions (Laubenbacher, Hower et al. 2009). In the context of cancer, such models would be helpful in deciphering the relevant biological network underlying particular type of malignancies.

Cancer is an outcome of the malfunctioning in the cell system with inherent genomic instability, thus it is identified as a systems biology disease. The progress in the treatment of cancer can be drastically accelerated, if the ongoing research is done in the systems biology perspective, instead of classic molecular biology based approaches (Hornberg, Bruggeman et al. 2006). Apart of intra-cellular interactions as represented by internal signal transduction in cell, inter-cellular interactions also play a vital role in the development of cancer. The horizon of system-level study of cancer should be expanded to the tissue level considering the role of inter-cellular interactions in the realistic comprehension of cancer. Challenges lie in various disciplines of sciences including mathematics, physics, chemistry and biology to develop models, generate precise data and calculate the interactions between the biological network components. Last but not the least, the contribution of informational technology is inevitable, not only for data management but also for the analyzing the simulated models of individual cancer cells and interactions within their proliferating society (cancer tissue).

2. Genome wide discrimination of normal and tumor samples

Rationale

The current study quantitatively estimates the relative importance of global and local features of gene expression regulation landscape in the process of tumor development. The work is based on the hypothesis that the cancer could be viewed as an attractor state.

Background

To date, most of the high-throughput studies of the gene expression studies are still focused on elucidation of the discriminatory gene signatures reflecting key regulatory processes participating in establishing cell phenotypes (J. Wang et al. 2003; Ben-Dor et al. 2000; Furge et al. 2004). On the other hand, a change in a cell phenotype requires coordinated interaction of a variety of genes that determine the functional identity of the cell within a population of cells (Bar-Yam et al. 2009). This notion implies an understanding that a given cell type could be represented as a dynamic system that can assume different states, thus, occupying a specific position in the multidimensional phase space spanned by the different genes (Tsuchiya, Piras et al. 2009; Tsuchiya, Selvarajoo et al. 2009; H. H. Chang et al. 2008).

The ability of gene regulatory circuits to assume multiple equilibrium states was first proposed by Max Delbruck in 1948 (cited according to S. Huang et al. 2009). In terms of dynamics, this specific position of equilibrium is called an 'attractor', i.e. a "stable" position characterized by a specific pattern of gene expression levels that determines the particular kind (differentiation pattern) of the cell population (S. Huang 2009). Multiple attractor states can exist. The current stable state of the cells depends on the history of the past states of cell, implicating the importance of epigenetic mechanisms in such a context. The attractor states are robust, distinct and possess self stabilizing properties. The gene expression pattern associated with a particular state could be maintained even after the original stimulus that placed the cell in the current attractor state has been removed (S. Huang 2009). Of course, the attractor state is a property of the cell population, so its location in the phase space corresponds to the average expression levels for the millions of single cells over thousands of genes. When individual gene expression levels are measured, cells could be different for each other, thus, demonstrating intra-population variance. In this sense, attractor state viewed as an analogy to the definition of the temperature in statistical mechanics that allows for evaluation of the intrinsic differences between the components of the system (Huang 2009).

Earlier studies have indicated that the differentiation destinies of the progenitor cells could be defined as high dimensional attractor states of the underlying molecular networks (H. H. Chang et al. 2008; S. Huang et al. 2007). Particularly, a study of the differentiation trajectories of blood stem cells demonstrated that specific differentiated cell types behave as attractors (A. C. Huang et al. 2009). The same group provided some evidences of an analogous behavior of the cancer cells that are to be considered as located at the ‘periphery’ of the correspondent normal cell attractor for the same kind of tissue (S. Huang & Ingber 2006; S. Huang et al. 2009; Yuchun Guo et al. 2006). Although cancer was proposed as an attractor state of a cell as early as 1971 (S Kauffman 1971), a path to verify such a notion has been paved only recently, with an advent of the genomic technologies.

Under “attractor” paradigm, cell population is considered as a dynamic system that could be attracted to one or another “stable” state by transition that implies extensive mutual regulation of all elements of cell’s genome. This is in striking contrast with the traditional idea of a division of the mRNA transcripts into those generated by ‘housekeeping’ and ‘tissue-specific genes’, where a set of the master genes is responsible for the switch between different phenotypes. In their seminal paper Bar-Yam and colleagues describe this dichotomy. Particularly, the definitions for a ‘democratic’ (no master genes, all genes act as mutual regulators going toward a global attractor state) and an ‘autocratic’ (few master genes drive the differentiation process) regulatory landscape were introduced (Bar-Yam et al. 2009).

A possible middle ground between “democratic” and ‘autocratic” regulatory landscapes may be described as a general attractor-like behavior of the regulatory machinery with some local ‘vantage points’ representing genes most sensitive to dynamical changes of the system. Recent study of Tsuchiya et al demonstrated biphasic nature of the cellular response to innate immune stimuli involving an acute-stochastic mode consisting of small number of sharply induced genes and a collective mode where a large number of weakly induced genes adjust their expression levels to novel "stable" state. We hypothesize that similar regulatory scenario takes place during tumor development.

Hypothesis

Cancer is an attractor state. Normal cell can became cancerous and progress toward malignant phenotype using an intermediate regulatory framework that combines both local and global regulatory features.

Here we propose to perform a quantitative estimation of the relative importance of global and local features of gene expression regulation landscape in the process of tumor development through an analysis of publicly available microarray data.

Materials and Methods

Microarray datasets were extracted from the NCBI Gene Expression Omnibus as raw data (.CEL files) by selecting the data using Oncomine browser (Barrett & Edgar 2006; Barrett et al. 2007; D. R. Rhodes et al. 2007). To exclude cross-platform variability factors, only the datasets profiled using Affymetrix oligonucleotide arrays were chosen. The chosen datasets were classified into the following three categories: 1) Two-point datasets describing paired normal and tumor tissue samples collected from the same individual (N=8); 2) Two-point datasets describing a group of normal and a group of tumor samples collected from the same tissue type across a number of subjects (N=9); 3) Multi-point datasets describing three or more physiological groups of normal and tumor samples collected from same subject or across a number of subjects (N=7). The detailed descriptions of these datasets are given in the tables 1, 2 and 3 for each of categories, respectively.

Table 1: The table describes the attributes of two-point datasets describing paired normal and tumor tissue samples collected from the same individual.

GEO ID	Sample source	Number of samples	Total Number of transcripts extracted; Total number of transcripts significant by MW test	Reference
GSE5764	Invasive ductal (IDC) and lobular breast (ILC) carcinomas in postmenopausal patients	IDC (N=5) Normal ductal (N =5)	54675; 2278	(Turashvili et al. 2007)
		ILC (N=5) Normal lobular (N=5)	54675; 988	
GSE2514	Pulmonary adenocarcinoma and adjacent lung tissue	Lung AdCa (N=20)	12625; 5857	(Stearman et al. 2005)
		Normal (N =19)		
GSE7670	pulmonary adenocarcinoma and adjacent lung tissue	Lung AdCa (N=27)	22283; 8599	(Su et al. 2007)
		Normal (N =27)		
GSE6344	Renal cell carcinoma (RCC)	Stage 1 tumor (N=5)	44760; 23701	(Gumz et al. 2007)
		Stage 2 tumor (N=5)		
		Stage 1 normal (N=5)		
		Stage 2 normal (N=5)		
GSE781	Renal clear cell carcinoma (RCC)	Tumor (N=7)	44760; 11119	(Lenburg et al. 2003)
		Normal (N=7)		
GSE6631	Head and neck squamous cell carcinoma (HNSCC)	Tumor (N=22)	12625; 2880	(Kuriakose et al. 2004)
		Normal (N =22)		
GDS1665	papillary thyroid carcinoma (PTC)	Tumor (N=9)	54675; 13985	(H. He et al. 2005)
		Normal (N =9)		

Analysis was performed by R data analysis packages of Bioconductor. Affy package was used for the data processing and normalization (Reimers & Carey 2006; Gregory Alvord et al. 2007; Gautier et al. 2004). Perl scripting has been used to automate the analysis pipeline. The gene expression data were background corrected, normalized and the summarized expression values were calculated using Robust Multichip Average (RMA) method that consists of three steps: a background adjustment, quantile normalization and, finally, summarization (R. A. Irizarry et al. 2003). The expression values for individual genes in each of the cancer and normal samples were subjected to non-parametric Mann-Whitney test that extracted the transcripts with significant ($P < 0.05$) differential expression (Mann & Whitney 1947). The global and specific expression distances (DGlobal and DSpecific) were calculated based on the whole transcripts on the chip and significantly differentially expressing transcripts as selected by Mann-Whitney test, respectively. The distance between two samples i and j corresponds to : $Dij = 1 - Rij$, where Rij is the Pearson correlation coefficient between the vectors correspondent to i and j samples and having as dimensions the entire set of transcripts (DGlobal) or only the gene with statistically significant expression differences (DSpecific). These distance metrics were previously adopted for the dynamical characterization of microarray data and statistical analysis of microarrays in other studies (Hayden et al. 2009; H. H. Chang et al. 2008; Tsuchiya, Selvarajoo et al. 2009).

Table 2: The table describes the two-point datasets comprised of normal and tumor samples collected from the same tissue type across a number of subjects

GEO ID	**Sample source**	**Number of samples**	**Total Number of transcripts extracted;**	**Reference**
			Total number of transcripts significant by MW test	
GSE6791	Gene Expression Profiles of HPV-Positive and -Negative Head/Neck Cancers	Normal Head/Neck (N=14)	54675; 35778	(Pyeon et al. 2007)
		Head/Neck Cancer (N=42)		
	Gene Expression Profiles of HPV-Positive and –Negative Cervical Cancers	Normal Cervix (N=8)	54675; 25098	
		Cervical Cancer (N=20)		
GSE3678	Papillary thyroid carcinoma	Normal Thyroid (N=7)	54675; 5617	--
		Papillary thyroid carcinoma (N=7)		
GSE3524	Oral squamous cell carcinoma (OSCC)	OSCC (N=16)	22283; 5757	(Toruner et al. 2004)
		Normal (N=4)		
GSE10797	Transcriptomes of breast epithelium and stroma in normal reduction mammoplasty and invasive breast cancer patients.	Normal breast epithelium (N=5)	22277; 2491	(Casey et al. 2009)
		Invasive breast cancer epithelium (N=28)		
		Normal breast stroma (N=5)	22277; 1190	
		Invasive breast cancer stroma (N=28)		
GSE12345	Global gene expression profiling of human pleural mesotheliomas	Normal pleural tissue (N=8)	54675; 5995	--
		Mesothelioma tissue (N=8)		
GSE12452	mRNA expression profiling of nasopharyngeal carcinoma	Normal nasopharyngeal tissue (N=10)	54675; 15383	(Dodd et al. 2006; Sengupta et al. 2006)
		nasopharyngeal carcinoma (N=31)		
GSE14762	Renal Cell Carcinoma: Hypoxia and Endocytosis	Normal renal tissue (N=12)	54675; 18501	(Y. Wang et al. 2009)
		Renal carcinoma (N=10)		

Table 3: The table describes the datasets with three or more physiological groups of normal and tumor samples collected across the same subject or a number of subjects

GEO ID	Sample source	Number of samples	Total Number of transcripts extracted;	Reference
			Total number of transcripts significant by MW test	
GSE1420	Barrett's esophagus, Barrett's-associated adenocarcinomas and normal esophageal epithelium	Normal (N=8)	22283; 6552	(Kimchi et al. 2005)
		Barrett' esophagus (N=8)		
		Barrett's-associated adenocarcinoma (N = 8)		
GSE3325	Benign prostate, primary and metastatic prostate cancer samples	Benign prostate (N=6)	54675; 20667	(Sooryanarayana Varambally et al. 2005)
		primary prostate cancer (N=7)		
		metastatic prostate (N=6)		
http://dot.ped.med.umich.edu:2000/pub/Panc_tumor/index.html	Normal pancreas, chronic pancreatitis and pancreatic adenocarcinoma (microdissected)	Normal pancreas (N=5)	7129; 2289	(Logsdon et al. 2003)
		Chronic pancreatitis (N=5)		
		Pancreatic adenocarcinomas (n=10)		
GSE3167	Normal Bladder, superficial transitional cell carcinoma(sTCC), STCC with carcinoma in situ, metastatic transitional cell carcinoma, normal cystectomy and cystectomy with CIS	Normal Bladder (N=9)	22283; 13861	(Dyrskjøt et al. 2004)
		sTCC (N=15)		
		sTCC with CIS (N=13)		
		mTCC (N=13)		
		Cystectomy Normal(N=5) CIS (N=5)		
GSE6919	The Normal Prostate Tissue free of any pathological alteration., Metastatic Prostate Tumor, Primary Prostate Tumor, Normal Prostate Tissue Adjacent to Tumor	Normal Prostate Tissue free of any pathological alteration (N=17)	37757; 18973	(Yan Ping Yu et al. 2004; Chandran et al. 2007)
		Metastatic Prostate(N=25)		
		Primary Prostate (P=59)		
		Normal Prostate Tissue Adj to Tumor (N=62)		
GSE6764	Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma	Normal liver (N=10)	54675; 19250	(Wurmbach et al. 2007)
		Dysplastic liver tissue (N=17)		
		Cirrhotic liver tissue (N=13)		
		Very early HCC (N=8)		
		Early HCC (N=10)		
		Advanced HCC (N=7)		
		Very Adv HCC (N=10)		
GSE10971	Gene expression data from non-malignant fallopian tube epithelium and high grade serous carcinoma.	Normal controls (N=12)	54675; 15988	(Tone et al. 2008)
		BRCA-1/2 mutation carriers (N=12)		
		High grade serous carcinoma (N=13)		

Principal Component Analysis (PCA) was performed on the cancer microarray expression datasets based on the distance parameters (Roden et al. 2006). In this analysis, each sample is described by four distance based descriptors reflecting the average distance of each sample from i) cancer sample space (DC) and ii) normal sample space (DN) in both global and specific frames, therefore, producing following variables: DCglobal, DNglobal, DCspecific and DNspecific. PCA was performed using R on each of the datasets separately, in the four dimensional space represented by these parameters. The structure of correlations emerging from the analysis of the variable loadings on the extracted components allowed for a straightforward quantification of some relevant topological features of the analyzed systems.

Results and Discussion

a) Modeling strategy

The discrimination between a tumor and a normal sample can be achieved using both a summed expression change involving the entire set of mRNAs (DGlobal) and a summed expression change of the functionally important genes specifically involved in the development of the tumor state (DSpecific). In the case of the "democratic" regulatory landscape (no preferred vantage points, or particular mRNAs, specifically responding to the change of the physiological state), the discrimination would be achieved by DGlobal, while gene signature-based (DSpecific) distances should better reflect "autocratic" landscape with a profound changes in expression of master (or signature) genes while the great portion of mRNAs remain unaffected. In the latter case, the correlation between genome-wide (DGlobal) and signature-based (DSpecific) distances should not be substantial.

In case of an intermediate scenario, - a middle ground between “democratic” and “autocratic” regulatory landscapes, - the discrimination between tumor and normal sample calculated using DSpecific should be consistently better than the discrimination achieved using by DGlobal. However the two metrics should correlate, thus, demonstrating both the existence of a global attractor correspondent to the cell phenotype and reflecting the change of entire genome expression and the most influential roles for a specific set of the tumorigenesis-related genes.

The most natural metrics for estimating the distance between expression profiles of two biological samples is based on the Pearson correlation coefficient: the level of concordance of any two expression vectors correspondent to two different biological samples, x and y with n dimension (n = genes) and mean values of expressions $\bar{x}$ and $\bar{y}$ corresponds to their mutual Pearson correlation, $r = (\boldsymbol{x}, \boldsymbol{y})$ defined as:

$$r(\boldsymbol{x},\boldsymbol{y})=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{\sum_{i=1}^{n}X_iY_i}{\sqrt{\sum_{i=1}^{n}X_i^2}\sqrt{\sum_{i=1}^{n}Y_i^2}}=\frac{\boldsymbol{X}\cdot\boldsymbol{Y}}{|\boldsymbol{X}||\boldsymbol{Y}|}=cos\,\theta\,, \quad \text{Eq. 1}$$

where $\boldsymbol{X} = (x_1 - \bar{x}, x_2 - \bar{x}, .., x_n - \bar{x})$, $\boldsymbol{Y} = (y_1 - \bar{y}, y_2 - \bar{y}, .., y_n - \bar{y})$ correspond to the differences from the mean expression of each gene in the X and Y sample respectively and θ is the angle between two expression vectors. Geometrically, Eq. 1 shows the correlation coefficient can be viewed as the cosine of the angle on n-dimensional space between the two vectors of data which have been shifted by the average to have mean zero. Angle θ is a measure of the differences between the two vectors and consequently of the difference in expression pattern of the two sample, when θ = 0 (and consequently r = 1.0) the two expression patterns are completely coincident, and the two vectors are parallel. In the case of r = 1 (and consequently θ = 90 degrees) the two expression vectors are orthogonal, i.e. the expression patterns of the two samples are each other independent.

The measure Dij = 1-Rij with R = Pearson correlation coefficient between i and j samples can be considered as a distance between samples. This distance could vary from 0 (R = 1) reflecting the perfect resemblance of the two samples to 1 corresponding to maximal possible distance between two states (absence of correlation). In the case when samples are picked from two different sub-groups -- normal (N) and cancer (C) -- for each sample j analyzed two different descriptors DCj and DNj can be computed corresponding to the average distance of sample j from the spaces occupied by cancer (DCj) and normal (DNj) samples. Thus if (i) corresponds to a cancer sample DCi will be the average of all the pairwise distances of (i) vector from all the other cancer samples vectors, and consequently DNi the average of all the distances of (i) from the non-cancer samples.

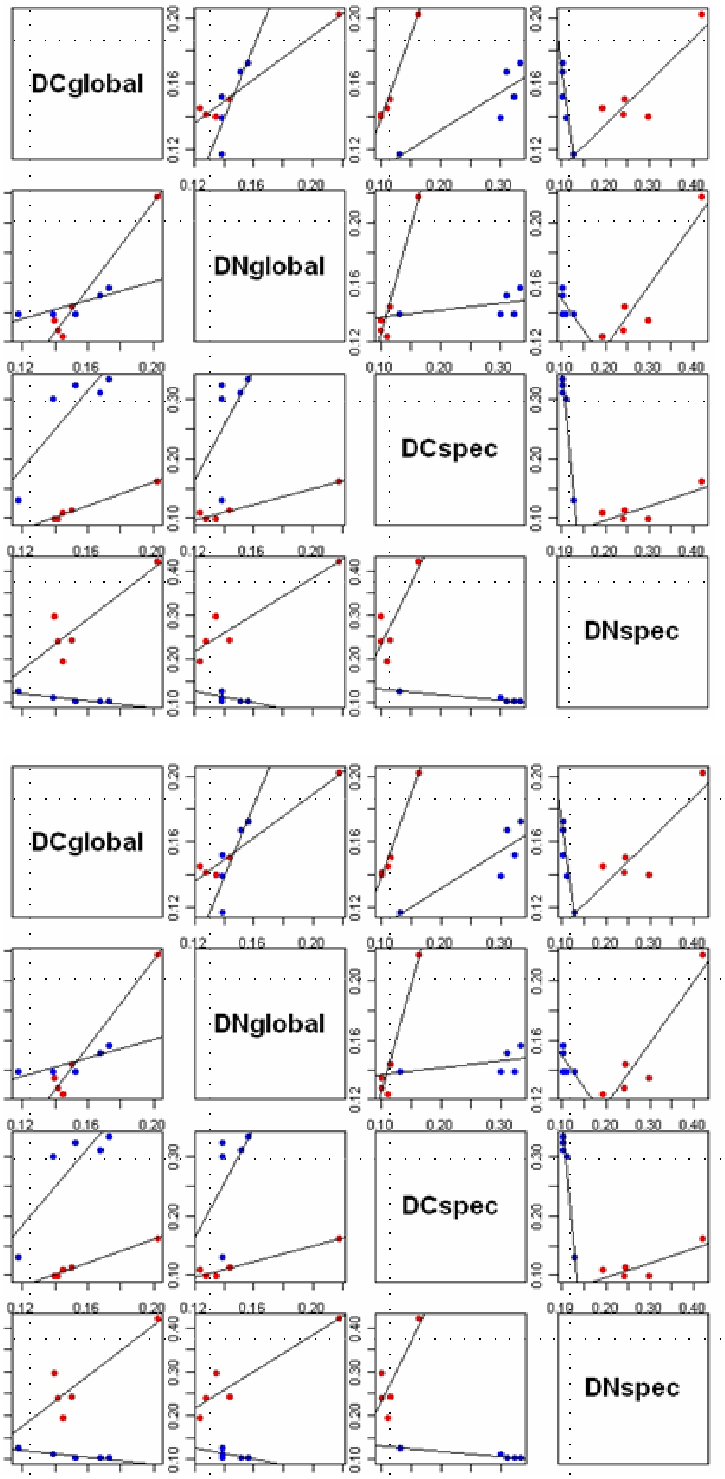
When the distance is computed only over the previously extracted differentiating gene signature defined as a set of genes with expression values significantly different between Cancer and normal subgroups by Mann-Whitney test, two similarly defined but gene signature-specific distance indexes (DCspecific, DNspecific) were obtained. In entirety, four descriptors were defined for every specific sample on each dataset:

- *DCglobal*: Genome-wide distance from cancer sample space to the particular sample
- DNglobal: Genome-wide distance from normal sample space to the particular sample
- DCspecific: Signature based distance from cancer sample space to the particular sample
- DNspecific: Signature based distance from normal sample space to the particular sample

b) Assessment of the global and signature-specific gene expression distances for two-point (Normal-Tumor) datasets

In this study we used a total of 17 two-point datasets represented by normal and tumor gene expression profiles. Paired datasets (tumor and normal samples derived form the same individual) and populational datasets (tumor and normal samples were collected across a number of subjects) were considered separately. Eight paired and nine population datasets profiled using the Affymetrix platforms were chosen for the two-point (normal-tumor) analysis (Tables 1, 2). For each dataset, the global and specific expression distances were calculated based either on the all probes present on the chip and passing the detection call (DNglobal and DCglobal) or on the genes highlighted as significantly differentially expressed according to Mann-Whitney test (DNspecific and DCspecific).

In both paired and population datasets, DC (global, specific) was greater than DN (global, specific) for most of the normal samples. The reverse was true, i.e. DC (global, specific) is less than DN (global, specific) for the tumor samples. Such a relation provides a basis an unbiased classification scheme, given a sufficiently relevant population of samples is achieved. Figure 1 depicts the four parameters as panels of paired plots for the lobular and ductal breast carcinoma dataset. The clear classification of the cancer and tumor samples using the complete chip data (global) using a simple metric like distance illustrate the differentiating power of the overall transcription. Moreover, ranking of the datasets based on global and specific distances of the tumor sample from the normal center were very similar, albeit not identical (Table 4). The conservation of global and specific distances across the datasets adds to the credibility of using this metric for diagnostic purpose.



Panel A

Panel B

Figure 1: The illustration depicts the distance parameters derived for the paired breast carcinoma dataset (GSE5764) using a paired-plot panels. Panel A and Panel B represent the distance parameters for the lobular and ductal carcinoma with in this dataset. Each point depicts individual cancer (red) or normal (blue) sample, line reflects a linear fit for each group of samples

Table 4: Rankings of the tumor malignancy potential according to the relative distance to the Normal Sample Space (two-point paired datasets) 1 – lowest; 9 – highest

DATASET	**Mean (DGlobal) from individual tumor samples to the Norma center**	**Mean (DSpecific) from individual tumor samples to the Normal center**
GSE2514 (pulmonary adenocarcinoma)	1	1
GDS1665 (papillary thyroid carcinoma)	2	2
GSE781 (RCC)	3	4
GSE6344 (RCC stage 2)	5	3
GDS2520 (HNSCC)	4	6
GSE6344 (RCC stage 1)	6	5
GSE7670 (pulmonary adenocarcinoma)	7	7
GSE5764 (ductal breast cancer subset)	8	8
GSE5764 (lobular breast cancer subset)	9	9

In case when distances were calculated using DNglobal, in all studied paired data, tumors were further away from the Normal Sample Space than the control samples with normal histology (Table 5). On average, for normal samples the distance to the Normal Space defined by DGlobal was 0.047+/-0.045 as compared to 0.080+/- 0.034 for Tumor samples ($P < 0.038$) in paired datasets. Distances between individual Normal samples and the Normal Space defined by DSpecific were also significantly different from that calculated for Tumor samples (Normal: 0.044+/-0.034; Tumor: 0.138 +/- 0.063, $P < 0.001$). All metrics were heavily correlated to each other. This correlation indicates strong attractor-like behavior; the discussion on this would be continued in the PCA results section. Here it is important to stress that signature-based and genome-wide approaches allow for the same level of discrimination efficiency of the data sets.

Table 5: Mean, Standard Deviation and Variance calculated for Global and Specific Distances from individual samples to the Normal Sample Space of the paired datasets

DATASET	Mean +/- SD variance (DNglobal) from individual normal samples to the Normal Sample Space	Mean +/- SD variance (DNglobal) from individual tumor samples to the Normal Sample Space	Mean +/- SD variance (DNspecific) from individual normal samples to the Normal Sample Space	Mean +/- SD variance (DNspecific) from individual tumor samples to the Normal Sample Space
GSE5764 (ductal breast cancer subset)	0.0989+/-0.0111 0.0001231	0.1134+/-0.0196 0.0003861	0.0634+/-0.00595 0.00003547	0.1827+/-0.02951 0.000870855
GSE5764 (lobular breast cancer subset)	0.1449+/-0.0084 0.0000704	0.1496+/-0.0389 0.00151395	0.1092+/-0.01037 0.00010758	0.2788+/-0.0873 0.00762137
GSE2514 (pulmonary adenocarcinoma)	0.0113+/-0.0015 0.0000023	0.0407+/-0.0199 0.000399112	0.0138+/-0.00211 0.0000044	0.0688+/-0.03296 0.001086227
GSE7670 (pulmonary adenocarcinoma)	0.0399+/-0.0104 0.000107128	0.0841+/-0.0285 0.000814786	0.0483+/-0.01129 0.000127647	0.1417+/-0.04826 0.002329823
GSE781 (RCC)	0.0187+/-0.008 0.0000646	0.0624+/-0.0087 0.0000751	0.0234+/-0.0128 0.0001639	0.1247+/-0.01585 0.0002513
GDS2520 (HNSCC)	0.0577+/-0.0151 0.000227314	0.0742+/-0.0141 0.000197704	0.0789+/-0.01866 0.000348429	0.1362+/-0.02979 0.000887682
GDS1665 (PTC)	0.0184+/-0.002 0.0000039	0.0407+/-0.0133 0.0001773	0.0168+/-0.002107 0.0000044	0.0785+/-0.0276 0.0007636
GSE6344 (RCC stage 1)	0.0216+/-0.0019 0.0000038	0.0802+/-0.0058 0.0000337	0.0219+/-0.00208 0.0000043	0.1213+/-0.00702 0.0000494
GSE6344 (RCC stage 2)	0.0196+/-0.0022 0.0000048	0.0758+/-0.0096 0.0000926	0.0201+/-0.00265 0.0000070	0.1098+/-0.01424 0.0002028

Similar to that in paired datasets, by DNglobal, tumors in all the population datasets were further away from the Normal Sample Space than the control samples with normal histology (Table 6). On average, for normal samples the distance to the Normal Space defined by DGlobal was 0.0520+/-0.021 as compared to 0.095+/- 0.032 for Tumor samples (P < 0.012). Distances between individual Normal samples and the Normal Space defined by DSpecific were also significantly smaller than that that calculated for Tumor samples (Normal: 0.054+/-0.018; Tumor: 0.154 +/- 0.029, P < 0.00078). The concordance between the populational and paired data sets allows us to exclude the hypothesis the ‘between distances’ correlation is driven by ‘individuality effects’, i.e. by the fact each single individual has a specific gene expression pattern accounting for the observed global/specific distance from tumor / distance from normal concordance.

C) Assessment of the global and signature-specific gene expression distances of Multi-stage (three or more stage) datasets

There were a total of 7 datasets describing tumor and normal samples collected from the same subject (1 dataset) or across a number of subjects (6 datasets). The development of the tumor usually involves its progression from the relatively benign to invasive and to metastatically aggressive phenotypes (Merlo et al. 2006). It is widely accepted that the gene expression signatures are able to discriminate between distinct stages of the tumor development. To explore the idea whether a summed expression change involving the entire set of mRNAs behaves similarly to the changes in signature-specific, “master” genes, we calculated DNGlobal and DNSpecific for 7 (six from NCBI GEO and one external) datasets representing normal and tumor samples that were comprised of three or more distinct physiological states of the underlying tissue.

Table 6: Mean, Standard Deviation and Variance calculated for Global and Specific distances from individual samples to the Normal Sample Space in the populational datasets

DATASET	DNGlobal		DNSpecific	
	From individual normal samples to the Normal Sample Space (Mean +/- SD; variance)	From individual tumor samples to the Normal Sample Space (Mean +/- SD; variance)	From individual normal samples to the Normal Sample Space (Mean +/- SD; variance)	From individual tumor samples to the Normal Sample Space (Mean +/- SD; variance)
GSE6791 (cervical cancer)	0.05721 +/- 0.01671; 0.000279167	0.13059 +/- 0.02493; 0.0006216929	0.064005 +/- 0.01937; 0.000375219	0.1585726 +/- 0.03052591; 0.000931831
GSE10797 (invasive breast cancer)	0.08878211 +/- 0.018546943; 0.0003439891	0.1480193 +/- 0.04274649; 0.0018272624	0.0475659 +/- 0.009414584; 0.0000886343	0.1545566 +/- 0.03623782; 0.0013131799
GSE12345 (pleural mesothelioma)	0.06323201 +/- 0.01296917; 0.0001681993	0.0871369 +/- 0.02157966; 0.0004656815	0.0789417 +/- 0.01844578; 0.0003402469	0.1960226 +/- 0.05226790; 0.0027319334
GSE12452 (nasopharyngeal carcinoma)	0.05510947 +/- 0.01769130; 0.0003129822	0.077843 +/- 0.013091253; 0.0001713809	0.0707538 +/- 0.01992132; 0.0003968591	0.1413587 +/- 0.027503111; 0.0007564211
GSE14762 (RCC)	0.02229638 +/- 0.004879693; 0.0000238114	0.1080666 +/- 0.09848668; 0.009699626	0.0302542 +/- 0.007646735; 0.0000584726	0.1875954 +/- 0.09617942; 0.009250482
GSE6791 (HNSCC)	0.05799383 +/- 0.01747614; 0.0003054155	0.0834743 +/- 0.01661218; 0.0002759646	0.0641543 +/- 0.01976013; 0.0003904628	0.1060674 +/- 0.02143241; 0.0004593481
GSE3678 (papillary thyroid carcinoma)	0.04147274 +/- 0.006705467; 0.00004496329	0.0560819 +/- 0.005507370; 0.0000303311	0.0493836 +/- 0.009431665; 0.0000889563	0.1582217 +/- 0.009836067; 0.0000967482
GSE3524 (oral squamous cell carcinoma)	0.02964479 +/- 0.006389468; 0.0000408253	0.0715533 +/- 0.01914830; 0.0003666572	0.0298668 +/- 0.005858421; 0.0000343211	0.1318646 +/- 0.03691830; 0.0013629609

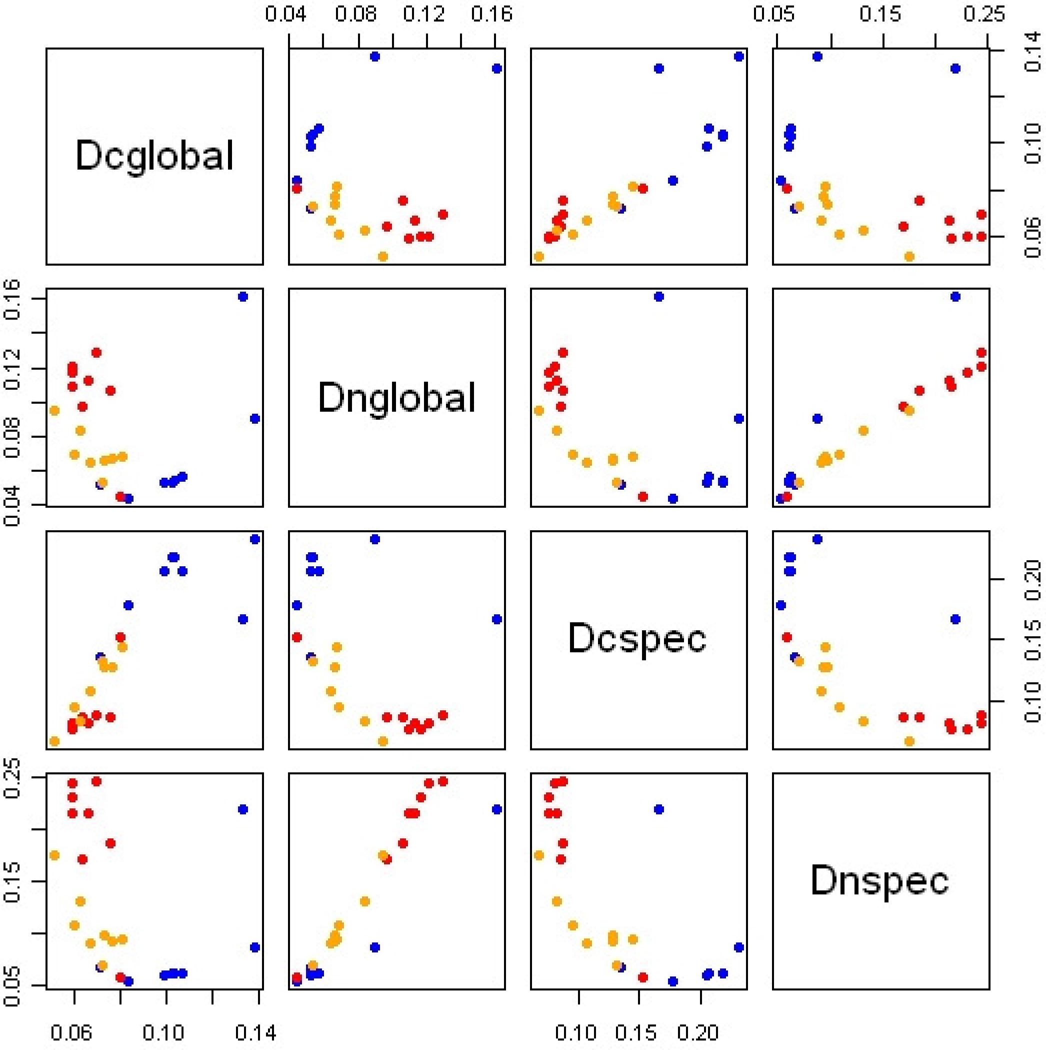


Figure 2: Distance parameters successfully separate samples in the esophageal sample (GSE1420) dataset representing normal esophagus (blue), Barrett's esophagus (orange) and esophagus carcinoma (red) samples

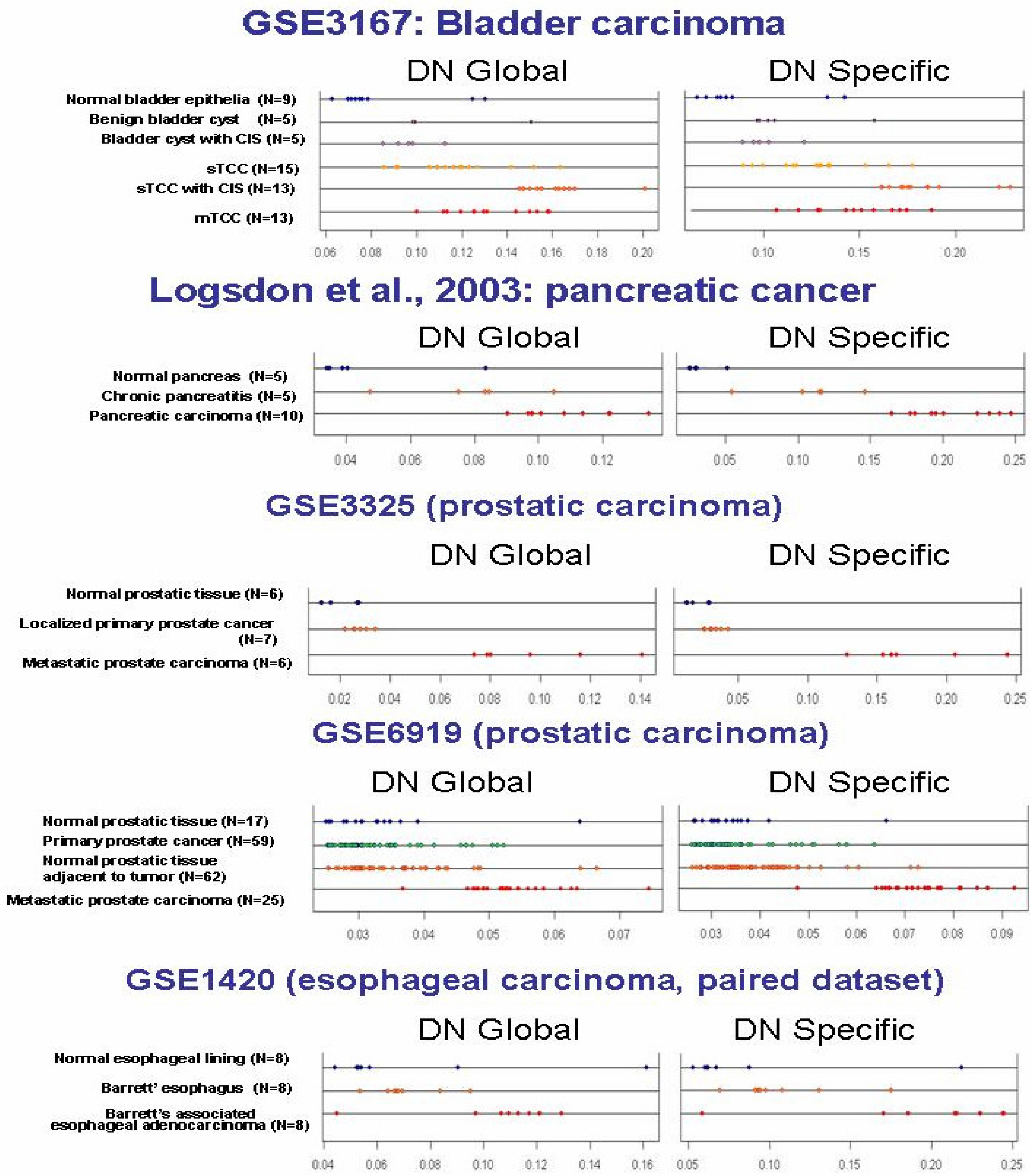
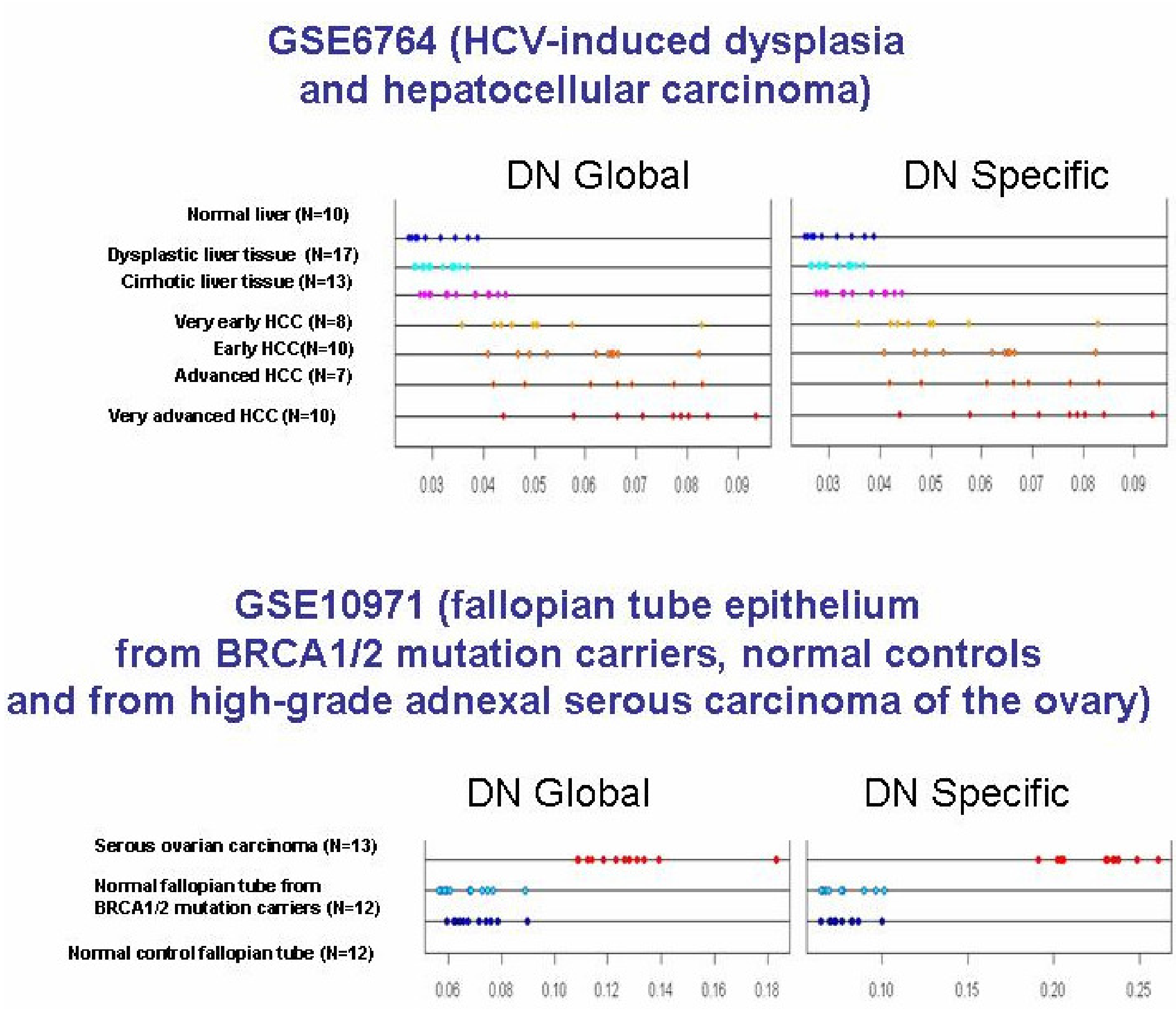


Figure 3(a): Linear graphs depicting the relative distance of every given sample to the Normal sample space as defined by DNGlobal and DNSpecific metrics in the multi-stage datasets.


GSE6764 (HCV-induced dysplasia and hepatocellular carcinoma)

GSE10971 (fallopian tube epithelium from BRCA1/2 mutation carriers, normal controls and from high-grade adnexal serous carcinoma of the ovary)

Figure 3(b): The figure illustrates the linear graphs of the DN metric for the multi-stage datasets GSE6764 and GSE10971. Various stages in the progression are depicted in each of these datasets.

As GEO database contains only one dataset, GSE1420 (Figure 2), that is represented by paired tissue samples profiled using Affimetrix platform, we added to this study 6 datasets comprised of the samples collected across a number of individuals and profiled using the same microarray platform (Table 3). For each dataset, the global and specific expression distances were calculated as described above. In all datasets, the progression of the disease was reflected in an increase of the distance of individual tumors from Normal Sample Space.

For each of these datasets linear graphs were generated. Each graph depicts the relative distance of every given sample to the Normal Sample Space as defined by DNGlobal and DNSpecific metrics (Figure 3). As could be seen at the Figure 1, both DNGlobal and DNSpecific place the most malignant tumors farther from the normal tissue control than the least malignant tumors or relatively benign tumors precursor states. The only case when metastatic tumors were less distant from the Normal Tissue Space than primary tumors, was the comparison of metastatic transitional cell carcinomas (TCC) of the bladder and superficial TCC with carcinoma in situ (TCC-CIS) (dataset GSE3167). This discrepancy might be explained by previous observations that the presence of concomitant CIS confers a worse prognosis in patients TCC (Shariat et al. 2007) et al., 2007). In all the cases when easy visual discrimination of the tumor and normal/benign samples could be achieved, the performances of DNGlobal and DNSpecific were comparable. These results suggest that the genome-wide metrics may help to assess the ‘degree of malignancy’ of the tumor cells.

d) Principal component analysis (PCA) of the distance spaces

In addition to the direct correlation between indexes, the degree of the mutual correlation between DNGlobal and DNSpecific distances could be quantified by the principal component analysis (PCA) on the four dimensional space spanned by these four indexes (DCglobal, DNglobal, DCspecific, DNspecific). PCA gives us an immediate quantitative appreciation of the relative importance of the architectural modes of gene regulation. Typical results of the PCA analysis of the two-point and multipoint (one for each type) datasets are reported in Table 7. The patterns of the component loading are remarkably consistent across all the 24 (including multi-stage) datasets analyzed. The proportion of the variation observed is also similar across the datasets. The variance data for the two-point data can be observed for paired and population datasets in Tables 8 and 9, respectively

In the four dimensional space, the PCA generated four components reflecting the variation in the data. The first component (PC1) is the largest one. In this component all the indexes enter with the same direction of correlation (loading sign). This component might reflect the presence of the attractor. The proportion of the variance it explains reflects the relative importance of attractor (cell type) driven dynamics in gene expression regulation. As all the distance indexes are positively correlated along this axis and as the distance from this attractor is equally measured by all the distance indexes adopted (DNglobal, DNspecific, DCglobal, DCspecific), this attractor corresponds to the center of distribution, and the PC1 (distance from the attractor) has the same sign as measured by any of the indexes. PC1 component explains by far major portion of information contained in the expression profiles and, given the homogeneity of signs, it reflects a topological ‘distance from a centre’ (here, a center of attractor) from which all the samples could have either lesser or higher distance independently of being cancer or normal samples.

Table 7: The table illustrates the relative importance of components and the actual loadings corresponding to the distances in the two-point datasets GDS1165 and GSE12345. The highlighted pattern of loadings is consistent across all the datasets. The results of the PCA analysis of all other datasets can be found in the Appendix.

Two-point dataset: Papillary thyroid carcinoma dataset (GDS1665)				
Relative importance	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
Standard deviation	0.0968	0.0398	0.00375	0.00120
Proportion of Variance explained by component	0.8542	0.1444	0.00128	0.00013
Cumulative Proportion	0.8542	0.9986	0.99987	1.00000
Component Loadings:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
DCGlobal	-0.4055784	0.1936700	-0.4882024	0.7481019
DNGlobal	-0.3365074	-0.2185903	-0.7043803	-0.5855164
DCSpecific	-0.6383106	0.6270132	0.3455199	-0.2828959
DNSpecific	-0.5610958	-0.7221944	0.3822601	0.1322270

Multi-stage dataset: Mesothelioma (GSE12345)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.244	0.0846	0.00892	0.00362
Proportion of Variance	0.892	0.1071	0.00119	0.0002
Cumulative Proportion	0.892	0.9986	0.9998	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.34642	0.145146	-0.63865	0.671609
DNGlobal	-0.34828	-0.08457	-0.59061	-0.72299
DCSpecific	-0.51471	0.780904	0.334231	-0.11643
DNSpecific	-0.70269	-0.60164	0.362763	0.112533

The second component (PC2) puts in opposition (opposite loading signs) the distances from cancer (DC) and normal (DN) poles. The topological structure described by PC2 corresponds to the fact that normal and cancer poles do in effect occupy distinct positions in the gene expression space and thus, as for this structure, there must be a component of the distances indexes reflecting the relatively higher (lower) distance of a sample from the Normal or Tumor pole (Figure 4). The modulation driven by Tumor/Normal relative distance is definitively less important than the cell-kind attractor, as is inferred from the observation that the portion of the variance explained by PC2 is considerably lower than the portion explained by PC1. Along this component, DNspecific-DNglobal indices enter with the same loading sign, while being in opposition to the DCspecific-DCglobal pair.

Table 8: PCA profiles of two-point paired datasets representing the proportion of variance observed by each component

Proportion of Variance / Dataset	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)
Ductal Breast Carcinoma (GSE5764)	0.908	0.0901	0.00107
Lobular Breast Carcinoma (GSE5764)	0.882	0.116	0.00199
Pulmonary adenocarcinoma (GSE2514)	0.8635	0.1361	0.00022
Pulmonary adenocarcinoma (GSE7670)	0.917	0.0815	0.00108
Renal cell carcinoma (GSE6344)	0.777	0.2231	0.00023
Renal cell carcinoma (GSE781)	0.781	0.219	0.00055
Head and neck squamous cell carcinoma (GSE6631)	0.954	0.0436	0.00252
Papillary thyroid carcinoma (GSE3467)	0.8542	0.1444	0.00128
Esophagus Carcinoma (GSE1420)	0.875	0.124	0.124

The third component (PC3) reflects the 'degree of autonomy' of the signature genes from the global behavior of the cell-kind attractor. Relative strength of PC3 tells us whether signature genes possess intrinsic difference from the components of the general expression landscape or simply represent transcription units most sensitive to the common regulatory signal. Latter behavior is registered by PC2, while purely 'democratic' behavior of gene expression profile is registered by PC1. Intuitively, the loading pattern of PC3 component (the loadings correspond to the correlation coefficient of the original variables with the components) should have the specific (DNspecific,DCspecific) and global (DNglobal,DNspecific) indexes entering with opposite signs.

Table 9: PCA profiles of two-point population datasets representing the proportion of variance observed by each component

Proportion of Variance / Dataset	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)
Invasive Breast (Epithelial) Carcinoma (GSE10797)	0.978	0.0192	0.00233
Invasive Breast (Stromal) Carcinoma (GSE10797)	0.986	0.012	0.0013
Cervical Carcinoma (GSE6791)	0.884	0.1153	0.00029
Head and Neck Carcinoma (GSE6791)	0.967	0.0319	0.00072
Mesothelioma (GSE12345)	0.892	0.1071	0.00119
Nasopharyngeal Carcinoma (GSE12452)	0.934	0.0655	0.00062
Oral Squamous Cell Carcinoma (GSE3524)	0.914	0.0857	0.00059
Renal Cell carcinoma(GSE14762)	0.891	0.1027	0.00669
Papillary thyroid carcinoma (GSE3678)	0.814	0.1847	0.00094

The proportion of the variation explained by fourth component (PC4) was negligible in all the cases compared to three previously discussed components. The PC4 might represent the 'background' noise generated by the stromovascular or other cells that may be present in the analyzed tissue samples. The PC4 would explain the smallest proportion of observed variation between sample sets. Its relatively small size reflects the strict quality controls used in the procedure of the selection of the published high-throughput datasets used in the current study.

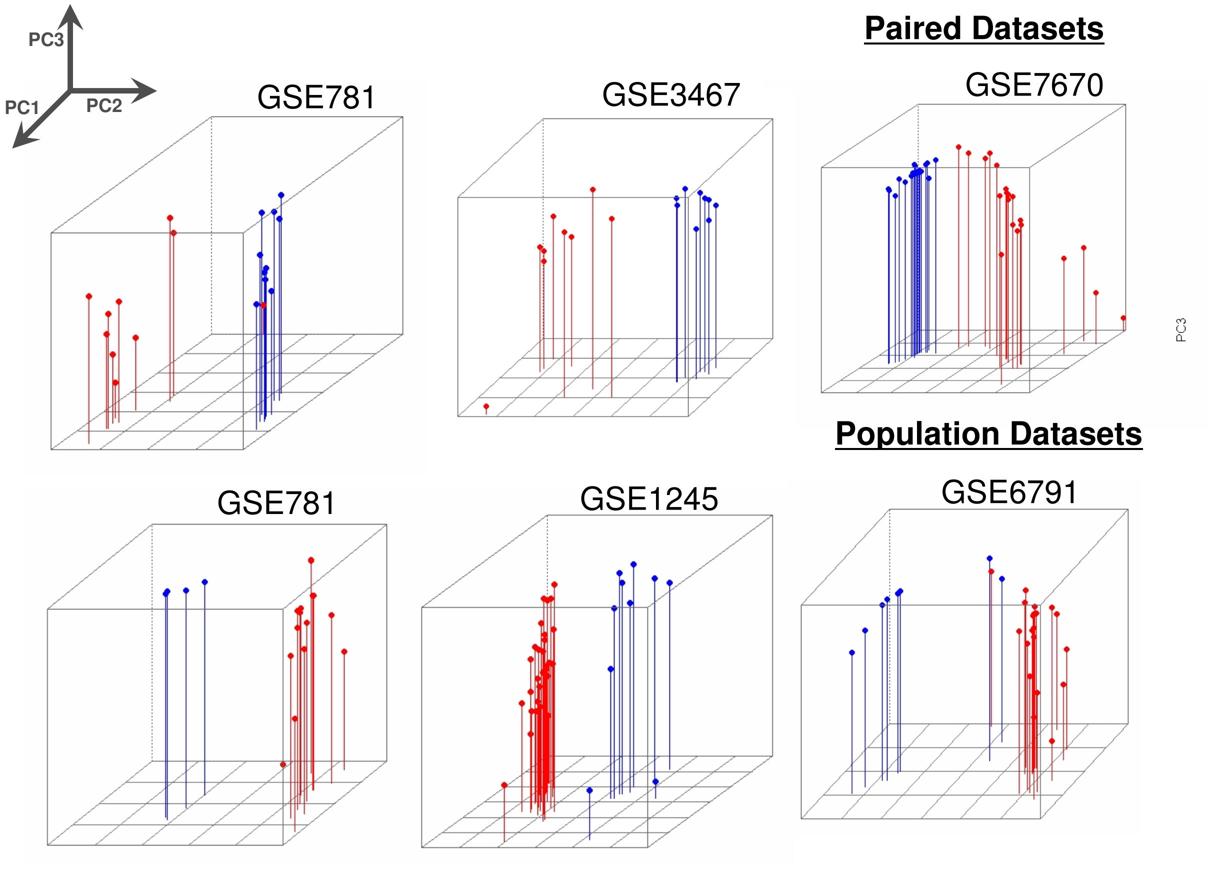


Figure 4: Three dimensional representation of the principal components PC1, PC2 and PC3 in the two-point paired and population datasets. Normal samples are shown in blue and tumor samples are shown in red. This figure specifically highlights the classification power of PC2 (Normal/Cancer classifier) that does not require selection or validation of the minimized expression signature.

In analyzed datasets, the relative importance of cell-kind driven gene expression regulation (PC1) was ranged from 77% to 98%, while the distinction between normal and cancer poles (PC2) was ranged from 22% to 1%. The 'degree of autonomy' (signature genes working independently of global attractor dynamics) was represented by smallest component (PC3) being less than 1% in all datasets with an exception of esophageal dataset (GSE1420).

e) Cancer – An attractor with intermediate regulatory framework

Results of the principal component analysis could be used to discern the topological structure of cancer and cell-kind attractors. Our observations support the hypothesis of cancer being a stable attractor state in the dynamic system with intermediate regulation architecture could be described as a midpoint between "democratic" and "autocratic" regulatory landscapes. The intermediate paradigm is illustrated through an analysis of PC2 that is able to "readily sense" the difference between Normal and Cancer samples using both specific and global distance measures. Despite the fact that specific indices (gene signatures) enter as higher loadings on PC2 as compared to global distance indexes, latter indices also play a substantial role. In the case of purely 'democratic' architecture, PC3 would be expected to accounts for only a very small portion of variation; otherwise, at least some degree of autonomy of signature, or 'master', genes shall be acknowledged. Thus, after analysis of the principal components we conclude the canalization of the tumor development towards the stabilization of the cell population in the cancer attractor state follows the intermediate paradigm [not fully "democratic" or not fully "autocratic"].

It is worth noting, that the use of the distances (instead of the differences in the expression levels for individual genes) allows for an unbiased estimation of the regulatory paradigm in the living system, as each descriptive parameter of the system (global, specific, normal, tumor) is described by numerical value and evaluated as such, being not affected by the number of genes that passed some arbitrary significance threshold chosen for individual dataset. The cancer attractor model arising from the results obtained in the present study is depicted in the Figure 5.

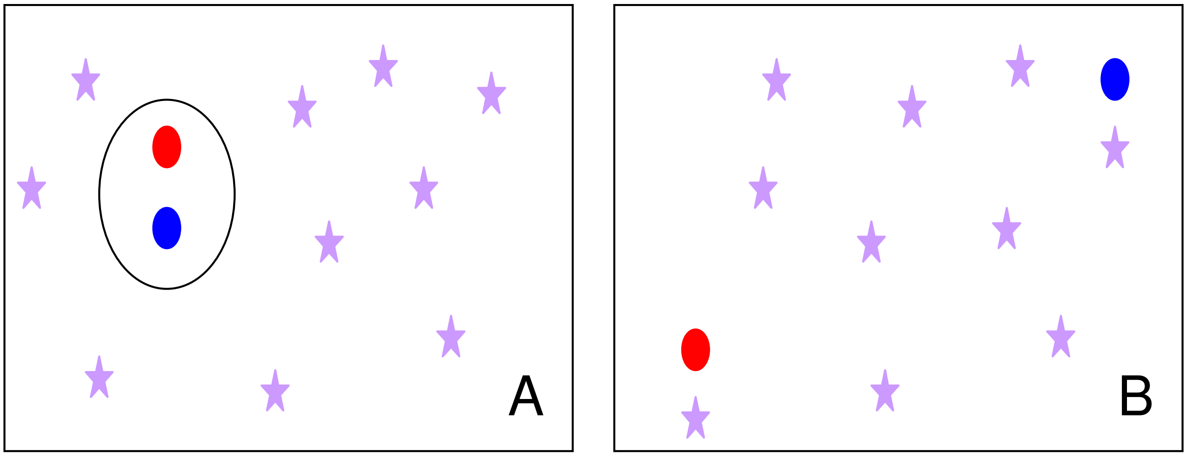


Figure 5: Panel A describes the topology of the cell-kind and tumor attractors supported by present study. Panel B reports the classical view of cancer. The red and blue circles represent the cancer and normal attractor states as distinct poles. The rectangle represents the phase space of possible gene expression profiles, the stars are the observed samples, while the ellipse represents the general cell-kind attractor. From this model we can derive that the cells that by one or another reason leave "stable state" and depart from the normal attractor may with relatively high probability be attracted to the road toward cancer attractor without the prerequisite of getting departed from relatively strong cell-kind attractor.

As could be seen from the Figure 5, the topology of the cell-kind and tumor attractors supported by present study closely follows the Huang's hypothesis stating that the cancer is a sub-attractor of the general cell kind attractor (S. Huang et al. 2009). The main component defining the location of the sample in the space occupied by all samples is its distance from the general cell-kind attractor, thus the samples far removed from the normal subattractor are also distant from the cancer subattractor(PC1 component). In case of PC1, DN and DC indices are correlated and enter with the same sign into the component. The second component, PC2, discriminates if a given sample is closer to the cancer or normal sub-attractors (PC2 has opposite signs for DN and DC). Therefore, the similarity between cancer and normal samples is greater than the difference between them. In other words, prostatic cancer cell remains a prostate cell after all. Notable, this view is substantially different from the "classical" understanding of the tumorigenesis, when tumor and normal cells occupy the opposite poles of the allowed expression space (Fig.5, Panel B). If the"classical" model was correct, PC1 should have DN and DC indices entering with opposite signs reflecting negative correlation values.

A case study performed on the breast carcinoma dataset (GSE10971) may serve as a good illustration for an attractor model. The multi-stage dataset comprises luteal phase fallopian tube epithelium from BRCA1/2 mutation carriers and from normal controls as well as the samples of the high-grade adnexal serous carcinoma of the ovary. Traditional analysis of this data collected using Affymetrix microarrays highlighted specific gene signature that passed multiple test correction places. This gene signature places fallopian tube epithelium from BRCA1/2 mutation carriers close to the high-grade serous carcinoma samples (Tone et al. 2008). Our analysis of both Global and Specific distance charcteristics indicated that the normal epithelial samples collected from the patients predisposed to ovarian carcinoma have not yet embarked on the travel toward “cancer” attractor (Figure 3b). Other three-point datasets also provided clear discrimination between normal and malignant states, while providing relatively poor discrimination for the true normal and pre-malignant samples (Figure 3a). The only case when surefooted discrimination was possible at the earliest stages of the carcinogenesis was a set of samples representing the progression of the hepatocellular carcinoma (dataset GSE6764, Figure 3b). All together, our observations point that that the shift toward cancer attractor either takes place relatively late in the process of carcinogenesis or requires some time to become substantial. This observation also goes well with the hypothesis that cancer-specific changes of the expression landscape are subject to intermediate regulatory pattern, representing the middle ground between “democratic” and “autocratic” regulatory landscapes.

Conclusion and Future Perspective

Here we presented quantitatively evidence supporting the structure of the cancer attractor previously suggested by Huang and the hypothesis that cancer-specific changes of the expression landscape are subject to intermediate regulatory pattern, representing the middle ground between “democratic” and “autocratic” regulatory landscapes.. The remarkable similarity of the observations made using multiple independent datasets, including these comprised of multiple types of samples demonstrates robustness of the genome-wide expression signatures as a mean to diagnose tumors. This study supports the view of the cell population as dynamic system. Moreover, the strong correlation between the ‘distance from normal’ and ‘distance from cancer’ poles for all the analyzed samples proves existence of a cell-kind-attractor, with cancer and normal poles representing two sub-attractors.

There are a number of immediate applications of the analyses performed. First, after initial sets of normal and tumors samples for each particular cancer are analyzed to define Normal and Cancer Spaces, the classification of any new sample to be diagnosed could be achieved by calculation sample specific distance from this sample to Normal Space (DN) and Cancer Sample (DC). If DN > DC sample will be classified as cancer, If DC > DN, sample will be classified as normal. An increase in the number of the initially profiled samples will provide for better definition of the Normal and Cancer Spaces and better classification of the subsequent samples. Second, for every sample to be diagnosed, the distance from the sample to the Normal Space could be plotted linearly, and the degree of the malignancy of the given sample will be proportional to the linear distance. Importantly, relative degree of the malignancy could be assigned to the sample using whole-genome patterns of the gene expression, without the need for specific biomarkers or gene signatures. Third, the principal component analysis (PCA) on the four dimensional space spanned by four indexes (DCGlobal, DNGlobal, DCSpecific, DNSpecific) could be used for diagnostic discrimination of the sampels. Each new sample to be diagnosed should be added to initial (reference) dataset of the cancer and normal tissues of the particular cell-type, PCA executed at whole dataset, then first three components (PC1, PC2, PC3) should be used for three dimensional graphing of the results. New samples will be co-classified with the group of the samples with similar degree of the malignancy.

Cell populations are collective dynamic systems living in a phase space where only very specific low energy states (cell kind attractors) are compatible with survival. These attractor states define cell differentiation. When cell departs from its cell-kind attractor, there are only three possible scenarios. One, cell could die as a result of a profound deregulation of its molecular networks incompatible with survival. Second, cell could be attracted back to the normal pole of the cell-kind attractor. Third, cell could randomly fall under the influence of the cancer pole of the cell-kind attractor, and acquire tumorigenic properties. We are still far from the exploiting a statistical mechanics of life, but our data suggest that, in principle, this can be done. The 'cell kind' barriers are energetically much higher than the normal/cancer one, thus, offering a possibility of the 'global reversion' of cancer phenotype. It might be possible to find the way to "kick" the cell out of equilibrium, and, therefore, out of the influence of cancer pole of cell-kind attractor. Being removed from low energy state, cell will be pushed to face three possible fates again: death, normalization or attracting back to the cancer pole. Of course, the molecular or other mean of the 'global reversion' therapy should be delivered specifically to the cancer cells. 'Global reversion' therapy cannot be based on the exploitation of 'master key genes', but should rely on more general means, for example, previously postulated morphogenetic fields sharing some similarities in embryonic and cancer cells.

3. Abundance based transcriptome analysis as a tool for automated discovery of the tumor biomarkers

Rationale

The purpose of the current study is to explore the composition of the human transcriptome over a wide range of normal tissues and tumors using EST abundance analysis. We hypothesize that analysis of EST abundance might help to identify novel biomarkers of cancer initiation and progression.

Background

The enormous scale in which cancer affected mankind in the past century emphasizes an importance of both prevention and early stage detection of this devastating disease. Indeed, biomarkers project a promising future for the early stage detection of cancer (Oluwadara & Chiappelli 2009; Cazzaniga et al. 2009; A. Scott & Salgia 2008). Additionally, prognostic biomarkers provide vital information influencing therapeutic decisions (Ludwig & John N Weinstein 2005; J. J. Liang et al. 2009; Hwa et al. 2008). A number of bioinformatics and machine learning methods for the detection of biomarkers have been developed and utilized previously (Phan et al. 2009).

Unlike genomes comprised of relatively stable, species-specific DNA, tissue transcriptomes are very dynamic in nature. The functional and structural landscape of a particular cell phenotype depends heavily on the relative frequency of the transcription of individual genes. These frequencies, usually described as expression levels, are prone to change under the influence of the environmental and internal stimuli (Martínez & Reyes-Valdés 2008). Gene expression can be quantitatively measured at the transcriptional level by a number of low- to high-throughput methods.

The inventory of the human transcripts has increased dramatically in recent years, to include large number of non-coding mRNAs. Accumulating data on non-protein-coding transcripts suggest that besides the regulatory machinery driven by proteins, another yet enigmatic regulatory network of RNA molecules operates and has considerable impact on cell functions (Széll et al. 2008; Pontius et al. 2007; Waterston et al. 2002). Indeed, exonic sequences cover only 1.1% of the Human genome; the majority of not-yet-processed and spliced transcripts are represented by intronic and intergenic sequences (J. C. Venter et al. 2001). Therefore, it is not surprising that aside from mutations and polymorphisms in protein-coding genes, much of the variation between individuals, including that which may affect our predispositions to cancer, is probably due to differences in the non-coding regions of the genome (Mattick JS 2003).

In the course of an analysis of tissue-specific transcriptomes, many non-coding transcripts have been identified. Additionally, the class of the transcripts with relatively low coding potential has been described. Later transcripts encode only short open reading frames (50–70 amino acids) and, in many cases, these ORFs lack Kozak sequences at translation start sites or are not evolutionarily conserved. Many of these non-coding transcripts RNAs with low coding potential were predicted using bioinformatic methods. For example, a study of Washietl et al, 2005 evaluated conserved genomic DNA sequences for signatures of structural conservation in base-pairing patterns with exceptional thermodynamic stability and predicted more than 30,000 structured RNA elements in the human genome (Washietl et al. 2005). Almost a 1,000 of these sequences were found to be conserved across all vertebrates (Washietl et al. 2005). Chromosome tiling experiments using DNA microarrays also demonstrated that most of the genome sequences are transcribed, and that many introns encode for the novel RNA species (Weile et al. 2007; Kapranov et al. 2005).

The role of non-coding RNAs, particularly miRNA, in the context of cancer has been recently established (Visone & Croce 2009; Conrad et al. 2006). In many cases non-coding RNA species has been shown to regulate the alternative splicing of essential proteins, a phenomenon that has great implication in inflammation, disease and cancer (Mallardo et al. 2008). Interestingly, the use of tiling microarrays revealed genome-wide hyper-transcription in mouse embryonic stem cells (ESCs), including expression of normally silent, non-coding regions. This hyper-transcription reflects the unusual "open" structure of ESC chromatin and contributes to the plasticity of the stem cells. Hyper-transcription points represent additional commonality between ESC and cancer (Efroni et al. 2008; B. M. Turner 2008). Thus, it is logical to assume that non-coding RNAs could become valuable source for novel prognostic and diagnostic biomarkers for human malignancies.

DNA Microarrays and protein arrays, with their inherent ability to capture the diverse aspects of cancer are the most commonly used data source for the discovery of novel biomarkers. Most often, these arrays are fabricated to cover only the protein encoding genes (Ambros 2001). In a number of previous studies, meta analysis was attempted using a compendium of microarrays allowing one to mine for novel markers of cancer (Xu et al. 2007; Wren 2009). However, these high-level analyses in most part were restricted to searches for the markers performing in the context of a particular subtype of cancer. Multi-cancer analysis attempts are less common. One example of this kind of studies is analysis of expression profiles covering 60 human cancer cell lines of NCI-60 panel spanning 9 different human tissues (Shankavaram et al. 2007).

Complete transcriptome studies of cancer cells that include non-coding transcripts are warranted. One barrier to such studies is relative difficulty of obtaining expression profiles that are comprehensive enough to cover low-abundancy intra and intergenic transcripts. However, these difficulties may be solved using publicly available data describing EST sequences. Expressed Sequence Tags (ESTs) were introduced during the initiation phase of high-throughput cDNA sequencing and quickly became useful in the identification of novel genes and mapping of the genomic sequences (M. D. Adams et al.

1991). On average, these sequences are relatively short, with a length ranging from 400 to 600 bases, and are relatively inaccurate (NCBI 2002). Many genome projects utilized the EST based approach to gene mapping, resulting in an ever-increasing number of ESTs in the databases maintained by NCBI. ESTs provide a quick and inexpensive roadmap not only to identify novel genes, but also to obtain data on gene expression and gene regulation in various tissues. The highly redundant nature of ESTs makes them an excellent material for the RNA expression quantification *in silico*. The study of cancer transcriptomes using both coding and non-coding ESTs would help to comprehend the cancer cell expression patterns down to a better degree than the classic gene-coding RNA studies.

The expression patterns of both tumor tissues and their normal counterparts may be profiled by analyzing the largest publicly available expression database, the UniGene. This system is developed, maintained and updated by the National Center for Biotechnological Information at National Institute of Health (NCBI, NIH). UniGene is comprised of non-redundant set of gene-centered clusters of the expressed sequences, including mRNAs, ESTs and high throughput cDNA (HTC) sequences (NCBI 2002; D. L. Wheeler et al. 2007). To avoid false alignment within the gene clusters, the repetitive stretches of the nucleotides are masked (Schneeberger et al. 2005). High quality reads on the templates of expressed sequences with a sequence length of at least 100 base pairs are used for subsequent clustering. UniGene forms the basis for other core NCBI resources utilized in the routine searches for protein similarities and putative evolutionary relationships and in comparisons of EST-based expression profiles of various tissues (NCBI 2002).

ESTs form the building blocks of UniGene and represent an integral part of nearly every gene cluster in the UniGene. The EST data in the UniGene database can be utilized to study the expression patterns of genes in normal tissues and tumor samples. This methodology enables a study of the abnormal expression of genes in cancer through the direct comparison of the diverse express patterns of genes within a wide range of human tissues and tumor samples and will possibly lead to the discovery of new human tumor marker candidates.

Hypothesis

Abundance based meta-analysis of tissue-specific transcriptomes may reveal novel diagnostic and prognostic biomarkers for human tumors. Here we attempted to profile the variation in the abundance of ESTs derived from normal tissues and tumors and available through UniGene database system and to extract candidate biomarker genes for various human malignancies.

Materials and Methods

To compare the patterns of the gene expression in the normal and tumor EST libraries, the UniGene build 210 of *Homo sapiens* was downloaded from the NCBI ftp portal of the Unigene database. Descriptions of the UniGene cluster data and the cDNA libraries associated with each of these clusters were deconvoluted. This particular version of the UniGene database contained 123,687 UniGene clusters associated with 8668 cDNA libraries used to generate all contributing ESTs. Perl scripts and MySQL database management system were used to automate the data extraction and to provide for the data storage, respectively. R data analysis package with Bioconductor was utilized to perform the statistical analysis on the downloaded dataset (Gregory Alvord et al. 2007; Mark Reimers & Carey 2006).

The abundance of ESTs was used as the fundamental quantitative unit describing the expression levels similarly to the hybridization intensity of the DNA microarray. The abundance of the ESTs belonging to a particular gene cluster in a given tissue of the Unigene was defined as the ratio of a number of ESTs that belong to the particular gene cluster and are expressed in a given tissue to the total number of ESTs captured from that particular tissue. The terms “UniGene cluster”, “cluster” and “unigene” are used interchangeably in the following text. Diversity coefficients of the tissue transcriptomes as represented by abundance measures for various UniGene clusters were calculated using Shannon diversity index (Shannon 1948). This index was formally defined in a recent study addressing diversity in the context of transcriptome (Martínez & Reyes-Valdés 2008). The richness (quantity of unigenes) and the evenness of each tissue were calculated using Pielou index. A t-test was used to compare the diversities in the normal and the corresponding tumor tissues (Magurran 1988). The significance of the variation between the paired normal and cancer human tissues was tested using non-parametric

Mann-Whitney procedure (H. Mann & Whitney 1947). Metastats, an expression variation analysis tool that handles sparsely sampled features was used to compute the false discovery rate (White et al. 2009). Gene cluster related information for the significant unigenes was extracted from the data provided in the current version of the UniGene database. Functional analysis was performed using a population of genes with expressions significantly skewed toward tumors. For this analysis we used the KEGG pathway painter (KPP) specially developed for this project using the KEGG API (Kawashima et al. 2003). The detailed description of the KPP tool is given in the Appendix.

Table 10: A list of broader tumor descriptors used in this study.

Cancer Type
Liver tumors
Leukemia
Lymphoma
Pancreatic tumors
Head and neck tumors
Germ cell tumors
Cervical tumors
Breast (mammary gland) tumors
Thymoma
Uterine tumors
Testicular tumors
Esophageal tumors
Adrenal tumors
Kidney tumors
Meningioma
Retinoblastoma
Multiple myeloma
Adenoid cystic carcinoma (unspecified tissue)
Gallbladder tumors
Prostate tumors
Lung tumors
Nerve tissue tumors
Bladder tumors
Bone tumors
Ovarian tumors
Gastrointestinal tumors
Schwannoma/glioma/astrocytoma
Skin tumors
Non-malignant tumors (benign tumors/cysts/benign proliferative disease)

Results and Discussion

a) Development of the standard vocabulary describing human tissues

The text descriptors annotating all the tissues and cancer types that served as a source for the production of the cDNA libraries comprising UniGene were extracted. For most of these libraries, data descriptions containing information about their tissue of origin and type of the tumor were available. The complete list of tissue descriptors includes a total of 64 tissue categories and 51 different types of cancers originated in these tissues.

These library descriptors were compressed into broader tissue descriptors. For example, terms “hepatic cancer”, “hepatic carcinoma”, “hepatic tumor”, “liver cancer”, “liver tumor”, “HCC” etc were co-classified as “Hepatic tumors”. Re-classification of the descriptors resulted in a final vocabulary comprised of 37 different tissues and 28 types of tumors derived form these tissues (Table 10). The descriptions of cDNA libraries stored in the MySQL database were automatically updated using these broadly defined classes that were utilized in further analysis.

b) Classification of the cDNA libraries used in the study

For the purpose of this study, all cDNA libraries were classified in four different broad categories: i) Normal, ii) Cancer, iii) Non-malignant diseases (e.g. Parkinson disease, ischemia) and iv) Libraries lacking proper tissue descriptions.

For each tissue, we have calculated number of normal and cancer libraries, numbers of gene clusters (unigenes) and total number of ESTs derived from normal and cancer libraries (Table 11), as well as libraries made of the diseased tissues and libraries lacking tissue descriptions. The normal and the cancer groups represented the largest part (87%) of the expression data represented by the ESTs. There were also a substantial number of ESTs lacking proper descriptions (12.5%), while the amount of ESTs from the diseased tissues and/or benign tumors was negligible compared to the other three groups. The diseased tissue samples were mostly restricted to the bone and muscle diseases (e.g. Duchenne dystrophy), illustrating an unevenness of efforts in characterization of transcriptome changes in non-malignant disorders (Table 12). ESTs from the non-malignant diseased tissues and tissues lacking proper descriptors were excluded from further analysis.

According to our observations, cumulative efforts aimed at the cancer transcriptomes, though, seems to be more advanced, as UniGene database contains more cancer cDNA libraries (N = 4642) than the libraries built using normal tissues (N = 2682). An overall excess of cancer cDNA libraries spans an entire dataset, with every tissue covered by twice as much of tumor-derived libraries than normal libraries. On the other hand, the total number of ESTs derived from each normal tissue, on average, is substantially higher compared to that derived from respective cancer. This paradox can be explained by the fundamental objective of the Human Genome project to catalogue all the genes in the normal human genome.

As could be seen from Table 11, the distribution of ESTs across tumor and normal tissues is far from being well-balanced. For example, categories “heart”, “vasculature”, “peritoneum”, “ear” and “pineal gland” lack tumor ESTs, while “olfactory mucosa”, “chest wall” and “penis” lack ESTs derived form normal tissues. This discrepancy was, to certain extent, normalized by taking into account per tissue abundance for each unigene.

c) Unique and Common Unigenes

After extraction the data describing ESTs and gene clusters, it became feasible to separately extract tissue-specific unigenes represented by both normal and tumor ESTs, tissue-specific unigenes built solely upon ESTs derived from tumor libraries and tissue-specific unigenes that contain only ESTs derived from normal libraries (Table 12).

For example, according to the Table 12, in the brain the number of “normal unigenes” represents the number of EST clusters where at least one EST was derived from normal brain samples; the number of “cancer unigenes” reflect the number of EST clusters, where at least one EST was derived from brain tumor samples; the number of “unique or exclusive normal unigenes” is a number of EST clusters, where all brain-specific ESTs were derived from normal brain samples; the number of “exclusive cancer unigenes” represent the number of EST clusters, where all brain-specific ESTs were derived from brain tumor samples; number of “common unigenes” is a number of EST clusters, where at least one EST was derived from normal brain samples and at least one EST was derived from brain tumor sample. Tissues with sparse representation by ESTs are removed from the table. EST clusters summarized in the Table 12 represents both the protein coding as well as non-coding genes.

Table 11: Statistics describing the libraries, unigenes (gene clusters) and ESTs derived from normal tissues and tumors samples. Normal unigenes are defined as gene clusters that contain at least one EST from the particular normal tissue; Cancer unigenes are defined as gene clusters that contain at least one EST from the cDNA library representing a particular tumor type.

Tissue	Normal libraries	Normal unigenes	Normal ESTs	Cancer libraries	Cancer unigenes	Cancer ESTs
Brain	497	34604	787007	87	18642	174529
Tissues lacking descriptors	59	39259	352400	174	24057	223929
Eye	56	24976	160932	3	8086	39123
eart	32	14619	77193	0	0	0
Liver	39	15068	99207	54	13687	107936
Kidney	31	19817	107596	87	17058	89069
Bone	9	6157	16172	22	13702	50358
Adrenal gland	12	6029	19762	14	6274	13221
Muscle	19	15397	75841	6	5592	23898
Testis tissue	170	25782	147432	73	13859	104177
Pregnancy tissue	472	31392	512435	9	7091	30351
Pancreas	19	14908	105643	18	14142	102394
Lung tissue	159	22985	158569	630	23543	195964
Lymphatic tissue	93	23380	220930	342	16707	148729
Vasculature	18	9876	46530	0	0	0
Lower gastrointestinal tissue	240	13571	83030	1040	24175	229598
Skin	27	13031	74232	33	13274	119604
Adipose tissue	13	5468	12417	1	558	720
Mammary gland	330	11156	47829	767	16565	106111
Upper gastrointestinal tissue	13	4075	12363	262	12470	72982
Peritoneum	5	180	291	0	0	0
Female reproductive tissue	27	12859	53761	318	24819	250839
Parathyroid	1	24	24	2	7045	20602
Prostate	144	16667	81107	165	14591	107355
Salivary gland	4	912	2514	3	2891	10355
Connective tissue	8	4770	16389	17	16599	82051
Thyroid	79	4700	12031	280	9706	33611
Pineal gland	4	3437	6353	0	0	0
Ear	6	5523	16378	0	0	0
Pituitary gland	7	5403	13430	2	923	1392
Nerve tissue	7	9582	25579	7	609	701
Bladder tissue	12	3304	8550	53	6611	18295
Olfactory mucosa	0	0	0	1	2	2
Chest wall	0	0	0	1	3	3
Penis	0	0	0	2	6	8

As could be seen from the Table 12, there is a substantial number of unigenes that include both normal and tumor ESTs derived from the same tissue ("common unigenes").

These common unigenes corresponds to the parts of the cellular machinery in one or another way essential for running both normal and tumor cells derived from the same tissue. Existence of a large number of such unigenes signifies the inevitable need to fulfill basic metabolic and other cellular function. However, it seems that the sizes of the gene sets supporting basic functions differ from tissue to tissue.

As inferred from the Table 12, the total number of unigenes in normal and cancer groups was classified as either exclusive or common unigenes. The portion of exclusive and common unigenes in the normal and cancer categories is illustrated as percentage of total unigenes in each category (Table 12).

The category “pregnancy-related tissues” corresponds to the largest number of exclusive normal unigenes. Our definition of “pregnancy-related tissues” includes umbilical cord, amniotic fluid and embryonic tissues. The high number of gene clusters that contain at least one EST from tissues of this category can be attributed to the fact that embryonic parts represent all tissues of the human body.

Next in the tissue list ranked by number of exclusively normal unigenes are brain and eye followed by testis. These patterns correspond to the diversity patterns previously reported for these tissues (Martínez & Reyes-Valdés 2008; Piatigorsky 1989). In case of exclusive cancer unigenes these numbers might be influenced both by an intrinsic versatility of expression and the variability among the assortment of cancer subtypes derived of the particular tissue. Largest number of unigenes that include at least one tumor EST derived from a given were found in the categories “lower gastrointestinal tissues”, followed by “connective tissue”, “uterus’ and “ovary”.

Table 12: The table summarizes tissue-specific distribution of EST clusters composed of both normal and tumor ESTs, tumor ESTs only and normal ESTs only.

Tissue	Normal unigenes	Cancer unigenes	Exclusive Normal unigenes (ENU)	Exclusive Cancer unigenes (ECU)	Common unigenes (CU)	% of ENU	% of CNU	% of ECU	% of CCU
Brain	34604	18642	19392	3430	15212	56.04	43.96	18.40	81.60
Pregnancy – related tissues	31392	7091	24706	405	6686	78.70	21.30	5.71	94.29
Testis tissue	25782	13859	15725	3802	10057	60.99	39.01	27.43	72.57
Eye	24976	8086	17787	897	7189	71.22	28.78	11.09	88.91
Lymphatic tissue	23380	16707	10960	4287	12420	46.88	53.12	25.66	74.34
Lung tissue	22985	23543	8662	9220	14323	37.69	62.31	39.16	60.84
Kidney	19817	17058	8300	5541	11517	41.88	58.12	32.48	67.52
Prostate	16667	14591	7091	5015	9576	42.55	57.45	34.37	65.63
Muscle	15397	5592	11343	1538	4054	73.67	26.33	27.50	72.50
Liver	15068	13687	6263	4882	8805	41.56	58.44	35.67	64.33
Pancreas	14908	14142	6068	5302	8840	40.70	59.30	37.49	62.51
Lower gastrointestinal tract	13571	24175	2539	13143	11032	18.71	81.29	54.37	45.63
Skin	13031	13274	4345	4588	8686	33.34	66.66	34.56	65.44
Mammary gland	11156	16565	3198	8607	7958	28.67	71.33	51.96	48.04
Uterus	11084	20800	2097	11813	8987	18.92	81.08	56.79	43.21
Nerve tissue	9582	609	9298	325	284	97.04	2.96	53.37	46.63
Bone	6157	13702	1997	9542	4160	32.43	67.57	69.64	30.36
Adrenal gland	6029	6274	3287	3532	2742	54.52	45.48	56.30	43.70
Ovary	5503	14698	1377	10572	4126	25.02	74.98	71.93	28.07
Adipose tissue	5468	558	5241	331	227	95.85	4.15	59.32	40.68
Pituitary gland	5403	923	4905	425	498	90.78	9.22	46.05	53.95
Connective tissue	4770	16599	1030	12859	3740	21.59	78.41	77.47	22.53
Thyroid	4700	9706	1764	6770	2936	37.53	62.47	69.75	30.25
Upper gastrointestinal tract	4075	12470	986	9381	3089	24.20	75.80	75.23	24.77
Bladder tissue	3304	6611	1646	4953	1658	49.82	50.18	74.92	25.08
Salivary gland	912	2891	644	2623	268	70.61	29.39	90.73	9.27
Parathyroid	24	7045	8	7029	16	33.33	66.67	99.77	0.23

The percentage of unigenes contributing to the exclusive and common pools by the normal and cancerous gene clusters is also summarized in the table. In the cases of tissues with substantial number of unigenes (Eg: Brain, testis, eye, lung, kidney), the cancer unigenes tend to show a lot of commonality with the normal pool of unigenes. Interestingly, the exclusive gene clusters of the cancerous tissue were underrepresented when compared to the pool of exclusive normal unigenes. It might push the notion that the spectrum of transcriptional activity in normal tissue is higher than that of the cancer tissue. If this is the case, then the increased transcription observed in cancer might be due to amplified transcription of already active genes rather than increased transcriptional diversity.

d) Estimation of the diversity of transcripts within normal and tumor tissues

An introduction of the tissue-specific measures of the diversity may shed light on the versatility of the transcription of human genome in particular tissue contexts. An estimation of the diversity has been was performed through quantification of the tissue labels attached to ESTs contributing to unigenes that are expressed in each tissue (or tumor) of interest.

Diversity is the composition of two fundamental components: a) Variety and b) Relative abundance. The tissue-specific "richness" coefficients (S) depict the diversity pattern in a way that reflects not only the number of unigenes *per se*, but also the expression levels (abundance) of these unigenes. Thus, the "richness" of each unigene contributes to the tissue-specific "richness". Although as a heterogeneity measure, Shannon's index takes into account the evenness of the abundance of species; it is possible to calculate a separate additional measure of evenness (Magurran 1988). Evenness is the normalized Shannon and falls in the range of 0 and 1. An evenness of 1 indicates that the all unigenes contributing to tissue transcriptome are equally abundant, while an evenness of 0 indicates that there is absolutely no commonality in the abundance of unigenes. Both the normal tissue and tumors demonstrate substantial evenness of the distribution of abundances of the ESTs contributing to the gene clusters. High values of the evenness factor indicate that the process of the normalization of EST abundance successfully reduced an initial bias of the UniGene database, where tissues were unevenly covered by EST libraries.

Comparison of the diversity measures calculated for normal and cancerous tissues as represented by the abundance of contributing ESTs is presented in Tables 13 and 14. Among human tumors, the most diverse pattern of expression was observed in the category: “lower gastrointestinal cancer” comprising gastric, colonic and intestinal malignancies. The lung cancer tissue group consisting of tumors associated with lung, bronchus, pleura, trachea, larynx and pharynx displayed second most highly diverse pattern. Observed diversity patterns observed also provided interesting data concerning expression patterns in normal human tissues. According to our observations, eye tops the list of the tissues with highly diverse expression patterns, thus, confirming a conclusion inferred after the quantification of exclusively normal unigenes (Table 12). Second place in the list is occupied by testis.

Table 13: Diversity statistics describing of diversity, richness, evenness and variance of diversity reflecting the abundance of ESTs in each of the normal human tissues.

Tissue	Number of Sequences	Diversity H'	Richness S	Evenness E	Variance of H'	Standard Error of H'
Eye	160932	9.31899	24976	0.920333	1.18E-05	0.003435
Testis tissue	147432	9.250885	25782	0.91075	1.63E-05	0.004042
Pregnancy tissue	512435	9.013833	31392	0.870539	5.31E-06	0.002305
Kidney	107596	8.950741	19817	0.904636	2.55E-05	0.005045
Lung tissue	158569	8.946467	22985	0.890852	1.80E-05	0.004248
Prostate	81107	8.883301	16667	0.913808	2.99E-05	0.005467
Female reproductive	53761	8.859325	12859	0.936326	3.19E-05	0.005646
Lymphatic tissue	220930	8.833965	23380	0.878159	1.33E-05	0.003644
Brain	787007	8.804831	34604	0.842429	4.39E-06	0.002095
Skin	74232	8.713065	13031	0.919576	2.42E-05	0.004923
Heart	77193	8.70495	14619	0.907704	3.26E-05	0.005707
Nerve tissue	25579	8.553971	9582	0.933061	7.60E-05	0.00872
Muscle	75841	8.449439	15397	0.876323	4.65E-05	0.006819
Mammary gland	47829	8.400288	11156	0.901344	6.12E-05	0.007825
Pancreas	105643	8.316743	14908	0.865457	4.31E-05	0.006565
LowerGastrointestinal tissue	83030	8.208109	13571	0.862587	4.55E-05	0.006743
Vascular	46530	8.08016	9876	0.878482	7.21E-05	0.008489
Adipose tissue	12417	8.06399	5468	0.936947	0.000141	0.011867
Bone	16172	7.874164	6157	0.902447	0.000206	0.014359
Ear	16378	7.80055	5523	0.905285	0.00017	0.013049
Pineal gland	6353	7.785438	3437	0.956165	0.000204	0.014297
Adrenal gland	19762	7.602619	6029	0.873429	0.000195	0.01397
Liver	99207	7.601805	15068	0.790181	7.87E-05	0.008873
Pituitary gland	13430	7.510406	5403	0.873841	0.000351	0.018742
Thyroid	12031	7.501853	4700	0.887235	0.000338	0.018382
Connective tissue	16389	7.355674	4770	0.868428	0.000208	0.014424
Bladder tissue	8550	7.272814	3304	0.897558	0.000324	0.018008
UpperGastrointestinal tissue	12363	7.166962	4075	0.862178	0.000325	0.018018
Salivary gland	2514	5.292427	912	0.776512	0.001936	0.044001
Peritoneum	291	4.893458	180	0.942326	0.003833	0.061913

The diversity patterns of the normal tissue and corresponding tumor tissue were compared by calculating the tissue specific T-statistics (Table 15). According to this measure, the diversity patterns in the tumor and normal lung tissue appears to be similar. With the exception on lung tissue, all other human tissue demonstrated significant changes in the diversity after malignization. The diversity pattern in the lung tumors might be overestimated due to the tumor set comprising of lung, bronchus, pleura, trachea, larynx and pharynx tissues. The lack of variation between the normal and tumor lung tissue diversity might be due to an overestimation in the diversity in the normal tissue due to large variety of the epithelial tissue components of this organ that is comprised of lung, bronchus, pleura, trachea, larynx and pharynx.

Table 14: Diversity statistics describing of diversity, richness, evenness and variance of diversity reflecting the abundance of ESTs in each of the human tumors.

Tissue	Number of Sequences	Diversity H'	Richness S	Evenness E	Variance of H'	Standard Error of H'
Lower Gastro-intestinal tissue	229598	9.028835	24175	0.894558	9.99E-06	0.003161
Lung tissue	195964	8.954499	23543	0.889527	1.44E-05	0.003793
Female reproductive	250839	8.924292	24819	0.881902	1.13E-05	0.003357
Connective tissue	82051	8.868648	16599	0.912685	2.56E-05	0.005058
Bone	50358	8.853011	13702	0.929421	3.74E-05	0.006112
Kidney	89069	8.834877	17058	0.906664	2.99E-05	0.005466
Mammary gland	106111	8.782686	16565	0.904029	2.19E-05	0.004683
Brain	174529	8.679998	18642	0.882726	1.71E-05	0.004138
Lymphatic tissue	148729	8.596777	16707	0.884116	1.86E-05	0.004308
Prostate	107355	8.504952	14591	0.887027	2.44E-05	0.004939
Testis tissue	104177	8.479037	13859	0.889096	2.44E-05	0.004942
Upper Gastro-Intestinal tissue	72982	8.407071	12470	0.891422	3.60E-05	0.006
Pancreas	102394	8.392749	14142	0.878187	3.09E-05	0.005563
Thyroid	33611	8.384825	9706	0.91333	7.34E-05	0.008569
Parathyroid	20602	8.383832	7045	0.946249	6.29E-05	0.007932
Skin	119604	8.377987	13274	0.882491	2.20E-05	0.004694
Liver	107936	8.286482	13687	0.870045	3.27E-05	0.005723
Adrenal gland	13221	8.232684	6274	0.941506	0.000129	0.011374
Bladder tissue	18295	8.165420	6611	0.928259	0.000104	0.010189
Pregnancy tissue	30351	7.966101	7091	0.898441	8.31E-05	0.009115
Eye	39123	7.833186	8086	0.870558	9.87E-05	0.009933
Muscle	23898	7.677075	5592	0.889674	0.00011	0.010511
Salivary gland	10355	7.2727	2891	0.912583	0.000192	0.013858
Nerve tissue	701	6.274774	609	0.978626	0.001336	0.036549
Pituitary gland	1392	6.217004	923	0.910566	0.002162	0.046495
Adipose tissue	720	6.033908	558	0.954074	0.002033	0.045086

Interestingly, “human tumor” categories were almost evenly split into those with increased expression diversity when compared to its normal tissue counterpart, and those with decreased diversity. As could be seen form tables 15(a) and 15(b), tumors with increased diversity mostly originate from epithelial cells (carcinomas), while tumors with decreased diversity were mostly represented by the malignancies of non-epithelial tissues, e.g. brain (gliomas and astrocytomas), pituitary gland, nerve, muscle. The only exception to this list was skin. However, in our classification skin tumors included melanomas that are derived of melanocytes (non-epithelial skin components). Importantly, the changes in the diversity patterns observed in this study cannot be explained by the change in the alternative splicing, as our datapoints represent EST clusters (unigenes) rather than ESTs.

Table 15(a): The following table illustrates the tissues having increased diversity in tumors compared to normal tissue. T-statistics reflects the diversity in the EST abundance data of the Unigene system in human normal tissues and the corresponding cancerous tissues.

Tissue	*Normal*			*Cancer*			T-stat	P-value
	Number of ESTs	**Diversity**	**Variance**	**Number of ESTs**	**Diversity**	**Variance**		
Lung tissue	158569	8.946467	1.80E-05	195964	8.954499	1.44E-05	-1.4110	0.158321
Female-reproductive	53761	8.859325	3.19E-05	250839	8.924292	1.13E-05	-9.8843	<0.00001
Mammary gland	47829	8.400288	6.12E-05	106111	8.782687	2.19E-05	-41.948	<0.00001
Pancreas	105643	8.316743	4.31E-05	102394	8.39275	3.09E-05	-8.8356	<0.00001
Lower GI tissue	83030	8.208109	4.55E-05	229598	9.028836	9.99E-06	-110.17	<0.00001
Bone	16172	7.874164	0.000206	50358	8.853011	3.74E-05	-62.717	<0.00001
Adrenal gland	19762	7.602619	0.000195	13221	8.232684	0.000129	-34.975	<0.00001
Liver	99207	7.601805	7.87E-05	107936	8.286482	3.27E-05	-64.869	<0.00001
Thyroid	12031	7.501853	0.000338	33611	8.384826	7.34E-05	-43.537	<0.00001
Connective tissue	16389	7.355674	0.000208	82051	8.868648	2.56E-05	-98.977	<0.00001
Bladder tissue	8550	7.272814	0.000324	18295	8.16542	0.000104	-43.141	<0.00001
Upper GI tissue	12363	7.166962	0.000325	72982	8.407071	3.60E-05	-65.300	<0.00001
Salivary gland	2514	5.292427	0.001936	10355	7.2727	0.000192	-42.926	<0.00001

Table 15(b): The following table illustrates the tissues having decreased diversity in tumors compared to normal tissue. T-statistics reflect the diversity in the EST abundance data of the Unigene system in human normal tissues and the corresponding cancerous tissues.

Tissue	*Normal*			*Cancer*			T-stat	P-value
	Number of ESTs	**Diversity**	**Variance**	**Number of ESTs**	**Diversity**	**Variance**		
Eye	160932	9.31899	1.18E-05	39123	7.833186	9.87E-05	141.345	<0.00001
Testis tissue	147432	9.250885	1.63E-05	104177	8.479038	2.44E-05	120.985	<0.00001
Pregnancy tissue	512435	9.013833	5.31E-06	30351	7.966101	8.31E-05	111.429	<0.00001
Kidney	107596	8.950741	2.55E-05	89069	8.834878	2.99E-05	15.5664	<0.00001
Prostate	81107	8.883301	2.99E-05	107355	8.504952	2.44E-05	51.3443	<0.00001
Lymphatic tissue	220930	8.833965	1.33E-05	148729	8.596777	1.86E-05	41.9949	<0.00001
Brain	787007	8.804831	4.39E-06	174529	8.679999	1.71E-05	26.9283	<0.00001
Skin	74232	8.713065	2.42E-05	119604	8.377987	2.20E-05	49.2975	<0.00001
Nerve tissue	25579	8.553971	7.60E-05	701	6.274775	0.001336	60.6583	<0.00001
Muscle	75841	8.449439	4.65E-05	23898	7.677075	0.00011	61.6460	<0.00001
Adipose tissue	12417	8.06399	0.000141	720	6.033909	0.002033	43.5439	<0.00001
Pituitary gland	13430	7.510406	0.000351	1392	6.217005	0.002162	25.8007	<0.00001

e) An analysis of the unigenes for putative tumor biomarkers

The tissue-specific unigenes exclusively represented by cancer or normal ESTs (Table 12) represent a unique mining resource that may be used to produce biomarker candidates for a particular cancer type or for a set of related cancers. Given an enormous share occupied by these exclusive gene clusters in a wide range of human tissues, extracting the gene clusters with differences between the normal and cancer tissues, poses interesting statistical problem.

A total of 27 human tumor tissues having a corresponding normal tissue were used to identify the biomarkers. Pair-wise Mann-Whitney test was performed on the normalized EST abundance data of each Unigene gene cluster. The gene clusters which show significant variation in this pair-wise non-parametric test across the 27 tissues were filtered for further analyses. The tissue groups are marked by only their pathological characterstic (normal or tumor) leaving their tissue identities in the filtered gene clusters. This data is scaled to illustrate EST abundance as frequency. Metastats is used on this grouped normal and tumor frequency data to extract differentially abundant features (gene clusters). The gene clusters with a significant degree of differential abundance were only considered as biomarkers.

Of the 123,687 gene clusters in the current Unigene database system, only 2863 were observed to have significant variation (P-values < 0.05) by pairwise non-parametric testing. These significant unigenes were ran through Metastats software to filter these unigenes differentiating normal and tumor samples in order to pinpoint ones associated with low false discovery rate (<0.05). This filter resulted in a final list of candidate unigenes (1751 gene clusters) comprised of 668 tumor-specific and 1083 normal-specific expressing gene clusters.

Heatmap of the EST abundance of these significant unigenes showed a clear difference between the normal and tumor groups (Figure 6). Selected unigenes were mostly represented by the protein-coding gene clusters and, to some extent, by non-coding unigenes, with 139 non-coding unigenes in tumor biomarker list and 157 unigenes in the normal (anti-cancer) list. The Genbank accession numbers for all listed biomarker candidate transcripts are presented in the Appendix.

There were 526 cancer specific and 826 normal specific protein-coding gene clusters whose expression patterns were statistically different between normal and tumor tissues.

These unigenes are potential candidate biomarkers that may help to identify various subtypes of human cancer. An entire list of these biomarkers is also given in an Appendix.

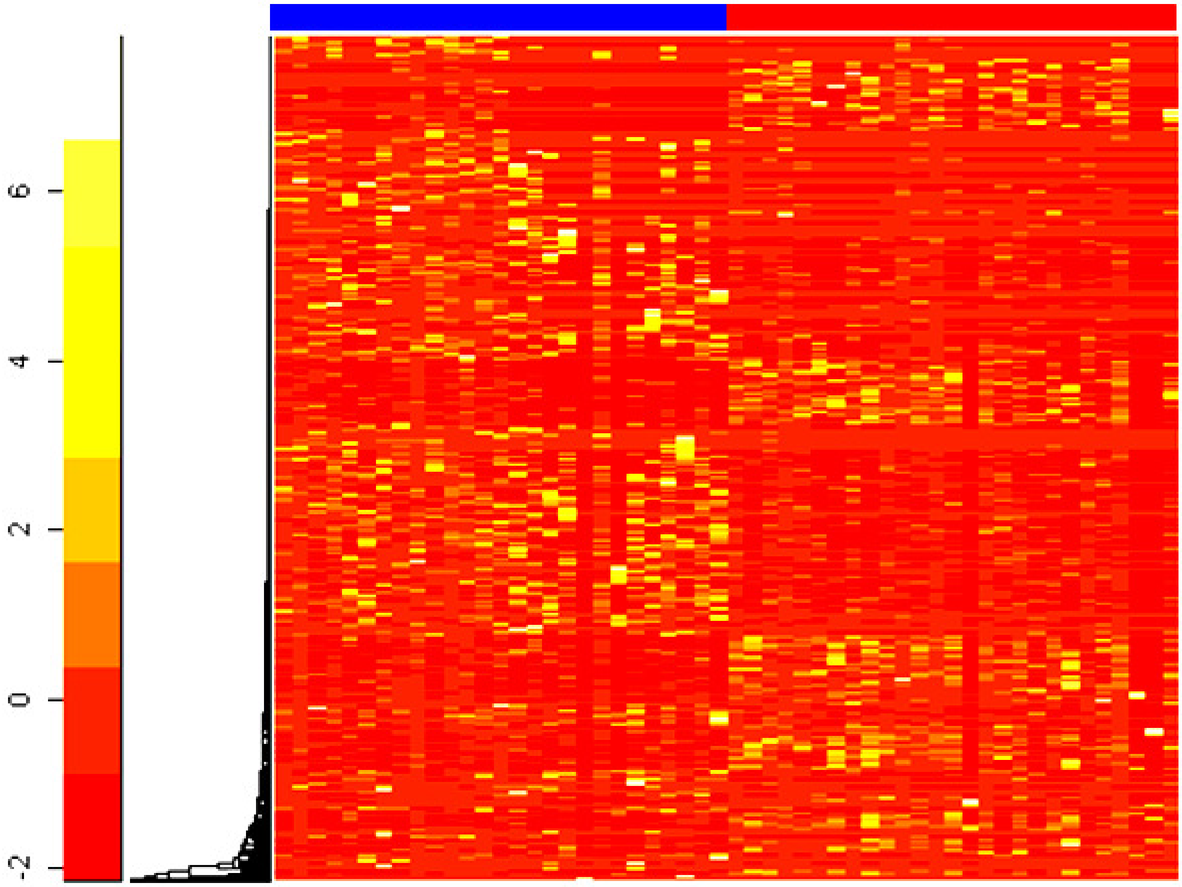


Figure 6: Heatmap of the EST abundance data from 27 different human tissues from the normal and corresponding cancer cDNA libraries of the unigene system. Blue bar at the top of heatmap identifies the normal tissues, while red bar marks their malignant counterparts.

f) Functional analysis of protein-coding unigenes identified as tumor biomarker candidates

Functional analysis of protein-coding unigenes identified as tumor biomarker candidates was performed using the molecular pathway information provided in the KEGG knowledge base (Kanehisa et al. 2008). A novel tool, KEGG Pathway painter (KPP) was built in-house with the purpose to automatically identify the function of the candidate biomarker unigenes through high-throughput analysis of the KEGG pathway maps. The methodology underlying KPP processing and the detailed description of KPP are available in the Appendix. Briefly, the complete sets of human molecular pathways were extracted by KPP separately for the cancer-specific and normal-specific significant unigenes. The enrichment of the pathways with the genes that belong to particular functional categories was assessed using the DAVID functional analysis framework (Dennis et al. 2003; Huang et al. 2009).

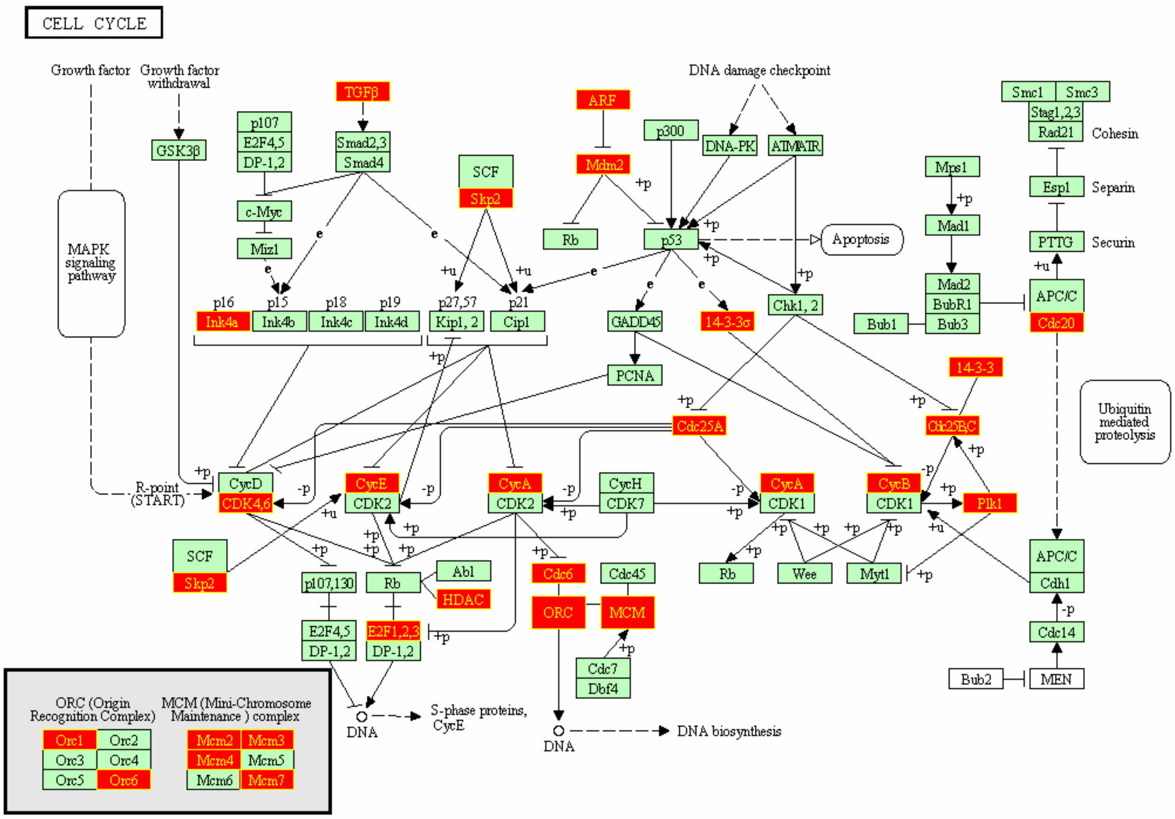


Figure 7: KPP representation of the KEGG cell cycle pathway map painted using genes differentially expressed in the malignant tumors (red background) as compared to normal human tissues.

The enrichment analysis revealed that sets of the pathways highlighted by the cancer and normal biomarker candidates are substantially different. In this context, the normal specific genes could be called anti-cancer markers as they are exclusively present in the normal tissue, and are devoid in cancerous tissue. The fundamental biological processes enriched by cancer specific unigenes as represented by KPP analysis include cell cycle regulation, p53 signaling, ubiquitin mediated proteolysis and apoptosis. The enhanced proliferation activity influenced by the set of tumor biomarker candidate unigenes in the perspective of cell cycle is illustrated in Figure 7.

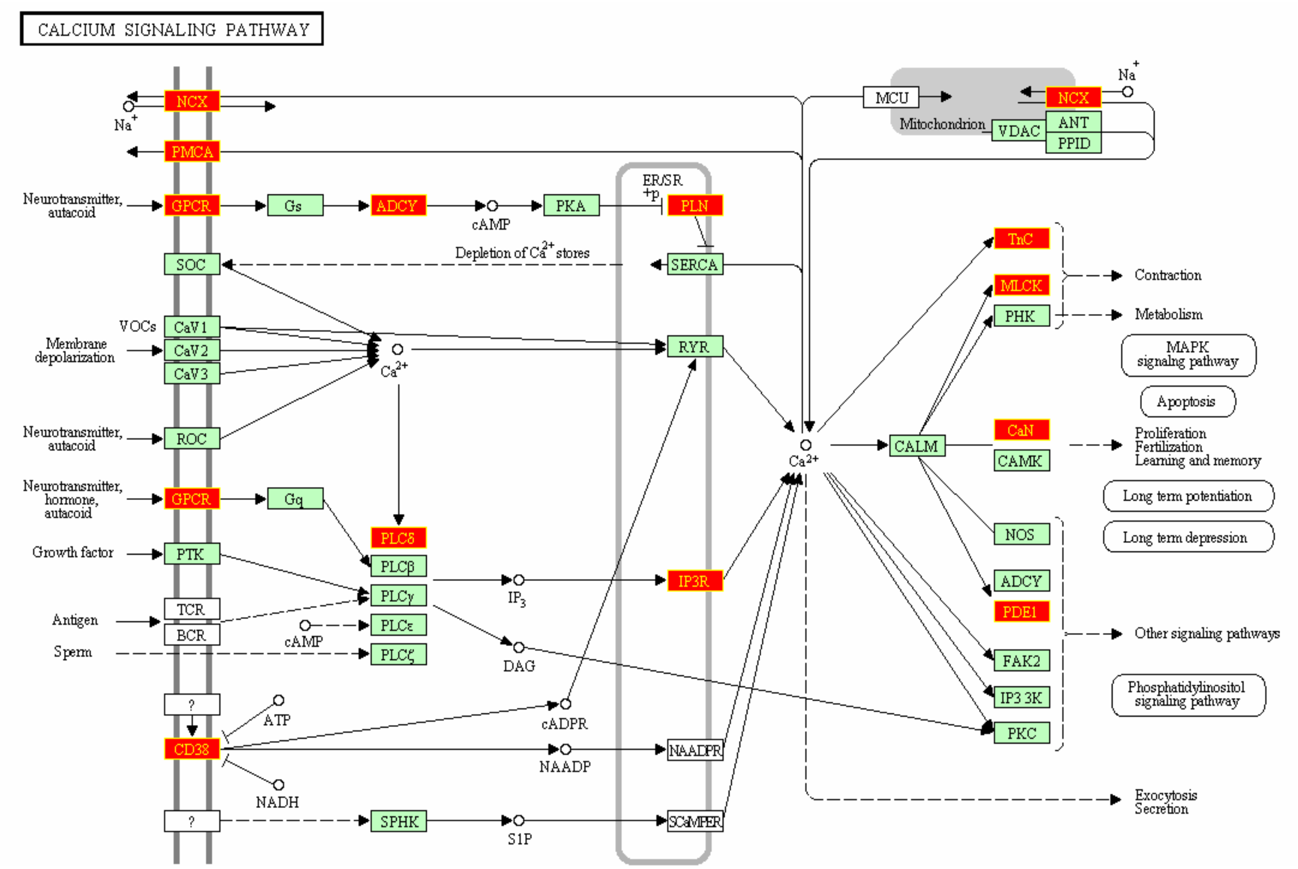


Figure 8: Illustration of the Genes expressing more abundantly in the normal tissue (red background) compared to the cancerous tissue on the KEGG Calcium signaling pathway.

Wide set of cancer specific pathways pertaining to chronic myeloid leukemia, melanoma, glioma, pancreatic cancer, prostate cancer, colorectal cancer, bladder cancer, small cell and non-small cell lung cancer was also shown as being enriched by these cancer-specific markers. The extensive range of cancer pathways enrichment supports the cancer-specificity of the unigenes associated in the context.

The anti-cancer markers represented by the unigenes abundant in normal tissues, on the other hand, demonstrated an enrichment of the broad variety of biological pathways and networks. The cascades highlighted by this analysis include calcium signaling, MAPK signaling, insulin signaling, gonadotropin-releasing hormone (GnRH) signaling and T-cell receptor signaling. The enrichment of calcium signaling pathway by the anti-cancer markers (Specific to normal tissue) is depicted visually in figure 8. The specific emphasis identified in the context of normal tissue in comparison to cancer tissue is signaling in a broad sense. Speaking generally, our analysis redefined cancer as the loss of tissue-specific signaling, while, at the same time, cancer demonstrated gains in the proliferating system. This definition is equivalent to the classic understanding of the tumor phenomenon. The full set of the pathways enriched by the cancer and anti-cancer biomarkers could be found in the appendix.

Conclusion

In this study, the diversity of the human tissue specific transcriptomes was studied. After normalization, relative abundancies of the ESTs derived from normal human tissues were compared to these collected using their malignant counterparts. An analysis of the abundance of EST in the UniGene database was proven useful as a tool for a search for candidate biomarkers for human tumors. Functional analysis of the protein-coding biomarker candidates described cancer as the loss of tissue-specific signaling with simultaneous gain in the proliferating system.

4. Effects of the tumor-specific telomere rearrangements on the adjacent gene expressions

Rationale

Telomere rearrangements may result in the disturbance of the expression levels of adjacent genes. Telomere position effect may provide a mechanism to incrementally alter phenotype of the cancer cells. Telomere-related cancer studies concentrated at the structural changes have to be complemented by studies of gene expression changes. Here we hypothesize that the telomere rearrangements commonly found in cancer cells perturb an expression of the genes adjacent to telomeres in human tumors.

Background

Telomeres are made of guanine (G) rich repetitive DNA, composed of TTAGGG motifs representing the ends of chromosomes. DNA loss incurred during the DNA replication leads to the formation of 3’ G-rich protrusions which transforms into G-quartet structure at both the chromosomal ends (Pommier, Lebeau et al. 1995). The G-rich single stranded overhangs, with the help of telomeric repeat binding factors (TRF1 and TRF2) bind to the double-stranded telomeric region, forming a lasso-like loop structure called the T(telomere)-loop (Shin, Hong et al. 2006; Verdun and Karlseder 2007). This T-loop structure is supposed to protect the chromosome ends from recognition by recombination and double strand DNA-repair machinery.

Telomeres also undergo length alterations during tumor development. These alterations have been shown to be independent markers of cancer prognosis (Bisoffi, Heaphy et al. 2006). Normal human somatic cells stop dividing after a finite number (~50) of replication cycles, a phenomenon known as the “Hayflick limit” (Hayflick 1965). This limitation of cell cycle is associated with the reduction of the number of telomere repeats at the chromosomal ends after each cycle. The reduction of the telomeres activates p53 which prevents the damage of cells by inducing DNA damage repair and inhibiting further cell division, causing cellular senescence (Artandi and Attardi 2005). However, some cells are exempt from the replicative senescence, particularly, the immortal germ cells and some tissue stem and progenitor cells expressing special telomere restoring enzyme called telomerase (Sherr and DePinho 2000).

Telomerase is a complex ribonucleoprotein that consists of the a DNA polymerase and a RNA template components and possesses reverse transcriptase activity at the telomeres (Cech 2004). In human cells, this enzyme stabilizes telomere length by adding TTAGGG repeats onto the telomeric ends during the S phase of cell cycle (Hug and Lingner 2006). The reactivation of telomerase in the somatic cells leads to the uncontrollable cell division, which might further pave the way to cancer. In the cells with p53 mutations, an expression of telomerase lifts the “Hayflick limit” and helps to maintain genomic instability (Finkel, Serrano et al. 2007).

Telomere maintenance is essential to the proliferation of tumor cells as expression of the mutant catalytic subunit of human telomerase results in complete inhibition of telomerase activity, reducing the telomere length, which ultimately lead to the death of tumor cells (Hahn, Stewart et al. 1999). Moreover, the minimal set of molecular events required for direct tumorogenic conversion of normal human epithelial and fibroblast cells requires an abnormal expression of the telomerase catalytic subunit (hTERT) in combination with overexpression of two oncogenes, the simian virus 40 large-T oncoprotein and an oncogenic allele of H-ras (Hahn, Counter et al. 1999). Indeed, a majority of human tumor cells acquire immortality through expression of the catalytic subunit of telomerase (hTERT) (Hiyama and Hiyama 2002). Approximately 10% of human cancers do not show evidence of telomerase activity, and a subset of these maintain telomere lengths by a recombination-based mechanism termed alternative lengthening of telomeres (ALT). The ALT phenotype, relatively common in certain sarcomas and germ cell tumors, is very rare in carcinomas. This alternative mechanism of telomere maintenance does not depend on the actions of telomerase (Stewart 2005).

In addition to the telomerase that is directly responsible for addition of telomeric sequences to the ends of the chromosomes, a plethora of other proteins also play important roles in the regulation of telomere length maintenance (de Lange 2005). Simultaneous and balanced upregulation of genes encoding telomere-associated proteins in cancer cell is an unlikely event. As active telomerase alone is not sufficient for preserving normal functionality of the telomere-associated protein complex in cancer cells, many telomerase-expressing tumors exhibit chromosomal instability triggered by dysfunctional telomeres (Gisselsson and Hoglund 2005; Calcagnile and Gisselsson 2007). Invasive tumors frequently demonstrate intra-tumoral heterogeneity of telomere lengths that include both an increase in telomere length over the normal range and telomere shortening (Meeker, Hicks et al. 2004; Hansel, Meeker et al. 2006; Maida, Kyo et al. 2006).

In eukaryotic yeast cells, particularly in *Saccharomyces cerevisiae,* genes located near telomeres undergo reversible silencing (Tham and Zakian 2002). This effect, termed as the telomere position effect (TPE), involve changes in the chromatin conformation and is dependent on both the distance from the telomere and the telomere length (Tham and Zakian 2002). The silencing effect of the genes near telomeres, and the spontaneous reactivation of these genes have been described in HeLa cells (Baur, Zou et al. 2001; Baur, Shay et al. 2004). However, the literature on possibility of TPE in mammalian cells is scarce and conclusions remain controversial. Some researchers failed to find evidence for TPE and concluded that in higher eukaryotes the gene expression is independent of telomere length (Bayne, Broccoli et al. 1994; Sprung, Sabatier et al. 1996). Others demonstrated an influence of telomeres on the expression of the adjacent transgenes within the human and mouse cell lines (Baur, Zou et al. 2001; Koering, Pollice et al. 2002; Pedram, Sprung et al. 2006). A study of the expression of endogenous genes located near telomeres in human fibroblasts revealed a discontinuous pattern of altered gene expression during senescence-associated telomere shortening (Ning, Xu et al. 2003).

Subtelomeric regions may buffer or facilitate the spreading of silencing that emanates from the telomere (Ottaviani, Gilson et al. 2008). As these regions are particularly prone to recombination in tumor cells, they may add an extra level of the complexity to the problem reviewed.

Telomere rearrangements observed in cancer cells may result in the disturbance of the expression levels of the adjacent genes, and this in genomic instability. So far, the studies of the telomeres in cancer cells were concentrated at the structural changes in the telomeres themselves and the quantification of the telomere-associated molecules, including telomerase and telomere-binding proteins TRF1, TRF2, TIN1, POT1, TPP1, Cdc13p (Hug and Lingner 2006). In our opinion, the characterization of expression levels of the genes located near telomeres in tumors and normal tissues may shed light on the effects of TPE in the physiology of cancer. Therefore, we assessed effects of the telomere position and its rearrangements in the cancer cell studying the perturbations of expression of the genes adjacent to telomeres in human tumors.

Hypothesis

Telomere rearrangements commonly found in cancer cells perturb an expression of the genes adjacent to telomeres in human tumors.

Materials and Methods

To investigate the possibility that rearrangement of telomeres in cancer cells may lead to the disturbance in the expression levels of adjacent genes, we retrieved the publicly available data describing changes in gene expression patterns in the prostatic carcinoma, a tumor with common findings of the telomerase reactivation and telomere erosion (Vukovic, Park et al. 2003; Meeker, Hicks et al. 2004; Meeker 2006). Prostatic carcinoma dataset GDS1439 was obtained from Gene Expression Omnibus (GEO) repository (Barrett and Edgar 2006). This dataset describes raw data collected in course of the microarray profiling of primary (N=7) and metastatic (N=6) prostate carcinoma samples along with the samples of normal prostate tissue (N = 6). The microarray platform employed for this study was the Affymetrix GeneChip U133 Plus 2.0 (GPL570), an oligonucleotide array covering an entire human genome with over 47,000 transcript-specific probes.

The GenBank annotated Human Genome (Build 36, Version 2) sequence was downloaded from the NCBI Genome repository (Benson, Karsch-Mizrachi et al. 2007). Contig information for each of the human chromosomes was obtained from the NCBI MapViewer (Wheeler, Barrett et al. 2007). R data analysis package with Bioconductor (Reimers and Carey 2006; Gregory Alvord, Roayaei et al. 2007) and Perl scripting language were used to perform the computational analysis.

Based on the GenBank annotation and the contig information, locations for each human gene on their respective chromosomes were calculated as absolute numeric coordinates defined as nucleotide position from the end of the available chromosome sequence to the gene. All genes were associated with their expression data based on their HUGO gene symbols. A total of 17522 out of 28479 genes annotated in the human genome at the time of the project were in this manner assigned to the human genome map and associated with their respective expression levels. These mapped transcripts were then categorized into telomere associated and non-telomere (body of the chromosome) associated fractions based on their position. Two computational experiments were performed (see Fig 9). In experiment I, 10% of genes on either extreme of each chromosome (except acrocentric chromosomes with only one telomere), accounting to a total of 20% per chromosome (10% for acrocentric chromosome) were defined as telomere associated genes, and remaining 80% were defined as centromere associated or non-telomeric genes. In experiment II the telomere associated gene definition was narrowed down, to include only 5% on each extreme of chromosome, accounting to a total of 10% telomere associated genes per chromosome.

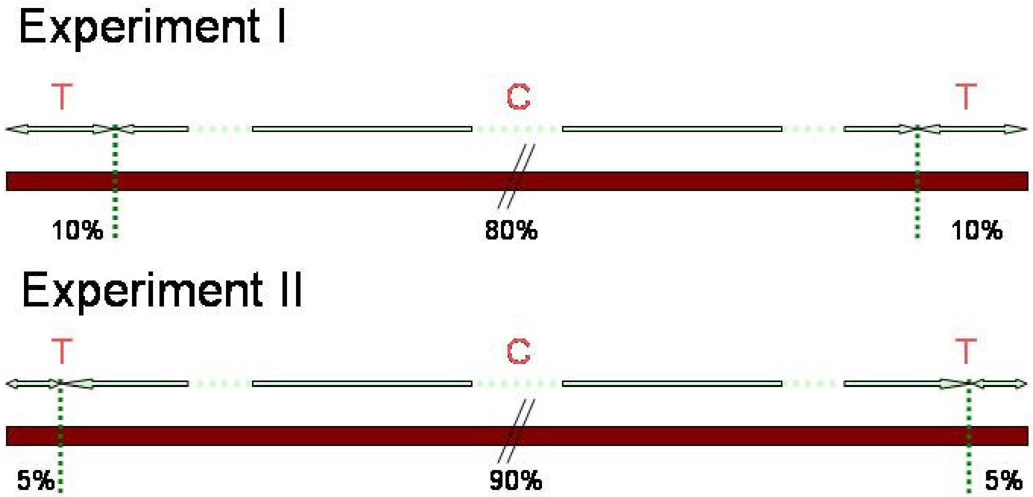


Figure 9: All the genes in each of the human chromosomes were subdivided into sum of their telomere associated (T) genes and rest as the centromere associated (C) genes based on the number of the genes as shown above in two experiments.

The gene expression data were background corrected, normalized and the individual gene expression values were calculated from the CEL files using Robust Multichip Average (RMA) method of background adjusted, normalized, and log transformed perfect match (PM) values (Irizarry, Bolstad et al. 2003). RMA software is a part of the R package affy that is a module of Bioconductor. It provides an interactive environment for data analysis and exploration of Affymetrix oligonucleotide array's probe level data (Gautier, Cope et al. 2004). The mean expression value of all the samples in each sample type (B-normal, P-primary and M-metastatic) was taken as the representative expression for that gene in the particular stage of the tumor development or in normal prostatic tissue.

Results and Discussion

a) The definition of the over/underexpressed genes

The dataset GDS1439 contains gene expression profiles for three types of prostate samples: primary (N=7) and metastatic (N=6) prostatic carcinoma samples along with the samples of normal prostate tissue (N = 6). Therefore, in this study over- and underexpressed genes could be derived in three different comparisons: metastatic (M) vs. primary (P) samples, metastatic (M) vs. normal (B) samples and primary (P) vs. normal (B) samples. For the purpose of this study we termed the genes that significantly change their expression in M/B comparison. If M/B was greater than or equal to 1.4, then the gene was considered to be overexpressed; else if M/B less than or equal to 0.71 (i.e. 1/1.4), then the gene was considered to be underexpressed. A total of 528 over/under expressed genes pertinent to the GEO dataset GDS1439 correspond to the 3% of the genes matched with GenBank annotation of the whole Human genome (Benson, Karsch-Mizrachi et al. 2007). In Experiment I, the number of over/under expressed genes present in telomere-associated regions covering 20% of coding human chromosome DNA was 15% of the total number of over/under expressed genes.

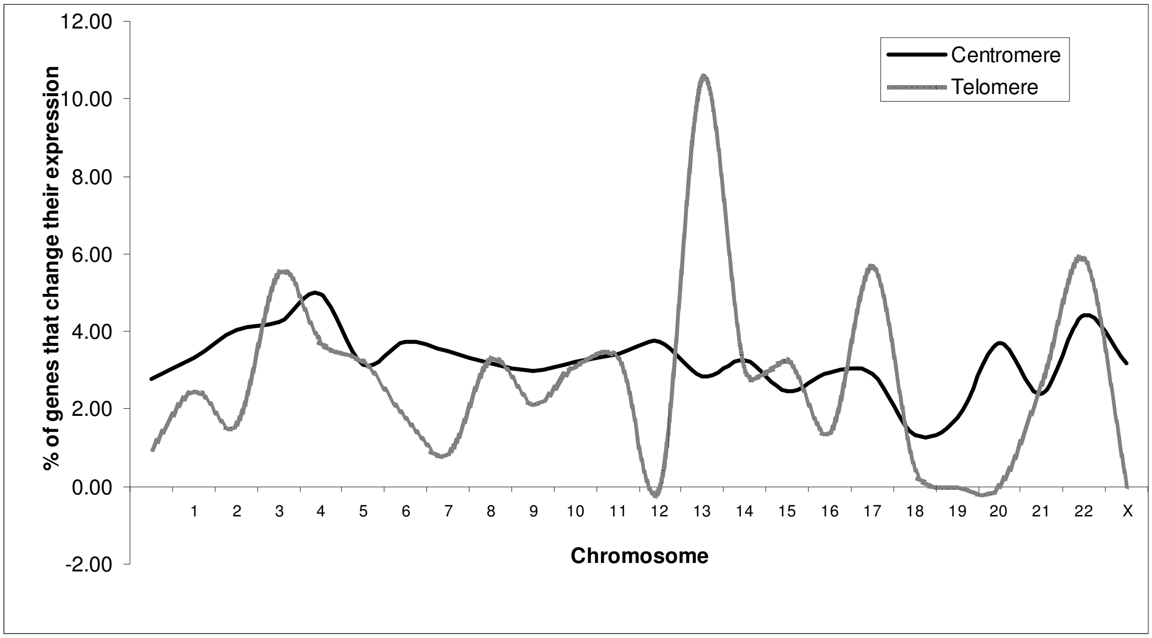


Figure 10: The graph depicts the percentage of genes that change their expression on each chromosome in the telomere and centromere associated regions. On average, the percentage of the over/under expressed genes varies by a greater degree in the telomere associated region compared to the body of chromosome, when individual chromosomes are taken into account.

The percentage of centromere associated genes that change their expression was approximately the same for every chromosome, while the percentage of telomere associated genes that change their expression was chromosome specific, as demonstrated by high variation in a number of over/under expressed genes on individual chromosomes (see Fig 10). Chromosome Y was devoid of overexpressed genes in the non-telomere associated regions. Telomeric regions of many other chromosomes were shown to lack either under or over expressed genes or both.

b) Correlation with tumor stage

The patterns of the changes of gene expression at different stages of tumor development might be different for the telomeric genes and genes in the body of chromosome. The metastatic to normal (M/B) gene expression ratio was initially introduced in order to estimate the variation of expression in telomeric as compared to centromere associated genes. We interpreted this measure as representation of expression change imbalance between telomeric and centromeric regions. Mann-Whitney test (Mann and Whitney 1947), a non-parametric testing procedure was employed for assessing the statistical significance of the difference in the gene expressions in the telomere and non-telomere associated regions in each of the chromosome.

The gene expression ratios reflecting the shift of the tissue between three studied states (M/B, M/P and P/B) were calculated as independent vectors for all genes associated with the telomeres and the body of chromosomes. The differences between the groups of corresponding ratios for telomere-associated and body of chromosome-associated genes were tested using the Mann-Whitney test under the null hypothesis that they are similar. Thus, the correlation between the location of the gene and the change in its expression was tested. The tests were performed for the whole human genome (i.e. all telomeres together vs. all bodies of chromosomes together) at once, and for chromosome separately.

The chromosomes that showed statistically significant variation in Experiment I, were again tested for variation in Experiment II. The chromosomes that revealed significant differences in both experiments were again tested for variation in the telomere and non-telomere associated regions using metastatic to primary (M/P) and primary to normal (P/B) gene expression ratios. These analyses allowed to correlate the changes in the expression for genes located in telomere and centromere associated regions with the stage of the tumor development.

In Experiment I, the variation of the M/B ratio across the telomere and centromere associated regions was significant by Mann-Whitney test (p-values < 0.005) in chromosomes 4, 5, 8, 9, 16 and X. There was a marginal significance (p-value < 0.05) in the variation for the chromosomes 2, 10 and Y as assessed by the same test. Mann-Whitney test results also showed the difference in the M/B ratio (p-values < 0.005) for chromosomes 1, 4, 5, 7, 8, 16 and X, and a marginal variation (p-value < 0.05) for chromosome 13 in Experiment II. The chromosomes that show significant variation in M/B ratio in both Experiments I and II (4, 5, 8, X, 16) were further tested for variation in the M/P ratio and P/B ratio in the telomere and non-telomere associated regions.

The expression change imbalance of the M/P ratios for the telomeric and body of chromosome associated was significant when all of these chromosomes were taken into account both in Experiment I ($P < 8.53e-14$) and Experiment II ($P < 1.05e-11$). In case of P/B ratios, expression change imbalance was not significant in almost all the cases, except for chromosomes 12, 15 and 16 using telomere definition of Experiment II (see

Table 16). Differences in P/B ratios were not significant even when all the chromosomes are compared together representing the genome.

Table 16: Telomere associated genes are more likely to change their level of expression as compared to non-telomere genes. When metastatic tumors were compared to normal prostates using (M/P) ratio, these changes were significant for 11 out of 24 chromosomes (Exp I) or 8 out of 24 chromosomes (Exp II), while in the primary tumor to normal tissue (P/B) comparisons only 3 out of 24 chromosomes demonstrated expression change imbalance. HS-Highly significant (p-value < 0.0005), S-Significant (p-value < 0.005), MS-Marginally significant (p-value < 0.05), NS-Not significant (p-value ≥ 0.05).

	Metastatic / Normal		Metastatic / Primary		Primary / Normal	
Chr	Experiment I	Experiment II	Experiment I	Experiment II	Experiment I	Experiment II
1	NS	S	NS	S	MS	NS
2	MS	NS	NS	NS	MS	NS
3	NS	NS	NS	NS	NS	NS
4	HS	HS	S	S	NS	NS
5	HS	S	HS	HS	NS	NS
6	NS	NS	NS	NS	NS	NS
7	NS	S	NS	S	NS	NS
8	HS	S	MS	MS	NS	NS
9	S	NS	MS	NS	NS	NS
10	MS	NS	NS	NS	NS	NS
11	NS	NS	NS	NS	NS	NS
12	NS	NS	NS	NS	NS	MS
13	NS	MS	NS	NS	NS	NS
14	NS	NS	NS	NS	NS	NS
15	MS	NS	NS	MS	NS	MS
16	S	HS	MS	S	NS	S
17	NS	NS	NS	NS	NS	NS
18	NS	NS	NS	NS	NS	NS
19	NS	NS	NS	NS	NS	NS
20	NS	NS	NS	NS	NS	NS
21	MS	NS	MS	NS	MS	NS
22	NS	NS	NS	NS	NS	NS
X	HS	S	HS	MS	NS	NS
Y	MS	NS	NS	NS	NS	NS

As could be seen from the Table 1, the variation in the ratios of the gene expression between the telomere and non-telomere (centromere) associated regions increases with the progression of the prostatic tumor and is the most pronounced for the metastatic tumors.

c) Variation in distribution of expression changes

To assess possible variation in distribution of individual gene expression values in each of the metastatic (M) and normal (B) samples, these values were subjected to pairwise t-test. The p-values obtained in the t-test were adjusted to control the false discovery rate (FDR) using Multtest module of R package (Pollard, Dudoit et al. Dec 2004) that utilizes Benjamini and Hochberg (BH) method (Benjamini and Hochberg 2000). The distributions of these adjusted p-values from telomere and non-telomere associated genes were also constructed using R. The overall variation in the distribution of p-values in the two regions (telomere and centromere associated) was tested using Kolmogorov-Smirnov test (Conover 1971). This variation in the p-values was also verified by t-test and rank test (Mann and Whitney 1947).

The paired t-test between the metastatic (M) and normal (B) expression data for every gene in the genome resulted in a total of 1521 significantly varying genes (p-value < 0.05) after controlling the false discovery rate. Interestingly, the distributions of these p-values in the telomere and centromere associated genes as defined in Experiment I were found to be more dense at lower cutoffs (p-value < 0.01), showing a significant variation between metastatic and normal samples. The variation between the telomere and centromere associated genes in the context of metastatic and normal samples was assessed by testing the variation between these p-values by a number of statistical tests (Table 17). A significant variation was seen in these p-values corresponding to telomere and centromere associated genes, returning the p-values in the order of 0.001. The rank test between the first 1000 p-values of telomere and centromere associated genes show even more substantial variation described by a p-value less than 3.18796e-05.

Table 17: The p-values obtained from the t-test between the metastatic and normal expression data were subjected to various statistical tests to find the variation between the telomere and centromere associated genes. The significant p-values (< 0.05) corresponding to genes in the telomere and centromere associated regions indeed show variation in all three tests.

Test	**p-values for all the genes, including not significantly changed**	**p-values < 0.05**
Kolmogorov-Smirnov test	0.06352	0.003997
t-test	NS	0.001201
Rank test (Mann-Whitney test)	NS	0.000886426

d) Gene Ontology analysis

For each of the the telomere and centromere associated genes with p-values remaining significant after FDR adjustement, the gene ontology (GO) term descriptions defining its relationship to certain cellular compartment, molecular function and/or biological process were retrieved. A comparative analysis of the GO terms distributions between the telomere and non-telomere associated genes was done using the goTools package of R (Paquet and Yang 2007) and presented on Fig. 11.

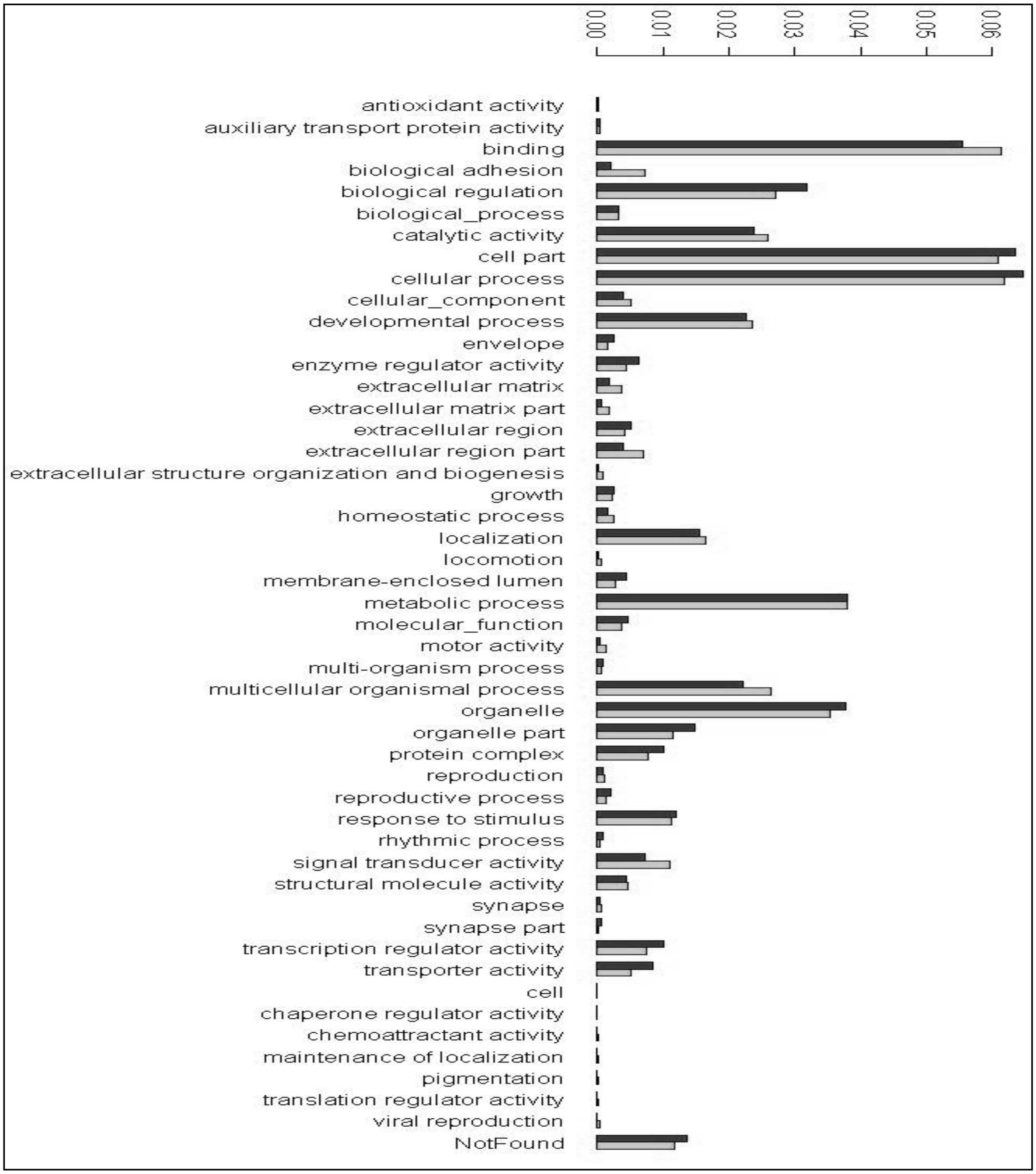


Figure 11: The bar graph in each GO category represents the ratio of the number of the direct GO identifiers associated with the telomere associated (shown in black) and centromere associated (shown in grey) genes to the number of GO identifiers associated with all the genes in the Affymetrix GeneChipU133 Plus 2.0 oligonucleotide array. As could be seen from the Figure, relative abundances of the GO terms in each GO category for the telomere and centromere associated genes were similar.

A comparative analysis of the GO terms in the telomere and centromere associated genes was performed with reference to the number of probe identifiers and GO identifiers associated with all the genes in the array. GO term analysis revealed that differentially expressed telomere and non-telomere associated genes are involved in a wide variety of biological processes, carrying various molecular functions in different cellular locations. The relative abundances of the GO terms and probe identifiers in each GO category for the telomere and centromere associated genes were almost equal (see Fig 11), revealing the functional similarity of the telomere and non-telomere associated genes based on GO terms. This functional similarity points out that the variation between the telomere and non-telomere associated genes was due to the effect of their adjacent to telomeres rather than any bias in their functional importance.

e) Defining the maximal length of the subtelomeric fragment that might be influenced by telomere rearrangements in cancer cells

In the Experiments I and II genes were called “telomeric” or “centromeric” (non-telomere associated) based on the overall number of genes on each chromosome (Fig 9). However, in the perspective of TPE, a physical distance from the gene to the telomere is important. We hypothesized that the plotting the gene expression ratios obtained in M/B (metastatic to normal), M/P (metastatic to primary) and P/B (primary to normal) comparisons across the chromosome length may help to define the distances with measurable TPE.

The gene locus table of the GenBank annotated Human genome (Build 36, Version 2) sequence was downloaded from the NCBI. Exact physical coordinates for each gene have been utilized to map M/B ratios for the genes present on Affymetrix chip across chromosomal lengths. These plots have not revealed any discernible pattern that might help to define the length of the telomeric fragment susceptible to TPE. However, the plots provided for the clear illustration of the concept that the scale of the changes in the gene expression dramatically increases with the progression of the prostatic tumors (see Fig 12 for an example of the plot).

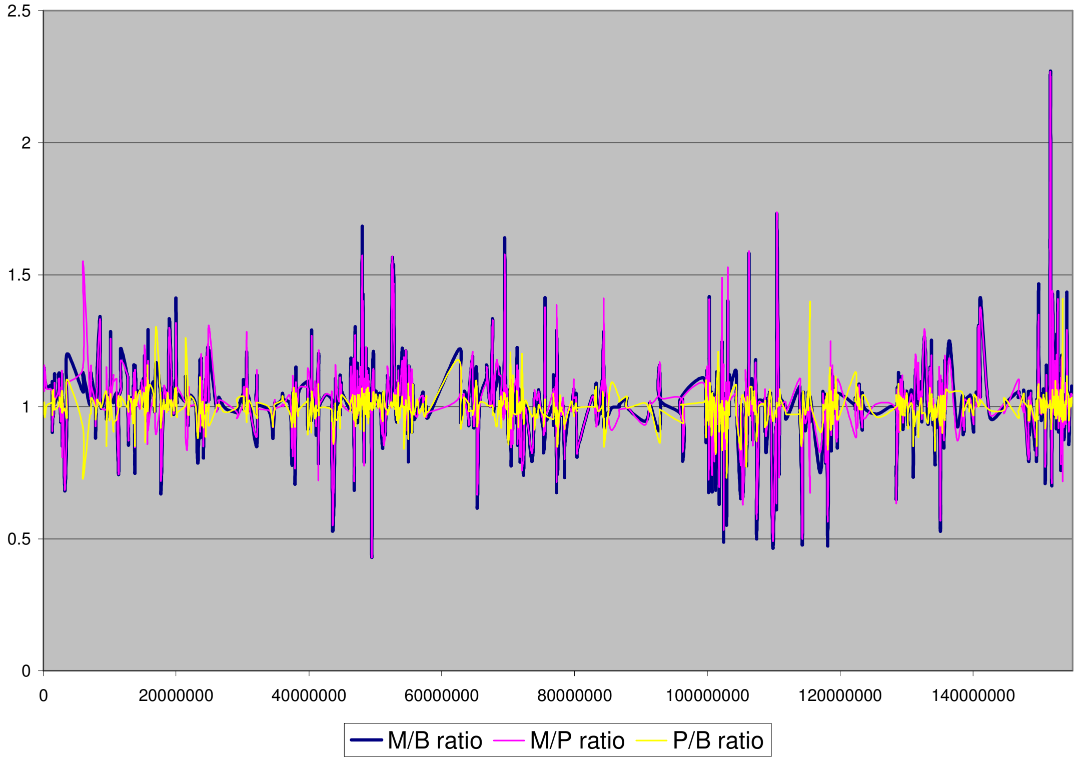


Figure 12: This graph plots M/B, M/P and P/B gene expression ratios (y-axis) to the physical distance from the telomere (x-axis) on human chromosome X. The variation in the M/B ratio (blue) is larger than the variation in M/P ratio (purple), while the P/B ratio (yellow), is characterized by smallest variation along the length of chromosome X.

An alternative approach to find the characteristic length of the subtelomeric region susceptible to TPE is to calculate gene expression changes in a variety of different sub-telomere lengths windows that were arbitrary chosen from the range of 0.2MB to 50MB. The tested arbitrary telomeric lengths were stepwise increased up to as the maximum physical length that can be assigned to telomere from the end of the chromosome to centromere of this chromosome in each of its arms. This telomere length window was defined globally for the whole genome, while chromosomes with shorter arms ended up with a smaller subtelomeric region. The genomic regions accounting for 5MB of physical length on either side of the centromere were truncated as the heterochromatic parts of the chromosomes surrounding its centromeres might also effect the gene expression levels via CPE (centromere positioning effect).

The metastatic to normal (M/B) expression ratios were used to find the variation between the telomere genes and genes in the body of chromosome for 20 gradually increasing subtelomeric windows. The analysis was performed by separately for a) all genes changing their expression, b) upregulated genes only and c) downregulated genes only. Cumulative changes in the expression levels of the genes located within given subtelomeric window and the rest of the chromosome were compared.

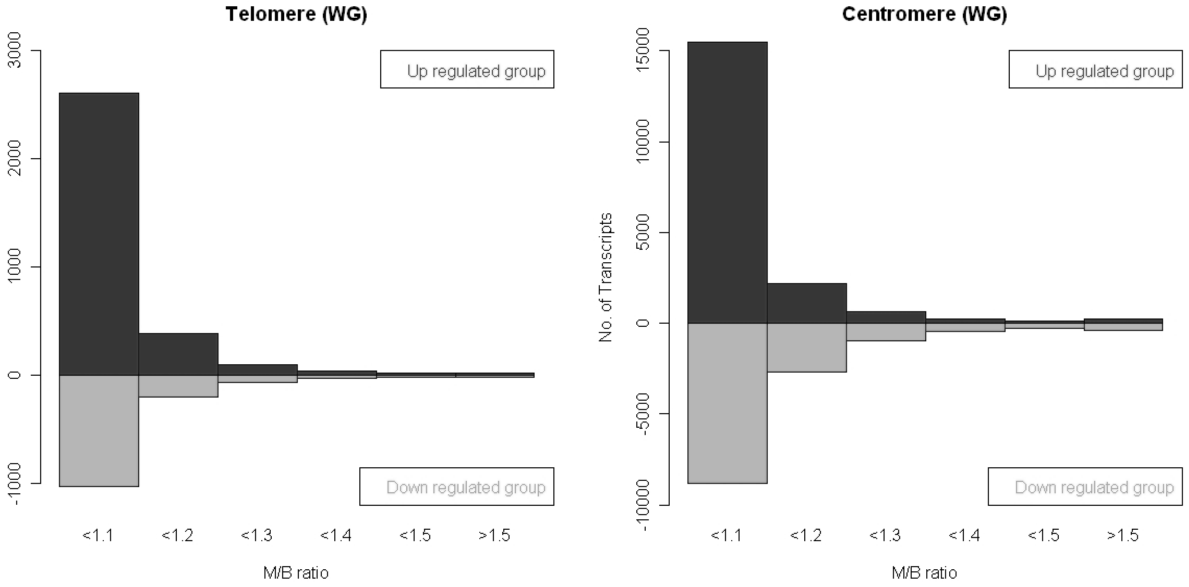


Figure 13: This bargraph plots the number of upregulated (black) and downregulated (grey) transcripts in each of the M/B ratio bins for the whole human genome (WG), considering a telomere length of 5 MB on either side of chromosome. Similar bargraphs were built for each subtelomeric window separately for a) all genes changing their expression, b) upregulated genes only and c) downregulated genes only. Genes located within given subtelomeric window and the rest of the chromosome were compared.

The distributions of the transcripts in the upregulated groups in the telomere and centromere associated regions were calculated by sorting them into bins of the range 1 - 1.1, 1.1 - 1.2, 1.2 - 1.3, 1.4 - 1.5, and > 1.5. Similar methodology was used for downregulated groups where the bins representing the ranges of 1 - 1/1.1, 1/1.1 – 1/1.2, 1/1.2 – 1/1.3, 1/1.3 – 1/1.4, 1/1.4 – 1/1.5 and < 1/1.5. The distribution plots were built for each of the subtelomere windows; Fig 13 represents the distribution of upregulated and downregulated genes for subtelomeric window of 5 Mb. Although these distributions were difficult to assess visually, statistical analysis performed on the frequency distribution help to dissect expression change imbalance between the telomere and centromere associated regions in specific subtelomeric windows. Chi-square test of homogeneity was used to find the extent of variation between the distributions in the telomere and centromere associated genes.

The homogeneity test was performed separately for a) all genes changing their expression, b) upregulated genes only and c) downregulated genes only. Genes located within given subtelomeric window and the rest of the chromosomes were compared. Differentially expressing groups DEG1, DEG2 and DEG3 comprising of transcripts with varying levels of changes in their expression were formed by excluding transcripts with less pronounced expression changes. M/B ratios in the range 1/1.1 to 1.1, 1/1.2 to 1.2 and 1/1.3 to 1.3 were excluded in DEG1, DEG2 and DEG3, respectively (Table 18). Almost all the tests seem to be significant when gene expression changes were considered wit no regards to their directionality, except for subtelomere windows of 2, 4, 5 and 6 Mb. The position effects for downregulated group of genes were significant when all expression ratios were considered, but lost significance when tests were performed only for genes in DEG1 and DEG2 groups. The upregulated group of genes demonstrated substantial position effect in telomeres with lengths ranging from 0.5 Mb to 5Mb. The smaller DEG2 and DEG1 groups with larger differences in expression levels seem pinpoint subtelomeric windows with length range of 2 to 4Mb, with largest effect at the subtelomeric window of 3Mb.

To study the correlation between the significance of gene expression changes and gene location in the subtelomere region, chi-square test using a 2x2 contingency table was performed. For each arbitary subtelomeric window, a 2x2 contingency table consisting of

Table 18: The table represents the p-values obtained from the chi-square test of homogeneity between the distribution of M/B ratios in the telomere and centromere associated regions for varying subtelomeric windows. P-values shown in grey are not significant (>=0.05). In the differentially expressing groups (DEG), the transcripts showing minimal changes of expression were excluded. In DEG1, DEG2 and DEG3 M/B ratios in the range 1/1.1 to 1.1, 1/1.2 to 1.2 and 1/1.3 to 1.3 were excluded respectively.

Telomere .Length...	All (Up + Down regulated)				Downregulated			Upregulated		
	All	**DEG 1**	**DEG 2**	**DEG 3**	**All**	**DEG 1**	**DEG 2**	**All**	**DEG 1**	**DEG 2**
50 MB	8.47E-21	0	7.37E-118	2.53E-18	0.00047	0.361	0.284	0.549	0.675	0.747
40 MB	4.33E-20	0	9.12E-114	1.42E-11	3.59E-07	0.017	0.05	0.22	0.319	0.245
30 MB	3.37E-30	0	3.53E-95	2.24E-10	9.93E-10	0.502	0.999	0.017	0.008	0.006
25 MB	8.01E-33	0	7.66E-88	1.47E-07	2.00E-08	0.099	0.84	0.095	0.055	0.045
20 MB	1.35E-51	0	1.47E-77	3.29E-08	2.07E-10	0.42	0.957	0.016	0.014	0.144
15 MB	1.60E-53	0	6.36E-67	0.002048	5.85E-14	0.121	0.414	0.148	0.15	0.105
10 MB	2.01E-47	0	4.33E-44	0.02297	1.80E-13	0.047	0.158	0.203	0.211	0.123
9 MB	4.81E-51	0	2.43E-42	0.025999	1.73E-10	0.049	0.156	0.444	0.346	0.215
8 MB	2.46E-45	0	2.62E-41	0.006398	7.64E-10	0.037	0.072	0.748	0.613	0.489
7 MB	8.34E-44	0	1.17E-33	0.017777	9.65E-11	0.064	0.109	0.616	0.472	0.466
6 MB	1.81E-48	4.36E-320	5.99E-32	0.133988	1.57E-11	0.12	0.283	0.534	0.439	0.406
5 MB	2.27E-54	5.38E-291	2.27E-28	0.211632	1.24E-10	0.289	0.291	0.184	0.11	0.18
4 MB	2.03E-57	1.17E-256	5.85E-28	0.161683	1.51E-08	0.223	0.44	0.021	0.011	0.013
3 MB	1.02E-45	4.21E-213	1.80E-28	0.043766	5.62E-06	0.241	0.387	0.002	8E-04	4E-04
2 MB	9.12E-31	7.90E-161	3.71E-17	0.234099	0.00053	0.451	0.505	0.05	0.025	0.029
1 MB	8.93E-16	7.60E-87	1.85E-11	0.000474	0.00027	0.051	0.431	0.045	0.053	0.023
0.75 MB	2.05E-11	1.14E-73	1.77E-08	5.48E-05	0.00264	0.015	0.088	0.023	0.039	0.02
0.5 MB	3.39E-06	4.22E-39	8.48E-05	2.78E-05	0.03309	0.15	0.3	0.021	0.027	0.008
0.25 MB	0.069175	3.71E-11	0.004005	0.00125	0.19165	0.14	0.503	0.125	0.104	0.142
0.2 MB	0.012106	2.12E-07	0.004893	0.000139	0.57665	0.424	0.477	0.012	0.007	0.008

the number of significantly changing genes within the subtelomere window, significantly changing genes outside of the subtelomere window, not significantly changing genes within the subtelomere window and not significantly changing genes outside of the subtelomere window was computed for two different significance levels, DEG2 and DEG3. The chi-square tests were performed on each of these tables using a) all genes changing their expression, b) upregulated genes only and c) downregulated genes only.

The combined (upregulated + downregulated) and downregulated groups seem to show good p-values (< 0.05) in both the significance levels, except for small telomere lengths 0.2 and 0.25 MB (Table 19).

Table 19: The table represents the p-values obtained by two significance levels (DEG2 and DEG3) by the chi-square test using 2x2 contingency table. The test is performed between the telomere and centromere associated genes with all the genes combined, as well as by dividing the genes into upregulated and downregulated groups. P-values shown in grey are not significant (>=0.05).

Telomere Length	**DEG2 = ratios < 1/1.2 and > 1.2**			**DEG3 = ratios < 1/1.3 and > 1.3**		
	All	Downregulated	Upregulated	All	Downregulated	Upregulated
50 MB	1.91E-06	0.002436953	0.119896351	0.000252	0.034614613	0.196168367
40 MB	1.24E-08	3.75E-06	0.152223813	2.05E-07	6.14E-05	0.092946198
30 MB	1.96E-09	5.87E-08	0.485674473	7.60E-07	8.66E-05	0.253873821
25 MB	6.26E-11	2.23E-08	0.264393533	7.33E-08	9.39E-06	0.306561668
20 MB	8.85E-16	2.52E-08	0.003617533	2.87E-10	2.84E-05	0.018070286
15 MB	7.24E-14	3.75E-10	0.178829797	1.89E-11	8.45E-08	0.09808447
10 MB	3.83E-12	9.36E-10	0.332284513	8.12E-09	5.09E-07	0.432338738
9 MB	3.22E-10	3.13E-08	0.670891063	1.57E-07	2.93E-05	0.42769367
8 MB	3.41E-09	1.97E-07	0.699547048	2.03E-07	3.40E-05	0.365660498
7 MB	5.83E-10	9.99E-08	0.42180974	4.35E-07	6.11E-05	0.35165262
6 MB	4.43E-10	1.88E-08	0.623543202	2.78E-07	5.08E-05	0.341613399
5 MB	5.26E-11	6.51E-07	0.127351618	4.71E-08	0.000369424	0.057675389
4 MB	1.28E-11	8.10E-07	0.063464319	4.00E-09	0.000354439	0.012018911
3 MB	5.75E-09	2.57E-05	0.125535116	8.55E-09	0.000464098	0.007000918
2 MB	3.86E-07	0.000728299	0.09450423	3.93E-06	0.009672502	0.024137083
1 MB	0.000132	2.91E-05	0.992908906	0.000681	0.000719066	0.841405461
0.75 MB	0.000219	0.000208505	0.723969426	0.001125	0.00138297	0.710549007
0.5 MB	0.006767	0.00354004	0.938509461	0.048815	0.010236132	0.759429064
0.25 MB	0.225291	0.01420092	0.367672817	0.29151	0.033364957	0.462961703
0.2 MB	0.380347	0.140937114	0.732884195	0.64757	0.115185791	0.331735767

The upregulated group opens up a window again in the higher differential expression level (DEG3). This window specifically corresponds to telomere lengths 2 to 4 MB, which seem to be significant in the chi-square test of homogeneity also. The ideal telomere length obtained (3MB) in the homogeneity test show a better p-value (0.007) than for telomere lengths 2MB (0.024) and 4MB (0.012), that are present in this window.

This test concludes that the most suitable telomere length to define subtelomeric genes for the current dataset as 3MB.

f) Functionally, the behavior of the genes located in the subtelomeric regions of human chromosomes is not different from that of the genes located in other parts of the human chromosomes

Using results of the *in silico* experiment described above the definition of subtelomeric region may be derived. According to our findings, TPE susceptible subtelomeric regions of human chromosomes are approximately 3Mb in length. To find out whether subtelomeric genes contribute to the cancer phenotype to larger extent than the genes located in the bodies of the chromosomes, we quantified the variation in both sets of genes by distance analysis. The distance (D) between the tissue states was calculated based on Pearson' correlation (R), as D = 1-R (Materials and Methods, Chapter 2). This distance might be computed over the entire genome (n = all the gene of chip) or over the particular set of the genes (i.e. subtelomeric genes). Fundamentally, two kinds of the global distances may be computed, one that measure the distance from the particular tissue sample to the metastatic tumor space and from the particular tissue sample to the normal tissue space, DMglobal and DBglobal, respectively.

- DMglobal: Genome-wide distance from Metastatic sample space
- DBglobal: Genome-wide distance from Normal (Benign) sample space

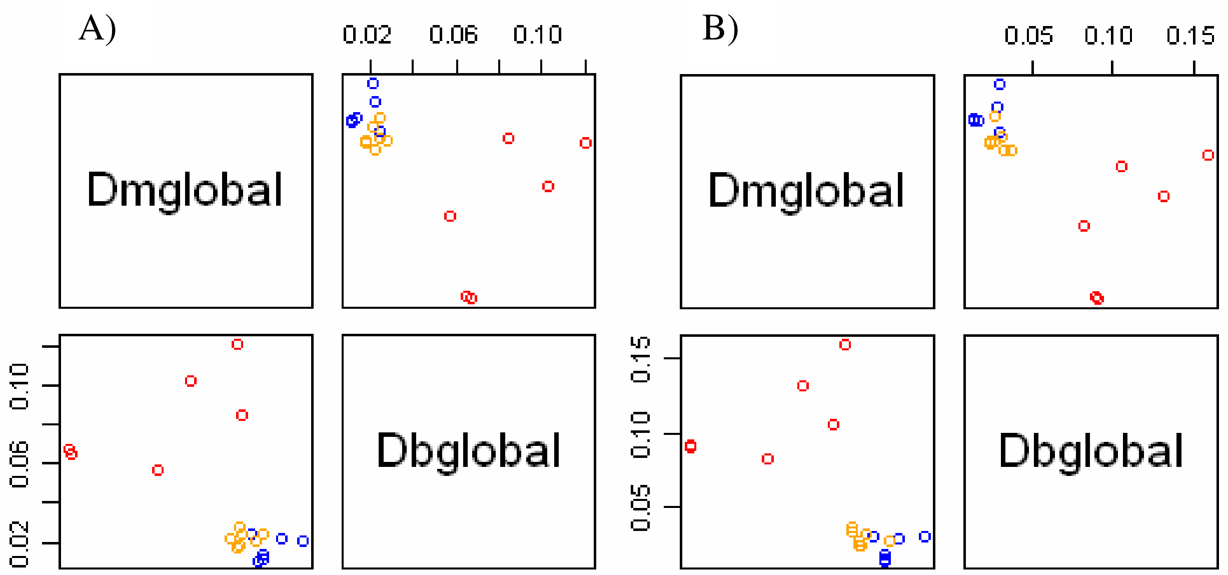


Figure 14: The Panel A represents the distance indices (DMglobal and DBglobal) for the telomeric genes prostate dataset. Panel B represents the distance indices for the non-telomeric or genes in the body of chromosome. On these panels metastatic tumors are represented by red circles, primary tumors by circles and normal prostate samples as blue circles.

The patterns of the distances between prostatic tumors and normal prostate samples look similar when DB and DM are calculated using subtelomeric and non-telomeric genes as illustrated in Figure 14, panels A) and B) respectively. According to our 3-Mb definition of human telomere, the telomeric genes account for 8.4% of all human genes. The Metastatic (red) are distant from the primary tumor (orange) and normal (blue) samples. The primary tumor and normal samples didn’t show much variation as compared to the metastatic tumors as the distribution of the distances between prostatic tumors and normal prostates are very similar when calculated using subtelomeric or non-telomeric gene sets, thus indicating that the changes of expression of subtelomeric genes are rather consequence than the cause of the tumor progression.

Conclusion

Here we demonstrate an *in silico* evidence proving telomere positioning effect (TPE) in human tumors. In studied dataset, TPE effect was found to spread over 3Mb of the subtelomeric distance. In microarray experiment profiling human prostatic tumors, a considerable increase in the number of up-regulated and down-regulated genes in telomeres compared to other regions of genome was observed. Importantly, an extent of this increase parallels progression of the prostatic tumors form normal tissue to primary tumor to metastatic carcinoma, with a steep increase at the late stage of cancer progression. The distances between prostatic tumors and normal prostates are very similar when calculated using subtelomeric or non-telomeric gene sets, thus indicating that the changes of expression of subtelomeric genes are rather consequence than the cause of the tumor progression.

Appendix A

KEGG Pathway Painter

High-throughput technologies became common tools to decipher genome-wide changes of gene expression (GE) patterns. Functional analysis of GE patterns is a daunting task as it requires often recourse to the public repositories of biological knowledge. On the other hand, in many cases researcher's inquiry can be served by a comprehensive glimpse. The KEGG PATHWAY database is a compilation of manually verified maps of biological interactions represented by the complete set of pathways related to signal transduction and other cellular processes. Rapid mapping of the differentially expressed genes to the KEGG pathways may provide an idea about the functional relevance of the gene lists provided by the high-throughput expression experiments Here web based graphic tool KEGG Pathway Painter (KPP) is described. KPP paints pathways from the KEGG database using large sets of the candidate genes accompanied by "overexpressed" or "underexpressed" marks, for example, those generated by microarrays or miRNA profilings. KPP will provide fast and comprehensive visualization of the global GE changes by consolidating a list of the color-coded candidate genes into the KEGG pathways. KPP is freely available and can be accessed at http://www.cos.gmu.edu/~gmanyam/kegg/

Introduction

High-throughput technologies became common tools to decipher genome-wide changes of gene expression (GE) patterns or relative protein abundance. Typical output of these large-scale studies is represented as a list comprised of hundreds of gene candidates with attached quantitative labels. Functional analysis of these gene lists is a daunting task as it requires often recourse to the public repositories of biological knowledge or use of expensive databases of manually curated biological annotation (Ganter and Giroux 2008). On the other hand, in many cases researcher's inquiry can be successfully served by a comprehensive glimpse.

Functional analysis of markers identified from large-scale datasets can be performed using a wide variety of bioinformatics tools. As microarrays became a common tool to decipher global gene expression, centralized systems like Gene Expression Omnibus (GEO), ArrayExpress were developed to congregate the valuable expression profile data (Barrett and Edgar 2006; Barrett, Troup et al. 2007; Parkinson, Kapushesky et al. 2007). The combined expression profile analysis fusing various microarray datasets (termed as the meta-analysis) is useful in the process of the development of the biomarker panels for various human diseases and specifically for various types of cancers (Rhodes and Chinnaiyan 2005; McShea, Marlatt et al. 2006). Meta-analysis lead to an increase of the complexity of microarray analysis pipeline and further sophistication of subsequent functional analysis is anticipated. Recently, Gene Ontology (GO) and Pathway-based analysis became the most important entry points to the functional analysis of expression data derived from high-throughput platforms (Werner 2008).

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a compendium of databases covering annotated genomes and protein interaction networks for all sequenced organisms. KEGG PATHWAY is a compilation of manually verified pathway maps displaying both the molecular interactions and the biochemical reactions (Kanehisa 2002). The recent version of this database includes a complete set of pathways related to signal transduction and other cellular processes (Kanehisa, Araki et al. 2008). The extensive collection of the pathways at KEGG can be utilized for the rapid graphical evaluation of the functional relevance of the observed changes in GE patterns. This will save the precious time of the expert biologists and bioinformatic specialists.

Pathways assembled into the KEGG database are displayed as semi-static objects that can be manipulated using tools like KGML and KEGG application programmable interface (API) (Kawashima, Katayama et al. 2003; Klukas and Schreiber 2007). KEGG API provides a routine that highlights specified genes within the particular metabolic pathway (http://www.genome.jp/kegg/tool/color_pathway.html).Gene set functional analysis tools like DAVID also returns the KEGG pathways by marking genes of interest (Dennis et al. 2003; Huang et al. 2009). The upregulated and downregulated marks can't be incorporated in such platforms. Similar task may be also executed using G-language Genome Analysis Environment (Arakawa, Kono et al. 2005). Both approaches work on the pathway by pathway basis. Another tool, Pathway express, calculates the pathway-wise impact of differentially expressed genes based on normalized fold change and depicts the pathways with differentially expressed genes (Khatri, Sellamuthu et al. 2005).

However, the fold-change approach and its associated standard t-test statistics usually produce severely over-fitted models. A number of recently developed approaches generate gene rankings dissociated from the fold change estimates (Hsiao, Worrall et al. 2004; Simon 2008). An analysis of these gene lists may benefit from the binary graphical mapping of upregulated and downregulated elements within the complete collection of pathway maps. Resulting graphical pictures may be helpful both as tool for a quick assessment of the functional relevance of a gene list and as a set of the snapshots easily convertible into the illustrative material for presentations or manuscript figures.

With this notion, here we present a web-based tool, KEGG Pathway Painter (KPP). KPP performs a batch painting of relevant pathways using the uploaded lists of up-regulated and down-regulated genes in KEGG. KPP returns a set of images that give a holistic perspective to the functional importance of the change in the GE patterns revealed by a given high-throughput experiment and facilitate the extraction of the biological insights.

Algorithm and Implementation

KPP accepts the up-regulated and down-regulated gene lists as two different text files containing the gene identifiers of any sequenced organism. Permitted identifiers include GenBank id, NCBI GENE id, NCBI GI accession, Unigene ID and Uniprot ID.

a) Algorithm

These gene identifiers are converted to KEGG identifiers and all the pathways associated with each of identified genes are extracted. Mapped genes are automatically consolidated within each pathway. The number of the KPP returned pathways could be filtered by either the total number of the painted genes in a given pathway or the ratio of painted genes to the total number of genes in a given pathway. The chosen pathways passing the criteria on filter are color coded according to users' preferences. Users can browse through these high-resolution pathway images along with gene information and an archive of the painted pathways can also be saved for future reference.

b) Implementation

KPP was implemented using PERL/CGI, and KEGG API was used to communicate with kegg/pathway database. The API allows access to the resources stored in KEGG system in a interactive and user-friendly way (http://www.genome.jp/kegg/soap/). Conversion of the gene identifiers, extraction of the corresponding pathway and their painting is performed by specific API routines. The KPP processes data through direct interface to the KEGG database, and therefore, the KPP painted pathways are always up-to date with reference to KEGG knowledgebase. Genes of interest can be also highlighted with user-specified forground and background colors for differentiating up and down regulated genes. The URL to the index of resulting output images is sent to the user by e-mail along with the job summary.

Discussion

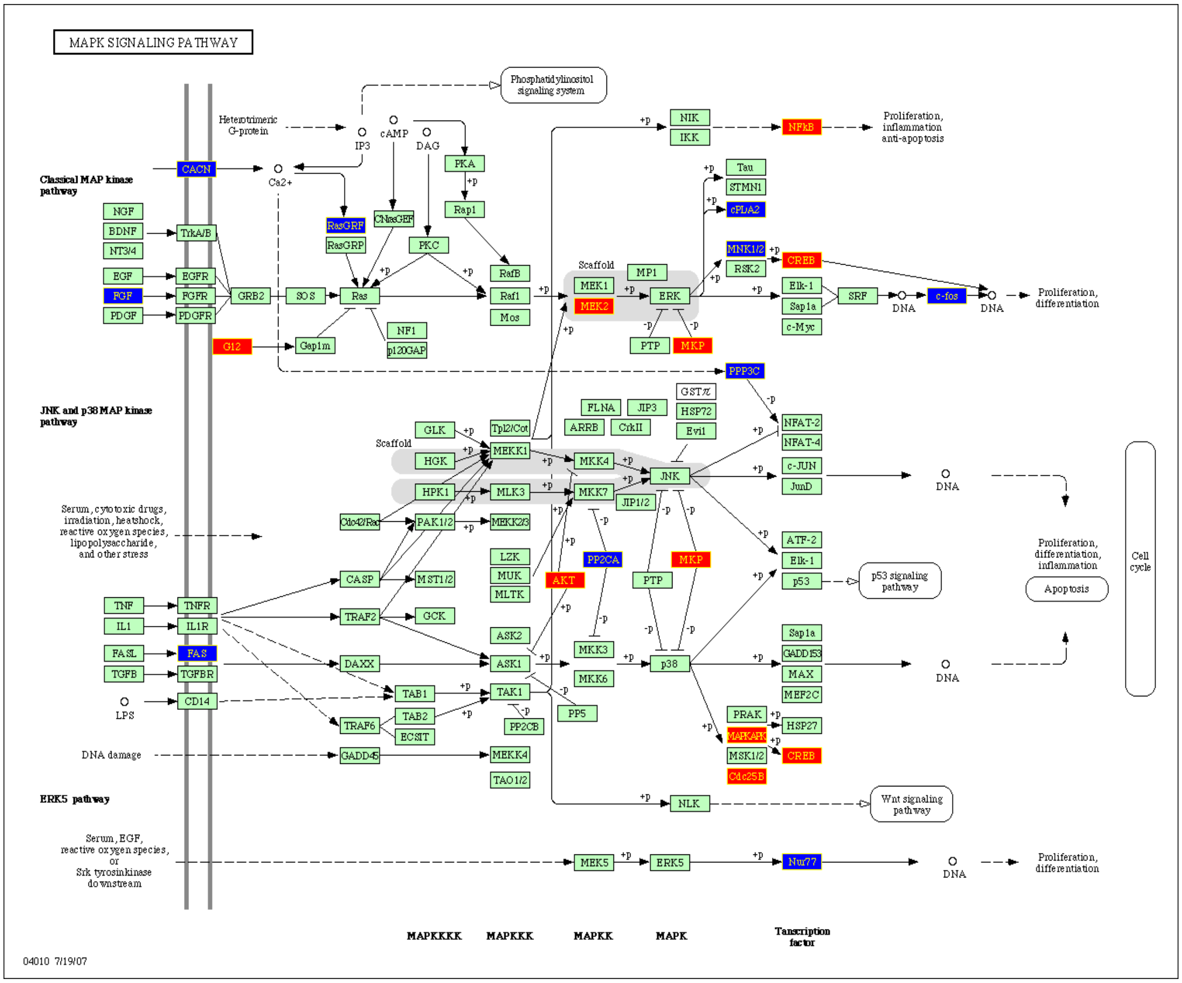


Figure 1: Image of the MAPK signaling pathway painted by KPP according to the imported list of genes differentially expressed in the prostatic carcinoma as compared to normal prostate. Red and blue boxes represent up- and down- regulated genes, respectively. The genes in green background represent the specifies specific genes (Homo sapiens, in this case) in the KEGG reference pathway

The motivation for the development of KPP came up from the idea to build a user-friendly, platform-independent and simple tool to visualize the genes in their associated pathways. The simplicity factor of KPP was due to the acceptance of gene identifiers instead of association with a microarray platform. This isolation would enhance its utility study the quantified transcript data from RT-PCR or even to validate varies hypothesis surrounding groups of genes that regulate patterns of gene expression in abnormal tissues. The utility is demonstrated using the publicly available prostate carcinoma dataset (GDS1439), from the NCBI GEO database (see Figure 1).

The major fetching point of the tool lies in its tight connection with the KEGG database, as this will allow for the pathway visualization of every sequenced organism. Although, this flexibility is at the cost of bottlenecks caused as a result of the delays during the data transfer. While the pathway painting can be performed for a given set of genes through the KEGG website (http://www.genome.jp/kegg/), the utility of KPP lies in generating the over-all glimpse up and down regulated genes of the dataset as a whole.

In summary, KPP provides fast and comprehensive visualization of the global gene expression changes by consolidating a list of the color-coded candidate genes into the KEGG pathways.

Appendix B

Enriched pathway plots generated by KPP for cancer-specific markers

The genes in light green background were human-specific.

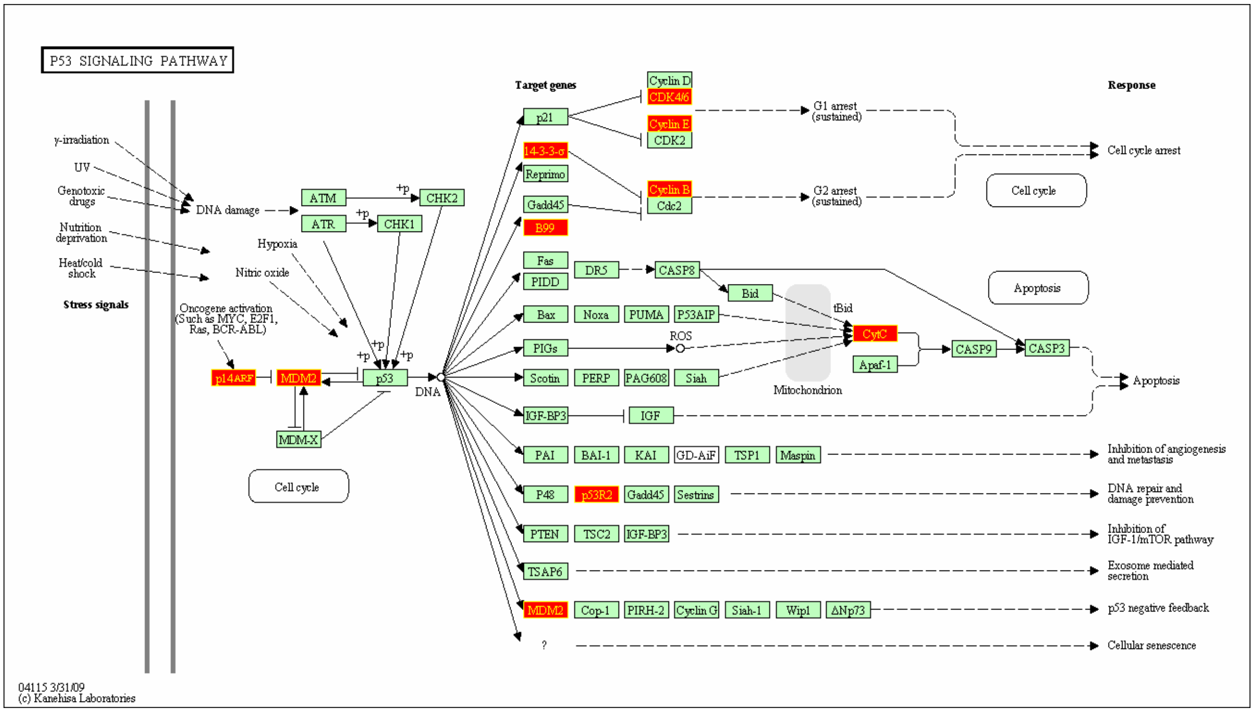


Figure 2: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG p53-signalling pathway.

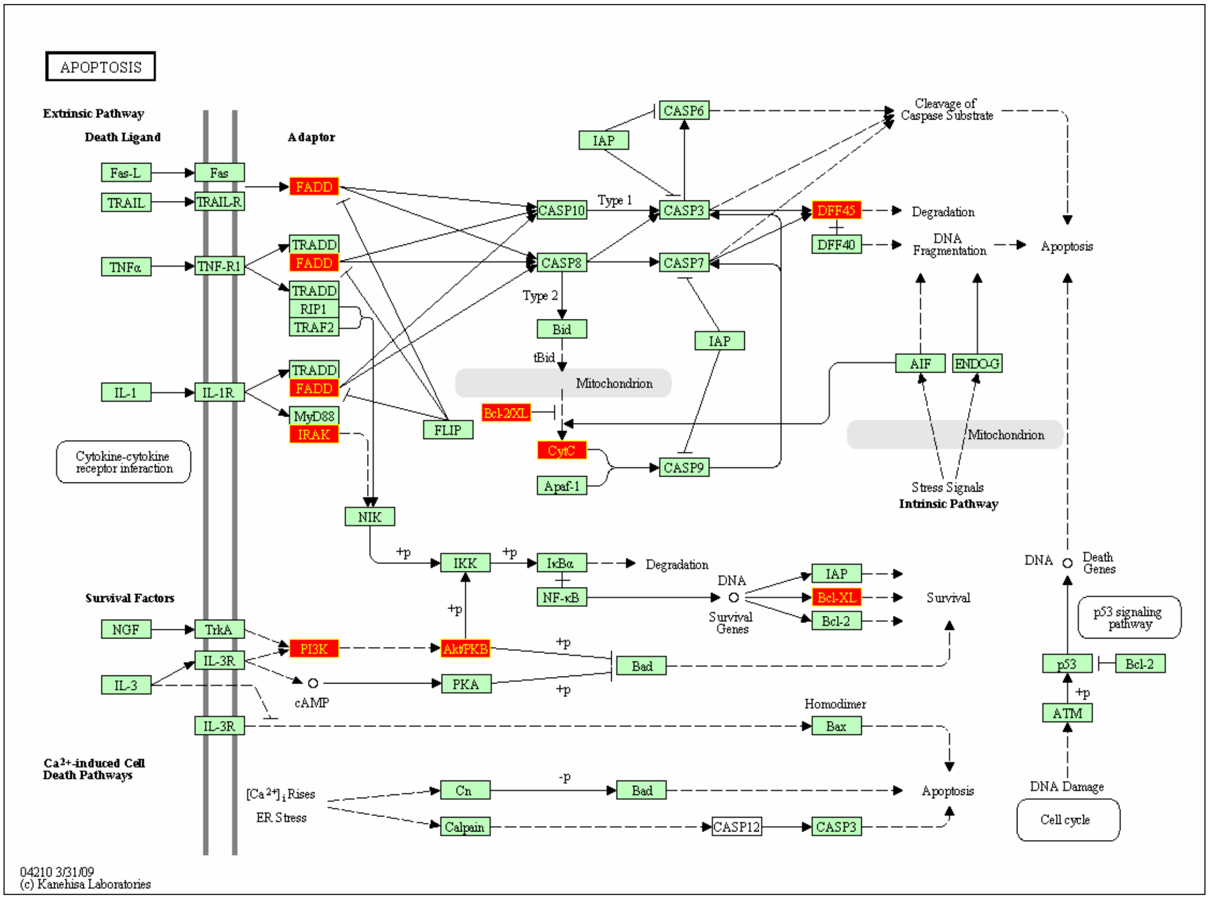


Figure 3: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG Apoptosis pathway.

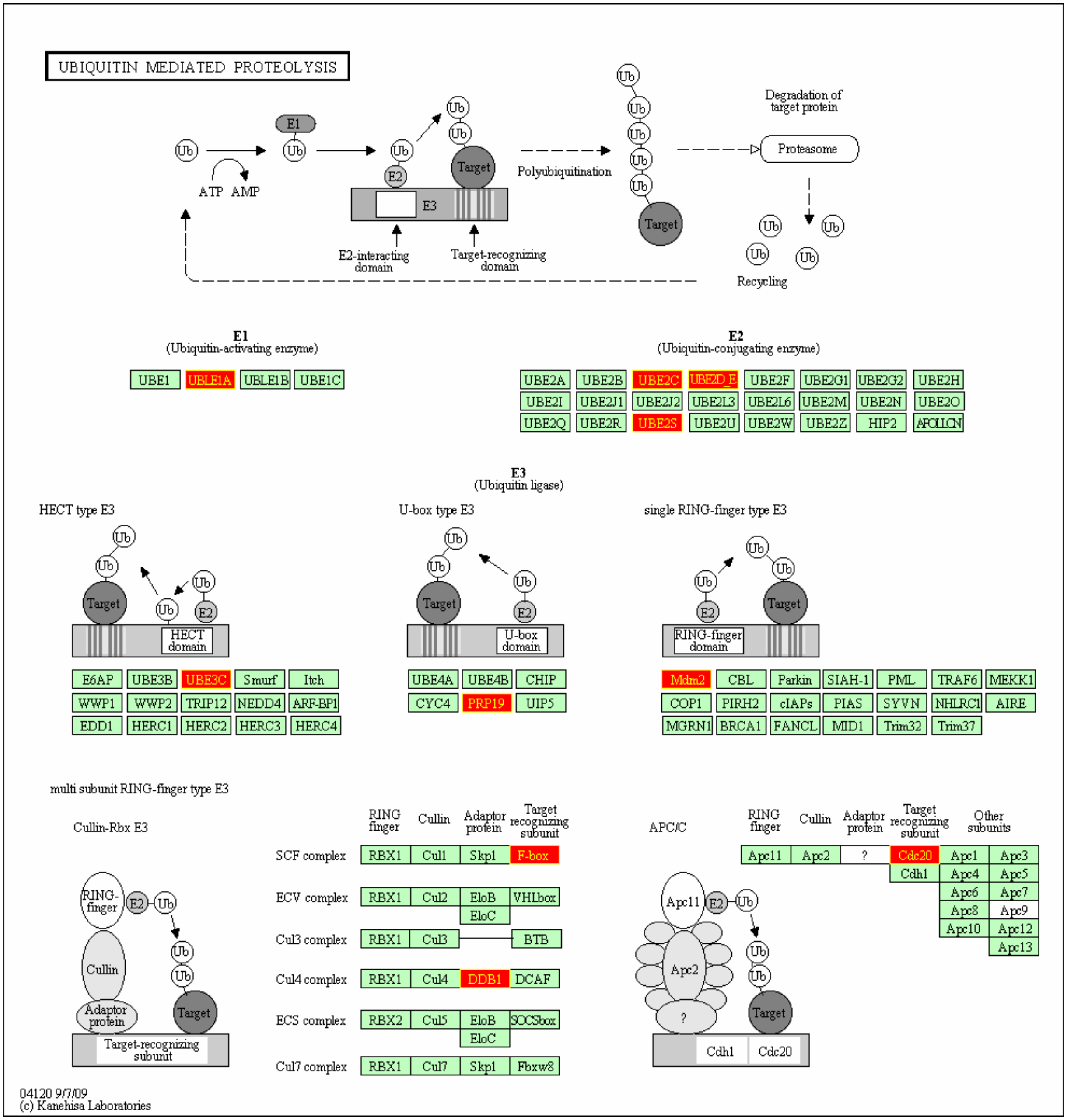


Figure 4: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG pathway representing the Ubiquitin mediated proteolysis.

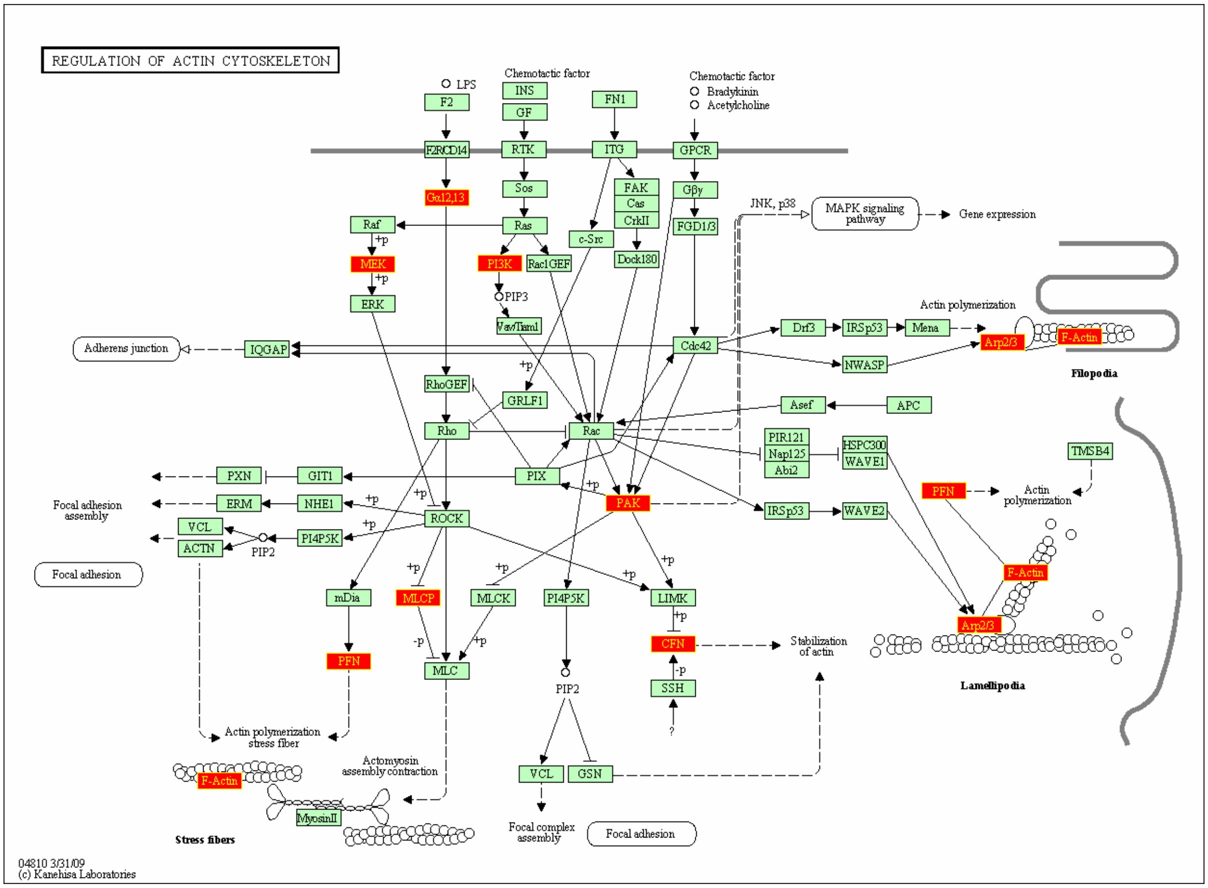


Figure 5: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG pathway representing the regulation of actin cytoskeleton.

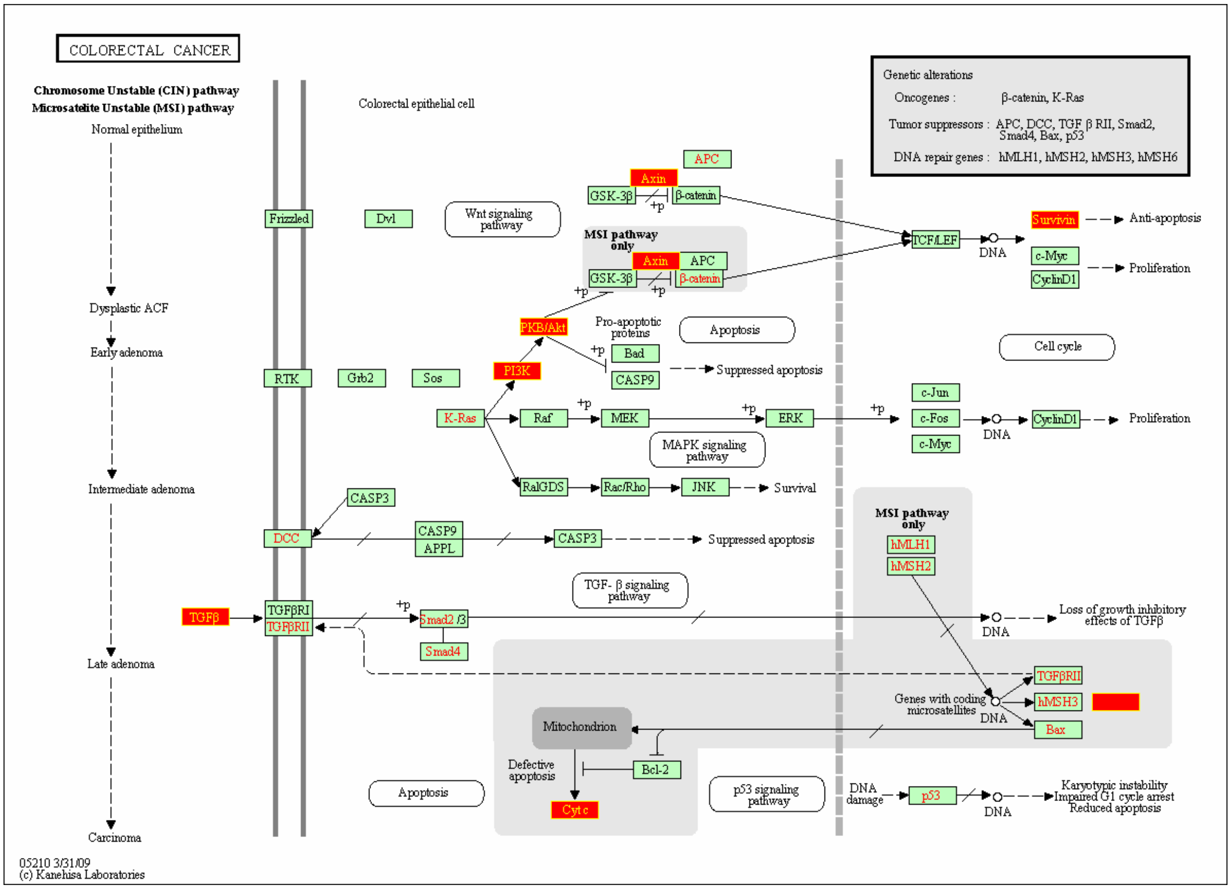


Figure 6: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG colorectal cancer pathway.

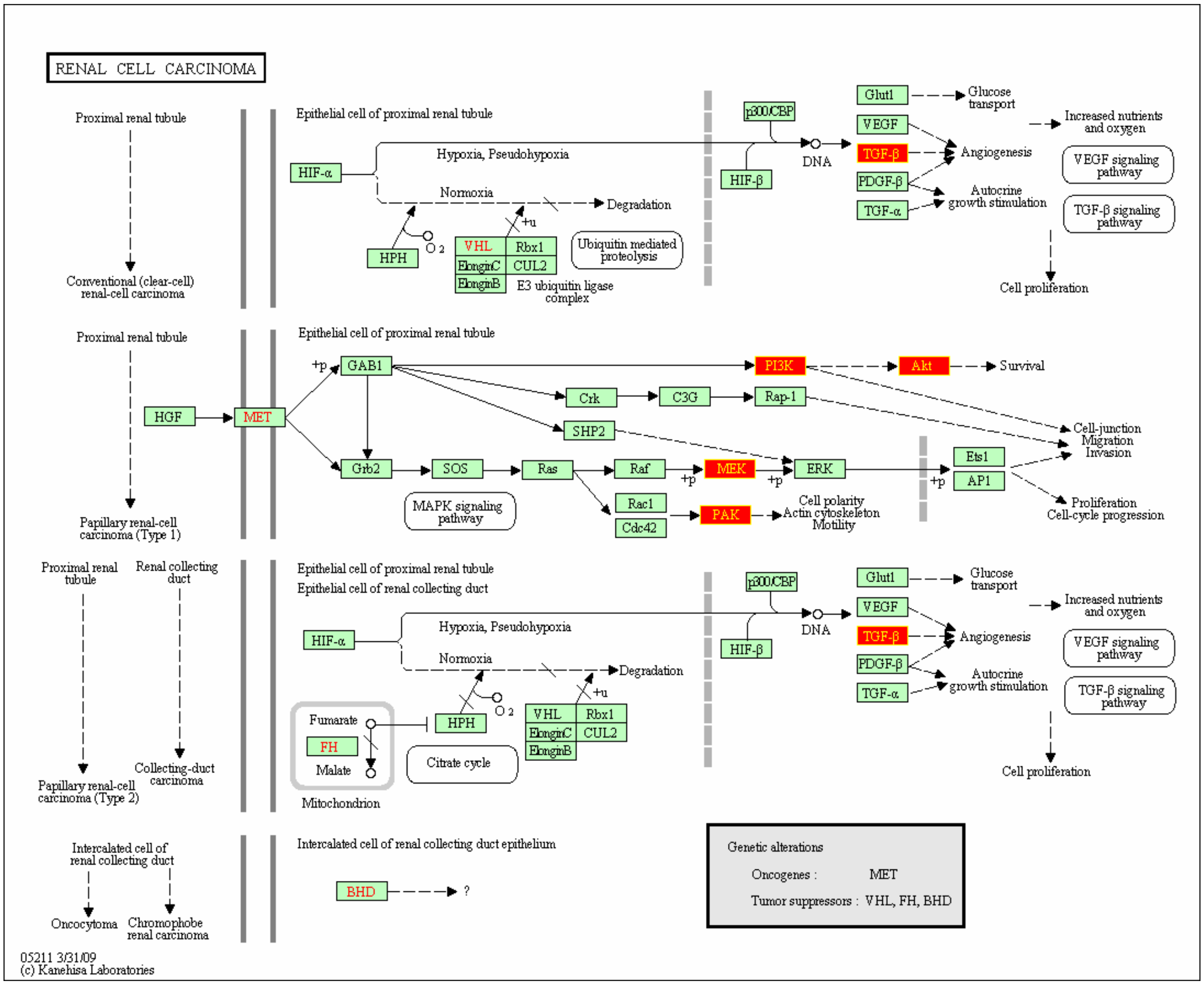


Figure 7: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG Renal cell carcinoma pathway.

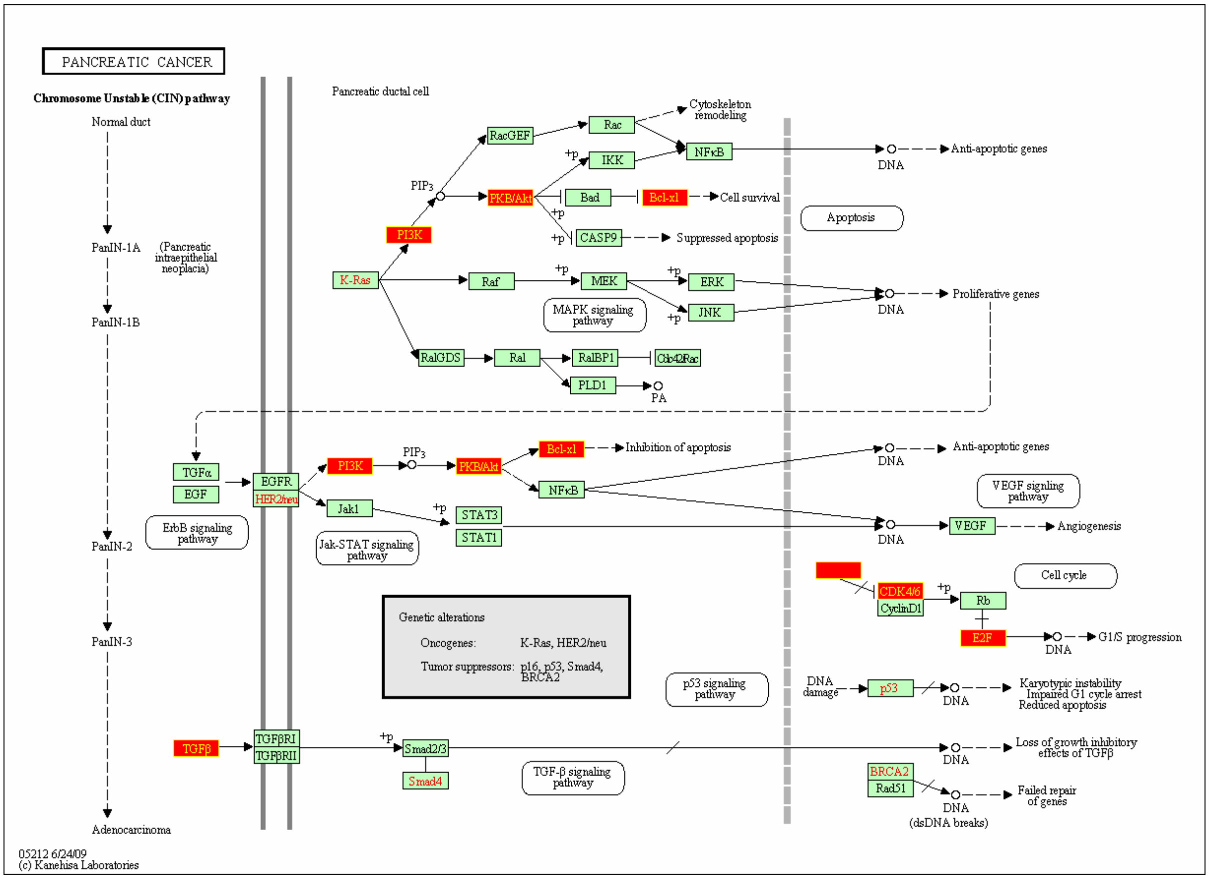


Figure 8: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG Pancreatic cancer pathway.

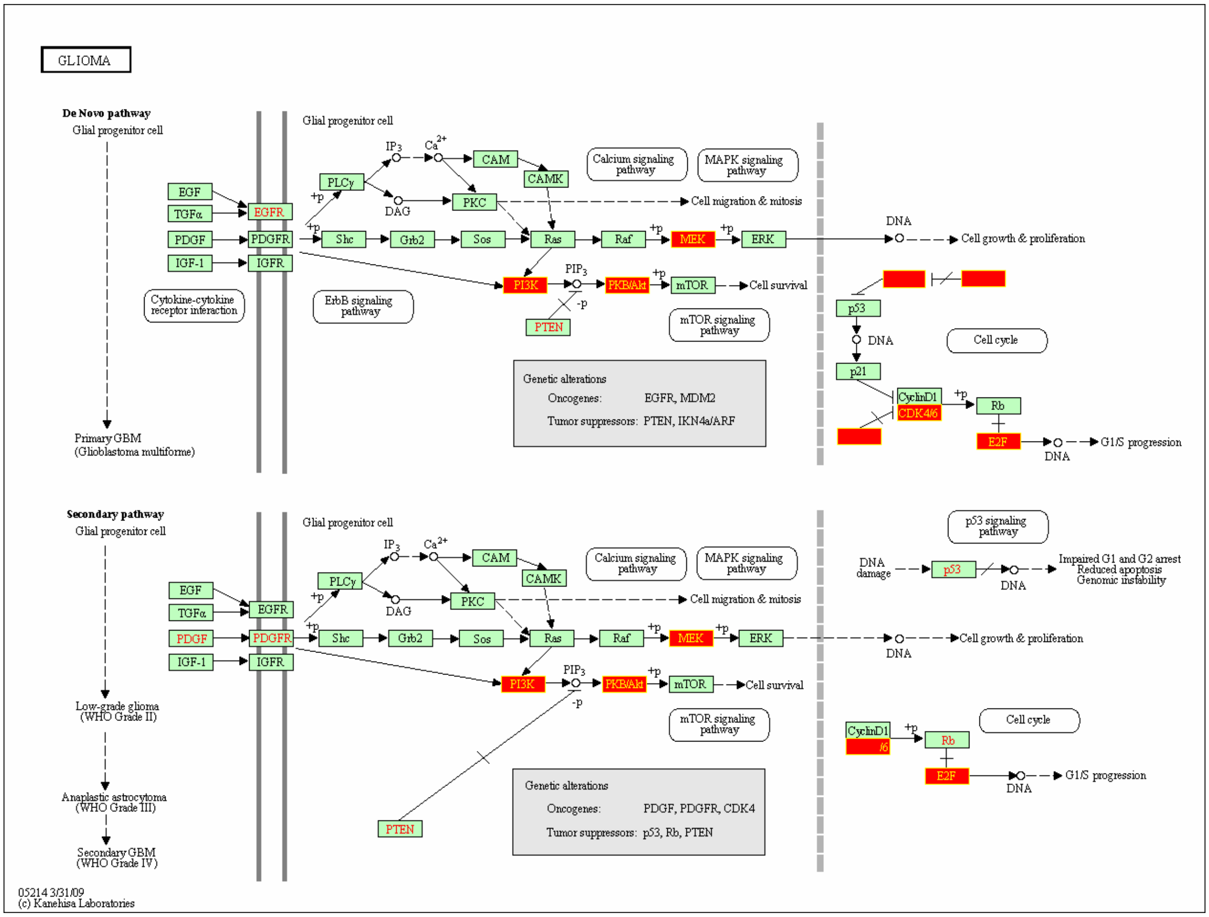


Figure 9: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG Human Glioma pathway.

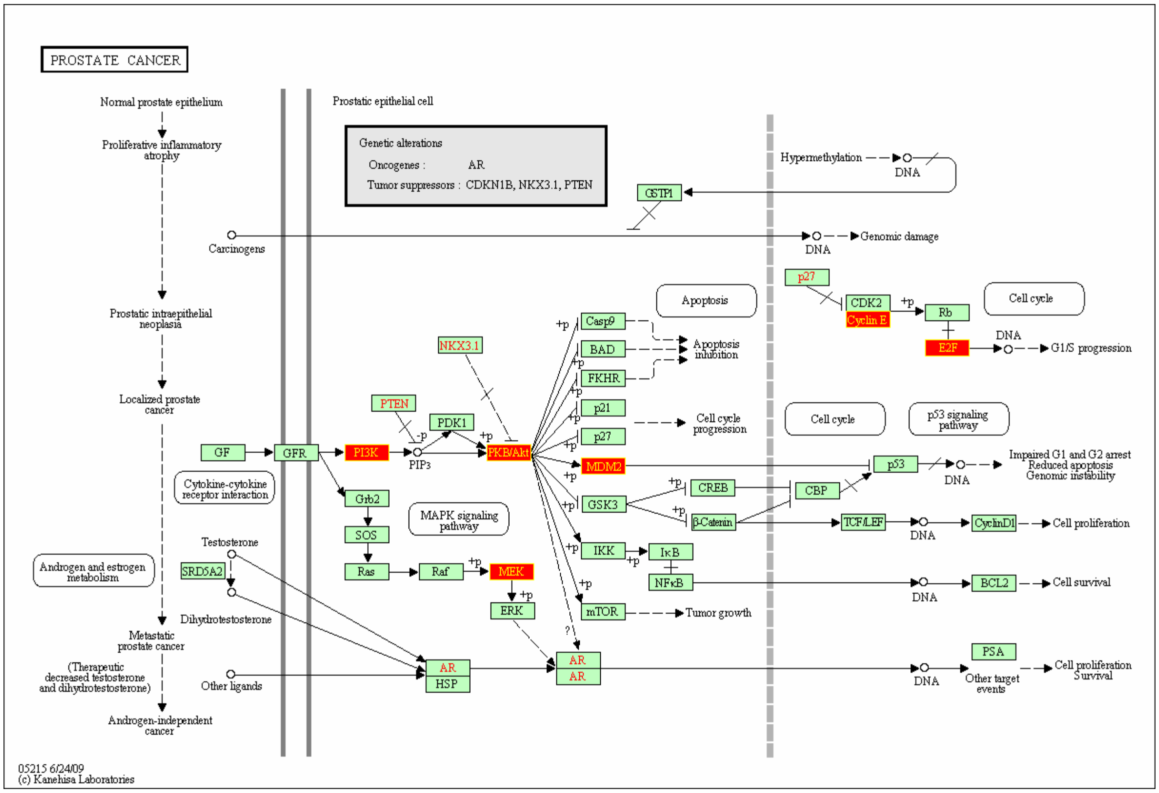


Figure 10: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG Prostate cancer pathway.

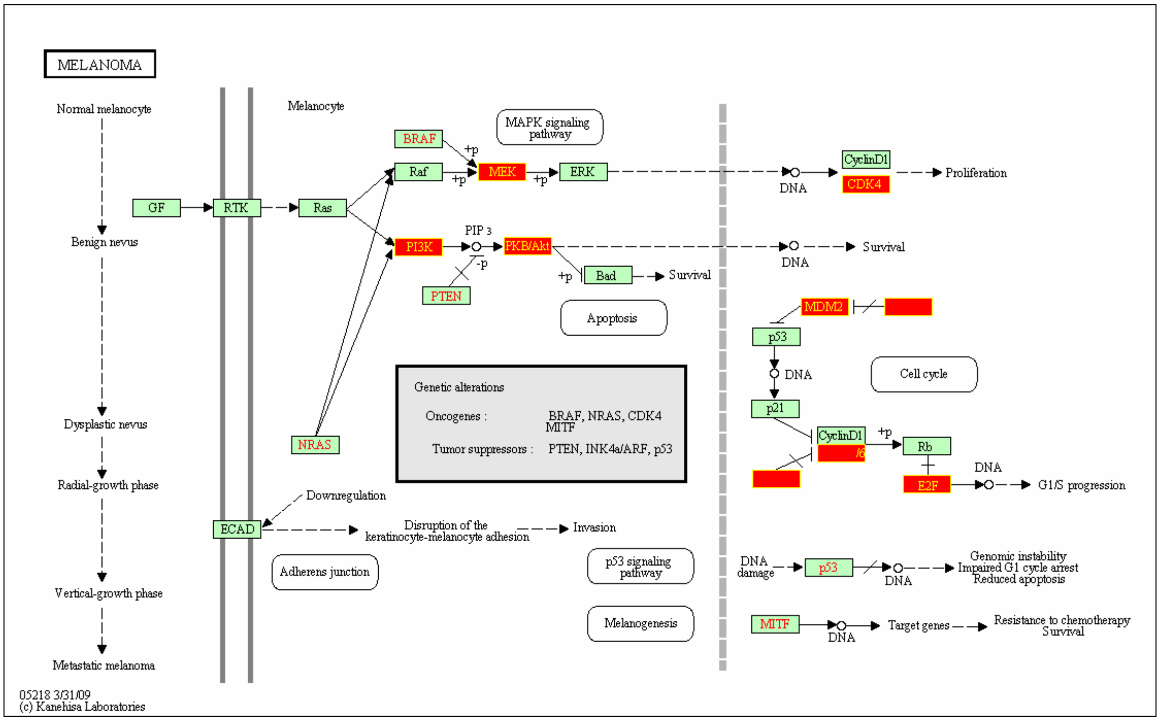


Figure 11: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG Human Melanoma pathway.

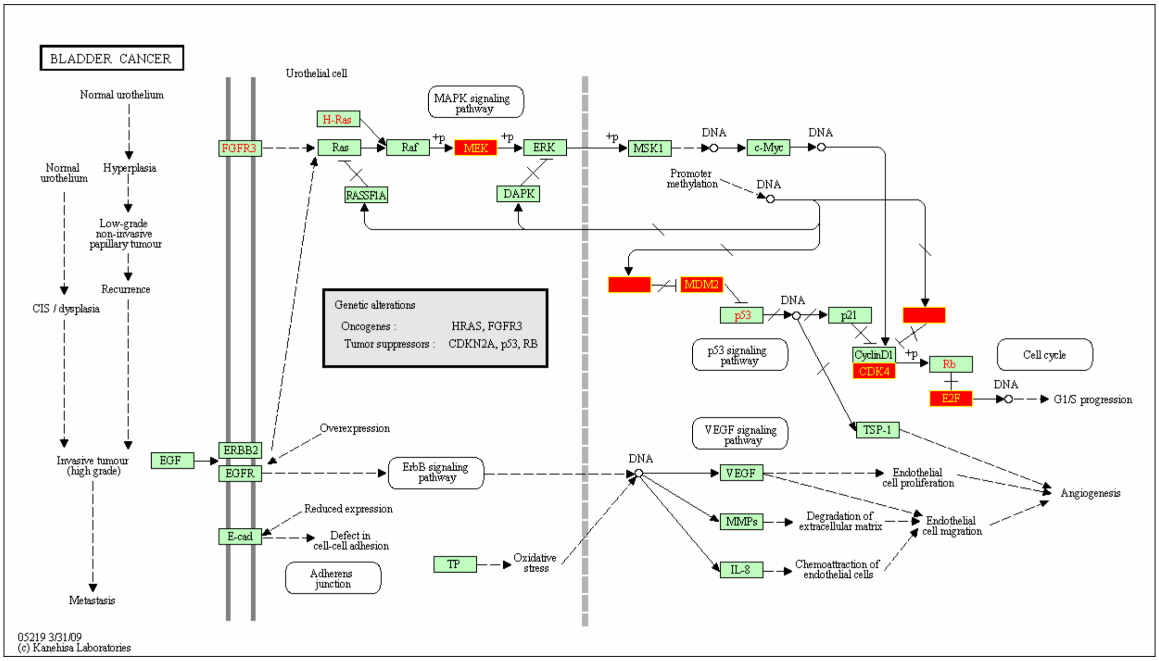


Figure 12: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG Bladder cancer pathway.

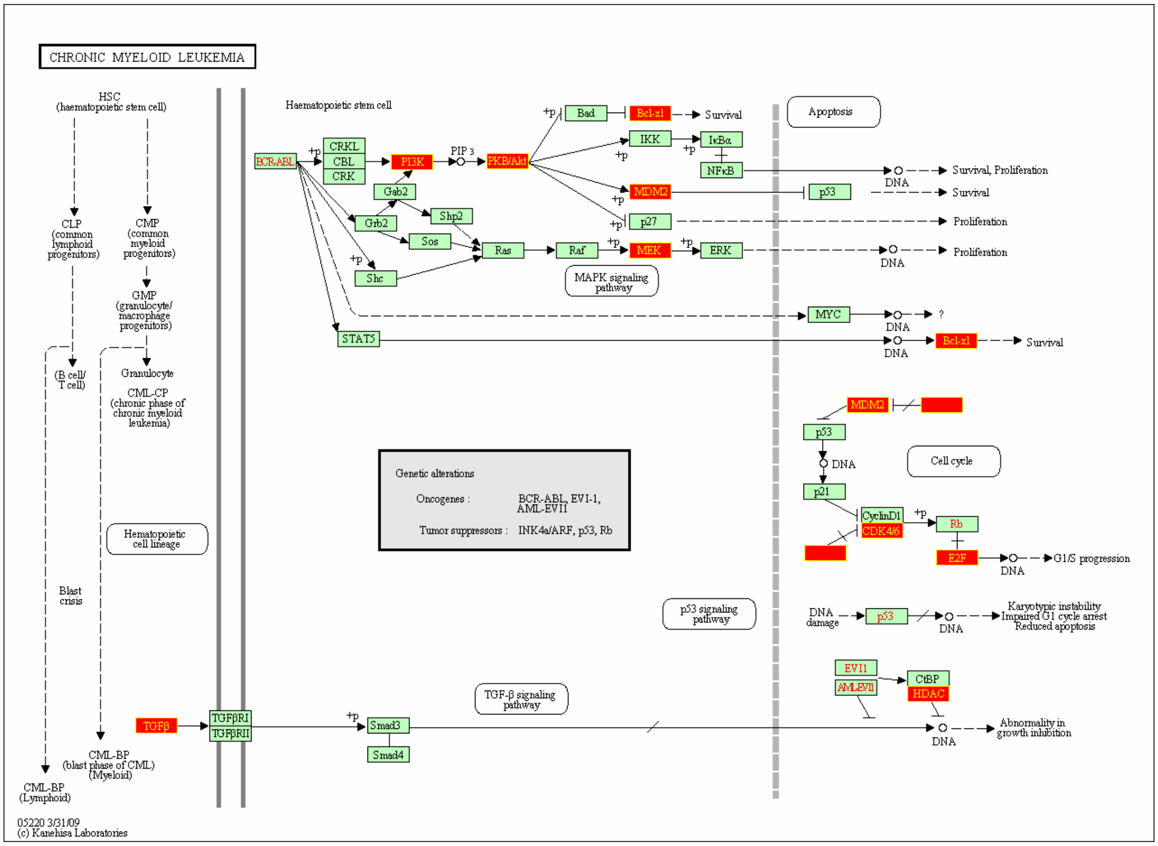


Figure 13: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG chronic myeloid leukemia pathway.

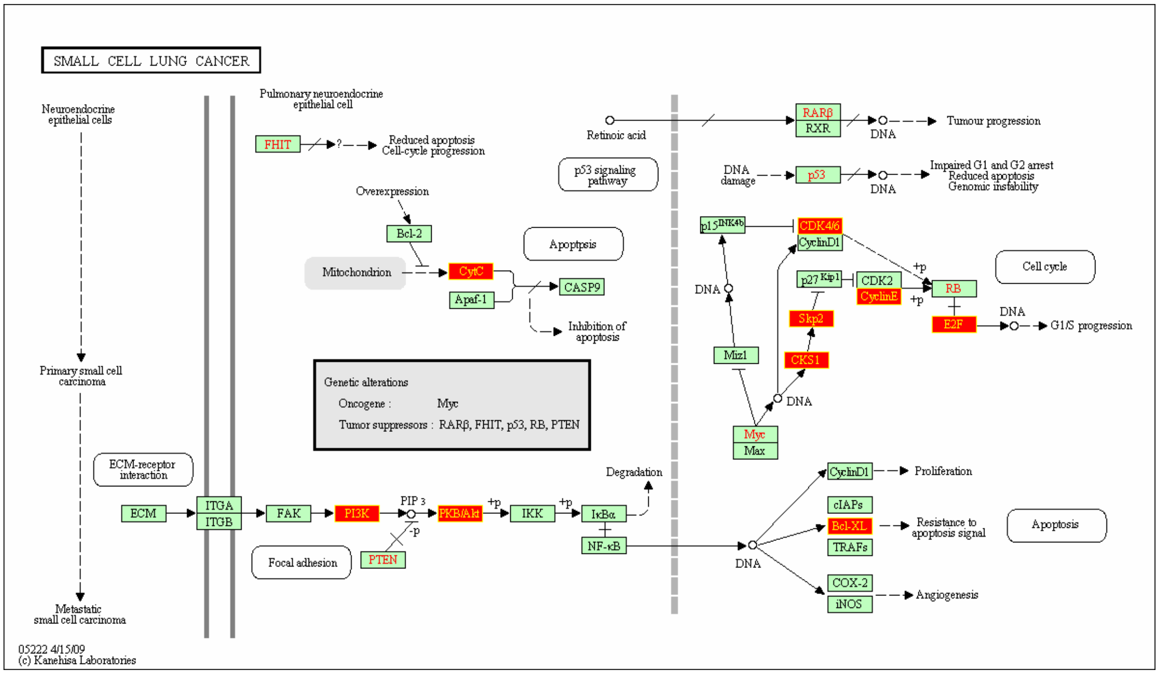


Figure 14: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG small cell lung cancer pathway.

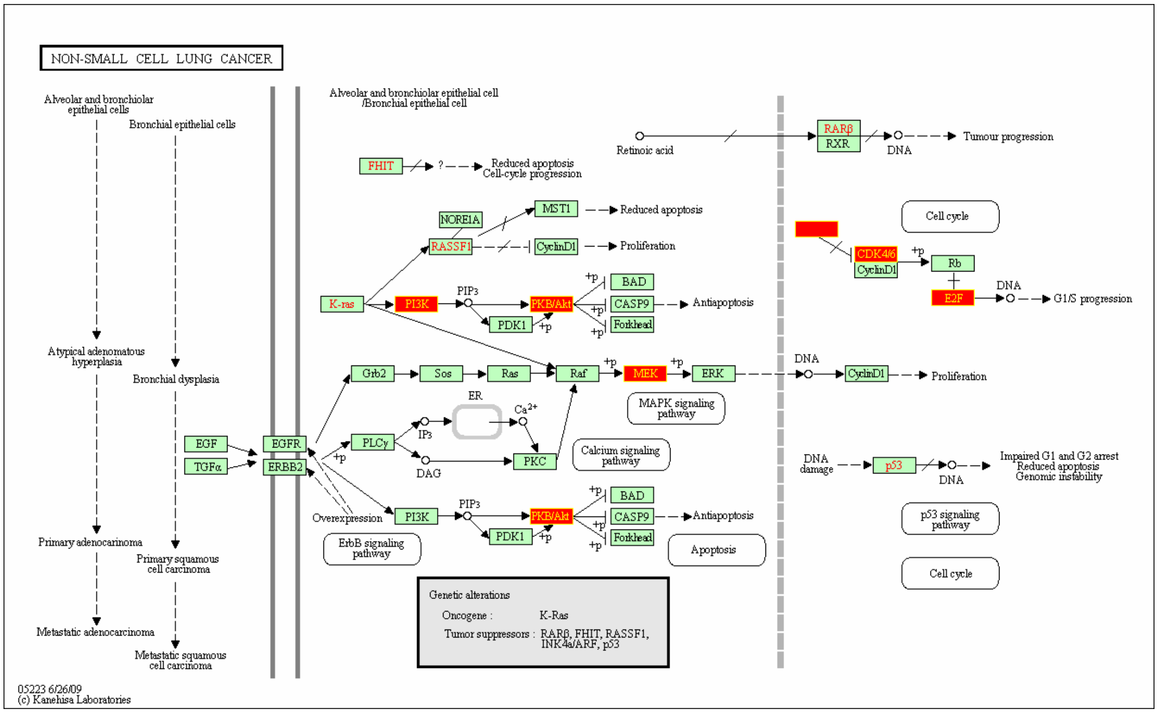


Figure 15: The figure illustrates abundant cancer-specific genes (in red background) in the KEGG Non-small cell lung cancer pathway.

Appendix C

Enriched pathway plots generated by KPP for normal-specific (anti-cancer) markers

The genes in light green background were human-specific.

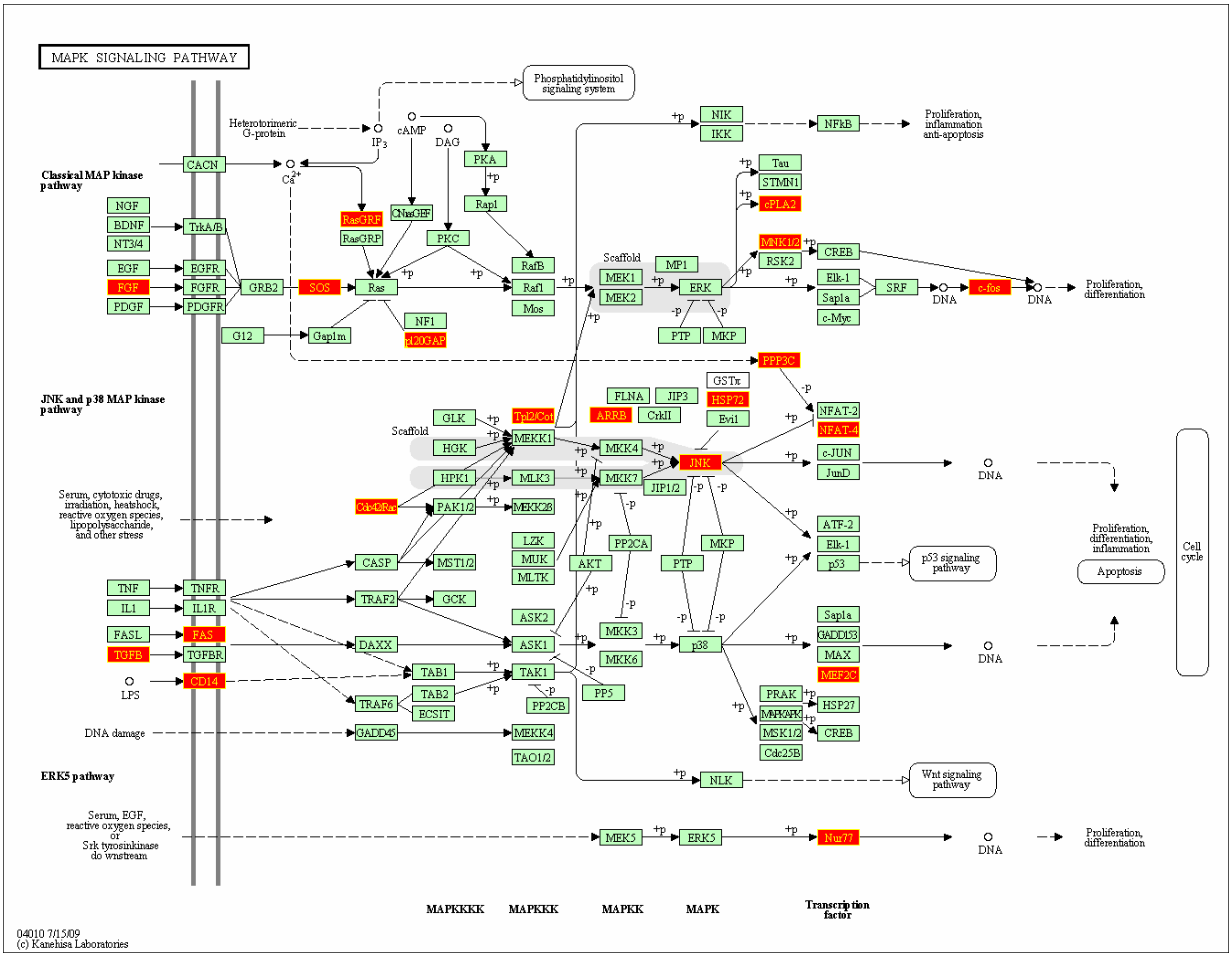


Figure 16: The figure illustrates abundant normal-specific genes (in red background) in the KEGG MAPK signaling pathway.

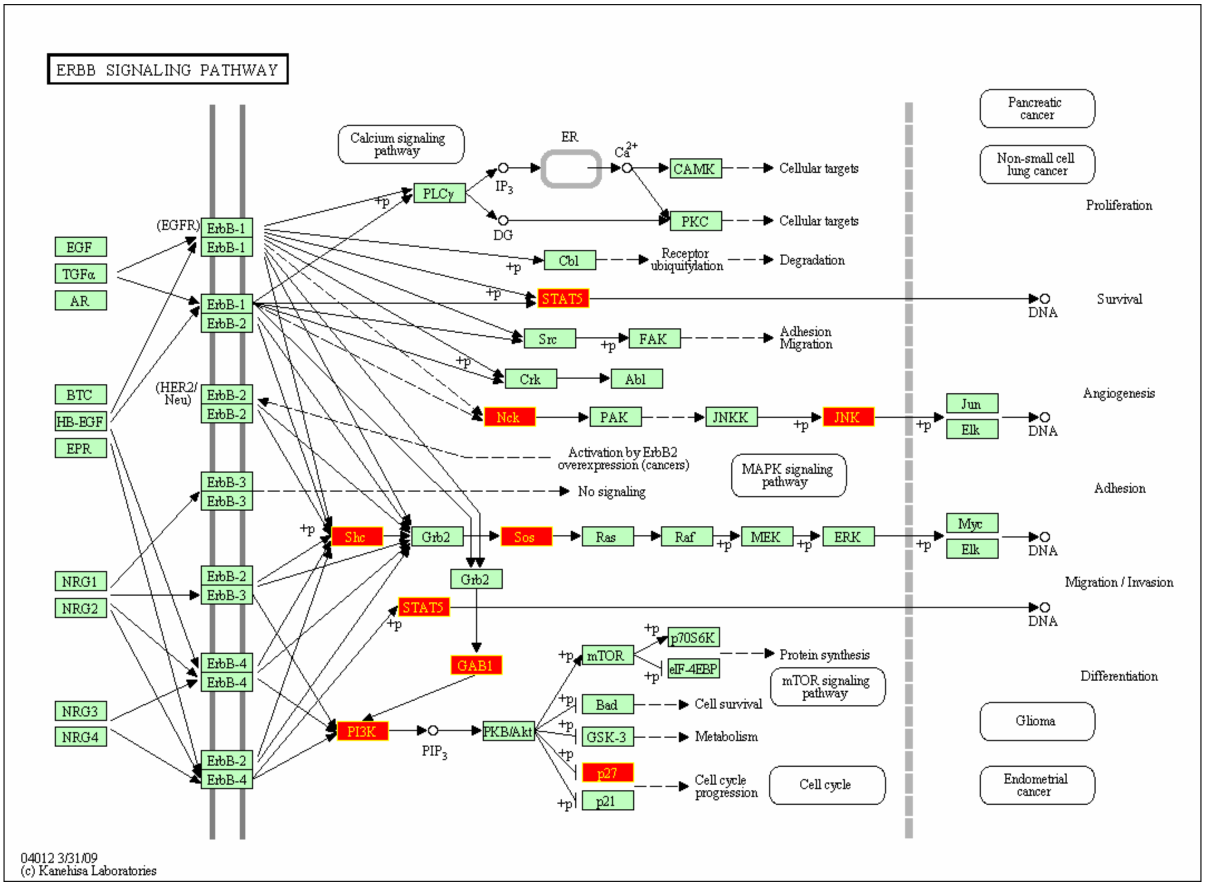


Figure 17: The figure illustrates abundant normal-specific genes (in red background) in the KEGG ERBB signaling pathway.

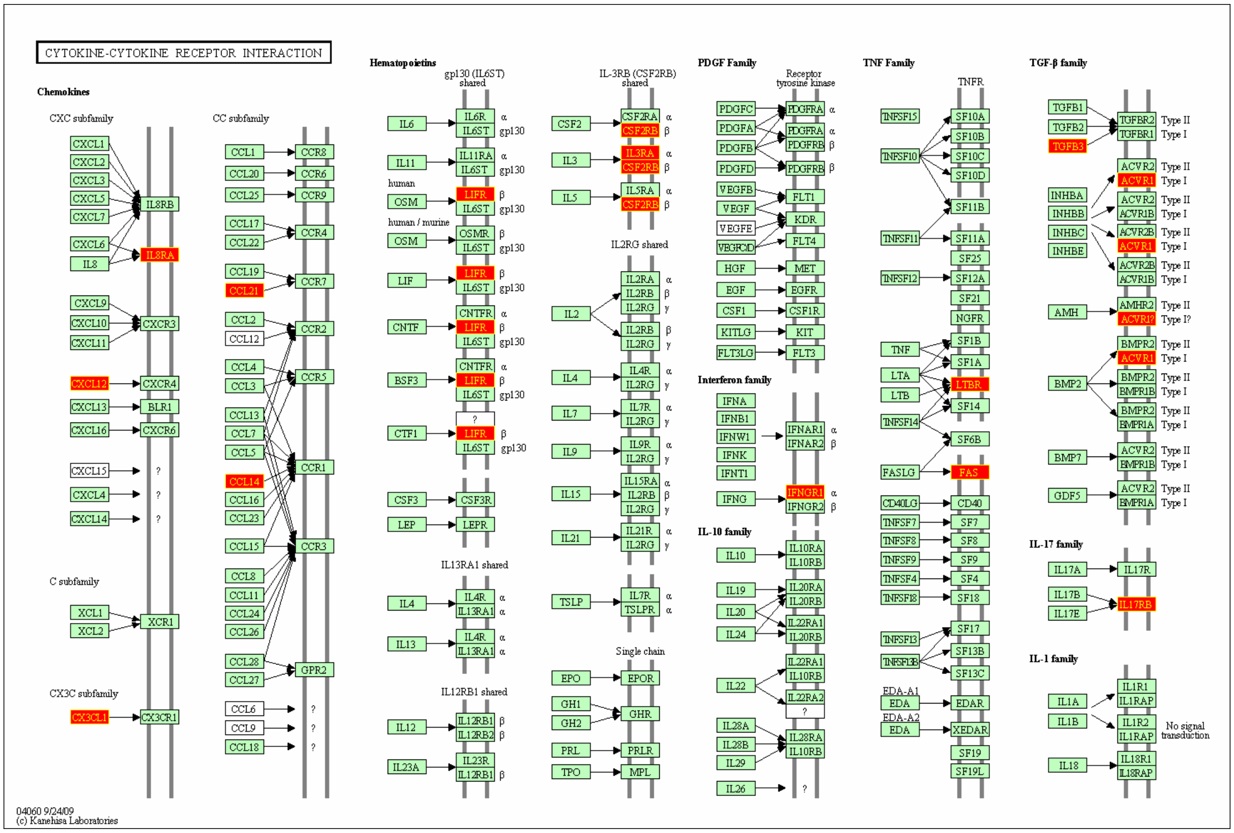


Figure 18: The figure illustrates abundant normal-specific genes (in red background) in the KEGG pathway representing cytokine-cytokine reception interaction genes.

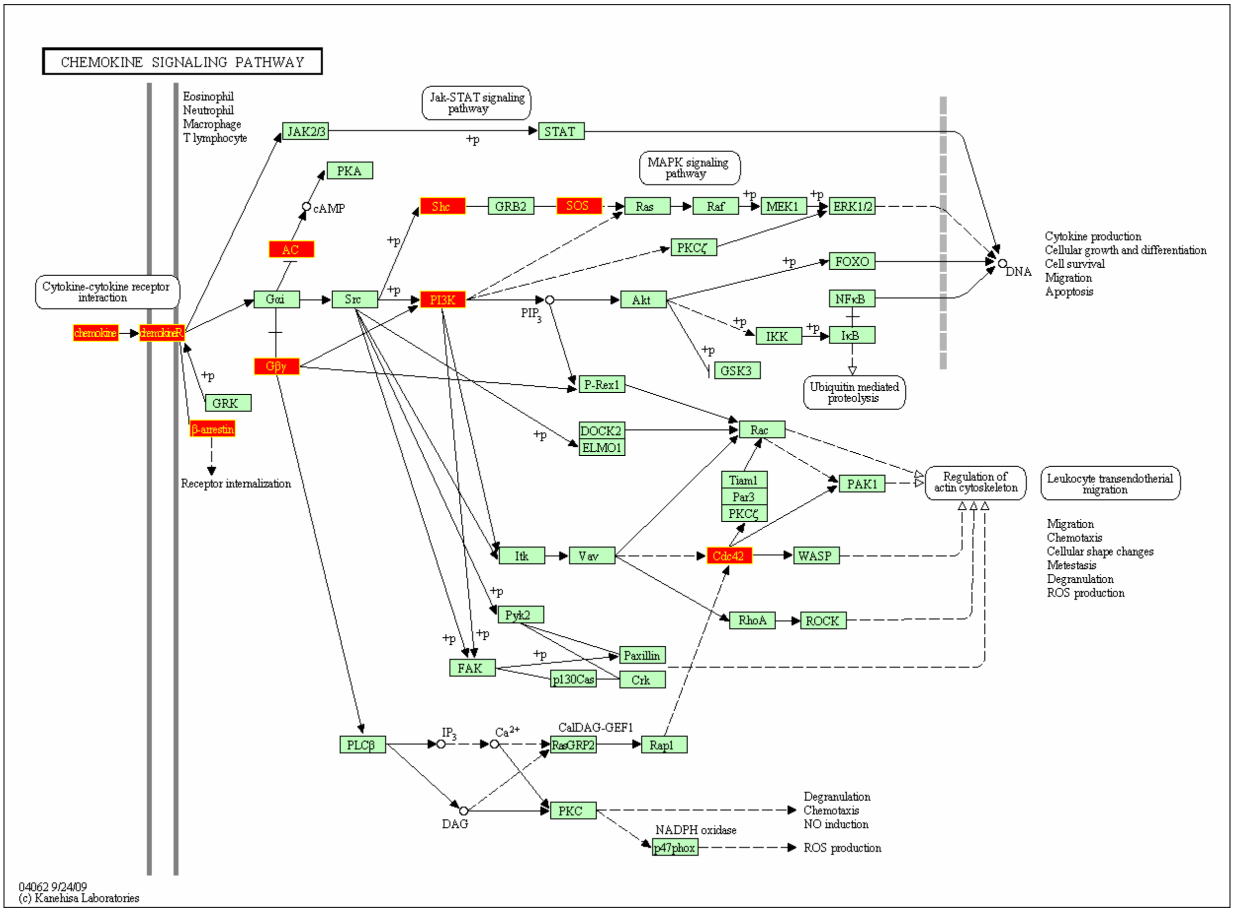


Figure 19: The figure illustrates abundant normal-specific genes (in red background) in the KEGG chemokine signaling pathway.

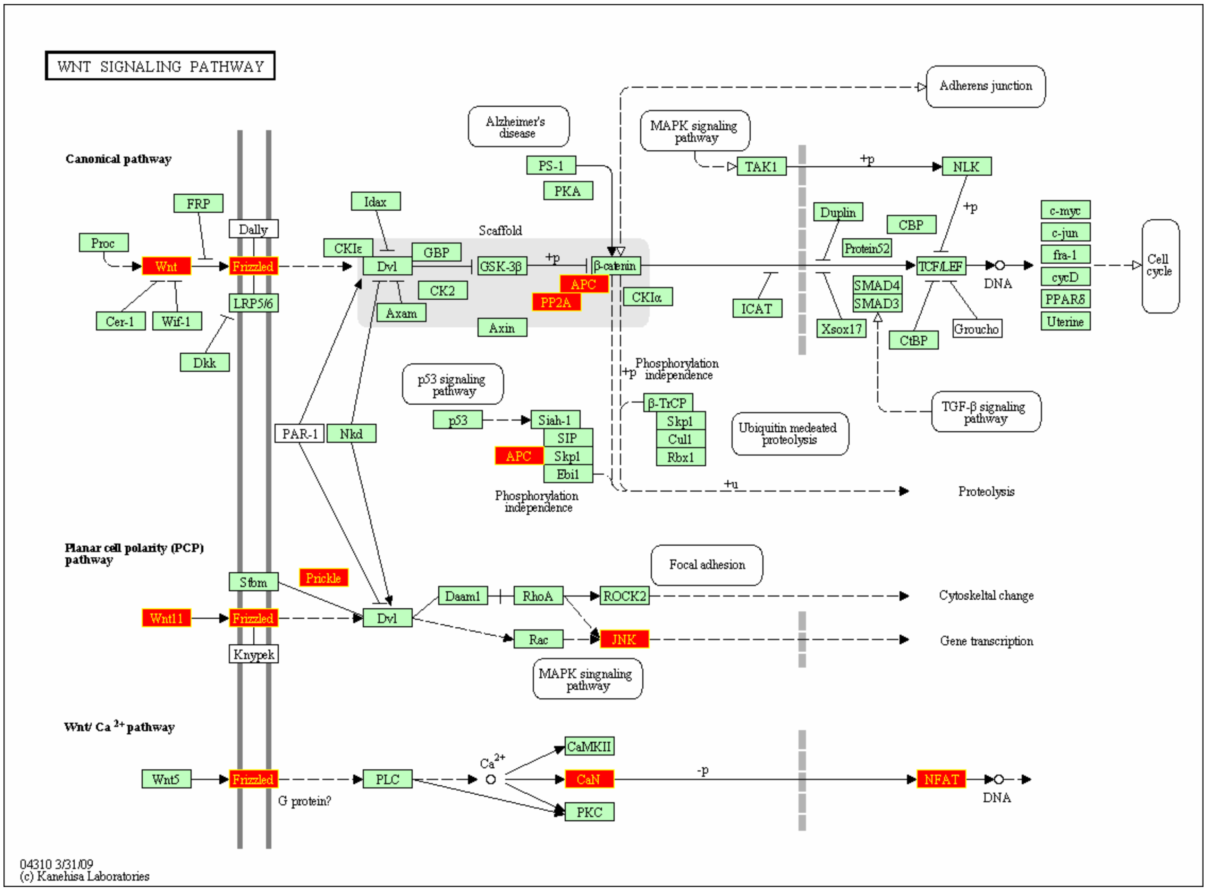


Figure 20: The figure illustrates abundant normal-specific genes (in red background) in the KEGG Wnt signaling pathway.

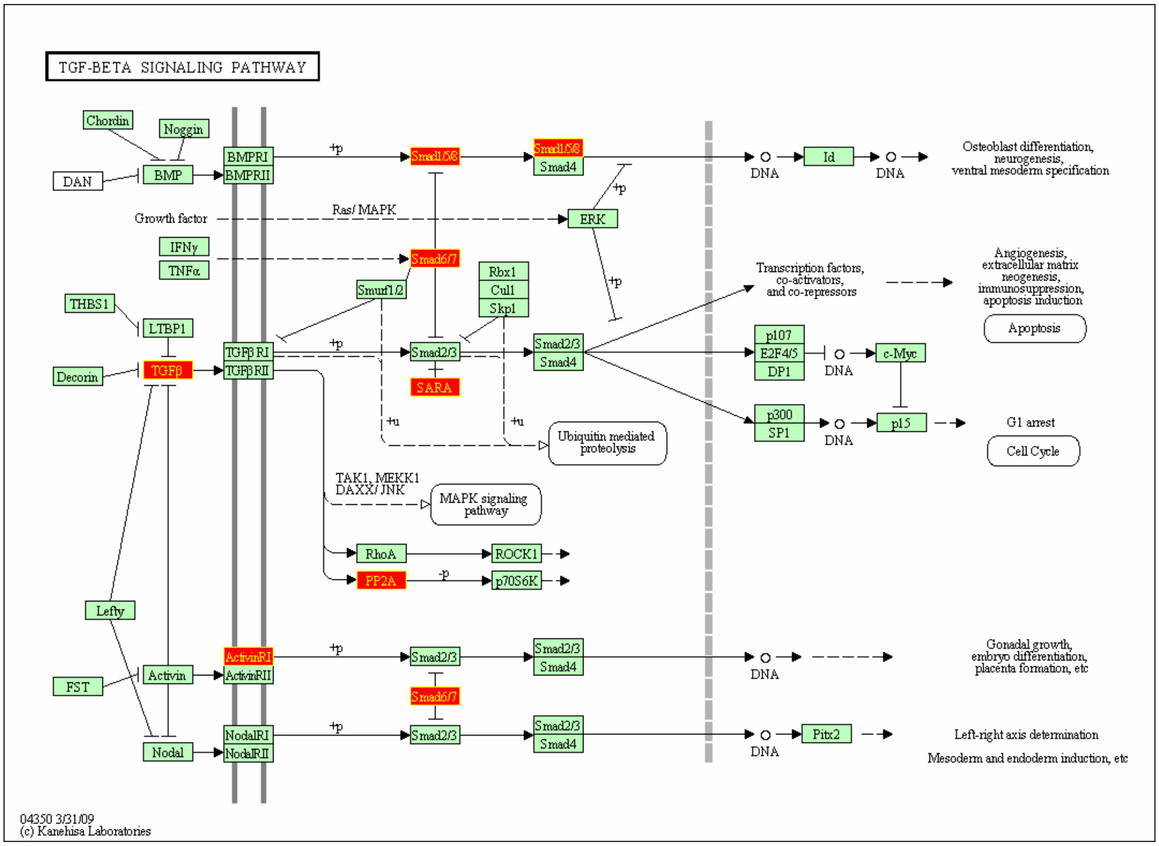


Figure 21: The figure illustrates abundant normal-specific genes (in red background) in the KEGG TGF-beta signaling pathway.

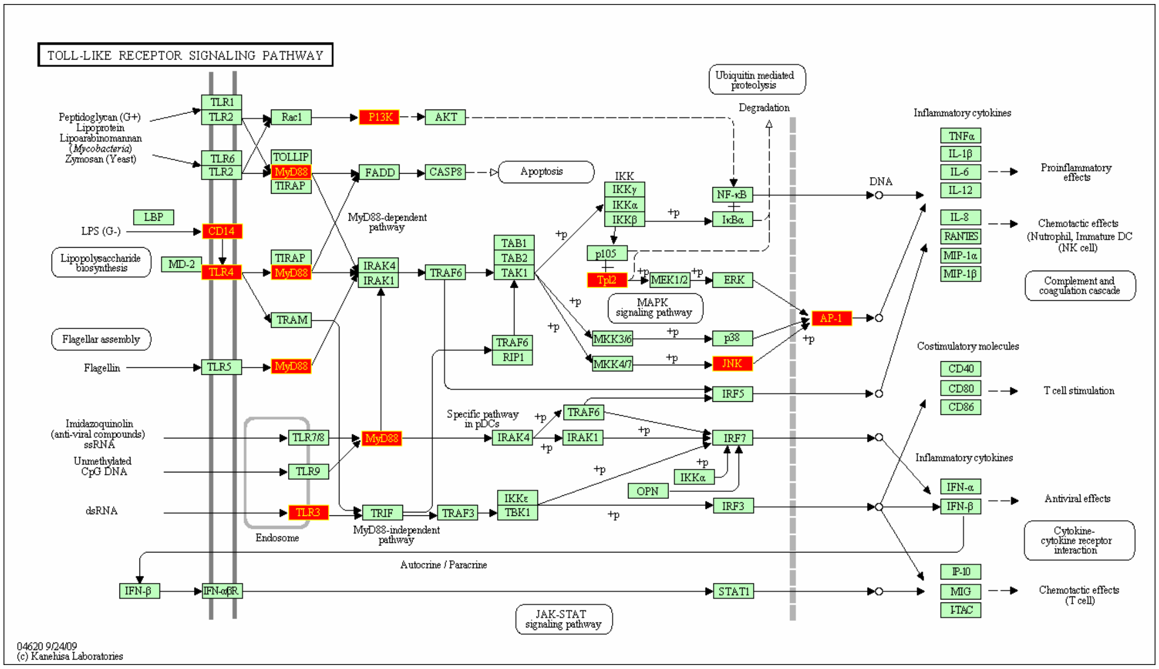


Figure 22: The figure illustrates abundant normal-specific genes (in red background) in the KEGG TOLL-like signaling pathway.

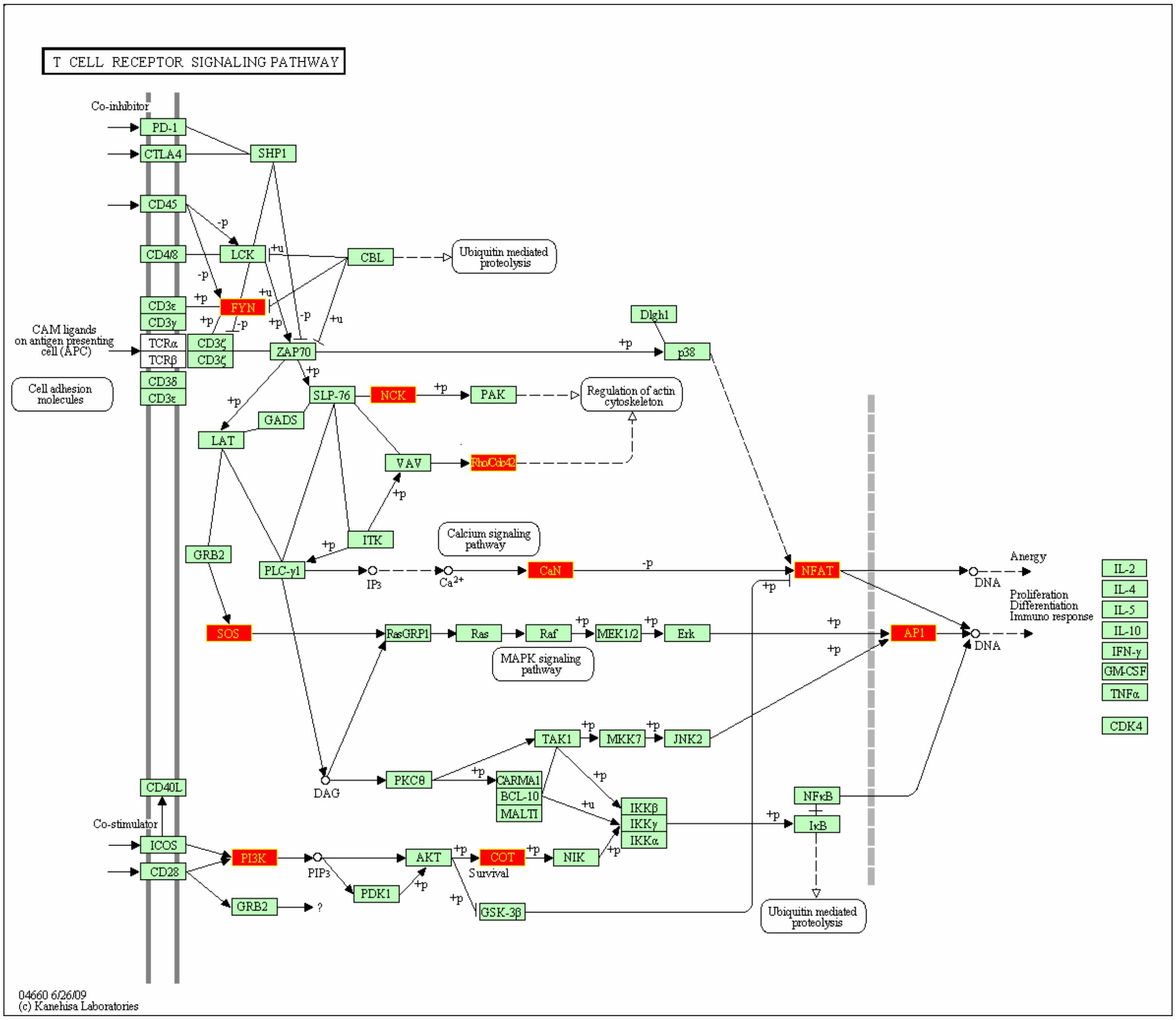


Figure 23: The figure illustrates abundant normal-specific genes (in red background) in the KEGG T-Cell receptor signaling pathway.

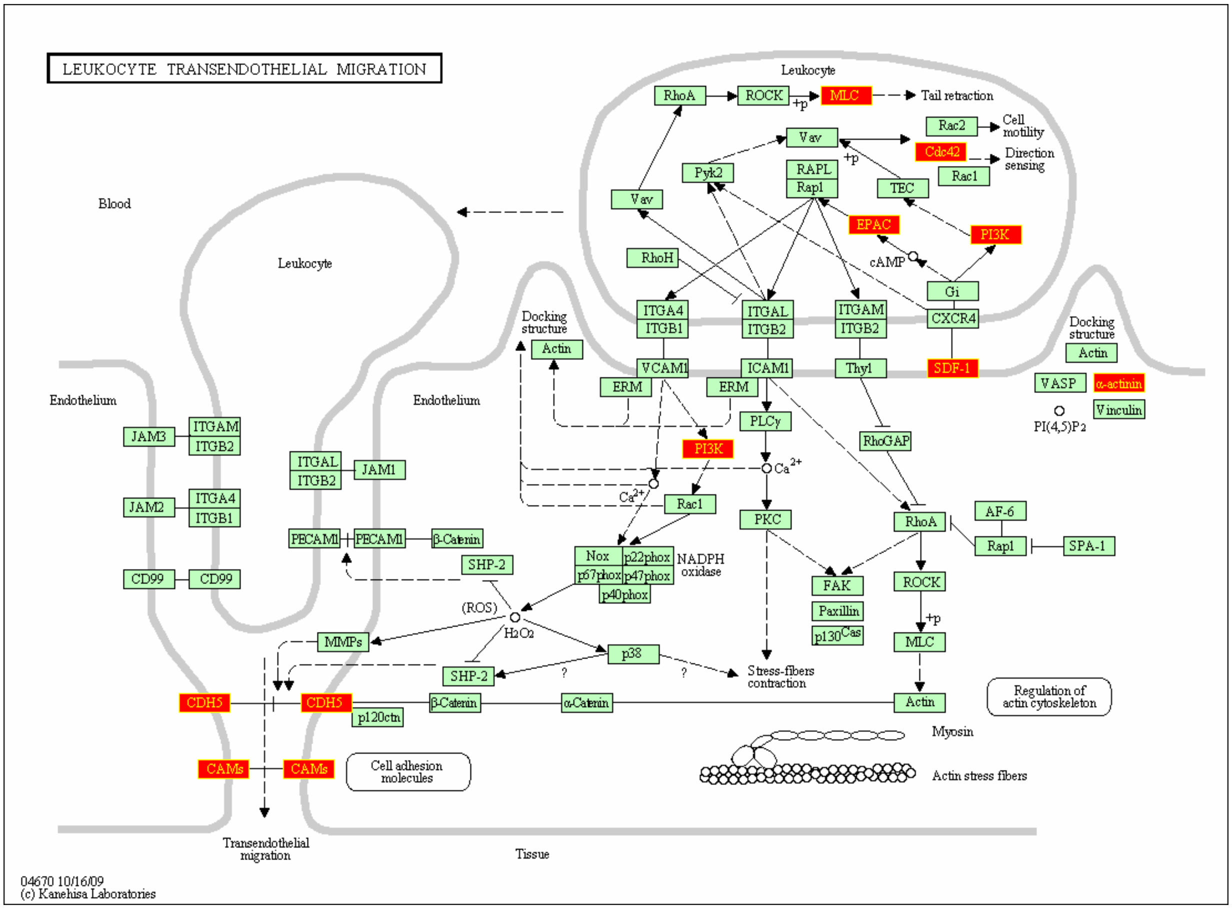


Figure 24: The figure illustrates abundant normal-specific genes (in red background) in the KEGG leukocyte trans-endothelial migration pathway.

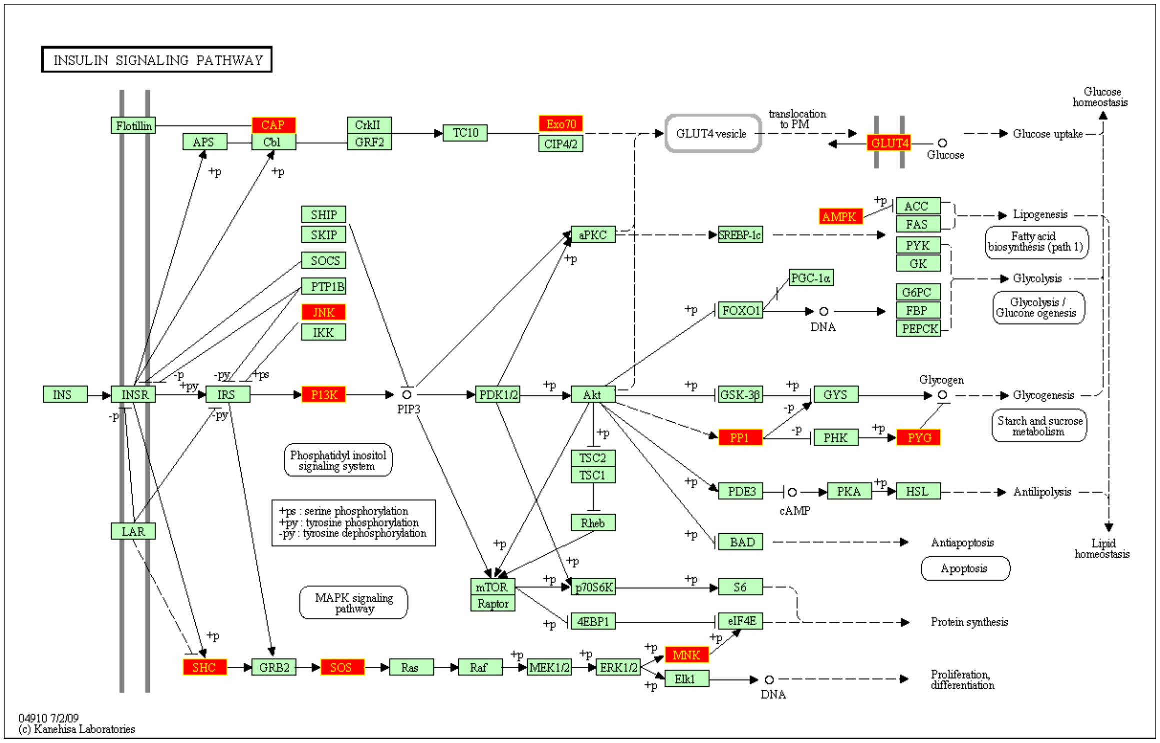


Figure 25: The figure illustrates abundant normal-specific genes (in red background) in the KEGG Insulin signalling pathway.

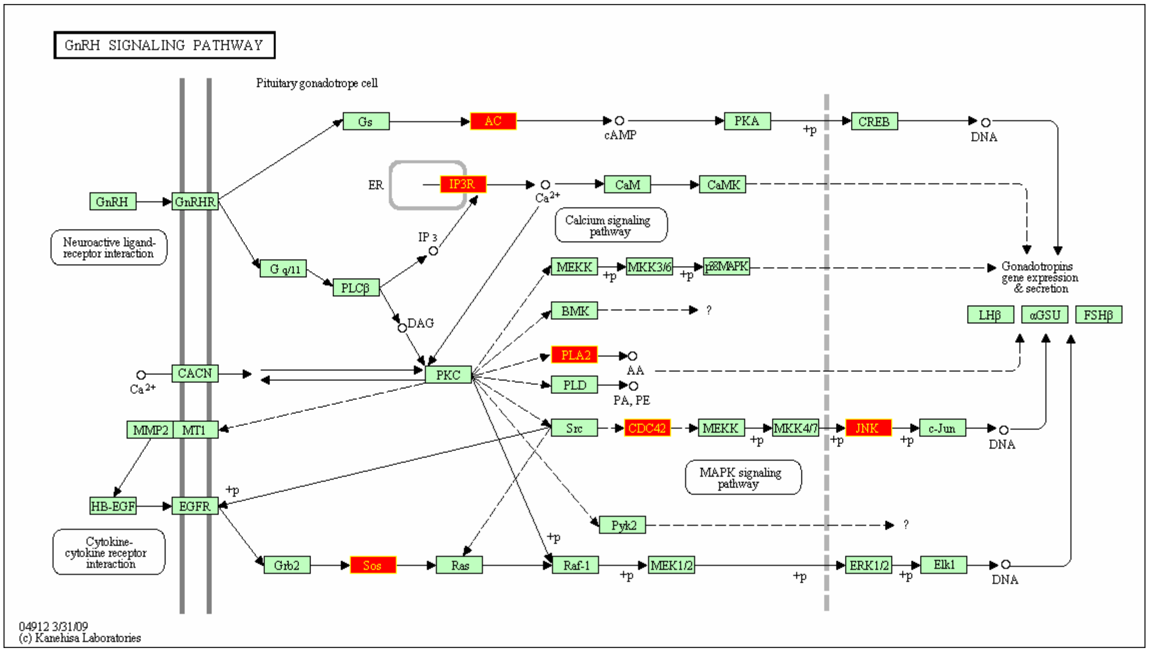


Figure 26: The figure illustrates abundant normal-specific genes (in red background) in the KEGG GnRH signalling pathway.

Appendix D

Results of the PCA analysis of the datasets comprised of two-point paired tumor and normal samples (Table 1)

Ductal Breast Carcinoma (GSE5764				
Relative importance:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
Standard deviation	0.261	0.0823	0.00898	0.00592
Proportion of Variance explained by component	0.908	0.0901	0.00107	0.00047
Cumulative Proportion	0.908	0.9985	0.99953	1.00000
Component Loadings:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
DCGlobal	-0.4792611	0.0347449	-0.3525736	0.8029903
DNGlobal	-0.4315327	-0.0642264	-0.7019314	-0.5629803
DCSpecific	-0.5729407	0.6708277	0.4351537	-0.1799179
DNSpecific	-0.5057938	-0.7380095	0.4400289	-0.0767414
Lobular Breast Carcinoma (GSE5764)				
Relative importance:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
Standard deviation	0.374	0.135	0.01778	0.00960
Proportion of Variance explained by component	0.882	0.116	0.00199	0.00058
Cumulative Proportion	0.882	0.997	0.99942	1.00000
Component Loadings:	PC1 “size”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
DCGlobal	-0.4336790	0.0032826	-0.5559545	0.7091025
DNGlobal	-0.4196183	-0.0321633	-0.5714445	-0.7045120
DCSpecific	-0.5651979	0.7156881	0.4093347	-0.0280528
DNSpecific	-0.5624894	-0.6976713	0.4436338	0.0070378

Pulmonary adenocarcinoma (GSE2514)				
Relative importance:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
Standard deviation	0.0901	0.0358	0.00143	0.000988
Proportion of Variance explained by component	0.8635	0.1361	0.00022	0.000100
Cumulative Proportion	0.8635	0.9997	0.99990	1.000000
Component Loadings:	PC1 “size”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
DCGlobal	-0.4223058	-0.2673878	0.3579391	0.7886958
DNGlobal	-0.3479712	0.3425144	0.7661334	-0.4178990
DCSpecific	-0.6195546	-0.6111058	-0.2421958	-0.4290024
DNSpecific	-0.5627841	0.6616172	-0.4756684	0.1388390

Pulmonary adenocarcinoma (GSE7670)				
Relative importance:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
Standard Deviation	0.193	0.0576	0.00663	0.00149
Proportion of Variance explained by component	0.917	0.0815	0.00108	0.00005
Cumulative Proportion	0.917	0.9989	0.99995	1.00000
Component Loadings:	PC1 “size”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
DCGlobal	-0.4241949	0.2328628	-0.4694148	0.7385685
DNGlobal	-0.3486490	-0.3041166	-0.6972629	-0.5475230
DCSpecific	-0.6355490	0.6064523	0.3269082	-0.3484596
DNSpecific	-0.5427522	-0.6967809	0.4319783	0.1825134

Esophageal Cancer (GSE1420)

Relative importance:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
Standard deviation	0.226	0.085	0.00990	0.00172
Proportion of Variance explained by component	0.875	0.124	0.00168	0.00005
Cumulative Proportion	0.875	0.998	0.99995	1.00000
Component Loadings:	PC1 “size”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
DCGlobal	-0.3634003	0.2365433	0.5838462	0.6863753
DNGlobal	-0.3856582	-0.2269184	0.6150850	-0.6491890
DCSpecific	-0.5892720	0.6806972	-0.3663376	-0.2349607
DNSpecific	-0.6098905	-0.6551386	-0.3828719	0.2285521

Renal cell carcinoma (GSE6344)

Relative importance:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
Standard deviation	0.133	0.0715	0.00230	0.000472
Proportion of Variance explained by component	0.777	0.2231	0.00023	0.000010
Cumulative Proportion	0.777	0.9998	0.99999	1.000000
Component Loadings:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
DCGlobal	-0.4536096	0.3104922	0.6132022	0.5672884
DNGlobal	-0.3913748	-0.3780481	0.5147935	-0.6624901
DCSpecific	-0.5869027	0.5997387	-0.4316633	-0.3309463
DNSpecific	-0.5446185	-0.6332359	-0.4154967	0.3602296

Renal cell carcinoma (GSE781)				
Relative importance:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
Standard deviation	0.138	0.073	0.00366	0.000972
Proportion of Variance explained by component	0.781	0.219	0.00055	0.000040
Cumulative Proportion	0.781	0.999	0.99996	1.000000
Component Loadings:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
DCGlobal	-0.4015367	0.1770823	-0.5931930	-0.6749313
DNGlobal	-0.3159439	-0.2905877	-0.6200694	0.6566979
DCSpecific	-0.6328689	0.6538785	0.3089273	0.2765569
DNSpecific	-0.5817428	-0.6757541	0.4101218	-0.1916557

Head and neck squamous cell carcinoma (GDS2520)				
Relative importance:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “Degree of autonomy”	PC4 “Noise”
Standard deviation	0.190	0.0407	0.00979	0.00126
Proportion of Variance explained by component	0.954	0.0436	0.00252	0.00004
Cumulative Proportion	0.954	0.9974	0.99996	1.00000
Component Loadings:	PC1 “Attractor”	PC2 “Normal/Cancer difference”	PC3 “degree of autonomy”	PC4 “noise”
DCGlobal	-0.3839816	0.1435169	0.5972208	0.6894116
DNGlobal	-0.3578240	-0.1809713	0.6062387	-0.6867940
DCSpecific	-0.6171703	0.6782271	-0.3589338	-0.1739981
DNSpecific	-0.5861919	-0.6976105	-0.3833646	0.1508320

Papillary thyroid carcinoma (GDS1665)

Relative importance:	PC1 "Attractor"	PC2 "Normal/Cancer difference"	PC3 "Degree of autonomy"	PC4 "Noise"
Standard deviation	0.0968	0.0398	0.00375	0.00120
Proportion of Variance explained by component	0.8542	0.1444	0.00128	0.00013
Cumulative Proportion	0.8542	0.9986	0.99987	1.00000
Component Loadings:	PC1 "Attractor"	PC2 "Normal/Cancer difference"	PC3 "Degree of autonomy"	PC4 "Noise"
DCGlobal	-0.4055784	0.1936700	-0.4882024	0.7481019
DNGlobal	-0.3365074	-0.2185903	-0.7043803	-0.5855164
DCSpecific	-0.6383106	0.6270132	0.3455199	-0.2828959
DNSpecific	-0.5610958	-0.7221944	0.3822601	0.1322270

\

Appendix E

Results of the PCA analysis of population datasets comprised of two-point tumor and normal samples (Table 2)

Invasive Breast (Epithelial) Carcinoma (GSE10797)

Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.298	0.0418	0.01457	0.00595
Proportion of Variance explained by component	0.978	0.0192	0.00233	0.00039
Cumulative Proportion	0.978	0.9973	0.99961	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.56631	-0.30366	0.196662	-0.74055
DNGlobal	-0.4949	0.21754	0.695137	0.473856
DCSpecific	-0.43919	-0.62278	-0.44098	0.474118
DNSpecific	-0.49141	0.687466	-0.53259	-0.04754

Cervical Cancer (GSE6791)

Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.223	0.0804	0.00402	0.00162
Proportion of Variance explained by component	0.884	0.1153	0.00029	0.00005
Cumulative Proportion	0.884	0.9997	0.99995	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.42173	-0.43508	0.605168	-0.51636
DNGlobal	-0.5185	0.346409	0.43942	0.646585
DCSpecific	-0.40695	-0.69282	-0.47056	0.364638
DNSpecific	-0.62265	0.459029	-0.46825	-0.42701

Head and Neck Cancer (GSE6791)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.169	0.0308	0.00463	0.0011
Proportion of Variance explained by component	0.967	0.0319	0.00072	0.00004
Cumulative Proportion	0.967	0.9992	0.99996	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.45602	-0.30736	0.542741	0.634831
DNGlobal	-0.47091	0.312987	0.525327	-0.63586
DCSpecific	-0.47903	-0.70074	-0.41236	-0.33083
DNSpecific	-0.58379	0.562613	-0.50934	0.288488

Mesothelioma (GSE12345)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.244	0.0846	0.00892	0.00362
Proportion of Variance explained by component	0.892	0.1071	0.00119	0.0002
Cumulative Proportion	0.892	0.9986	0.9998	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.34642	0.145146	-0.63865	0.671609
DNGlobal	-0.34828	-0.08457	-0.59061	-0.72299
DCSpecific	-0.51471	0.780904	0.334231	-0.11643
DNSpecific	-0.70269	-0.60164	0.362763	0.112533

Nasopharyngeal Carcinoma (GSE12452)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.184	0.0488	0.00475	0.00121
Proportion of Variance explained by component	0.934	0.0655	0.00062	0.00004
Cumulative Proportion	0.934	0.9993	0.99996	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.33784	0.289965	-0.70708	0.549388
DNGlobal	-0.4066	-0.1413	-0.46484	-0.77372
DCSpecific	-0.47991	0.752777	0.42756	-0.14215
DNSpecific	-0.70016	-0.57383	0.318053	0.281654

Oral Squamous Cell Carcinoma (GSE3524)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.171	0.0522	0.00434	0.00237
Proportion of Variance explained by component	0.914	0.0857	0.00059	0.00018
Cumulative Proportion	0.914	0.9992	0.99982	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.36222	-0.29688	0.725488	-0.50431
DNGlobal	-0.40249	0.162528	0.419678	0.797156
DCSpecific	-0.43109	-0.78102	-0.42128	0.163367
DNSpecific	-0.72177	0.524838	-0.3465	-0.28901

Renal Cell carcinoma(GSE14762)

Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.271	0.0922	0.02352	0.00252
Proportion of Variance explained by component	0.891	0.1027	0.00669	0.00008
Cumulative Proportion	0.891	0.9932	0.99992	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.48269	0.14762	-0.27471	-0.81838
DNGlobal	-0.33537	-0.40744	-0.75998	0.379411
DCSpecific	-0.64799	0.602155	0.178277	0.430968
DNSpecific	-0.4844	-0.67053	0.561411	-0.02369

Papillary thyroid carcinoma (GSE3678)

Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.17	0.0807	0.00574	0.00121
Proportion of Variance explained by component	0.814	0.1847	0.00094	0.00004
Cumulative Proportion	0.814	0.999	0.99996	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.3045	0.081804	0.496542	-0.80872
DNGlobal	-0.2992	-0.09796	0.759593	0.569125
DCSpecific	-0.64121	0.702814	-0.27129	0.145953
DNSpecific	-0.63766	-0.69983	-0.32073	-0.02762

Appendix F

Results of the PCA analysis of the datasets comprised of multi-stage tumor and normal samples collected across population (Table 3)

Ovarian (fallopian tube) carcinoma (GSE10971)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.271	0.0956	0.00778	0.00111
Proportion of Variance explained by component	0.889	0.1107	0.00073	0.00001
Cumulative Proportion	0.889	0.9992	0.99999	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.4319740	0.1653989	0.5648672	-0.6833496
DNGlobal	-0.3371510	-0.2908749	0.6148158	0.6509397
DCSpecific	-0.6792346	0.5755423	-0.3745092	0.2591029
DNSpecific	-0.4882294	-0.7461810	-0.4033230	-0.2053693
Bladder carcinoma (GSE3167)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.228	0.0694	0.00589	0.00160
Proportion of Variance explained by component	0.914	0.0850	0.00061	0.00005
Cumulative Proportion	0.914	0.9993	0.99995	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.4219051	0.4471904	-0.6546958	-0.4397615
DNGlobal	-0.5576904	-0.3487430	-0.3327277	0.6757603
DCSpecific	-0.3752885	0.7052464	0.5163564	0.3084836
DNSpecific	-0.6083880	-0.4254722	0.4405017	-0.5047729

Esophagus carcinoma (GSE1420)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.246	0.101	0.0119	0.00201
Proportion of Variance explained by component	0.854	0.144	0.0020	0.00006
Cumulative Proportion	0.854	0.998	0.9999	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.3604909	0.2376977	0.5853858	0.6861993
DNGlobal	-0.3849471	-0.2241357	0.6157891	-0.6499097
DCSpecific	-0.5851186	0.6857592	-0.3637739	-0.2346041
DNSpecific	-0.6160343	-0.6503824	-0.3818320	0.2273955

Hepato Cellular carcinoma (GSE6764)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.134	0.0425	0.00439	0.000683
Proportion of Variance explained by component	0.908	0.0911	0.00097	0.000020
Cumulative Proportion	0.908	0.9990	0.99998	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.4871506	0.1924781	-0.5238308	0.6717423
DNGlobal	-0.3563782	-0.3680381	-0.5963258	-0.6180114
DCSpecific	-0.6539801	0.5581170	0.3891397	-0.3307351
DNSpecific	-0.4560580	-0.7183352	0.4675117	0.2396632

Pancreas carcinoma (Logsdon et al.)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.213	0.106	0.0098	0.00296
Proportion of Variance explained by component	0.801	0.198	0.0017	0.00016
Cumulative Proportion	0.801	0.998	0.9998	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.3106436	0.3413867	-0.5844092	-0.6673991
DNGlobal	-0.4308919	-0.1495312	-0.6043367	0.6532609
DCSpecific	-0.4633882	0.7489053	0.4070460	0.2423340
DNSpecific	-0.7092984	-0.5479383	0.3571507	-0.2628743

Prostate carcinoma (GSE6919)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.094	0.0218	0.00202	0.000289
Proportion of Variance explained by component	0.948	0.0512	0.00044	0.000010
Cumulative Proportion	0.948	0.9996	0.99999	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.4868723	0.2591513	-0.5374330	0.6379355
DNGlobal	-0.3937599	-0.3927989	-0.5607571	-0.6133626
DCSpecific	-0.6342255	0.5425107	0.4349970	-0.3379612
DNSpecific	-0.4535047	-0.6958676	0.4555160	0.3203231

Prostate carcinoma (GSE3325)				
Relative importance:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
Standard deviation	0.185	0.0962	0.00555	0.00142
Proportion of Variance explained by component	0.787	0.2127	0.00071	0.00005
Cumulative Proportion	0.787	0.9992	0.99995	1
Component Loadings:	PC1 (Attractor)	PC2 (Normal/Cancer difference)	PC3 (Degree of autonomy)	PC4 (Noise)
DCGlobal	-0.4656498	0.1456634	-0.6308057	0.6033544
DNGlobal	-0.2653653	-0.3860657	-0.5525927	-0.6893300
DCSpecific	-0.7344495	0.4844237	0.3763689	-0.2902828
DNSpecific	-0.4163357	-0.7714072	0.3937917	0.2766288

Appendix G

Protein-coding genes extracted using the EST abundance analysis of the Human Unigene data as potential tumor biomarkers (Table 4).

S.No	Unigene_ID	Gene Symbol	NCBI Gene_ID	Genomic loci
1	Hs.655285	ABCF1	23	6p21.33
2	Hs.387567	ACLY	47	17q12-q21
3	Hs.514581	ACTG1	71	17q25
4	Hs.467125	AP2A1	160	19q13.33
5	Hs.388004	AHCY	191	20cen-q13.1
6	Hs.525622	AKT1	207	14q32.32|14q32.32
7	Hs.531682	ALDH3A1	218	17p11.2
8	Hs.511605	ANXA2	302	15q21-q22
9	Hs.517969	APEH	327	3p21.31
10	Hs.514527	BIRC5	332	17q25
11	Hs.502659	RHOC	389	1p13.1
12	Hs.465985	ASNA1	439	19q13.3
13	Hs.707979	ATP1B2	482	17p13.1
14	Hs.406510	ATP5B	506	12q13.13
15	Hs.514870	ATP5F1	515	1p13.2
16	Hs.516966	BCL2L1	598	20q11.21
17	Hs.106880	BYSL	705	6p21.1
18	Hs.555866	C1QBP	708	17p13.3
19	Hs.377010	CAD	790	2p22-p21
20	Hs.515162	CALR	811	19p13.3-p13.2
21	Hs.516155	CAPG	822	2p11.2
22	Hs.58974	CCNA2	890	4q25-q31
23	Hs.23960	CCNB1	891	5q12
24	Hs.244723	CCNE1	898	19q12
25	Hs.82916	CCT6A	908	7p11.2
26	Hs.501497	CD70	970	19p13
27	Hs.405958	CDC6	990	17q21.3
28	Hs.524947	CDC20	991	1p34.1
29	Hs.437705	CDC25A	993	3p21
30	Hs.153752	CDC25B	994	20p13
31	Hs.95577	CDK4	1019	12q14
32	Hs.512599	CDKN2A	1029	9p21
33	Hs.1594	CENPA	1058	2p24-p21
34	Hs.516855	CENPB	1059	20p13
35	Hs.75573	CENPE	1062	4q24-q25
36	Hs.170622	CFL1	1072	11q13

37	Hs.706874	CTSC	1075	11q14.1-q14.3
38	Hs.374378	CKS1B	1163	1q21.2
39	Hs.83758	CKS2	1164	9q22
40	Hs.563509	AP1S1	1174	7q22.1
41	Hs.414565	CLIC1	1192	6p22.1-p21.2
42	Hs.522114	CLTA	1211	9p13
43	Hs.132370	CSTF2	1478	Xq22.1
44	Hs.289271	CYC1	1537	8q24.3
45	Hs.706840	DCN	1634	12q21.33
46	Hs.290758	DDB1	1642	11q12-q13
47	Hs.484782	DFFA	1676	1p36.3-p36.2
48	Hs.4747	DKC1	1736	Xq28
49	Hs.632398	DPH2	1802	1p34
50	Hs.591664	DUSP7	1849	3p21
51	Hs.654393	E2F1	1869	20q11.2
52	Hs.703174	E2F3	1871	6p22
53	Hs.518299	ECT2	1894	3q26.1-q26.2
54	Hs.586423	EEF1A1	1915	6q14.1
55	Hs.520703	EEF1A1	1915	6q14.1
56	Hs.333388	EEF1D	1936	8q24.3
57	Hs.144835	EEF1G	1937	11q12.3
58	Hs.129673	EIF4A1	1973	17p13
59	Hs.707977	EIF4A2	1974	3q28
60	Hs.433750	EIF4G1	1981	3q27-qter
61	Hs.534314	EIF5A	1984	17p13-p12
62	Hs.522823	EMD	2010	Xq28
63	Hs.517145	ENO1	2023	1p36.3-p36.2
64	Hs.2913	EPHB3	2049	3q21-qter
65	Hs.299002	FBL	2091	19q13.1
66	Hs.110849	ESRRA	2101	11q13
67	Hs.434059	ETV4	2118	17q21
68	Hs.444082	EZH2	2146	7q35-q36
69	Hs.302003	FANCE	2178	6p22-p21
70	Hs.335918	FDPS	2224	1q22
71	Hs.409065	FEN1	2237	11q12
72	Hs.524183	FKBP4	2288	12p13.33
73	Hs.239	FOXM1	2305	12p13
74	Hs.524910	FTH1	2495	11q13
75	Hs.461047	G6PD	2539	Xq28
76	Hs.544577	GAPDH	2597	12p13
77	Hs.479728	GAPDH	2597	12p13
78	Hs.708288	GJA1	2697	6q21-q23.2
79	Hs.487341	GNA12	2768	7p22.2
80	Hs.185172	GNB2	2783	7q21.3-q22.1|7q22
81	Hs.523718	SFN	2810	1p36.11
82	Hs.594634	GRINA	2907	8q24.3

83	Hs.466828	GSK3A	2931	19q13.2
84	Hs.445052	MSH6	2956	2p16
85	Hs.75782	GTF3C2	2976	2p23.3
86	Hs.477879	H2AFX	3014	11q23.2-q23.3
87	Hs.171280	HSD17B10	3028	Xp11.2
88	Hs.88556	HDAC1	3065	1p34
89	Hs.707995	HMGB1	3146	13q12
90	Hs.181163	HMGN2	3151	1p36.1
91	Hs.518805	HMGA1	3159	6p21
92	Hs.703764	HMGA1	3159	6p21
93	Hs.569017	SLC29A2	3177	11q13
94	Hs.436181	HOXB7	3217	17q21.3
95	Hs.463350	HOXB9	3219	17q21.3
96	Hs.580427	HPCAL1	3241	2p25.1
97	Hs.20521	PRMT1	3276	19q13.3
98	Hs.707984	IDUA	3425	4p16.3
99	Hs.654400	IMPDH2	3615	3p21.2
100	Hs.522819	IRAK1	3654	Xq28
101	Hs.654848	EIF6	3692	20q12
102	Hs.2722	ITPKA	3706	15q14-q21
103	Hs.3100	KARS	3735	16q23-q24
104	Hs.436912	KIFC1	3833	6p21.3
105	Hs.532793	KPNB1	3837	17q21.32
106	Hs.533782	KRT8	3856	12q13
107	Hs.406013	KRT18	3875	12q13
108	Hs.449909	RPSA	3921	3p22.2
109	Hs.446149	LDHB	3945	12p12.2-p12.1
110	Hs.514535	LGALS3BP	3959	17q25
111	Hs.89497	LMNB1	4001	5q23.3-q31.1
112	Hs.706751	LTBP1	4052	2p22-p21
113	Hs.521903	LY6E	4061	8q24.3
114	Hs.546264	NBR1	4077	17q21.31
115	Hs.417816	MAGEA3	4102	Xq28
116	Hs.441113	MAGEA6	4105	Xq28
117	Hs.169246	MAGEA12	4111	Xq28
118	Hs.23650	MAZ	4150	16p11.2
119	Hs.477481	MCM2	4171	3q21
120	Hs.179565	MCM3	4172	6p12
121	Hs.460184	MCM4	4173	8q11.2
122	Hs.438720	MCM7	4176	7q21.3-q22.1
123	Hs.567303	MDM2	4193	12q14.3-q15
124	Hs.423348	MEN1	4221	11q13
125	Hs.80976	MKI67	4288	10q25-qter
126	Hs.391464	ABCC1	4363	16p13.1
127	Hs.179718	MYBL2	4605	20q13.1
128	Hs.524599	NAP1L1	4673	12q21.2

129	Hs.81469	NUBP1	4682	16p13.13
130	Hs.277677	NDUFA10	4705	2q37.3
131	Hs.148340	RPL10A	4736	6p21.3-p21.2
132	Hs.675285	NFKBIL2	4796	8q24.3
133	Hs.557550	NPM1	4869	5q35
134	Hs.473583	YBX1	4904	1p34
135	Hs.446427	OAZ1	4946	19p13.3
136	Hs.708130	OGN	4969	9q22
137	Hs.17908	ORC1L	4998	1p32
138	Hs.524498	PA2G4	5036	12q13.2
139	Hs.180909	PRDX1	5052	1p34.1
140	Hs.546271	PCBP2	5094	12q13.12-q13.13
141	Hs.255093	PFKL	5211	21q22.3
142	Hs.494691	PFN1	5216	17p13.3
143	Hs.290404	SLC25A3	5250	12q23
144	Hs.371344	PIK3R2	5296	19q13.2-q13.4
145	Hs.534770	PKM2	5315	15q22
146	Hs.154104	PLAGL2	5326	20q11.21
147	Hs.591953	PLCB3	5331	11q13
148	Hs.592049	PLK1	5347	16p12.1
149	Hs.279413	POLD1	5424	19q13.3
150	Hs.306791	POLD2	5425	7p13
151	Hs.356331	PPIA	5478	7p13
152	Hs.183994	PPP1CA	5499	11q13
153	Hs.533308	PPP2R5D	5528	6p21.1
154	Hs.516948	PRCC	5546	1q21.1
155	Hs.465627	MAP2K2	5605	19p13.3
156	Hs.523004	PSAP	5660	10q21-q22
157	Hs.89545	PSMB4	5692	1q21
158	Hs.77060	PSMB6	5694	17p13
159	Hs.250758	PSMC3	5702	11p12-p13
160	Hs.211594	PSMC4	5704	19q13.11-q13.13
161	Hs.518464	PSMD2	5708	3q27.1
162	Hs.459927	PTMA	5757	2q35-q36
163	Hs.458332	PYCR1	5831	17q25.3
164	Hs.368157	PYGB	5834	20p11.2-p11.1
165	Hs.647062	RFC2	5982	7q11.23
166	Hs.461925	RPA1	6117	17p13.3
167	Hs.79411	RPA2	6118	1p35
168	Hs.644628	RPL4	6124	15q22
169	Hs.186350	RPL4	6124	15q22
170	Hs.571841	RPL7	6129	8q21.11
171	Hs.499839	RPL7A	6130	9q34
172	Hs.178551	RPL8	6132	8q24.3
173	Hs.546285	RPLP0	6175	12q24.2
174	Hs.109059	MRPL12	6182	17q25

175	Hs.370895	RPN2	6185	20q12-q13.1
176	Hs.506997	RPS2	6187	16p13.3
177	Hs.498569	RPS2	6187	16p13.3
178	Hs.356366	RPS2	6187	16p13.3
179	Hs.546286	RPS3	6188	11q13.3-q13.5
180	Hs.356572	RPS3A	6189	4q31.2-q31.3
181	Hs.446628	RPS4X	6191	Xq13.1
182	Hs.378103	RPS5	6193	19q13.4
183	Hs.381126	RPS14	6208	5q31-q33
184	Hs.433427	RPS17	6218	15q
185	Hs.226390	RRM2	6241	2p25-p24
186	Hs.436687	SET	6418	9q34
187	Hs.97616	SH3GL1	6455	19p13.3
188	Hs.23348	SKP2	6502	5p13
189	Hs.631582	SLC1A5	6510	19q13.3
190	Hs.187946	SLC20A1	6574	2q11-q14
191	Hs.118400	FSCN1	6624	7p22
192	Hs.83753	SNRPB	6628	20p13
193	Hs.707993	SOX9	6662	17q24.3-q25.1
194	Hs.301540	SPR	6697	2p14-p12
195	Hs.443258	SREBF2	6721	22q13
196	Hs.511425	SRP9	6726	1q42.12
197	Hs.523680	SSRP1	6749	11q12
198	Hs.250822	AURKA	6790	20q13.2-q13.3
199	Hs.481860	TARS	6897	5p13.2
200	Hs.519672	TCOF1	6949	5q32-q33.1
201	Hs.708025	TEGT	7009	12q12-q13
202	Hs.513305	TFAP4	7023	16p13
203	Hs.645227	TGFB1	7040	19q13.2|19q13.1
204	Hs.78769	THOP1	7064	19q13.3
205	Hs.515122	TK1	7083	17q23.2-q25.3
206	Hs.707975	TNFRSF1A	7132	12p13.2
207	Hs.524219	TPI1	7167	12p13
208	Hs.654421	TPM3	7170	1q21.2
209	Hs.12084	TUFM	7284	16p11.2
210	Hs.170107	UQCRFS1	7386	19q12-q13.1
211	Hs.78601	UROD	7389	1p34
212	Hs.520943	EIF4H	7458	7q11.23
213	Hs.707878	ZFP36	7538	19q13.1
214	Hs.662176	ZNF90	7643	19p13.1-p12
215	Hs.234521	MAPKAPK3	7867	3p21.3
216	Hs.695957	DEK	7913	6p22.3
217	Hs.283565	FOSL1	8061	11q13
218	Hs.631661	USP5	8078	12p13
219	Hs.524214	MLF2	8079	12p13
220	Hs.513797	SLC7A5	8140	16q24.3

221	Hs.115232	SF3A2	8175	19p13.3-p13.2
222	Hs.75238	CHAF1B	8208	21q22.13
223	Hs.401509	RBM10	8241	Xp11.23
224	Hs.592082	AXIN1	8312	16p13.3
225	Hs.134999	HIST1H2AM	8336	6p22-p21.3
226	Hs.706783	RAD54L	8438	1p32
227	Hs.405046	CBX4	8535	17q25.3
228	Hs.708050	PPAP2B	8613	1pter-p22.1
229	Hs.371001	EIF3B	8662	7p22.2
230	Hs.492599	EIF3H	8667	8q24.11
231	Hs.86131	FADD	8772	11q13.3
232	Hs.212680	TNFRSF18	8784	1p36.3
233	Hs.412842	CDC123	8872	10p13
234	Hs.591942	ZNF259	8882	11q23.3
235	Hs.118631	TIMELESS	8914	12q12-q13
236	Hs.446522	RPL14	9045	3p22-p21.2
237	Hs.567385	PRC1	9055	15q26.1
238	Hs.54277	FAM50A	9130	Xq28
239	Hs.194698	CCNB2	9133	15q22.2
240	Hs.514590	HGS	9146	17q25
241	Hs.498248	EXO1	9156	1q42-q43
242	Hs.576875	DDX21	9188	10q21
243	Hs.442658	AURKB	9212	17p13.1
244	Hs.151787	EFTUD2	9343	17q21.31
245	Hs.350265	LONP1	9361	19p13.2
246	Hs.31442	RECQL4	9401	8q24.3
247	Hs.5258	MAGED1	9500	Xp11.23
248	Hs.522810	CSAG2	9598	Xq28
249	Hs.118351	UBE3C	9690	7q36.3
250	Hs.81892	KIAA0101	9768	15q22.31
251	Hs.292579	PTDSS1	9791	8q22
252	Hs.5719	NCAPD2	9918	12p13.3
253	Hs.656243	AMMECR1	9949	Xq22.3
254	Hs.654972	RCE1	9986	11q13
255	Hs.497353	MED6	10001	14q24.2
256	Hs.79018	CHAF1A	10036	19p13.3
257	Hs.58992	SMC4	10051	3q26.1
258	Hs.515500	SAE1	10055	19q13.32
259	Hs.654958	ABCF2	10061	7q36
260	Hs.489284	ARPC1B	10095	7q22.1
261	Hs.73625	KIF20A	10112	5q31
262	Hs.522817	BCAP31	10134	Xq28
263	Hs.467408	TRIM28	10155	19q13.4
264	Hs.696342	NUTF2	10204	16q22.1
265	Hs.516160	SF3B4	10262	1q12-q21
266	Hs.522087	OPRS1	10280	9p13.3

267	Hs.20447	PAK4	10298	19q13.2
268	Hs.173162	COX4NB	10328	16q24
269	Hs.707307	PCGF3	10336	4p16.3
270	Hs.524390	TUBA1B	10376	12q13.12
271	Hs.705373	TUBA1B	10376	12q13.12
272	Hs.433615	TUBB2C	10383	9q34
273	Hs.5662	GNB2L1	10399	5q35.3
274	Hs.435850	LYPLA1	10434	8q11.23
275	Hs.655909	TOMM40	10452	19q13
276	Hs.104019	TACC3	10460	4p16.3
277	Hs.386390	TADA3L	10474	3p25.3
278	Hs.470804	UBE2E3	10477	2q32.1
279	Hs.371416	CARM1	10498	19p13.2
280	Hs.532851	RNASEH2A	10535	19p13.13
281	Hs.494604	ANP32B	10541	9q22.32
282	Hs.518774	PAICS	10606	4q12
283	Hs.514033	SPAG5	10615	17q11.2
284	Hs.144936	IGF2BP1	10642	17q21.32
285	Hs.520794	YKT6	10652	7p15.1
286	Hs.348308	C1orf2	10712	1q21
287	Hs.241517	POLQ	10721	3q13.33
288	Hs.263812	NUDC	10726	1p35-p34
289	Hs.101937	SIX2	10736	2p16-p15
290	Hs.435120	KIF1C	10749	17p13.2
291	Hs.655012	GIPC1	10755	19p13.1
292	Hs.469030	MTHFD2	10797	2p13.1
293	Hs.458598	UTP14A	10813	Xq25
294	Hs.705916	CD3EAP	10849	19q13.3
295	Hs.310809	WDR3	10885	1p13-p12
296	Hs.337295	STIP1	10963	11q13
297	Hs.74405	YWHAQ	10971	2p25.1
298	Hs.15591	COPS6	10980	7q22.1
299	Hs.298716	GCN1L1	10985	12q24.2
300	Hs.528834	NUDT21	11051	16q13
301	Hs.93002	UBE2C	11065	20q13.12
302	Hs.397638	WDR5	11091	9q34
303	Hs.708055	HNRPUL1	11100	19q13.2
304	Hs.591363	ZWINT	11130	10q21-q22
305	Hs.160958	CDC37	11140	19p13.2
306	Hs.478150	PDCD10	11235	3q26.1
307	Hs.519993	NRM	11270	6p21.33
308	Hs.567419	MGAT4B	11282	5q35
309	Hs.504620	PHB2	11331	12p13
310	Hs.632296	PDAP1	11333	7q22.1
311	Hs.531563	DOLK	22845	9q34.11
312	Hs.515610	SAPS1	22870	19q13.42

313	Hs.244580	TPX2	22974	20q11.2
314	Hs.586219	KIAA0194	22993	5q33.1
315	Hs.246112	ASCC3L1	23020	2q11.2
316	Hs.155829	TBC1D9B	23061	5q35.3
317	Hs.517670	TTLL12	23170	22q13.31
318	Hs.535901	BOP1	23246	8q24.3
319	Hs.645279	BOP1	23246	8q24.3
320	Hs.583391	NOMO1	23420	16p13.11
321	Hs.513484	QPRT	23475	16p11.2
322	Hs.436329	SCRIB	23513	8q24.3
323	Hs.32018	SNAPIN	23557	1q21.3
324	Hs.49760	ORC6L	23594	16q12
325	Hs.173464	FKBP8	23770	19p12
326	Hs.648326	KIF4A	24137	Xq13.1
327	Hs.31334	PRPF6	24148	20q13.33
328	Hs.50915	KLK5	25818	19q13.3-q13.4
329	Hs.706873	SIN3A	25942	15q24.2
330	Hs.11314	C20orf4	25980	20pter-q12
331	Hs.708014	SERBP1	26135	1p31
332	Hs.532129	TSPAN17	26262	5q35.3
333	Hs.279529	PRELID1	27166	5q35.3
334	Hs.396393	UBE2S	27338	19q13.43
335	Hs.502705	PRPF19	27339	11q12.2
336	Hs.534041	CTA-126B4.3	27341	22q13.2-q13.31
337	Hs.504249	DCPS	28960	11q24.2
338	Hs.18349	MRPL15	29088	8q11.2-q13
339	Hs.5199	UBE2T	29089	1q32.1
340	Hs.157351	OLA1	29789	2q31.1
341	Hs.279877	FTSJ2	29960	7p22
342	Hs.436341	DONSON	29980	21q22.1
343	Hs.3439	STOML2	30968	9p13.1
344	Hs.645463	THAP4	51078	2q37.3
345	Hs.463797	MRTO4	51154	1p36.13
346	Hs.381256	GLTP	51228	12q24.11
347	Hs.584908	MRPL37	51253	1p32.1
348	Hs.475387	CHCHD8	51287	11q13.4
349	Hs.558499	CD320	51293	19p13.3-p13.2
350	Hs.69499	TRIAP1	51499	12q24.31
351	Hs.386189	GTSE1	51512	22q13.2-q13.3
352	Hs.706898	MTP18	51537	22q
353	Hs.528641	SIRT7	51547	17q25
354	Hs.514216	SLC25A39	51629	17q12
355	Hs.655138	WBP11	51729	12p12.3
356	Hs.193326	FGFRL1	53834	4p16
357	Hs.437060	CYCS	54205	7p15.2
358	Hs.592116	FAM64A	54478	17p13.2

359	Hs.481836	MTMR12	54545	5p13.3
360	Hs.654762	DDX56	54606	7p13
361	Hs.485449	GTPBP2	54676	6p21-p12
362	Hs.700127	QPCTL	54814	19q13.32
363	Hs.481526	NSUN2	54888	5p15.31
364	Hs.572318	TIPIN	54962	15q22.31
365	Hs.330663	C12orf48	55010	12q23.2
366	Hs.34045	CDCA4	55038	14q32.33
367	Hs.524571	CDCA8	55143	1p34.3
368	Hs.14559	CEP55	55165	10q23.33
369	Hs.655253	ATAD3A	55210	1p36.33
370	Hs.513126	FANCI	55215	15q26.1
371	Hs.267446	FLJ11184	55319	4q32.3
372	Hs.7570	CNO	55330	4p16.1
373	Hs.532968	HJURP	55355	2q37.1
374	Hs.518265	CDV3	55573	3q22.1
375	Hs.516450	SMPD4	55627	2q21.1
376	Hs.27222	NOLA2	55651	5q35.3
377	Hs.707116	ZNF446	55663	19q13.43
378	Hs.567803	C9orf86	55684	9q34.3
379	Hs.55028	CENPN	55839	16q23.2
380	Hs.656063	BAIAP2L1	55971	7q21.3
381	Hs.472667	CTNNBL1	56259	20q11.23-q12
382	Hs.22678	C10orf2	56652	10q23.3-q24.3
383	Hs.283739	UBQLN4	56893	1q21
384	Hs.250456	DHX33	56919	17p13.2
385	Hs.663740	ARNTL2	56938	12p12.2-p11.2
386	Hs.283734	MRPL47	57129	3q26.33
387	Hs.268488	LRRC47	57470	1p36.32
388	Hs.158381	AARS2	57505	6p21.1
389	Hs.107382	DHX37	57647	12q24.31
390	Hs.444173	PHF12	57649	17q11.2
391	Hs.325838	KIAA1542	57661	11p15.5
392	Hs.465829	ZNF317	57693	NA
393	Hs.516826	TRIB3	57761	20p13-p12.2
394	Hs.654720	KIAA1967	57805	8p22
395	Hs.107003	CCNB1IP1	57820	14q11.2
396	Hs.708201	CXCL16	58191	17p13
397	Hs.477498	EEFSEC	60678	3q21.3
398	Hs.175613	CLSPN	63967	1p34.2
399	Hs.604789	MCCC2	64087	5q12-q13
400	Hs.706966	HERPUD2	64224	7p14.2
401	Hs.527989	NXN	64359	17p13.3
402	Hs.15825	NOM1	64434	7q36.3
403	Hs.436102	ISG20L1	64782	15q26.1
404	Hs.18946	MRPS26	64949	20p13

405	Hs.103832	UPF3B	65109	Xq25-q26
406	Hs.157160	MRPS34	65993	16p13.3
407	Hs.437059	C17orf53	78995	17q21.31
408	Hs.209979	LRFN4	78999	11q13.1
409	Hs.208912	CENPM	79019	22q13.2
410	Hs.596726	TMEM106C	79022	12q13.1
411	Hs.240170	OBFC2B	79035	12q13.2
412	Hs.632191	XTP3TPA	79077	16p11.2
413	Hs.79625	C20orf149	79144	20q13.33
414	Hs.465374	EFHD2	79180	1p36.21
415	Hs.412304	THOC6	79228	16p13.3
416	Hs.661128	KREMEN2	79412	16p13.3
417	Hs.521168	GCC1	79571	7q32.1
418	Hs.418233	MRPL24	79590	1q21-q22
419	Hs.371642	ADIPOR2	79602	12p13.31
420	Hs.59425	NOL9	79707	1p36.31
421	Hs.411865	IPO4	79711	14q12
422	Hs.163754	ZNF669	79862	1q44
423	Hs.496501	CXorf34	79979	Xq22.1
424	Hs.302051	MYO19	80179	17q12
425	Hs.567594	CXXC6	80312	10q21
426	Hs.440899	TTYH3	80727	7p22
427	Hs.534492	PRR7	80758	5q35.3
428	Hs.458390	INTS5	80789	11q12.3
429	Hs.373741	HM13	81502	20q11.21
430	Hs.150837	TXNDC5	81567	6p24.3
431	Hs.495229	URM1	81605	9q34.11
432	Hs.656466	ANP32E	81611	1q21.2
433	Hs.122908	CDT1	81620	16q24.3
434	Hs.444046	NETO2	81831	16q11
435	Hs.374421	CEP78	84131	9q21.2
436	Hs.643537	TRAF7	84231	16p13.3
437	Hs.19673	MAF1	84232	8q24.3
438	Hs.124015	HAGHL	84264	16p13.3
439	Hs.76662	ZDHHC16	84287	10q24.1
440	Hs.462913	MRPL45	84311	17q21.2
441	Hs.548868	THOC3	84321	5q35.2
442	Hs.631633	ELOF1	84337	19p13.2
443	Hs.631506	MCM8	84515	20p12.3
444	Hs.405925	PSRC1	84722	1p13.3
445	Hs.538286	LMNB2	84823	19p13.3
446	Hs.133122	ZDHHC12	84885	9q34.11
447	Hs.102971	ALG10	84920	12p11.1
448	Hs.71574	TIGD5	84948	8q24.3
449	Hs.334713	UBL7	84993	15q24.1
450	Hs.655493	UNK	85451	17q25.1

451	Hs.495240	WDR34	89891	9q34.11
452	Hs.19322	C9orf140	89958	9q34.3
453	Hs.533597	PYGO2	90780	1q21.3
454	Hs.696333	LOC90784	90784	2p11.2
455	Hs.708161	C17orf72	92340	17q24.2
456	Hs.590956	TIMM50	92609	19q13.2
457	Hs.638652	LOC92755	92755	8p12
458	Hs.531614	BTBD14B	112939	19p13.13
459	Hs.101742	RPUSD1	113000	16p13.3
460	Hs.380094	PLCD3	113026	17q21.31
461	Hs.434886	CDCA5	113130	11q12.1
462	Hs.591998	SAAL1	113174	11p15.1
463	Hs.534521	TMEM54	113452	1p35-p34
464	Hs.406840	SLC35A4	113829	5q31.3
465	Hs.201083	MAL2	114569	8q23
466	Hs.347524	C16orf75	116028	16p13.13
467	Hs.705716	TLCD1	116238	17q11.2
468	Hs.368934	C17orf45	125144	17p11.2
469	Hs.708197	PODN	127435	1p32.3
470	Hs.416375	E2F7	144455	12q21.2
471	Hs.135094	LOC146909	146909	17q21.31
472	Hs.632255	RUNDC1	146923	17q21.31
473	Hs.631760	DHRS13	147015	17q11.2
474	Hs.164324	TRIM16L	147166	17p11.2
475	Hs.381225	SPC24	147841	19p13.2
476	Hs.105153	SGOL1	151648	3p24.3
477	Hs.377830	MBOAT1	154141	6p22.3
478	Hs.567739	LOC158345	158345	9p24.1
479	Hs.657472	GPR180	160897	13q32.1
480	Hs.406461	FAM86A	196483	16p13.3
481	Hs.189823	TRIM65	201292	17q25.1
482	Hs.380920	LOC201725	201725	4q32.1
483	Hs.636480	TUBB	203068	6p21.33
484	Hs.706772	TUBB	203068	6p21.33
485	Hs.533655	TYSND1	219743	10q22.1
486	Hs.165607	C11orf82	220042	11q14.1
487	Hs.448226	RPLP0-like	220717	2p22.1
488	Hs.706964	RPLP0-like	220717	2p22.1
489	Hs.88523	C13orf3	221150	13q12.11
490	Hs.72363	C14orf80	283643	14q32.33
491	Hs.633835	LOC284072	284072	17q25.1
492	Hs.378885	PIGW	284098	17q12
493	Hs.381204	METTL2A	339175	17q23.2
494	Hs.103939	C1orf174	339448	1p36.32
495	Hs.459311	ZNF710	374655	15q26.1
496	Hs.512492	RAB15	376267	14q23.3

497	Hs.500561	LOC387703	387703	10q23.33
498	Hs.433791	TMEM46	387914	13q12.13
499	Hs.560655	LOC388524	388524	19p12
500	Hs.647251	LOC390183	390183	11q12.1
501	Hs.355210	LOC400019	400019	12p11.21
502	Hs.620821	VMAC	400673	19p13.3
503	Hs.516290	LOC400963	400963	2p11.2
504	Hs.590999	RPL23AP2	401904	19p13.12
505	Hs.73105	PMS2CL	441194	7p22.1
506	Hs.536395	DUXAP10	503639	14q11.2
507	Hs.523097	EIF5AL3	642592	10q22.3
508	Hs.645558	LOC642784	642784	Xq21.31
509	Hs.646686	LOC642909	642909	5q21.1
510	Hs.507343	LOC643446	643446	4p15.33
511	Hs.647694	LOC643586	643586	1p13.2
512	Hs.531200	HMGN2P6	643872	14q12
513	Hs.614453	LOC643873	643873	Xq23
514	Hs.647368	LOC644035	644035	16q22.1
515	Hs.632240	FAM83G	644815	17p11.2
516	Hs.632598	LOC645018	645018	4q31.21
517	Hs.632537	LOC645691	645691	2q33.1
518	Hs.647919	LOC646612	646612	3q22.3
519	Hs.571791	LOC647150	647150	1q31.2
520	Hs.289232	LOC648232	648232	NA
521	Hs.654748	TMEM183B	653659	3q25.1
522	Hs.444467	LOC654007	654007	7q32.3
523	Hs.646673	LOC728301	728301	5q22.3
524	Hs.535769	LOC728554	728554	5q35.3
525	Hs.594117	HMGN2P3	728632	16p12.1
526	Hs.596312	hCG_1988300	728638	8q12.3
527	Hs.535464	EIF3CL	728689	16p11.2
528	Hs.512314	LOC728891	728891	NA
529	Hs.652172	LOC729859	729859	2p11.2

Appendix H

Protein-coding genes extracted using the EST abundance analysis of the Human Unigene data as potential biomarkers of normal tissues (Table 5).

S.No	Unigene_ID	Gene Symbol	NCBI Gene_ID	Genomic loci
1	Hs.506908	AADAC	13	3q21.3-q25.2
2	Hs.647097	ABP1	26	7q34-q36
3	Hs.445040	ACADM	34	1p31
4	Hs.532492	ACP2	53	11p11.2-p11.11|11p12-p11
5	Hs.498178	ACTN2	88	1q42-q43
6	Hs.591026	ACVRL1	94	12q11-q14
7	Hs.474018	ADARB1	104	21q22.3
8	Hs.481545	ADCY2	108	5p15.3
9	Hs.593293	ADCY5	111	3q13.2-q21
10	Hs.4	ADH1B	125	4q21-q23
11	Hs.654537	ADH1C	126	4q21-q23
12	Hs.197029	ADORA2A	135	22q11.23
13	Hs.249159	ADRA2A	150	10q24-q26
14	Hs.522666	ALAS2	212	Xp11.21
15	Hs.76392	ALDH1A1	216	9q21.13
16	Hs.2533	ALDH9A1	223	1q23.1
17	Hs.111256	ALOX15B	247	17p13.1
18	Hs.102	AMT	275	3p21.2-p21.1
19	Hs.283749	ANG	283	14q11.1-q11.2
20	Hs.620557	ANK2	287	4q25-q27
21	Hs.1239	ANPEP	290	15q25-q26
22	Hs.422986	ANXA4	307	2p13
23	Hs.412117	ANXA6	309	5q32-q34
24	Hs.406238	AOX1	316	2q33
25	Hs.158932	APC	324	5q21-q22
26	Hs.244139	FAS	355	10q24.1
27	Hs.76152	AQP1	358	7p14
28	Hs.455323	AQP7	364	9p13
29	Hs.502876	RHOB	388	2p24
30	Hs.6838	RND3	390	2q23.3
31	Hs.503284	ARRB1	408	11q13
32	Hs.88251	ARSA	410	22q13.31-qter|22q13.33
33	Hs.24976	ART3	419	4p15.1-p14|4p15.1-p14|4p15.1-p14
34	Hs.460	ATF3	467	1q32.3
35	Hs.343522	ATP2B4	493	1q32.1

36	Hs.492280	ATP7B	540	13q14.3
37	Hs.333738	BBS2	583	16q21
38	Hs.410026	BCL2L2	599	14q11.2-q12
39	Hs.821	BGN	633	Xq28
40	Hs.169998	BST1	683	4p15
41	Hs.150557	KLF9	687	9q13
42	Hs.384598	SERPING1	710	11q12-q13.1
43	Hs.8986	C1QB	713	1p36.12
44	Hs.458355	C1S	716	12p13
45	Hs.591148	C3AR1	719	12p13.31
46	Hs.78065	C7	730	5p13
47	Hs.149363	C18orf1	753	18p11.2
48	Hs.155097	CA2	760	8q22
49	Hs.82129	CA3	761	8q13-q22
50	Hs.59093	CACNB2	783	10p12
51	Hs.440961	CAST	831	5q15
52	Hs.458426	CCK	885	3p22-p21.3
53	Hs.292524	CCNH	902	5q13.3-q14
54	Hs.163867	CD14	929	5q22-q32|5q31.1
55	Hs.374990	CD34	947	1q32
56	Hs.120949	CD36	948	7q11.2
57	Hs.633085	CD36	948	7q11.2
58	Hs.278573	CD59	966	11p13
59	Hs.191346	7-Sep	989	7p14.3-p14.1
60	Hs.690198	CDC42	998	1p36.1
61	Hs.76206	CDH5	1003	16q22.1
62	Hs.238990	CDKN1B	1027	12p13.1-p12
63	Hs.106070	CDKN1C	1028	11p15.5
64	Hs.442378	CDO1	1036	5q22-q23
65	Hs.517106	CEBPB	1051	20q13.1
66	Hs.479867	CENPC1	1060	4q12-q13.3
67	Hs.657385	RCBTB2	1102	13q14.3
68	Hs.535891	CHRM2	1129	7q31-q35
69	Hs.334347	CKM	1158	19q13.2-q13.3
70	Hs.628393	CLN3	1201	16p12.1
71	Hs.30213	CLN5	1203	13q21.1-q32
72	Hs.465929	CNN1	1264	19p13.2-p13.1
73	Hs.368921	COL16A1	1307	1p35-p34
74	Hs.4055	KLF6	1316	10p15
75	Hs.432453	MAP3K8	1326	10p11.23
76	Hs.584750	CREB1	1385	2q34
77	Hs.200250	CREM	1390	10p11.21
78	Hs.115617	CRHBP	1393	5q11.2-q13.3
79	Hs.408767	CRYAB	1410	11q22.3-q23.1
80	Hs.592192	CSF2RB	1439	22q13.1
81	Hs.108080	CSRP1	1465	1q32

82	Hs.75262	CTSO	1519	4q31-q32
83	Hs.154654	CYP1B1	1545	2p21
84	Hs.516700	CYP27A1	1593	2q33-qter
85	Hs.481980	DAB2	1601	5p13
86	Hs.279806	DDX5	1655	17q21
87	Hs.594952	DES	1674	2q35
88	Hs.155597	CFD	1675	19p13.3
89	Hs.700572	CYB5R3	1727	22q13.31-qter|22q13.2-q13.31
90	Hs.80552	DPT	1805	1q12-q23
91	Hs.173381	DPYSL2	1808	8p22-p21
92	Hs.522074	TSC22D3	1831	Xq22.3
93	Hs.117060	ECM2	1842	9q22.3
94	Hs.784	EBI2	1880	13q32.3
95	Hs.126667	EDG2	1902	9q31.3
96	Hs.183713	EDNRA	1909	4q31.23
97	Hs.82002	EDNRB	1910	13q22
98	Hs.647061	ELN	2006	7q11.23
99	Hs.76753	ENG	2022	9q33-q34.1
100	Hs.253903	STOM	2040	9q34.1
101	Hs.473819	ERG	2078	21q22.3
102	Hs.155729	ETFDH	2110	4q32-q35
103	Hs.361463	F10	2159	13q34
104	Hs.591133	FBN1	2200	15q21.1
105	Hs.76224	EFEMP1	2202	2p16
106	Hs.58367	GPC4	2239	Xq26.1
107	Hs.7636	FES	2242	15q26.1
108	Hs.567268	FGF7	2252	15q15-q21.1
109	Hs.435369	FHL1	2273	Xq26
110	Hs.144912	FMO2	2327	1q23-q25
111	Hs.103183	FMR1	2332	Xq27.3
112	Hs.25647	FOS	2353	14q24.3
113	Hs.370858	FUCA1	2517	1p34
114	Hs.390567	FYN	2534	6q21
115	Hs.80720	GAB1	2549	4q31.21
116	Hs.75335	GATM	2628	15q21.1
117	Hs.62661	GBP1	2633	1p22.2
118	Hs.656774	GBP3	2635	1p22.2
119	Hs.2171	GDF10	2662	10q11.22
120	Hs.97469	GGTA1	2681	9q33.2-q34.11
121	Hs.437156	GGTLA1	2687	22q11.23
122	Hs.296310	GJA4	2701	1p35.1
123	Hs.83381	GNG11	2791	7q21
124	Hs.524418	GPD1	2819	12q12-q13
125	Hs.122926	NR3C1	2908	5q31.3
126	Hs.75652	GSTM5	2949	1p13.3

127	Hs.449630	HBA1	3039	16p13.3
128	Hs.654744	HBA2	3040	16p13.3
129	Hs.705371	HBG1	3047	11p15.5
130	Hs.302145	HBG2	3048	11p15.5
131	Hs.363396	CFH	3075	1q32
132	Hs.233325	HFE	3077	6p21.3
133	Hs.118651	HHEX	3087	10q23.33
134	Hs.196952	HLF	3131	17q22
135	Hs.524430	NR4A1	3164	12q13
136	Hs.632828	HNRPH2	3188	Xq22
137	Hs.436885	HRC	3270	19q13.3
138	Hs.195040	HSD11B1	3290	1q32-q41
139	Hs.406861	HSD17B4	3295	5q21
140	Hs.520028	HSPA1A	3303	6p21.3
141	Hs.97013	HSPB2	3316	11q22-q23
142	Hs.75619	HYAL1	3373	3p21.3-p21.2
143	Hs.654563	ICAM3	3385	19p13.3-p13.2
144	Hs.312485	CFI	3426	4q25
145	Hs.47338	IFIT3	3437	10q24
146	Hs.520414	IFNGR1	3459	6q23.3
147	Hs.8867	CYR61	3491	1p31-p22
148	Hs.632790	IL3RA	3563	Xp22.3 or Yp11.3
149	Hs.194778	IL8RA	3577	2q35
150	Hs.513022	ISLR	3671	15q23-q24
151	Hs.699822	ISLR	3671	15q23-q24
152	Hs.512235	ITPR2	3709	12p11
153	Hs.121495	KCNE1	3753	21q22.1-q22.2|21q22.12
154	Hs.591606	KCNJ3	3760	2q24.1
155	Hs.182971	KPNA5	3841	6q22.2
156	Hs.444414	AFF3	3899	2q11.2-q12
157	Hs.572535	LAIR1	3903	19q13.4
158	Hs.133421	LIFR	3977	5p13-p12
159	Hs.438236	ABLIM1	3983	10q25
160	Hs.65436	LOXL1	4016	15q24-q25|15q22
161	Hs.661130	LOXL2	4017	8p21.3-p21.2
162	Hs.1116	LTBR	4055	12p13
163	Hs.406475	LUM	4060	12q21.3-q22
164	Hs.656534	SMAD1	4086	4q31
165	Hs.465087	SMAD7	4092	18q21.1
166	Hs.446125	MAK	4117	6p24
167	Hs.102788	MAN1A1	4121	6q22
168	Hs.599039	MCAM	4162	11q23.3
169	Hs.387262	MCF2	4168	Xq27
170	Hs.6790	DNAJB9	4189	7q31|14q24.2-q24.3
171	Hs.699175	MEF2C	4208	5q14
172	Hs.170355	MEOX2	4223	7p22.1-p21.3

173	Hs.61418	MFAP1	4236	15q15-q21
174	Hs.432818	MFAP3	4238	5q32-q33.2
175	Hs.3745	MFGE8	4240	15q25
176	Hs.163924	NR3C2	4306	4q31.1
177	Hs.293970	ALDH6A1	4329	14q24.3
178	Hs.357128	MOCS1	4337	6p21.3
179	Hs.79015	CD200	4345	3q12-q13
180	Hs.396566	MPP3	4356	17q21.31
181	Hs.371225	MSH5	4439	6p21.3
182	Hs.349110	MST1	4485	3p21
183	Hs.471991	MTF1	4520	1p33
184	Hs.498187	MTR	4548	1q43
185	Hs.654589	MYBPC1	4604	12q23.2
186	Hs.82116	MYD88	4615	3p22
187	Hs.440895	MYH3	4621	17p13.1
188	Hs.278432	MYH7	4625	14q12
189	Hs.517939	MYL3	4634	3p21.3-p21.2
190	Hs.463300	MYL4	4635	17q21-qter
191	Hs.556600	MYLK	4638	3q21
192	Hs.436037	MYOC	4653	1q23-q24
193	Hs.444403	PPP1R12B	4660	1q32.1
194	Hs.66180	NAP1L2	4674	Xq13
195	Hs.477693	NCK1	4690	3q21
196	Hs.522615	NDP	4693	Xp11.4
197	Hs.699288	NEDD9	4739	6p25-p24
198	Hs.191911	NFIA	4774	1p31.3-p31.2
199	Hs.77810	NFATC4	4776	14q11.2
200	Hs.656450	NINJ2	4815	12p13
201	Hs.529509	NKTR	4820	3p23-p21
202	Hs.436100	NOTCH4	4855	6p21.3
203	Hs.529006	NPC1	4864	18q11-q12
204	Hs.237028	NPR3	4883	5p14-p13
205	Hs.268788	NRAP	4892	10q24-q26
206	Hs.410969	NTRK3	4916	15q25
207	Hs.563344	NR4A2	4929	2q22-q23
208	Hs.380271	OGG1	4968	3p26.2
209	Hs.31595	CLDN11	5010	3q26.2-q26.3
210	Hs.510334	SERPINA5	5104	14q32.1
211	Hs.680373	PDE1A	5136	2q32.1
212	Hs.530871	PDE1B	5153	12q13
213	Hs.256667	PDK2	5164	17q21.33
214	Hs.8364	PDK4	5166	7q21.3
215	Hs.190977	ENPP2	5168	8q24.1
216	Hs.532768	SERPINF1	5176	17p13.1
217	Hs.36473	PEPD	5184	19q12-q13.2
218	Hs.119316	PET112L	5188	4q27-q28

219	Hs.137415	VIT	5212	2p22-p21
220	Hs.307835	PGM5	5239	9q13
221	Hs.159628	SERPINA4	5267	14q31-q32.1
222	Hs.132225	PIK3R1	5295	5q13.1
223	Hs.99949	PIP	5304	7q34
224	Hs.75813	PKD1	5310	16p13.3
225	Hs.181272	PKD2	5311	4q21-q23
226	Hs.466804	PLA2G2A	5320	1p35
227	Hs.444975	PLAGL1	5325	6q24-q25
228	Hs.80776	PLCD1	5333	3p22-p21.3
229	Hs.442498	FXYD1	5348	19q13.1
230	Hs.170839	PLN	5350	6q22.1
231	Hs.372031	PMP22	5376	17p12-p11.2
232	Hs.292996	PMS2L2	5380	7q11-q22
233	Hs.632368	EXOSC10	5394	1p36.22
234	Hs.702224	PRRX1	5396	1q24
235	Hs.287518	4-Sep	5414	17q22-q23
236	Hs.1897	POMC	5443	2p23.3
237	Hs.530077	PON2	5445	7q21.3
238	Hs.505662	PPP1R1A	5502	12q13.2
239	Hs.303090	PPP1R3C	5507	10q23-q24
240	Hs.467192	PPP2R1A	5518	19q13.33
241	Hs.280604	PPP3R1	5534	2p15
242	Hs.125503	MAPK10	5602	4q22.1-q23
243	Hs.632287	PRKY	5616	Yp11.2
244	Hs.89983	MASP1	5648	3q27-q28
245	Hs.154658	PSD	5662	10q24
246	Hs.458324	PTGIR	5739	19q13.3
247	Hs.154084	PYGM	5837	11q12-q13.2
248	Hs.377992	RABGGTA	5875	14q11.2
249	Hs.695926	RASA1	5921	5q13.3
250	Hs.591111	RASGRF1	5923	15q24
251	Hs.50223	RBP4	5950	10q23-q24
252	Hs.235069	RECQL	5965	12p12
253	Hs.78944	RGS2	5997	1q31
254	Hs.657266	RPL3L	6123	16p13.3
255	Hs.287749	SC5DL	6309	11q23.3
256	Hs.272493	CCL14	6358	17q11.2
257	Hs.57907	CCL21	6366	9p13
258	Hs.531668	CX3CL1	6376	16q13
259	Hs.598247	SDC2	6383	8q22-q23
260	Hs.522891	CXCL12	6387	10q11.1
261	Hs.275775	SEPP1	6414	5q31
262	Hs.309090	SFRS7	6432	2p22.1
263	Hs.591727	SGCD	6444	5q33-q34
264	Hs.380691	SLC2A4	6517	17p13

265	Hs.530003	SLC2A5	6518	1p36.2
266	Hs.1964	SLC5A1	6523	22q13.1|22q12.3
267	Hs.149098	SMTN	6525	22q12.2
268	Hs.468274	SLC8A1	6546	2p23-p22
269	Hs.337696	SLC8A3	6547	14q24.1
270	Hs.518270	SLCO2A1	6578	3q21
271	Hs.152292	SMARCA1	6594	Xq25
272	Hs.2420	SOD3	6649	4p15.3-p15.1
273	Hs.654397	SOS1	6654	2p22-p21
274	Hs.167535	SRP54	6729	14q13.2
275	Hs.117715	ST5	6764	11p15
276	Hs.80642	STAT4	6775	2q32.2-q32.3
277	Hs.437058	STAT5A	6776	17q11.2
278	Hs.25590	STC1	6781	8p21-p11.2
279	Hs.479898	SULT1E1	6783	4q13.1
280	Hs.558403	SUOX	6821	12q13.2
281	Hs.2563	TAC1	6863	7q21-q22
282	Hs.632099	TAGLN	6876	11q23.2
283	Hs.644653	TCF4	6925	18q21.1
284	Hs.124503	ZEB1	6935	10p11.2
285	Hs.446392	DYNLT3	6990	Xp21
286	Hs.89640	TEK	7010	9p21
287	Hs.592317	TGFB3	7043	14q24
288	Hs.482390	TGFBR3	7049	1p33-p32
289	Hs.657724	TLR3	7098	4q35
290	Hs.174312	TLR4	7099	9q32-q33
291	Hs.267632	TMF1	7110	3p21-p12
292	Hs.494595	TMOD1	7111	9q22.3
293	Hs.505337	CLDN5	7122	22q11.21
294	Hs.182421	TNNC2	7125	20q12-q13.11
295	Hs.73454	TNNT3	7140	11p15.5
296	Hs.471381	TNS1	7145	2q35-q36
297	Hs.133892	TPM1	7168	15q22.1
298	Hs.108301	NR2C1	7181	12q22
299	Hs.159003	TRPC6	7225	11q21-q22
300	Hs.486292	TSPYL1	7259	6q22-q23
301	Hs.654592	TTN	7273	2q31
302	Hs.520348	UBC	7316	12q24.3
303	Hs.21899	SLC35A2	7355	Xp11.23-p11.22
304	Hs.516217	UGP2	7360	2p14-p13
305	Hs.409662	COL14A1	7373	8q23
306	Hs.133135	UTRN	7402	6q24
307	Hs.440848	VWF	7450	12p13.3
308	Hs.326420	WNT9B	7484	17q21
309	Hs.78919	XK	7504	Xp21.1
310	Hs.326801	ZNF711	7552	Xq21.1-q21.2

311	Hs.502127	ZNF155	7711	19q13.2-q13.32
312	Hs.157883	ZNF187	7741	6p21.31
313	Hs.8198	ZNF204	7754	6p21.3
314	Hs.406096	ZFAND5	7763	9q13-q21
315	Hs.279567	ZNF225	7768	19q13.2
316	Hs.371823	PRDM2	7799	1p36.21
317	Hs.406050	DNALI1	7802	1p35.1
318	Hs.512842	MFAP5	8076	12p13.1-p12.3
319	Hs.183428	SSPN	8082	12p11.2
320	Hs.187376	IFT88	8100	13q12.1
321	Hs.185910	HDHD1A	8226	Xp22.32
322	Hs.80358	JARID1D	8284	Yq11|Yq11
323	Hs.655309	USP9Y	8287	Yq11.2
324	Hs.591968	FZD4	8322	11q14.2
325	Hs.534371	PIP5K1B	8395	9q13
326	Hs.62886	SPARCL1	8404	4q22.1
327	Hs.484918	CMAH	8418	6p21.32
328	Hs.466766	LTBP4	8425	19q13.1-q13.2
329	Hs.388918	RECK	8434	9p13-p12
330	Hs.694819	TPST2	8459	22q12.1
331	Hs.655143	SORBS2	8470	4q35.1
332	Hs.442180	CILP	8483	15q22
333	Hs.158237	ITGA10	8515	1q21
334	Hs.171311	ITGA8	8516	10p13
335	Hs.558423	CYP4F2	8529	19pter-p13.11
336	Hs.40582	CDC14B	8555	9q22.33
337	Hs.514146	TCAP	8557	17q12
338	Hs.371594	MKNK1	8569	1p33
339	Hs.631562	PLA2G4C	8605	19q13.3
340	Hs.233552	CDC2L5	8621	7p13
341	Hs.198241	AOC3	8639	17q21
342	Hs.76873	HYAL2	8692	3p21.3
343	Hs.534375	B3GALT4	8705	6p21.3
344	Hs.520313	CD164	8763	6q21
345	Hs.511149	SNAP23	8773	15q15.1
346	Hs.546323	SUCLA2	8803	13q12.2-q13.3
347	Hs.509780	WDR22	8816	14q23-q24.1
348	Hs.390736	CFLAR	8837	2q33-q34
349	Hs.654371	STK19	8859	6p21.3
350	Hs.58756	PER2	8864	2q37.3
351	Hs.109590	STBD1	8987	4q24-q25
352	Hs.632460	SELENBP1	8991	1q21-q22
353	Hs.47357	CH25H	9023	10q23
354	Hs.71215	DOK2	9046	8p21.3
355	Hs.524491	PAPSS2	9060	10q23-q24
356	Hs.534377	CLDN10	9071	13q31-q34

357	Hs.23748	LDB2	9079	4p16
358	Hs.448851	USP6	9098	17p13
359	Hs.77854	RGN	9104	Xp11.3
360	Hs.625674	MTMR7	9108	8p22
361	Hs.216226	SYNGR1	9145	22q13.1
362	Hs.443683	MYOM2	9172	8p23.3
363	Hs.66	IL1RL1	9173	2q12
364	Hs.478031	SLC33A1	9197	3q25.31
365	Hs.533986	ZMYM6	9204	1p34.2
366	Hs.612814	CCPG1	9236	15q21.1
367	Hs.654558	CD83	9308	6p23
368	Hs.647113	ACCN3	9311	7q35
369	Hs.376206	KLF4	9314	9q31
370	Hs.632339	TRIP11	9321	14q31-q32
371	Hs.95243	TCEAL1	9338	Xq22.1
372	Hs.66708	VAMP3	9341	1p36.23
373	Hs.696554	ITGBL1	9358	13q33
374	Hs.80485	ADIPOQ	9370	3q27
375	Hs.667720	CYP7B1	9420	8q21.3
376	Hs.656823	RASAL2	9462	1q24
377	Hs.180871	PICK1	9463	22q13.1
378	Hs.643357	ADAMTS1	9510	21q21.2
379	Hs.459940	LITAF	9516	16p13.13
380	Hs.437075	CREB5	9586	7p15.1
381	Hs.371240	AKAP12	9590	6q24-q25
382	Hs.594708	SH3PXD2A	9644	10q24.33
383	Hs.468426	SOCS5	9655	2p21
384	Hs.654651	PDE4DIP	9659	1q12
385	Hs.655934	ZNF432	9668	19q13.33
386	Hs.518138	KIAA0040	9674	1q24-q25
387	Hs.168762	ULK2	9706	17p11.2
388	Hs.478868	KIAA0226	9711	3q29
389	Hs.559459	C6orf32	9750	6p22.3-p21.32
390	Hs.634856	TOX	9760	8q12.1
391	Hs.482660	ZFYVE16	9765	5q14
392	Hs.79276	KIAA0232	9778	4p16.1
393	Hs.31720	HEPH	9843	Xq11-q12
394	Hs.170999	LBA1	9881	3p22.2
395	Hs.524692	NUAK1	9891	12q23.3
396	Hs.5333	KBTBD11	9920	8p23.3
397	Hs.434951	USP15	9958	12q14
398	Hs.282735	NR1H4	9971	12q23.1
399	Hs.527105	HNRPDL	9987	4q13-q21
400	Hs.13967	NAALADL1	10004	11q12
401	Hs.508148	ABI1	10006	10p11.2
402	Hs.556496	TANK	10010	2q24-q31

403	Hs.474705	TOM1	10043	22q13.1
404	Hs.306412	SH2D3C	10044	9q34.11
405	Hs.20136	MAMLD1	10046	Xq28
406	Hs.490745	DNAJB6	10049	7q36.3
407	Hs.593923	DNAJB6	10049	7q36.3
408	Hs.498720	OPTN	10133	10p13
409	Hs.332706	OPTN	10133	10p13
410	Hs.123464	P2RY5	10161	13q14
411	Hs.655248	MBOAT5	10162	12p13
412	Hs.648603	LHFP	10186	13q12
413	Hs.470882	CALCRL	10203	2q32.1
414	Hs.559259	NBR2	10230	17q21
415	Hs.25691	RAMP3	10268	7p13-p12
416	Hs.13351	LANCL1	10314	2q33-q35
417	Hs.471619	NMUR1	10316	2q37.1
418	Hs.50282	RRAGB	10325	Xp11.21
419	Hs.504687	MYL9	10398	20q11.23
420	Hs.54403	CLEC10A	10462	17p13.1
421	Hs.584851	TRIM38	10475	6p21.3
422	Hs.523739	NXF1	10482	11q12-q13
423	Hs.332708	FBLN5	10516	14q32.1
424	Hs.448664	DEAF1	10522	11p15.5
425	Hs.696027	SORBS1	10580	10q23.3-q24.1
426	Hs.522449	POMT1	10585	9q34.1
427	Hs.369574	CDC42EP3	10602	2p21
428	Hs.480311	PDLIM5	10611	4q22
429	Hs.472227	POLR3F	10621	20p11.23
430	Hs.533977	TXNIP	10628	1q21.1
431	Hs.309288	CUGBP2	10659	10p13
432	Hs.59106	CGRRF1	10668	14q22.2
433	Hs.515048	AP4B1	10717	1p13.2
434	Hs.314246	ZNF271	10778	18q12
435	Hs.635221	WASF3	10810	13q12
436	Hs.519694	C5orf4	10826	5q31-q32
437	Hs.920	ARID5A	10865	2q11.2
438	Hs.654480	HCP5	10866	6p21.3
439	Hs.655332	LYVE1	10894	11p15
440	Hs.425144	MTMR11	10903	1q12-q21
441	Hs.486357	SMPDL3A	10924	6q22.31
442	Hs.523774	EHD1	10938	11q13
443	Hs.272168	SERINC3	10955	20q13.1-q13.3
444	Hs.75969	PNRC1	10957	6q15
445	Hs.509343	FERMT2	10979	14q22.2
446	Hs.532824	MAPRE2	10982	18q12.1
447	Hs.268551	RIPK3	11035	14q11.2
448	Hs.659219	C10orf10	11067	10q11.21

449	Hs.470646	RAPGEF4	11069	2q31-q32
450	Hs.521651	STMN2	11075	8q21.13
451	Hs.13852	DNAJB4	11080	1p31.1
452	Hs.666782	PRDM5	11107	4q25-q26
453	Hs.191510	BTN3A1	11119	6p22.1
454	Hs.159028	BTN2A1	11120	6p22.1
455	Hs.43670	KIF3A	11127	5q31
456	Hs.293411	AP4S1	11154	14q12
457	Hs.657271	LDB3	11155	10q22.3-q23.2
458	Hs.43666	PTP4A3	11156	8q24.3
459	Hs.506357	FAM107A	11170	3p21.1
460	Hs.647118	ABCB8	11194	7q36
461	Hs.105105	AKAP11	11215	13q14.11
462	Hs.373857	KLF12	11278	13q22
463	Hs.646614	KLF8	11279	Xp11.21
464	Hs.7884	SLCO2B1	11309	11q13
465	Hs.483909	GPR182	11318	12q13.3
466	Hs.8904	VSIG4	11326	Xq12-q13.3
467	Hs.439199	NLGN4Y	22829	Yq11.221
468	Hs.445030	RHOBTB3	22836	5q15
469	Hs.470457	COBLL1	22837	2q24.3
470	Hs.508010	FNDC3A	22862	13q14.2
471	Hs.188495	WDR37	22884	10p15.3
472	Hs.443109	ARHGEF15	22899	17p13.1
473	Hs.7972	RUFY3	22902	4q13.3
474	Hs.268107	MMRN1	22915	4q22
475	Hs.182982	GOLGA8A	23015	15q11.2
476	Hs.151220	PALLD	23022	4q32.3
477	Hs.98259	SAMD4A	23034	14q22.2
478	Hs.167115	ENDOD1	23052	11q21
479	Hs.155995	CLUAP1	23059	16p13.3
480	Hs.591221	MYCBP2	23077	13q22
481	Hs.584867	CDC2L6	23097	6q21
482	Hs.440414	SPG20	23111	13q13.3
483	Hs.301989	STAB1	23166	3p21.1
484	Hs.464585	ANKRD12	23253	18p11.22
485	Hs.633454	EXOC7	23265	17q25.1
486	Hs.193133	SASH1	23328	6q24.3
487	Hs.655410	DNAJC16	23341	1p36.1
488	Hs.12967	SYNE1	23345	6q25
489	Hs.343334	TENC1	23371	12q13.13
490	Hs.4014	ZDHHC17	23390	12q21.2
491	Hs.369779	SIRT1	23411	10q21.3
492	Hs.580782	MACF1	23499	1p32-p31
493	Hs.409081	OPN3	23596	1q43
494	Hs.271341	RABGAP1	23637	9q33.2

495	Hs.517617	MAFF	23764	22q13.1
496	Hs.533710	FLRT2	23768	14q24-q32
497	Hs.252839	IFIT5	24138	10q23.31
498	Hs.591976	PANX1	24145	11q21
499	Hs.30965	SHC2	25759	19p13.3
500	Hs.519075	LMOD1	25802	1q32
501	Hs.653847	METTL7A	25840	12q13.13
502	Hs.105460	DKFZP564O0823	25849	4q13.3-q21.3
503	Hs.591288	C3orf17	25871	3q13.2
504	Hs.655272	LETMD1	25875	12q13.13
505	Hs.477015	ABI3BP	25890	3q12
506	Hs.55044	DKFZP586H2123	25891	11p13
507	Hs.12844	EGFL6	25975	Xp22
508	Hs.466539	CLIP3	25999	19q13.12
509	Hs.431317	GORASP2	26003	2q31.1-q31.2
510	Hs.654657	SIPA1L1	26037	14q24.2
511	Hs.446017	WSB1	26118	17q11.1
512	Hs.279580	KIAA1279	26128	10q21.3
513	Hs.655108	ZBTB20	26137	3q13.2
514	Hs.693802	ZBTB20	26137	3q13.2
515	Hs.494985	FBXW2	26190	9q34
516	Hs.696160	PITPNC1	26207	17q24.2
517	Hs.643433	FBXL5	26234	4p15.33
518	Hs.76917	FBXO8	26269	4q34.1
519	Hs.400095	HSPB8	26353	12q24.23
520	Hs.78960	LATS2	26524	13q11-q12
521	Hs.491172	NBEA	26960	13q13
522	Hs.352656	GHITM	27069	10q23.1
523	Hs.657015	SDCBP2	27111	20p13
524	Hs.652367	PDE7B	27115	6q23-q24
525	Hs.502612	HSPB7	27129	1p36.23-p34.3
526	Hs.310893	CSDC2	27254	22q13.2-q13.31
527	Hs.306339	SRPX2	27286	Xq21.33-q23
528	Hs.419800	C10orf28	27291	10q24.2
529	Hs.696468	RBMS3	27303	3p24-p23
530	Hs.530053	RABGEF1	27342	7q11.21
531	Hs.523230	POLL	27343	10q23
532	Hs.696057	MAT2B	27430	5q34-q35
533	Hs.279819	MAGEH1	28986	Xp11.21
534	Hs.371856	ZC3H7A	29066	16p13-p12
535	Hs.699226	TRA2A	29896	7p15.3
536	Hs.470369	BAZ2B	29994	2q23-q24
537	Hs.109672	ST6GALNAC6	30815	9q34.11
538	Hs.278694	CD209	30835	19p13
539	Hs.631554	EHD2	30846	19q13.3
540	Hs.432132	G0S2	50486	1q32.2-q41

541	Hs.647182	CUZD1	50624	10q26.13
542	Hs.44685	RNF141	50862	11p15.4
543	Hs.241545	FAM26B	51063	10pter-q26.12
544	Hs.16606	CUTC	51076	10q24.2
545	Hs.579828	FCF1	51077	14q24.3
546	Hs.136309	SH3GLB1	51100	1p22
547	Hs.178170	DUSP13	51207	10q22.2
548	Hs.631730	C1RL	51279	12p13.31
549	Hs.27018	RASL12	51285	15q11.2-q22.33
550	Hs.439474	PCDH12	51294	5q31
551	Hs.274309	ERAF	51327	16p11.2
552	Hs.428147	CDC40	51362	6q21
553	Hs.279815	CSAD	51380	12q13.11-q14.3
554	Hs.647072	PRKAG2	51422	7q36.1
555	Hs.191213	SNX9	51429	6q25.1-q26
556	Hs.163776	UBE2J1	51465	6q15
557	Hs.371563	RAB14	51552	9q32-q34.11
558	Hs.705444	RAB14	51552	9q32-q34.11
559	Hs.696104	TTRAP	51567	6p22.3-p22.1
560	Hs.515890	YPEL5	51646	2p23.1
561	Hs.534458	TPPP3	51673	16q22.1
562	Hs.510327	ASB2	51676	14q31-q32
563	Hs.152913	EMCN	51705	4q23
564	Hs.656794	ZNF44	51710	19p13.2
565	Hs.125300	TRIM34	53840	11p15
566	Hs.547009	SLC37A1	54020	21q22.3
567	Hs.54725	C21orf49	54067	21q22.11
568	Hs.702188	CLIC6	54102	21q22.12
569	Hs.62880	MOV10L1	54456	22q13.33
570	Hs.353022	ETAA1	54465	2p13-p15
571	Hs.431081	USP53	54532	4q26
572	Hs.524121	ROBO4	54538	11q24.2
573	Hs.567513	WDR5B	54554	3q21.1
574	Hs.591901	EPB41L4B	54566	9q31-q32
575	Hs.558570	MXRA8	54587	1p36.33
576	Hs.591145	MANSC1	54682	12p13.2
577	Hs.370522	BTN2A3	54718	6p22.1
578	Hs.270851	OTUD4	54726	4q31.21
579	Hs.269654	MPHOSPH8	54737	13q12.11
580	Hs.441975	XAF1	54739	17p13.2
581	Hs.386684	AHI1	54806	6q23.3
582	Hs.435655	ASPN	54829	9q22
583	Hs.356216	FAM46C	54855	1p12
584	Hs.406223	CCDC49	54883	17q12
585	Hs.30141	MTMR10	54893	15q13.3
586	Hs.439894	CASZ1	54897	1p36.22

587	Hs.371210	C1orf27	54953	1q25
588	Hs.413123	C2orf42	54980	2p14
589	Hs.265018	FAM118A	55007	22q13
590	Hs.591900	STX17	55014	9q31.1
591	Hs.409352	PID1	55022	2q36.3
592	Hs.655317	C19orf60	55049	19p13.11
593	Hs.567523	DET1	55070	15q25.3
594	Hs.504597	TAPBPL	55080	12p13.31
595	Hs.159066	C10orf118	55088	10q25.3
596	Hs.168241	ATG2B	55102	14q32.2
597	Hs.353454	BSDC1	55108	1p35.1
598	Hs.476319	ECHDC2	55268	1p32.3
599	Hs.435933	PHF10	55274	6q27
600	Hs.24545	ZNF444	55311	19q13.42
601	Hs.567532	NIPSNAP3B	55335	9q31.1
602	Hs.647079	GIMAP5	55340	7q36.1
603	Hs.525589	MEG3	55384	14q32
604	Hs.700471	YOD1	55432	1q32.1
605	Hs.654970	IL17RB	55540	3p21.1
606	Hs.525163	ANKRD10	55608	13q34
607	Hs.515169	TRMT1	55621	19p13.13
608	Hs.259605	PIGV	55650	1p36.11
609	Hs.377705	ZNF692	55657	1q44
610	Hs.561954	CDC37L1	55664	9p24.1
611	Hs.469881	LIMS2	55679	2q14.3
612	Hs.584933	ZNF334	55713	20q13.12
613	Hs.696152	RCOR3	55758	1q32.2-q32.3
614	Hs.467210	ZNF83	55769	19q13.3
615	Hs.655166	ChGn	55790	8p21.3
616	Hs.4865	SCN3B	55800	11q23.3
617	Hs.435741	IQWD1	55827	1q24.2
618	Hs.32148	SELS	55829	15q26.3
619	Hs.187635	C20orf19	55857	20pter-q11.23
620	Hs.446438	GPRC5C	55890	17q25
621	Hs.507025	MYNN	55892	3q26.2
622	Hs.699209	ZNF395	55893	8p21.1
623	Hs.193406	C1orf183	55924	1p13.2
624	Hs.529100	LIN37	55957	19q13.1
625	Hs.168799	METTL3	56339	14q11.1
626	Hs.164144	EIF5A2	56648	3q26.2
627	Hs.499960	SAR1A	56681	10q22.1
628	Hs.29106	DUSP22	56940	6p25.3
629	Hs.9315	OLFML3	56944	1p13.2
630	Hs.118241	CABC1	56997	1q42.13
631	Hs.4859	CCNL1	57018	3q25.32
632	Hs.126035	RPGRIP1	57096	14q11

633	Hs.371794	ZNFX1	57169	20q13.13
634	Hs.645966	FAM91A2	57234	1q21.1
635	Hs.655636	KIAA0508	57244	1p36.32
636	Hs.333958	SLC4A10	57282	2q23-q24
637	Hs.656339	RHOJ	57381	14q23.2
638	Hs.525205	NDRG2	57447	14q11.2
639	Hs.21035	GALNTL1	57452	14q24.1
640	Hs.551552	ZNF512B	57473	20q13.33
641	Hs.705876	ZNF608	57507	5q23.2
642	Hs.7946	MTUS1	57509	8p22
643	Hs.657263	CDGAP	57514	3q13.32-q13.33
644	Hs.156352	KIAA1377	57562	11q22.1
645	Hs.211520	KIAA1432	57589	9p24.1
646	Hs.42586	GPAM	57678	10q25.2
647	Hs.438482	WDR19	57728	4p14
648	Hs.270869	ZNF410	57862	14q24.3
649	Hs.511251	SQRDL	58472	15q15
650	Hs.516994	TP53INP2	58476	20q11.22
651	Hs.655066	ZBED5	58486	11p15.3
652	Hs.201034	NTN4	59277	12q22-q23
653	Hs.501624	SIGIRR	59307	11p15.5
654	Hs.407694	ZNF350	59348	19q13.33
655	Hs.269764	BACH2	60468	6q15
656	Hs.42194	SPCS3	60559	4q34.2
657	Hs.187284	PAPPA2	60676	1q23-q25
658	Hs.348342	BRUNOL6	60677	15q24
659	Hs.463035	FKBP10	60681	17q21.2
660	Hs.372309	C10orf84	63877	10q26.11
661	Hs.567562	CIDEC	63924	3p25.3
662	Hs.198158	PBLD	64081	10pter-q25.3
663	Hs.487200	SMOC2	64094	6q27
664	Hs.24719	MOAP1	64112	14q32
665	Hs.591605	TMBIM1	64114	2p24.3-p24.1
666	Hs.199368	TINAGL1	64129	1p35.2
667	Hs.525597	DIO3OS	64150	14q32.31
668	Hs.319171	NFKBIZ	64332	3p12-q12
669	Hs.420830	HIF3A	64344	19q13.32
670	Hs.501289	IKZF5	64376	10q26
671	Hs.511143	ZFP106	64397	15q15.1
672	Hs.380897	AKTIP	64400	16q12.2
673	Hs.592982	TPSB2	64499	16p13.3
674	Hs.159430	FNDC3B	64778	3q26.31
675	Hs.71912	LMF1	64788	16p13.3
676	Hs.112981	DEPDC6	64798	8q24.12
677	Hs.471162	RAPH1	65059	2q33
678	Hs.235390	ZSCAN18	65982	19q13.43

679	Hs.363558	GRAMD3	65983	5q23.2
680	Hs.8035	RASL11B	65997	4q12
681	Hs.632772	SLC2A11	66035	22q11.2
682	Hs.8719	DUSP26	78986	8p12
683	Hs.241576	DERL1	79139	8q24.13
684	Hs.591453	MMEL1	79258	1p36
685	Hs.181173	GLB1L	79411	2q35
686	Hs.211511	TCTN1	79600	12q24.11
687	Hs.655660	RIC3	79608	11p15.4
688	Hs.90250	C4orf31	79625	4q27
689	Hs.459652	TMEM204	79652	16p13.3
690	Hs.655162	ZCCHC6	79670	9q21
691	Hs.458973	ZFHX4	79776	8q21.11
692	Hs.115497	RERGL	79785	12p12.3
693	Hs.524479	MMRN2	79812	10q23.2
694	Hs.665354	ASAM	79827	11q24.1
695	Hs.180402	ZNF671	79891	19q13.43
696	Hs.183390	ZNF613	79898	19q13.33
697	Hs.513296	FLJ14154	79903	16p13.3
698	Hs.694119	GRRP1	79927	1p36.11
699	Hs.189652	C7orf58	79974	7q31.31
700	Hs.522334	SVEP1	79987	9q32
701	Hs.309849	C14orf159	80017	14q32.12
702	Hs.390817	MYO15B	80022	17q25.1
703	Hs.286194	SLC24A6	80024	12q24.13
704	Hs.513343	ATF7IP2	80063	16p13.13
705	Hs.654967	ZNF606	80095	19q13.4
706	Hs.469561	UXS1	80146	2q12.2
707	Hs.288382	CENPT	80152	16q22.1
708	Hs.632527	SLC35F5	80255	2q14.1
709	Hs.167805	EPC1	80314	10p11
710	Hs.127126	CPEB4	80315	5q21
711	Hs.173716	ADAM33	80332	20p13
712	Hs.147434	TRAF3IP3	80342	1q32.3-q41
713	Hs.221597	SLC19A3	80704	2q37
714	Hs.120267	TSGA10	80705	2q11.2
715	Hs.42217	ITFG1	81533	16q12.1
716	Hs.372123	NDEL1	81565	17p13.1
717	Hs.656389	PLA2G12A	81579	4q25
718	Hs.356061	MAP1LC3B	81631	16q24.2
719	Hs.132599	DOCK8	81704	9p24.3
720	Hs.169333	TIGD6	81789	5q33.1
721	Hs.657508	ADAMTS10	81794	19p13.3-p13.2
722	Hs.444229	ARHGAP24	83478	4q21.23-q21.3
723	Hs.513779	CRISPLD2	83716	16q24.1
724	Hs.696346	SYT15	83849	10q11.1

725	Hs.631789	SETDB2	83852	13q14
726	Hs.136901	FSD1L	83856	9q31
727	Hs.655273	TBC1D10A	83874	22q12.1-qter
728	Hs.512773	USHBP1	83878	19p13
729	Hs.132121	SGIP1	84251	1p31.3
730	Hs.59486	HSDL2	84263	9q32
731	Hs.43125	C2orf40	84417	2q12.2
732	Hs.480848	USP38	84640	NA
733	Hs.126706	ACCS	84680	11p11
734	Hs.50334	C9orf24	84688	9p13.3
735	Hs.632528	PLCD4	84812	2q35
736	Hs.129959	IL17RC	84818	3p25.3|3p25.3-p24.1
737	Hs.564188	TMEM25	84866	11q23.3
738	Hs.135254	RSPO3	84870	6q22.33
739	Hs.655177	MFSD2	84879	1p34.2
740	Hs.522520	C9orf37	85026	9q34.3
741	Hs.655626	DIXDC1	85458	NA
742	Hs.376289	ZC3H12C	85463	11q22.3
743	Hs.512805	CCDC65	85478	12q13.12
744	Hs.149540	SEC16B	89866	1q25.2
745	Hs.655123	KLC4	89953	6p21.1
746	Hs.654661	CCDC32	90416	15q15.1
747	Hs.31917	C6orf176	90632	6q27
748	Hs.348390	IL33	90865	9p24.1
749	Hs.173840	ESAM	90952	11q24.2
750	Hs.145061	UBXD5	91544	1p36.11
751	Hs.380906	MYADM	91663	19q13.41
752	Hs.36859	WDR20	91833	14q32.31
753	Hs.49599	TMEM132C	92293	12q24.32
754	Hs.514402	LYK5	92335	17q23.3
755	Hs.651480	LOC92482	92482	10q25.2
756	Hs.89029	ANUBL1	93550	10q11.21
757	Hs.567641	MYOCD	93649	17p11.2
758	Hs.135167	ACRC	93953	Xq13.1
759	Hs.515417	EGLN2	112398	19q13.2
760	Hs.26670	PIK3IP1	113791	22q12.2
761	Hs.410388	LACTB	114294	15q22.1
762	Hs.514071	LRRC37B	114659	NA
763	Hs.593159	MRFAP1L1	114932	4p16.1
764	Hs.308480	PCMTD1	115294	8q11.23
765	Hs.656731	GPR146	115330	7p22.3
766	Hs.348350	DHRS1	115817	14q12
767	Hs.253247	OSR2	116039	8q22.2
768	Hs.516854	HSPA12B	116835	20p13
769	Hs.410316	HRASLS5	117245	11q13.2
770	Hs.17253	IHPK3	117283	6p21.31

771	Hs.529984	TAGAP	117289	6q25.3
772	Hs.162963	ANTXR2	118429	4q21.21
773	Hs.656887	CPXM2	119587	10q26.13
774	Hs.657163	TARSL2	123283	15q26.3
775	Hs.371690	C18orf51	125704	18q22.3
776	Hs.534538	HSPB6	126393	19q13.12
777	Hs.524767	ZNF684	127396	1p34.2
778	Hs.44277	LRRC39	127495	1p21.2
779	Hs.269546	FLJ40298	129852	2p16.2
780	Hs.233398	BBS5	129880	2q31.1
781	Hs.591615	RFTN2	130132	2q33.1
782	Hs.123933	OSR1	130497	2p24.1
783	Hs.534540	ZFAND2B	130617	2q35
784	Hs.40808	TMEM178	130733	2p22.1
785	Hs.230601	DNAJC19	131118	3q26.33
786	Hs.390823	IL17RE	132014	3p25.3
787	Hs.656937	CPEB2	132864	4p15.33
788	Hs.297814	ENPP6	133121	4q35.1
789	Hs.661876	LOC134466	134466	5q33.1
790	Hs.368203	DOCK11	139818	Xq24
791	Hs.591712	ASB5	140458	4q34.2
792	Hs.27453	RAB40A	142684	Xq22.1
793	Hs.210586	C13orf31	144811	13q14.11
794	Hs.144696	C14orf50	145376	14q23.3
795	Hs.658619	TMED6	146456	16q22.1
796	Hs.11782	MGC45438	146556	16p13.3
797	Hs.657197	C18orf18	147525	18p11.31
798	Hs.651111	ZNF565	147929	19q13.12
799	Hs.511848	ZNF569	148266	19q13.12
800	Hs.591445	SAMD13	148418	1p31.1
801	Hs.177744	PM20D1	148811	1q32.1
802	Hs.593721	LOC149448	149448	1q43
803	Hs.116254	C22orf15	150248	22q11.23
804	Hs.368312	FAM109B	150368	22q13.2
805	Hs.516176	SMYD1	150572	2p11.2
806	Hs.655700	ANKAR	150709	2q32.2
807	Hs.493819	C9orf19	152007	9p13-p12
808	Hs.484195	C5orf41	153222	5q35.2
809	Hs.289293	C8orf42	157695	8p23.3
810	Hs.190043	MOSPD2	158747	Xp22.2
811	Hs.42572	ALDH1L2	160428	12q23.3
812	Hs.13854	PPTC7	160760	12q24.11
813	Hs.525307	CLEC14A	161198	14q21.1
814	Hs.522545	ZNF791	163049	19p13.2-p13.13
815	Hs.65256	LGI4	163175	19q13.12|19q13.11
816	Hs.681239	CLEC4F	165530	2p13.3

817	Hs.699317	PRICKLE2	166336	3p14.1
818	Hs.132087	KLHDC6	166348	3q21.3
819	Hs.159006	ZNF800	168850	7q31.33
820	Hs.385493	C10orf128	170371	10q11.22
821	Hs.655519	SYNPO2	171024	4q26
822	Hs.527874	CCDC131	196441	12q21.1
823	Hs.532469	PAOX	196743	10q26.3
824	Hs.443935	TTC21A	199223	3p22.2
825	Hs.301243	TIGD4	201798	4q31.3
826	Hs.133337	RWDD4A	201965	4q35.1
827	Hs.199777	RANBP3L	202151	5p13.2
828	Hs.482625	CMYA5	202333	5q14.1
829	Hs.585069	STK32A	202374	5q32
830	Hs.28780	ZNF449	203523	Xq26.3
831	Hs.42400	USP12	219333	13q12.13
832	Hs.204947	PLAC9	219348	10q22.3
833	Hs.118513	MRGPRF	219928	11q13.2
834	Hs.668747	FLJ32682	220081	13q14.12
835	Hs.607594	FAM13C1	220965	10q21.1
836	Hs.147440	ZNF485	220992	10q11.21
837	Hs.427449	LOC221091	221091	11q12.3
838	Hs.412103	EFHA1	221154	13q12.11
839	Hs.25391	PI16	221476	6p21.2
840	Hs.519904	RBM24	221662	6p22.3
841	Hs.596587	C7orf38	221786	7q22.1
842	Hs.587427	HOXA11S	221883	7p15.2
843	Hs.368944	JAZF1	221895	7p15.2-p15.1
844	Hs.200100	C7orf41	222166	7p15.1
845	Hs.522863	CYorf15A	246126	Yq11.222
846	Hs.407926	RICTOR	253260	5p13.1
847	Hs.339024	MSRB3	253827	12q14.3
848	Hs.693749	ZDHHC20	253832	13q12.11
849	Hs.346575	C19orf26	255057	19p13.3
850	Hs.435515	LOC255167	255167	5p15.31
851	Hs.163451	LOC255275	255275	17q25.3
852	Hs.22575	BCL6B	255877	17p13.1
853	Hs.591401	KANK3	256949	19p13.2
854	Hs.503500	OLFML1	283298	11p15.4
855	Hs.560343	LOC283666	283666	15q21.3
856	Hs.569669	LOC283901	283901	16p12.1
857	Hs.664267	FLJ36208	283948	16p13.3
858	Hs.437191	PTRF	284119	17q21.31
859	Hs.400688	IZUMO1	284359	19q13.33
860	Hs.303669	SLC25A42	284439	19p13.11
861	Hs.567816	FAM126B	285172	2q33.1
862	Hs.518059	C3orf64	285203	3p14.1

863	Hs.559386	LOC285286	285286	3p14.2
864	Hs.476399	CCDC66	285331	3p14.3
865	Hs.449206	LOC285359	285359	3q12.3
866	Hs.588682	SUMF1	285362	3p26.2
867	Hs.399980	LOC285550	285550	4p15.33-p15.32
868	Hs.480371	LOC285556	285556	4q23
869	Hs.403594	EFHA2	286097	8p22
870	Hs.496530	MGC39900	286527	Xq22.2
871	Hs.146059	TUSC5	286753	17p13.3
872	Hs.700799	LOC339290	339290	18p11.31
873	Hs.471067	LOC339483	339483	1p35.1
874	Hs.146730	KY	339855	3q22.2
875	Hs.379754	TMEM173	340061	5q31.2
876	Hs.444834	RSPO2	340419	8q23.1
877	Hs.21249	ZC3H12B	340554	Xq11.1
878	Hs.369380	MGC40069	348035	14q11.2
879	Hs.208673	NMNAT3	349565	3q23
880	Hs.595458	MAST4	375449	5q12.3
881	Hs.704486	LOC387647	387647	10p11.23
882	Hs.32478	FIBIN	387758	11p14.2
883	Hs.131035	GLTPD2	388323	17p13.2
884	Hs.657260	RP13-401N8.2	388358	20p11.21-p11.1
885	Hs.204449	ZNF470	388566	19q13.43
886	Hs.435013	VGLL3	389136	3p12.1
887	Hs.658041	MGC21881	389741	9q21.11
888	Hs.497573	FLJ45244	400242	14q32.13
889	Hs.153827	C14orf180	400258	14q32.33
890	Hs.187134	LOC400464	400464	15q26.3
891	Hs.641441	LOC400604	400604	17q21.33
892	Hs.61508	LOC400657	400657	18q22.3
893	Hs.668085	C1orf220	400798	1q25.2
894	Hs.173705	LOC401152	401152	4q26
895	Hs.131064	KLHL31	401265	6p12.1
896	Hs.561708	LOC401320	401320	7p15.1
897	Hs.461247	MRC1L1	414308	10p12.33
898	Hs.530380	FAM116B	414918	22q13.33
899	Hs.536319	IQSEC3	440073	12p13.33
900	Hs.512963	ALG11	440138	13q14.2
901	Hs.449880	LOC440434	440434	17q12
902	Hs.641142	FLJ46875	440918	2q24.1
903	Hs.507676	FLJ12993	441027	4q21.22
904	Hs.510098	C6orf217	441171	6q23.3
905	Hs.559067	ARMETL1	441549	10p13
906	Hs.647105	GIMAP6	474344	NA
907	Hs.596537	C2orf64	493753	2q11.2
908	Hs.661883	PGA4	643847	11q12.2

909	Hs.380698	WIPF3	644150	7p15.1
910	Hs.58690	LOC644192	644192	15q26.2
911	Hs.575741	LOC644431	644431	NA
912	Hs.683930	LOC644554	644554	19q13.12-q13.13
913	Hs.693822	LOC645431	645431	14q23.3
914	Hs.463652	LOC645638	645638	17q23.1
915	Hs.444950	LOC652968	652968	22q12
916	Hs.632434	LOC653513	653513	1q21.1
917	Hs.659982	LOC654841	654841	2q36-q37
918	Hs.655534	LOC728190	728190	10q23.2
919	Hs.535639	D2HGDH	728294	2q37.3
920	Hs.662127	RP11-592B15.4	728407	10q11.23
921	Hs.559428	SERF1B	728492	5q13.2
922	Hs.559827	LOC729096	729096	10q22.2
923	Hs.557608	LOC729178	729178	6q24.3
924	Hs.591387	KIAA1881	729359	19p13.3
925	Hs.371576	FAM18A	780776	16p13.13
926	Hs.1581	DDTL	100037417	22q11.23

REFERENCES

REFERENCES

ACS (2009). Cancer Facts & Figures 2009. Cancer Facts & Figures. Atlanta, American Cancer Society.

Adams, M.D. et al., (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project." Science 252: 1651-6.

Ahn, A. C., M. Tewari, et al. (2006). "The limits of reductionism in medicine: could systems biology offer an alternative?" PLoS Med 3(6): e208.

Alles, M. C., M. Gardiner-Garden, et al. (2009). "Meta-analysis and gene set enrichment relative to er status reveal elevated activity of MYC and E2F in the "basal" breast cancer subgroup." PLoS One 4(3): e4710.

Ambros, V., (2001). "microRNAs: tiny regulators with great potential." Cell 107(7): 823-826.

An, W. (2007). "Histone acetylation and methylation: combinatorial players for transcriptional regulation." Subcell Biochem 41: 351-369.

Anders, C. K., D. S. Hsu, et al. (2008). "Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression." J Clin Oncol 26(20): 3324-3330.

Aoki-Kinoshita, K. F. and M. Kanehisa (2007). "Gene annotation and pathway mapping in KEGG." Methods Mol Biol 396: 71-91.

Artandi, S. E. and L. D. Attardi (2005). "Pathways connecting telomeres and p53 in senescence, apoptosis, and cancer." Biochem Biophys Res Commun 331(3): 881-90.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet 25(1): 25-29.

Axelsen, J. B., J. Lotem, et al. (2007). "Genes overexpressed in different human solid cancers exhibit different tissue-specific expression profiles." Proc Natl Acad Sci U S A 104(32): 13122-13127.

Barrett, J. C., M. Oshimura, et al. (1986). "Role of oncogenes and tumor suppressor genes in a multistep model of carcinogenesis." Symp Fundam Cancer Res 39: 45-56.

Barrett, T. and R. Edgar (2006). "Gene expression omnibus: microarray data storage, submission, retrieval, and analysis." Methods Enzymol 411: 352-369.

Barrett, T., D. B. Troup, et al. (2007). "NCBI GEO: mining tens of millions of expression profiles--database and tools update." Nucleic Acids Res 35(Database issue): D760-765.

Bar-Yam, Y., Harmon, D. & de Bivort, B., (2009). "Systems biology. Attractors and democratic dynamics." Science 323(5917): 1016-1017.

Baur, J. A., J. W. Shay, et al. (2004). "Spontaneous reactivation of a silent telomeric transgene in a human cell line." Chromosoma 112(5): 240-6.

Baur, J. A., Y. Zou, et al. (2001). "Telomere position effect in human cells." Science 292(5524): 2075-7.

Baylin, S. B., M. Esteller, et al. (2001). "Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer." Hum Mol Genet 10(7): 687-692.

Bayne, R. A., D. Broccoli, et al. (1994). "Sandwiching of a gene within 12 kb of a functional telomere and alpha satellite does not result in silencing." Hum Mol Genet 3(4): 539-46.

Ben-Dor, A. et al., (2000). "Tissue classification with gene expression profiles." Journal of Computational Biology 7(3-4): 559-583.

Benjamini, Y. and Y. Hochberg (2000). "On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics." 25(1): 60-83.

Benson, D. A., I. Karsch-Mizrachi, et al. (2007). "GenBank." Nucleic Acids Res 35(Database issue): D21-5.

Biocarta. "Biocarta molecular pathway repository." from http://www.biocarta.com/genes/index.asp.

Bisoffi, M., C. M. Heaphy, et al. (2006). "Telomeres: prognostic markers for solid tumors." Int J Cancer a) all genes changing their expression, b) upregulated genes only and c) downregulated genes only. (10): 2255-60.

Bliss, M. (1982). The Discovery of Insulin. Chicago, IL, University of Chicago Press.

Buness, A., R. Kuner, et al. (2007). "Identification of aberrant chromosomal regions from gene expression microarray studies applied to human breast cancer." Bioinformatics 23(17): 2273-2280.

Butte, A. J., V. J. Dzau, et al. (2001). "Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues"." Physiol Genomics 7(2): 95-96.

Calcagnile, O. and D. Gisselsson (2007). "Telomere dysfunction and telomerase activation in cancer--a pathological paradox?" Cytogenet Genome Res 118(2-4): 270-6.

Campagna, D., L. Cope, et al. (2008). "Gene expression profiles associated with advanced pancreatic cancer." Int J Clin Exp Pathol 1(1): 32-43.

Casey, T. et al., (2009). "Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer." Breast Cancer Research and Treatment 114(1): 47-62.

Cazzaniga, M. et al., (2009). "Biomarkers for risk assessment and prevention of breast cancer." Current Cancer Drug Targets 9(4): 482-499.

Cech, T. R. (2004). "Beginning to understand the end of the chromosome." Cell 116(2): 273-9.

Chandran, U.R. et al., (2007). "Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process." BMC Cancer 7: 64.

Chang, H.H. et al., (2008). "Transcriptome-wide noise controls lineage choice in mammalian progenitor cells." Nature 453(7194):544-547.

Chatterjee, A., E. Mambo, et al. (2006). "Mitochondrial DNA mutations in human cancer." Oncogene 25(34): 4663-4674.

Cimino, J. J., T. F. Hayamizu, et al. (2009). "The caBIG terminology review process." J Biomed Inform 42(3): 571-580.

Clark, S. J. (2007). "Action at a distance: epigenetic silencing of large chromosomal regions in carcinogenesis." Hum Mol Genet 16 Spec No 1: R88-95.

Conover, W. J. (1971). "Practical nonparametric statistics." New York, John Wiley & Sons.

Conrad, R., Barrier, M. & Ford, L.P., (2006). "Role of miRNA and miRNA processing factors in development and disease." Birth Defects Research. 78(2): 107-117.

Croce, C. M. (2008). "Oncogenes and cancer." N Engl J Med 358(5): 502-511.

Curwen, V., E. Eyras, et al. (2004). "The Ensembl automatic gene annotation system." Genome Res 14(5): 942-950.

Dave, S. S., G. Wright, et al. (2004). "Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells." N Engl J Med 351(21): 2159-2169.

de Lange, T. (2005). "Shelterin: the protein complex that shapes and safeguards human telomeres." Genes Dev 19(18): 2100-10.

Dennis, G., Jr., B. T. Sherman, et al. (2003). "DAVID: Database for Annotation, Visualization, and Integrated Discovery." Genome Biol 4(5): P3.

Dodd, L.E. et al., (2006). "Genes involved in DNA repair and nitrosamine metabolism and those located on chromosome 14q32 are dysregulated in nasopharyngeal carcinoma." Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology 15(11): 2216-2225.

Driouch, K., T. Landemaine, et al. (2007). "Gene arrays for diagnosis, prognosis and treatment of breast cancer metastasis." Clin Exp Metastasis 24(8): 575-585.

Dyrskjøt, L. et al., (2004). "Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification." Cancer Research 64(11): 4040-4048.

Efferth, T. (2005). "Microarray-based prediction of cytotoxicity of tumor cells to cantharidin." Oncol Rep 13(3): 459-463.

Efroni, S. et al., (2008). "Global transcription in pluripotent embryonic stem cells." Cell Stem Cell 2(5): 437-447.

Ehrlich, M., M. A. Gama-Sosa, et al. (1982). "Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells." Nucleic Acids Res 10(8): 2709-2721.

Ellis, M. J. (2003). "Breast cancer gene expression analysis--the case for dynamic profiling." Adv Exp Med Biol 532: 223-234.

Esteller, M. (2006). "The necessity of a human epigenome project." Carcinogenesis 27(6): 1121-1125.

Finkel, T., M. Serrano, et al. (2007). "The common biology of cancer and ageing." Nature 448(7155): 767-74.

Frigola, J., J. Song, et al. (2006). "Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band." Nat Genet 38(5): 540-549.

Furge, K.A. et al., (2004). "Robust classification of renal cell carcinoma based on gene expression data and predicted cytogenetic profiles." Cancer Research 64(12): 4117-4121.

Ganter, B. and C. N. Giroux (2008). "Emerging applications of network and pathway analysis in drug discovery and development." Curr Opin Drug Discov Devel 11(1): 86-94.

Gautier, L., L. Cope, et al. (2004). "affy--analysis of Affymetrix GeneChip data at the probe level." Bioinformatics 20(3): 307-15.

Gisselsson, D. and M. Hoglund (2005). "Connecting mitotic instability and chromosome aberrations in cancer--can telomeres bridge the gap?" Semin Cancer Biol 15(1): 13-23.

Golub, T. R., D. K. Slonim, et al. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." Science 286(5439): 531-537.

Gronbaek, K., C. Hother, et al. (2007). "Epigenetic changes in cancer." Apmis 115(10): 1039-1059.

Grosso, A. R., A. Q. Gomes, et al. (2008). "Tissue-specific splicing factor gene expression signatures." Nucleic Acids Res 36(15): 4823-4832.

Gregory Alvord, W. et al., (2007). "A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R." Brief Bioinform 8: 415-31.

Gumz, M.L. et al., (2007). "Secreted frizzled-related protein 1 loss contributes to tumor phenotype of clear cell renal cell carcinoma." Clinical Cancer Research 13(16): 4740-4749.

Guo, Y. et al., (2006). "Towards a holistic, yet gene-centered analysis of gene expression profiles: a case study of human lung cancers." Journal of Biomedicine & Biotechnology 2006(5): 69141.

Hahn, W. C., C. M. Counter, et al. (1999). "Creation of human tumour cells with defined genetic elements." Nature 400(6743): 464-8.

Hahn, W. C., S. A. Stewart, et al. (1999). "Inhibition of telomerase limits the growth of human cancer cells." Nat Med 5(10): 1164-70.

Hanahan, D. and R. A. Weinberg (2000). "The hallmarks of cancer." Cell 100(1): 57-70.

Hansel, D. E., A. K. Meeker, et al. (2006). "Telomere length variation in biliary tract metaplasia, dysplasia, and carcinoma." Mod Pathol 19(6): 772-9.

Hayden, D., Lazar, P. & Schoenfeld, D., (2009). "Assessing statistical significance in microarray experiments using the distance between microarrays." PloS One 4(6): e5838.

Hayflick, L. (1965). "The Limited in Vitro Lifetime of Human Diploid Cell Strains." Exp Cell Res 37: 614-36.

He, H. et al., (2005). "The role of microRNA genes in papillary thyroid carcinoma." Proceedings of the National Academy of Sciences of the United States of America 102(52): 19075-19080.

Henrickson, S. E., E. M. Hartmann, et al. (2007). "Gene expression profiling in malignant lymphomas." Adv Exp Med Biol 593: 134-146.

Hitchins, M. P., V. A. Lin, et al. (2007). "Epigenetic inactivation of a cluster of genes flanking MLH1 in microsatellite-unstable colorectal cancer." Cancer Res 67(19): 9107-9116.

Hiyama, E. and K. Hiyama (2002). "Clinical utility of telomerase in cancer." Oncogene 21(4): 643-9.

Hormaeche, I. and J. D. Licht (2007). "Chromatin modulation by oncogenic transcription factors: new complexity, new therapeutic targets." Cancer Cell 11(6): 475-478.

Hornberg, J. J., F. J. Bruggeman, et al. (2006). "Cancer: a Systems Biology disease." Biosystems 83(2-3): 81-90.

Hsiao, A., D. S. Worrall, et al. (2004). "Variance-modeled posterior inference of microarray data: detecting gene-expression changes in 3T3-L1 adipocytes." Bioinformatics 20(17): 3108-3127.

Huang, A.C. et al., (2009). "Using cell fate attractors to uncover transcriptional regulation of HL60 neutrophil differentiation." BMC Systems Biology 3: 20.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature Protocols 4(1), 44-57.

Huang, S. & Ingber, D.E., (2006). "A non-genetic basis for cancer progression and metastasis: self-organizing attractors in cell regulatory networks." Breast Disease 26: 27-54.

Huang, S. et al., (2007). "Bifurcation dynamics in lineage-commitment in bipotent progenitor cells." Developmental Biology 305(2): 695-713.

Huang, S., (2009). "Reprogramming cell fates: reconciling rarity with robustness." BioEssays 31(5): 546-560.

Huang, S., Ernberg, I. & Kauffman, S., (2009). "Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective." Seminars in Cell & Developmental Biology 20(7): 869-876.

Huang da, W., B. T. Sherman, et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc 4(1): 44-57.

Hug, N. and J. Lingner (2006). "Telomere length homeostasis." Chromosoma 115(6): 413-425.

Hwa, H. et al., (2008). "Prediction of breast cancer and lymph node metastatic status with tumour markers using logistic regression models." Journal of Evaluation in Clinical Practice 14(2): 275-280.

Ikediobi, O. N., H. Davies, et al. (2006). "Mutation analysis of 24 known cancer genes in the NCI-60 cell line set." Mol Cancer Ther 5(11): 2606-2612.

International Human Genome Sequencing Consortium, (2004). "Finishing the euchromatic sequence of the human genome." Nature 431(7011): 931-945.

Irizarry, R. A., B. M. Bolstad, et al. (2003). "Summaries of Affymetrix GeneChip probe level data." Nucleic Acids Res 31(4): e15.

Jiang, Z., B. A. Woda, et al. (2004). "Discovery and clinical application of a novel prostate cancer marker: alpha-methylacyl CoA racemase (P504S)." Am J Clin Pathol 122(2): 275-289.

Jongeneel, C. V., M. Delorenzi, et al. (2005). "An atlas of human gene expression from massively parallel signature sequencing (MPSS)." Genome Res 15(7): 1007-1014.

Kakazu, K. K., L. W. Cheung, et al. (2004). "The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research." Hawaii Med J 63(9): 273-275.

Kanehisa, M. (2002). "The KEGG database." Novartis Found Symp 247: 91-101; discussion 101-103, 119-128, 244-152.

Kanehisa, M., M. Araki, et al. (2008). "KEGG for linking genomes to life and the environment." Nucleic Acids Res 36(Database issue): D480-484.

Kapranov, P. et al., (2005). "Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays." Genome Research 15(7): 987-997.

Kauffman, S., (1971). "Differentiation of malignant to benign cells." Journal of Theoretical Biology 31(3): 429-451.

Kawashima, S., T. Katayama, et al. (2003). "KEGG API: A Web Service Using SOAP/WSDL to Access the KEGG System." Genome Informatics 14: 673--674.

Kennedy, R. D. and A. D. D'Andrea (2006). "DNA repair pathways in clinical practice: lessons from pediatric cancer susceptibility syndromes." J Clin Oncol 24(23): 3799-3808.

Khatri, P., S. Sellamuthu, et al. (2005). "Recent additions and improvements to the Onto-Tools." Nucleic Acids Res 33(Web Server issue): W762-765.

Kimchi, E.T. et al., (2005). "Progression of Barrett's metaplasia to adenocarcinoma is associated with the suppression of the transcriptional programs of epidermal differentiation." Cancer Research 65(8): 3146-3154.

Kimmins, S., N. Kotaja, et al. (2004). "Testis-specific transcription mechanisms promoting male germ-cell differentiation." Reproduction 128(1): 5-12.

Kitano, H. (2002). "Systems biology: a brief overview." Science 295(5560): 1662-1664.

Klukas, C. and F. Schreiber (2007). "Dynamic exploration and editing of KEGG pathway diagrams." Bioinformatics 23(3): 344-350.

Koering, C. E., A. Pollice, et al. (2002). "Human telomeric position effect is determined by chromosomal context and telomeric chromatin integrity." EMBO Rep 3(11): 1055-61.

Konopka, J. B., S. M. Watanabe, et al. (1985). "Cell lines and clinical isolates derived from Ph1-positive chronic myelogenous leukemia patients express c-abl proteins with a common structural alteration." Proc Natl Acad Sci U S A 82(6): 1810-1814.

Kopnin, B. P. (2000). "Targets of oncogenes and tumor suppressors: key for understanding basic mechanisms of carcinogenesis." Biochemistry (Mosc) 65(1): 2-27.

Kuriakose, M.A. et al., (2004). "Selection and validation of differentially expressed genes in head and neck cancer." Cellular and Molecular Life Sciences 61(11): 1372-1383.

Lagarde, S. M., P. E. Ver Loren van Themaat, et al. (2008). "Analysis of gene expression identifies differentially expressed genes and pathways associated with lymphatic dissemination in patients with adenocarcinoma of the esophagus." Ann Surg Oncol 15(12): 3459-3470.

Landry, J. R., D. L. Mager, et al. (2003). "Complex controls: the role of alternative promoters in mammalian genomes." Trends Genet 19(11): 640-648.

Laubenbacher, R., V. Hower, et al. (2009). "A systems biology view of cancer." Biochim Biophys Acta.

Lee, J. S. and S. S. Thorgeirsson (2004). "Genome-scale profiling of gene expression in hepatocellular carcinoma: classification, survival prediction, and identification of therapeutic targets." Gastroenterology 127(5 Suppl 1): S51-55.

Lenburg, M.E. et al., (2003). "Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data." BMC Cancer 3: 31.

Liang, J.J. et al., (2009). "Diagnostic and prognostic biomarkers in pancreatic carcinoma." International Journal of Clinical and Experimental Pathology 2(1): 1-10.

Liu, E. T. (2004). "Expression genomics and cancer biology." Pharmacogenomics 5(8): 1117-1128.

Logsdon, C.D. et al., (2003). "Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer." Cancer Research 63(10): 2649-2657.

Ludwig, J.A. & Weinstein, J.N., (2005). "Biomarkers in cancer staging, prognosis and treatment selection." Nature Reviews. Cancer 5(11): 845-856.

Maida, Y., S. Kyo, et al. (2006). "Distinct telomere length regulation in premalignant cervical and endometrial lesions: implications for the roles of telomeres in uterine carcinogenesis." J Pathol 210(2): 214-23.

Mallardo, M., Poltronieri, P. & D'Urso, O.F., (2008). "Non-protein coding RNA biomarkers and differential expression in cancers: a review." Journal of Experimental & Clinical Cancer Research 27: 19.

Maglott, D., J. Ostell, et al. (2007). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Res 35(Database issue): D26-31.

Magurran, A.E., (1988). "Ecological diversity and its measurement" Taylor & Francis.

Mann, H. B. and D. R. Whitney (1947). "On a test of whether one of two random variables is stochastically larger than the other." Annals of Mathematical Statistics 18: 50-60.

Martínez, O. & Reyes-Valdés, M.H., (2008). "Defining diversity, specialization, and gene specificity in transcriptomes through information theory." Proceedings of the National Academy of Sciences of the United States of America 105(28): 9709-9714.

Mathew, J. P., B. S. Taylor, et al. (2007). "From bytes to bedside: data integration and computational biology for translational cancer research." PLoS Comput Biol 3(2): e12.

Matlin, A. J., F. Clark, et al. (2005). "Understanding alternative splicing: towards a cellular code." Nat Rev Mol Cell Biol 6(5): 386-398.

Mattick, J.S., (2003). "Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms." BioEssays 25(10): 930-939.

McShea, A., M. W. Marlatt, et al. (2006). "The application of microarray technology to neuropathology: cutting edge tool with clinical diagnostics potential or too much information?" J Neuropathol Exp Neurol 65(11): 1031-1039.

Meeker, A. K. (2006). "Telomeres and telomerase in prostatic intraepithelial neoplasia and prostate cancer biology." Urol Oncol 24(2): 122-30.

Meeker, A. K., J. L. Hicks, et al. (2004). "Telomere length abnormalities occur early in the initiation of epithelial carcinogenesis." Clin Cancer Res 10(10): 3317-26.

Merlo, L.M.F. et al., (2006). "Cancer as an evolutionary and ecological process." Nature Reviews. Cancer 6(12): 924-935.

Modrek, B. and C. Lee (2002). "A genomic view of alternative splicing." Nat Genet 30(1): 13-19.

NCBI, (2002). "NCBI handbook." Available at: http://www.ncbi.nlm.nih.gov/entrez/qery.fcgi?db=Books.

Nicolson, G. L. (1991). "Gene expression, cellular diversification and tumor progression to the metastatic phenotype." Bioessays 13(7): 337-342.

Ning, Y., J. F. Xu, et al. (2003). "Telomere length and the expression of natural telomeric genes in human fibroblasts." Hum Mol Genet 12(11): 1329-36.

Oluwadara, O. & Chiappelli, F., (2009). "Biomarkers for early detection of high risk cancers: From gliomas to nasopharyngeal carcinoma." Bioinformation 3(8): 332-339.

Ottaviani, A., E. Gilson, et al. (2008). "Telomeric position effect: from the yeast paradigm to human pathologies?" Biochimie 90(1): 93-107.

Paquet, A. and Y. Yang (2007). "Getting started with goTools package." Unpublished Manuscript.

Parkinson, H., M. Kapushesky, et al. (2007). "ArrayExpress--a public database of microarray experiments and gene expression profiles." Nucleic Acids Res 35(Database issue): D747-750.

Pedram, M., C. N. Sprung, et al. (2006). "Telomere position effect and silencing of transgenes near telomeres in the mouse." Mol Cell Biol 26(5): 1865-78.

Pegtel, D. M., A. Subramanian, et al. (2005). "Epstein-Barr-virus-encoded LMP2A induces primary epithelial cell migration and invasion: possible role in nasopharyngeal carcinoma metastasis." J Virol 79(24): 15430-15442.

Phan, J.H. et al., (2009). "Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment." Trends in Biotechnology 27(6): 350-358.

Piatigorsky, J., (1989). "Lens crystallins and their genes: diversity and tissue-specific expression." The FASEB Journal 3(8): 1933-1940.

Pollard, K. S., S. Dudoit, et al. (Dec 2004). "Multiple Testing Procedures: R multtest Package and Applications to Genomics." U.C. Berkeley Division of Biostatistics Working Paper Series Working Paper 164.

Pommier, J. P., J. Lebeau, et al. (1995). "Chromosomal instability and alteration of telomere repeat sequences." Biochimie 77(10): 817-25.

Pontius, J.U. et al., (2007). "Initial sequence and comparative analysis of the cat genome." Genome Research 17(11): 1675-1689.

Pyeon, D. et al., (2007). "Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers." Cancer Research 67(10): 4605-4619.

Radman, M., P. Jeggo, et al. (1982). "Chromosomal rearrangement and carcinogenesis." Mutat Res 98(3): 249-264.

Reimers, M. & Carey, V.J., (2006). "Bioconductor: an open source framework for bioinformatics and computational biology." Methods in Enzymology 411: 119-134.

Rhodes, D. R. and A. M. Chinnaiyan (2005). "Integrative analysis of the cancer transcriptome." Nat Genet 37 Suppl: S31-37.

Rhodes, D. R., S. Kalyana-Sundaram, et al. (2007). "Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles." Neoplasia 9(2): 166-180.

Rhodes, D. R., M. G. Sanda, et al. (2003). "Multiplex biomarker approach for determining risk of prostate-specific antigen-defined recurrence of prostate cancer." J Natl Cancer Inst 95(9): 661-668.

Rideout, W. M., 3rd, G. A. Coetzee, et al. (1990). "5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes." Science 249(4974): 1288-1290.

Roden, J.C. et al., (2006). "Mining gene expression data by interpreting principal components." BMC Bioinformatics 7: 194.

Romualdi, C., C. De Pitta, et al. (2006). "Defining the gene expression signature of rhabdomyosarcoma by meta-analysis." BMC Genomics 7: 287.

Safran, M., I. Solomon, et al. (2002). "GeneCards 2002: towards a complete, object-oriented, human gene compendium." Bioinformatics 18(11): 1542-1543.

Scanlan, M. J., A. J. Simpson, et al. (2004). "The cancer/testis genes: review, standardization, and commentary." Cancer Immun 4: 1.

Schneeberger, K. et al., (2005). "Masking repeats while clustering ESTs." Nucleic Acids Res 33: 2176-80.

Scott, A. & Salgia, R., (2008). "Biomarkers in lung cancer: from early detection to novel therapeutics and decision making." Biomarkers in Medicine 2(6): 577-586.

Sengupta, S. et al., (2006). "Genome-wide expression profiling reveals EBV-associated inhibition of MHC class I expression in nasopharyngeal carcinoma." Cancer Research 66(16): 7999-8006.

Shankavaram, U. T., W. C. Reinhold, et al. (2007). "Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study." Mol Cancer Ther 6(3): 820-832.

Shannon, C., (1948). "A Mathematical Theory of Communication." CSLI Publications. Available at: http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html [Accessed October 17, 2009].

Shariat, S.F. et al., (2007). "Concomitant carcinoma in situ is a feature of aggressive disease in patients with organ-confined TCC at radical cystectomy." European Urology 51(1): 152-160.

Sherr, C. J. and R. A. DePinho (2000). "Cellular senescence: mitotic clock or culture shock?" Cell 102(4): 407-10.

Shin, J. S., A. Hong, et al. (2006). "The role of telomeres and telomerase in the pathology of human cancer and aging." Pathology 38(2): 103-13.

Shoemaker, R. H. (2006). "The NCI60 human tumour cell line anticancer drug screen." Nat Rev Cancer 6(10): 813-823.

Shyamsundar, R., Y. H. Kim, et al. (2005). "A DNA microarray survey of gene expression in normal human tissues." Genome Biol 6(3): R22.

Simon, R. (2008). "Microarray-based expression profiling and informatics." Curr Opin Biotechnol 19(1): 26-29.

Singh, R. and J. Valcarcel (2005). "Building specificity with nonspecific RNA-binding proteins." Nat Struct Mol Biol 12(8): 645-653.

Smith, D. D., P. Saetrom, et al. (2008). "Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation." BMC Bioinformatics 9: 63.

Sonnenschein, C. and A. M. Soto (2008). "Theories of carcinogenesis: an emerging perspective." Semin Cancer Biol 18(5): 372-377.

Sprung, C. N., L. Sabatier, et al. (1996). "Effect of telomere length on telomeric gene expression." Nucleic Acids Res 24(21): 4336-40.

Stearman, R.S. et al., (2005). "Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model." The American Journal of Pathology 167(6): 1763-1775.

Stewart, S. A. (2005). "Telomere maintenance and tumorigenesis: an "ALT"ernative road." Curr Mol Med 5(2): 253-7.

Stransky, N., C. Vallot, et al. (2006). "Regional copy number-independent deregulation of transcription in cancer." Nat Genet 38(12): 1386-1396.

Stratton, M. R., P. J. Campbell, et al. (2009). "The cancer genome." Nature 458(7239): 719-724.

Strauss, B. S. (1998). "Hypermutability in carcinogenesis." Genetics 148(4): 1619-1626.

Su, L. et al., (2007). "Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme." BMC Genomics 8: 140.

Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A 102(43): 15545-15550.

Széll M, Bata-Csörgo Z & Kemény L., (2008). "The enigmatic world of mRNA-like ncRNAs: their role in human evolution and in human diseases." Semin Cancer Biol 18(2):141-8.

Talbot, S. J. and D. H. Crawford (2004). "Viruses and tumours--an update." Eur J Cancer 40(13): 1998-2005.

Tham, W. H. and V. A. Zakian (2002). "Transcriptional silencing at Saccharomyces telomeres: implications for other organisms." Oncogene 21(4): 512-21.

Thomassen, M., Q. Tan, et al. (2009). "Gene expression meta-analysis identifies chromosomal regions and candidate genes involved in breast cancer metastasis." Breast Cancer Res Treat 113(2): 239-249.

Tone, A.A. et al., (2008). "Gene expression profiles of luteal phase fallopian tube epithelium from BRCA mutation carriers resemble high-grade serous carcinoma." Clinical Cancer Research 14(13): 4067-4078.

Toruner, G.A. et al., (2004). "Association between gene expression profile and tumor invasion in oral squamous cell carcinoma." Cancer Genetics and Cytogenetics 154(1): 27-35.

Tsuchiya, M., Piras, V. et al., (2009). "Emergent genome-wide control in wildtype and genetically mutated lipopolysaccarides-stimulated macrophages." PloS One 4(3): e4905.

Tsuchiya, M., Selvarajoo, K. et al., (2009). "Local and global responses in complex gene regulation networks." Physica A: Statistical Mechanics and its Applications 388(8): 1738-1746.

Tsujimoto, Y., J. Gorham, et al. (1985). "The t(14;18) chromosome translocations involved in B-cell neoplasms result from mistakes in VDJ joining." Science 229(4720): 1390-1393.

Turashvili, G. et al., (2007). "Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis." BMC Cancer 7: 55.

Turner, B.M., (2008). "Open chromatin and hypertranscription in embryonic stem cells." Cell Stem Cell 2(5): 408-410.

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A 98(9): 5116-5121.

Umar, A. (2004). "Applications of bioinformatics in cancer detection: a lexicon of bioinformatics terms." Ann N Y Acad Sci 1020: 263-276.

van de Vijver, M. J., Y. D. He, et al. (2002). "A gene-expression signature as a predictor of survival in breast cancer." N Engl J Med 347(25): 1999-2009.

Varambally, S. et al., (2005). "Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression." Cancer Cell 8(5): 393-406.

Varambally, S., S. M. Dhanasekaran, et al. (2002). "The polycomb group protein EZH2 is involved in progression of prostate cancer." Nature 419(6907): 624-629.

Varmus, H. (2006). "The new era in cancer research." Science 312(5777): 1162-1165.

Venter, J.C. et al., (2001). "The sequence of the human genome. " Science 291(5507): 1304-1351.

Verdun, R. E. and J. Karlseder (2007). "Replication and protection of telomeres." Nature 447(7147): 924-31.

Visone, R. & Croce, C.M., (2009). "MiRNAs and cancer." The American Journal of Pathology 174(4): 1131-1138.

Vogelstein, B. and K. W. Kinzler (2004). "Cancer genes and the pathways they control." Nat Med 10(8): 789-799.

Vukovic, B., P. C. Park, et al. (2003). "Evidence of multifocality of telomere erosion in high-grade prostatic intraepithelial neoplasia (HPIN) and concurrent carcinoma." Oncogene 22(13): 1978-87.

Wang, J. et al., (2003). "Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data." BMC Bioinformatics 4: 60.

Wang, Y. (2005). "Gene expression-driven diagnostics and pharmacogenomics in cancer." Curr Opin Mol Ther 7(3): 246-250.

Wang, Y. et al., (2009). "Regulation of endocytosis via the oxygen-sensing pathway." Nature Medicine 15(3): 319-324.

Washietl, S. et al., (2005). "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome." Nature Biotechnology 23(11): 1383-1390.

Waterston, R.H. et al., (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature 420(6915): 520-562.

Weile, C. et al., (2007). "Use of tiling array data and RNA secondary structure predictions to identify noncoding RNA genes." BMC Genomics 8: 244.

Weinberg, R. A. (2007). The biology of Cancer. New York, Garland Science, Taylor & Francis Group LLC.

Werner, T. (2008). "Bioinformatics applications for pathway analysis of microarray data." Curr Opin Biotechnol 19(1): 50-54.

Wheeler, D. L., T. Barrett, et al. (2007). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res 35(Database issue): D5-12.

White, J.R., Nagarajan, N. & Pop, M., (2009). "Statistical methods for detecting differentially abundant features in clinical metagenomic samples." PLoS Computational Biology 5(4): e1000352.

Wishart, D. S. (2005). "Bioinformatics in drug development and assessment." Drug Metab Rev 37(2): 279-310.

Worm, J. and P. Guldberg (2002). "DNA methylation: an epigenetic pathway to cancer and a promising target for anticancer therapy." J Oral Pathol Med 31(8): 443-449.

Wren, J. D. (2009). "A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide." Bioinformatics 25(13): 1694-1701.

Wu, T. D. (2001). "Bioinformatics in the post-genomic era." Trends Biotechnol 19(12): 479-480.

Wurmbach, E. et al., (2007). "Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma." Hepatology 45(4): 938-947.

Xu, L., Geman, D. & Winslow, R.L., (2007). "Large-scale integration of cancer microarray data identifies a robust common cancer signature." BMC Bioinformatics 8: 275.

Yeo, G., D. Holste, et al. (2004). "Variation in alternative splicing across human tissues." Genome Biol 5(10): R74.

Yi, M., J. D. Horton, et al. (2006). "WholePathwayScope: a comprehensive pathway-based analysis tool for high-throughput data." BMC Bioinformatics 7: 30.

Yu, Y.P. et al., (2004). "Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy." Journal of Clinical Oncology 22(14): 2790-2799.

Zhang, L. and W. H. Li (2004). "Mammalian housekeeping genes evolve more slowly than tissue-specific genes." Mol Biol Evol 21(2): 236-239.

Zhu, J., F. He, et al. (2008). "On the nature of human housekeeping genes." Trends Genet 24(10): 481-484.

CURRICULUM VITAE

Ganiraju Manyam was born and raised in the state of Andhra Pradesh in southern India. He graduated from Gandhi Centenary high school and then completed his higher secondary school at Pragathi Junior college in Kakinada, India. He attended Bharatiar University in Coimbatore where he received B.E degree in Computer Science & Engineering. He received his MS in Bioinformatics from International Institute of Information Technology, Hyderabad, India, in 2004. He was a Junior Research Fellow at the Centre for Cellular and Molecular Biology, India for a period of two years. He is now working on his PhD in Biosciences at Molecular and Microbiology Department, College of Science, George Mason University, USA. His research interests are in the areas of cancer genomics, data mining, machine learning and Bioinformatics.