IDENTIFICATION OF NOVEL EPIGENETIC BIOMARKERS FOR EARLY DETECTION IN VARIOUS CANCER TYPES

by

Santosh Mahadevana Goud

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Biosciences

Committee:

_____	Dr. Serguei G. Popov, Dissertation Chair
_____	Dr. Raja Mazumder, Committee Member
_____	Dr. Alessandra Luchini, Committee Member
_____	Dr. Barney Bishop, Committee Member
_____	Dr. Iosif Vaisman, Department Chair
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs and Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, School of Systems Biology
Date: _____	Spring Semester 2017 George Mason University Fairfax, VA

Identification of Novel Epigenetic Biomarkers for early detection in various Cancer types

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

By

Santosh Mahadevana Goud
Master of Science
University of Hartford, USA, 2007
Bachelor of Science
Bangalore University - Bangalore, India, 1999

Dissertation Chair: Dr. Serguei G. Popov
Department of School of Systems Biology

Spring Semester 2017
George Mason University
Fairfax, VA

Dedication

This research effort is dedicated to my parents Mr. Mahadevana Goud, Mrs. Shashi Rekha Goud and to my sister Mrs. Seema Maregoudra.

Acknowledgments

This work was supported, in part, by my Pre-Doctoral Graduate Teaching Fellowship Award from the Department of Biology.

I would like to thank my mentors, Dr. Raja Mazumder and Dr. Serguei G. Popov. Their patience and guidance has helped me become a more thoughtful and analytical researcher. I would like to thank my thesis committee members, Dr. Bishop Barney and Dr. Alessandra Luchini for their invaluable support as well. I have greatly benefited from every interaction with my committee, be it advice on a particular analysis or about research direction. Their help has been my guiding light towards being a better researcher.

I would like to thank my parents, Mr. Mahadevana Goud and Mrs. Shashi Rekha Goud, my sister, Mrs. Seema Maregoudra, my brother-in-law Mr. Rakesh Wastrad, my mentor, Mr. Rashid Gill, my friends, particularly John Rodriguez, Azad Naik and co-workers for their support, interactions and invaluable moral guidance.

Table of Contents

List of Tables

List of Figures

Abstract

IDENTIFICATION OF NOVEL EPIGENETIC BIOMARKERS FOR EARLY DETECTION IN VARIOUS CANCER TYPES

Santosh Mahadevana Goud, PhD

George Mason University, 2017

Dissertation Chair: Dr. Serguei G. Popov

Epigenetic landscape of cancer cells undergoes profound and significant changes during the development of human malignancies. In fact, global changes in the epigenetic landscape are a hallmark of cancer. Histone modifications and DNA methylation are prominent among such changes. The genome undergoes a large-scale DNA methylation changes alongside other alterations in a collective events of post-translational chromatin modifications being observed. Such aberrant epigenetic changes have a high impact at various stages of tumorigenesis. Identification of such epigenetic aberrations for their use as predictive and prognostic biomarkers has been the focus of cancer genomics research recently. We have selected five genes of interest (ITPKA, GDF15, BLCAP, PIWIL4 and DMRT1) based on literature search and identified each of them as novel epigenetic biomarkers in certain cancer types. Such identified novel epigenetic methylation biomarker gene is subjected to SNV identification by querying it against BioMuta and analyzing its relevant phenotypic effects. BioMuta is a curated single-nucleotide variation (SNV) and disease association database. Here, the variations are mapped to the genome/protein/gene. Such query helps to identify variations and since the database is compiled from various sources through bio curation, it paves ways for prioritizing variations for further experimental evaluations. Furthermore,

such identified epigenetic methylation biomarker gene is subjected to gene expression analysis by querying it against BioXpress database. BioXpress is a curated gene expression and disease association database. Here, the expression levels are mapped to genes. BioXpress is useful in identifying differences between expression levels in disease and normal pairs and to discover differential expression for a gene. It also helps in identification of potential biomarkers or pathways that lead to tumor formation or to explore the overall expression of specific genes across multiple cancer types. Upon additional validations, these findings on novel epigenetic methylation biomarker gene will possibly open new avenues in translation medicine and can be utilized as a novel prognostic biomarker for early stage cancer detection.

Chapter 1: INTRODUCTION

1.1 Hypothesis, rationale, and specific aims

The literal meaning of the term epigenetics is in addition to changes in genetic sequence. In other words, any process that has the capability to alter gene activity without any accompanying changes in DNA sequence, leading to modifications that are transmitted to daughter cells can be defined as an epigenetic change. However, it is been shown that some epigenetic changes can be reversed. The exact definition and/or meaning of the term epigenetic is still debatable and undergoing constant changes.

Epigenetic processes identified till date involves/ includes: methylation, acetylation, phosphorylation, ubiquitylation and sumolyation. Epigenetic processes are natural and in many cases are essential to normal organism functioning. However, in certain cases, it seems to show some major adverse health and behavioral effects.

DNA methylation is a well-studied and well-documented epigenetic process. It involves addition or removal of a methyl group (CH3), predominantly at sites where cytosine bases occur consecutively. This event was first documented in 1983 and has been continuously monitored and found in many disease and health disorders.

Chromatin modification is yet another epigenetic process. Chromatin is DNA and proteins (histones) complex which is highly compacted into the nucleus. Chromatin complex is modified by processes like acetylation (addition of acetyl groups), enzymes and certain RNAs (micro and small interfering RNAs). These modifications can influence gene expression by directly altering the chromatin structure. Highly compacted and condensed chromatin does not allow expression, whereas unfolded or open chromatin structure is functional and allows or facilitates gene expression to take place.

Recent studies have shown a strong link between epigenetic processes and cancer. It has now been established that epigenetic mechanisms and/or processes are one of the most significant considerations in cancer research and accounts for one-third to one-half of known genetic alterations.

This research has three specific aims:

Aim 1: To establish that, alteration or aberrations in DNA methylation and subsequent gene expression, specifically in the promoter or regulatory region of the five genes of our interest (ITPKA, GDF15, BLCAP, PIWIL4 and DMRT1) can be utilized to identify them as Novel epigenetic methylation biomarkers and for their use in as predictive and prognostic biomarkers in certain cancer types.

Our research goal is derived from the following established biological concept. Aberrant DNA methylation is now established as a central/ key feature in carcinogenesis. It is known to be responsible for defective gene expression, faulty condensation and chromosomal instability. Also, it is a hallmark of cellular defenses acting to silence foreign DNA. Specific DNA methylation patterns is often observed to correlate with clinical parameters (cancer stage, survival time and chemotherapy resistance).

Secondly, it is now established that changes in methylation at specific CpG positions in the human genome can turn genes on or off. This has been linked to a wide variety of important normal and impaired molecular pathways. Therefore, DNA methylation is one of the most significant and fertile platform for new biomarker discovery.

Thirdly, epigenetic processes amplify mutational effects and can pave way for disease development and progression in the absence of any detectable relevant genetic changes. Epigenetic pathways are susceptible and are affected by environmental stimuli and insults to a greater extent compared to classical genetic pathways. It is known that certain cancers have a CpG island methylator phenotype. These can arise early and can substantially drive carcinogenesis forward. CpG island methylator phenotypes can vary in different malignancies and may confer poor prognosis. Our five genes of interest (ITPKA, GDF15, BLCAP,

PIWIL4 and DMRT1) that are shortlisted based on literature search will be primarily subjected to novel DNA biomarker discovery strategy. Such discovery will give rise to new opportunities for informed treatment decisions and survival prognosis, thus enabling more personalized cancer therapy.

Aim 2: To identify Single Nucleotide Variations (SNVs) in the genes of our interest (ITPKA, GDF15, BLCAP, PIWIL4 and DMRT1) and to identification of cancer driver genes within our list of genes and exploring their implications in the cancer genomic perspective and establishing their biological significance. To achieve this, each of the above mentioned genes will be subjected to SNV identification by querying it against BioMuta database and its subsequent phenotypic effects will be analyzed.

Our research goal is derived from the following established biological concept. Alterations in genes which encodes for cellular signaling molecules, especially protein kinases, can result in cancers. Sensitivity of drugs that target mutant kinases depends on the genetic makeup of individual tumors. Therefore, mutational profiles of tumor DNA help prioritize anti-cancer therapy and direct patient management.

Gene alterations are a common occurrence in cancer. One such alteration is Single Nucleotide Variation (SNV). SNVs (also referred to as point mutations) results from a base substitution at one nucleotide. Such a substitution may result in one of the following: A change in the amino acid sequence of the encoded protein (missense mutation) or a premature truncation of the protein (nonsense mutation).

Rapid progress in high-throughput sequencing technology has made it easy to identify single nucleotide variants (SNVs) in the genome or exome. Such identification of SNVs have far exceeded our capacity to experimentally validate their impact on disease phenotypes. In this context, bioinformatics and computational methods that can predict the biological impact of non-synonymous SNVs (nsSNVs) on protein function have attained very high popularity. Methods are being developed to distinguish disease-related nsSNVs from neutral polymorphisms. Also, the relevance of nonsynonymous somatic variants in cancer emergence needs to be assessed. In principle, functional somatic mutations can only be a causative

agent, provided they affect cancer driver genes, which upon mutation confer a distinct selective advantage or a newly acquired capability to the cell.

Methylation biomarker discovery platform has benefited tremendously from the rapidly developing sequencing technology in the last few years. Hundreds and thousands of variations are being associated with diseases from single studies.

We plan to identify SNVs in our selected genes of interest. For this purpose, we have chosen BioMuta. BioMuta is a curated single-nucleotide variation (SNV) and disease association database. Here, variations are mapped to the genome/protein/gene. Such query helps to identify variations and since the database is compiled from various cancer centered sources through bio curation, it paves ways for prioritizing variations for further experimental evaluations. This will help in identification of cancer driver genes within our list of genes and exploring their implications in the cancer genomic perspective and establishing their biological significance.

Aim 3: To identify differences between expression levels in disease and normal pairs of the five genes of our interest (ITPKA, GDF15, BLCAP, PIWIL4 and DMRT1) and also to discover differential expression for a gene. To achieve this, each of the above mentioned genes will be subjected to gene expression analysis by querying it against BioXpress database.

Our research goal is derived from the following established biological concept. Epigenetics studies have shown that mechanisms associated or involving them provides an "extra" layer of transcriptional control that regulates how genes are expressed. Although such mechanisms are utmost essential in normal development and growth of cells, their abnormalities are causative factors for cancer, genetic disorders, pediatric syndromes and auto-immune diseases.

Epigenetic mechanisms exhibit two prominent features: DNA methylation and histone modifications. DNA methylation and changes to histone proteins orchestrate DNA organization and gene expression. Histone-modifying enzymes are recruited for one of the two purposes: either to ensure that a receptive DNA region is either accessible / available for

transcription or that DNA is targeted for silencing. It is now established that active regions of chromatin have unmethylated DNA and have high levels of acetylated histones, whereas inactive regions of chromatin contain methylated DNA and deacetylated histones. Therefore, it is now believed that an epigenetic tag is placed on targeted DNA, providing or marking it with a special status that specifically activates or silences genes. Also, since epigenetic mechanisms are reversible, these reversible modifications ensure that specific genes can be expressed or silenced depending on specific developmental or biochemical cues (hormone levels, dietary components or drug exposures).

Cancer development and progression involves a complex multistep process in which genetic and epigenetic errors accumulate and transform a normal cell into an invasive or metastatic tumor cell. It has been established that altered or aberrant DNA methylation patterns have a direct influence or can change the expression of cancer-associated genes. Additionally, it has been observed that DNA hypo methylation activates oncogenes and initiates chromosome instability, whereas DNA hyper methylation initiates silencing of tumor suppressor genes. The incidence of hyper methylation, particularly in sporadic cancers, varies with respect to the gene involved and the tumor type in which the event occurs. Such epigenetic changes is utilized by the research community in investigating or molecular diagnosis of a variety of cancers.

We aim to identify differences between expression levels in disease and normal pairs of the five genes of our interest and also to discover differential expression for a gene. We will achieve this by querying each of the above mentioned genes against BioXpress database. BioXpress is a curated gene expression and disease association database where the expression levels are mapped to genes. Such an investigation or query will also helps in identification of potential biomarkers or pathways that lead to tumor formation or to explore the overall expression of specific genes across multiple cancer types.

1.2 Introduction to Epigenetics

Epigenetics are referred to those heritable alterations that are not associated with changes in DNA sequence itself. Epigenetic modifications are sometimes referred to as molecular tags. These tags include DNA methylation and histone modifications which can alter DNA accessibility and chromatin structure. By doing so, they can regulate the patterns of gene expression. Normal development and differentiation of distinct cell lineages in adult organisms depend on precisely orchestrated normal gene regulation/expression mechanisms that are still susceptible for epigenetic mechanisms. Such exogenous epigenetic influence can result in environmental alterations of phenotype or patho-phenotypes. More importantly, regulations of pluripotency genes are also regulated by epigenetic mechanisms. These genes are inactivated during differentiation [1].

1.2.1 Introduction to Epigenetic Tags: their acquisition, maintenance, and inheritance

Chromatin Domains

Heterochromatin: transcriptional inactive, densely packed nucleosomes.

Constitutive: highly repetitive DNA sequences, such as centromeric and telomeric domains, hypoacetylated nucleosomes, H3K9me[1]

Facultative: includes silenced genes, such as inactive X chromosome or imprinted regions, or transcriptionally repressed genes, hypoacetylated nucleosomes, H3K27me

Euchromatin: transcriptional permissive chromatin, less densely packed. Accessible to nuclear factors and nuclear repressors, acetylated nucleosomes, H3K4me, H3L36me

Chromatin complex is chromosomal DNA and its associated proteins in nucleus [2]. Nucleosomes are usually referred to those units of DNA packaged around histone proteins in chromatin. Normally, DNA of around 147bp in association with octomeric core of histone proteins (two H3-H4 dimers of histone surrounded by two H2A-H2B dimers) are often

[1]histone methylation sites are listed in abbreviated forms, for example H3K9me, histone lysine 9 methylation [2]

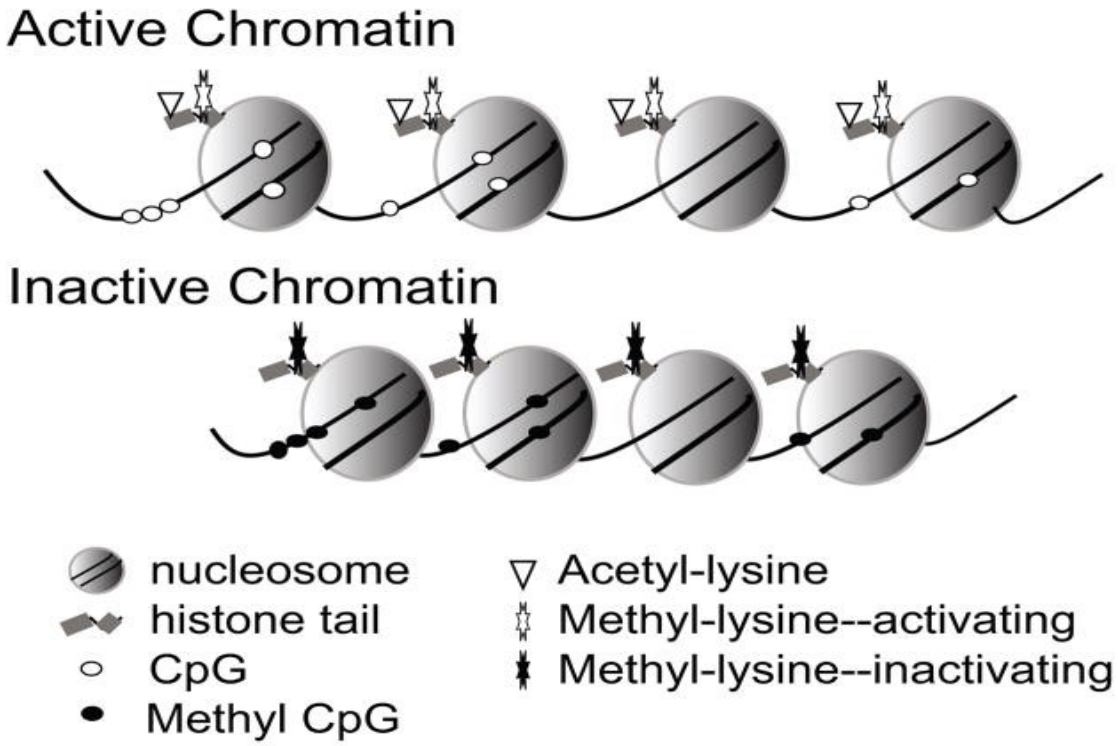


Figure 1.1: Epigenetic tags and chromatin structure

referred to as a Nucleosome. The N-terminal histone tails are often observed protruding into nuclear lumen from their respective nucleosomes. H1 histones are seen associating with linker DNA found between the nucleosomes. Chromatin structure is strongly dependent on nucleosome spacing. Chromatin structure is broadly divided into heterochromatin and euchromatin (Table 1).

Transcriptional machinery depends on chromatin structure and gene accessibility for its functioning and is regulated by both DNA and histone tail modifications (Figure 1.1) [2].

Chromosomal DNA is packaged around histone cores to form nucleosomes. Nucleosome spacing in open structure that is accessible to nuclear factors is maintained, in part, by post-translational modification of histone tails, including lysine acetylation and specific lysine methylation. CpG dinucleotides are unequally distributed throughout chromosomal DNA, and may be concentrated in regions called CpG islands that can overlap gene promoters. Methylation of cytosine in CpG dinucleotides is overall associated with inactive, condensed

states of the chromosome. Inactive chromatin is also maintained by specific histone lysine modifications [2].

1.2.2 Epigenetic mechanism: DNA methylation

Previously it was believed that covalently attached methyl group at C5 position of cytosine residues in CpG dinucleotide sequences (CpG or CpG islands) are the principle epigenetic tags found in differentiated mammalian cells [3]. However, recent findings indicate that even in undifferentiated stem cells, cytosines, other than those found in CpG sites can be methylated as well. Such methylations of non-CpG cytosines have proved vital for gene regulation in embryonic stem cells [4]. CpG methylation is observed to play a vital role in imprinting and X-chromosome inactivation and is also found to be necessary for transcription repression of transposons and repetitive elements [5]. CpG methylation can also be involved in transcriptional gene silencing and thereby restricts the expression of certain tissue-specific developmental genes and differentiation by suppressing them in non-expressing cells.

CpG methylation follows a predictable pattern of changes during development. Also, in early embryogenesis, methylation is nullified genome-wide and re-established in all except CpG islands (high density in genome found to have CpG residues). These CpG islands show consistency in being hypo methylated till late developmental stages and some of them become methylated [6, 7]. CpG islands that are subsequently methylated at cytosine and at other CpG dinucleotides are often associated with transcriptional repression, especially when the methylation sites involves a promoter or a gene regulatory regions [6, 7]. However, DNA methylation may activate transcriptional repressors if it prevents binding or limits expression. The degree to which methylation occurs in mammalian promoters is observed in a small percentage of CpG dinucleotides and inhibits transcription in just a small genes subset in differentiated cell types. Such repressed genes are usually germ-line specific which may include pluripotency genes [8]. This suggests methylation is an important mechanism in suppressing some key genes during differentiation.

Transcription is suppressed by CpG methylation by multiple mechanisms. Mostly, the methyl group at a specific CpG site may directly interfere or block DNA recognition and as well as its binding to transcription factors. One such example involves the direct inhibition of transcriptional activation at GC-boxes by methylation. This excludes Sp1 and Sp3 transcription factors binding in the context of promoter regions [9, 10]. Alternatively, methylation can block nuclear factor, Hif1, in hypoxic conditions by inducing erythropoietin transcription [11]. Furthermore, certain other factors may exhibit preferential binding to methylated DNA and block access to transcription factors. Examples include MeCP2 and related protein families binding to methyl CpG and inducing transcriptional repression. This is achieved by recruitment of histone-modifying proteins like histone deacetylases (HDAC) [12]. Histone deacetylation further promotes condensation of chromatin and thereby represses transcription [13, 14]. Such mechanisms clearly indicates as to how DNA methylation and histone modifications come together in function to contribute to gene transcriptional on or off state, subject to epigenetic modifications.

DNA methyltransferase enzymes (DNMTs), are family of enzymes responsible for de novo DNA methylation and its maintenance. During developmental embryogenesis, de novo methylation is carried out by DNMT3A and DNMT3B [15]. Although, DNMT3A and DNMT3B is indicted in maintaining methylation in certain cell types, the ubiquitously expressed DNMT1 is primarily responsible for maintaining CpG methylation in most cell types [16,17]. It is observed that alternative promoter induced transcription yields truncated oocyte- specific DNMT1 isoform (DNMT1o) which is essential for early embryogenesis to occur [18]. DNMT1 along with a complex can recognize a hemi-methylated DNA and adds a methyl groups to the non-methylated daughter strand formed during replication [19]. This is aided by the CpG base pairing which helps in reciprocal maintenance of methylation in the next subsequent replication cycles. Such processes help in a non-genetic trait (DNA methylation) being passed from cell to cell with associated contextual effects on gene expression. By considering such evidences, we can come to an understanding that methylation is a long-term, relatively stable, epigenetic trait whose effects help maintain

cellular phenotypes.

1.2.3 DNA methylation and single nucleotide polymorphisms (SNPs) or single nucleotide variations (SNVs)

SNPs may create CpG sites that can be potential targets for epigenetic modifications and potential loss of such sites will inhibit DNA methylation. Polymorphism that yields CpG in the promoter region of the gene NDUFB6 exhibits or provides a platform for cross-talk between genetic and epigenetic regulation. NDUFB6 protein expression is suppressed in Type 2 diabetes. This is a respiratory chain protein. In geriatric population, NDUFB6 expression and DNA methylation levels are inversely correlated. This infers the presence of a CpG site conferring high risk for decreased expression along with associated disease risk, compared to loss of this site [20]. These findings suggests that epigenetic modifications can increase or influence complex diseases.

1.2.4 Biomarkers of genome instability and cancer epigenetics

Genetic and epigenetic alterations together constitute a multistep process leading to tumorigenesis. Such a process drives somatic evolution from normal cells to malignant derivatives. Researchers can take advantage of this fact by combining the genetic and epigenetic alterations into biomarkers for risk assessment, early stage tumor detection, and accurate tumor characterization for treatment. Application of mass sequencing has provided systematic approaches to study cancer genomics. It has broadly led to identification of two platforms: genome instability and epigenetics. Ability of cancer to develop, evolve, adapt and spread through genetic and epigenetic lesions of varying sizes and quality. These include point mutations, small insertions/deletions, large scale chromosomal rearrangements, whole chromosome copy number alterations, predisposition or preferential allelic expression of cancer risk alleles and processes that increase mutation rates in tumor. There also exists epigenetic mechanisms that inhibit tumor adaptation. These include DNA methylation, histone

modifications, remodeling of nucleosome, transcription factor activity, and small non coding RNAs. Two biggest challenges that remain elusive: 1) to interpret essentially different signals (non-comparable) across numerous genes and summarize them into diagnostic value. 2) Identification of epigenetic processes that induces increased cancer rates due to increased exposure of toxic environmental stress and pollution in an organisms developmental stages [21].

1.2.5 Relationship between DNA methylation and Gene expression

It is generally found that housekeeping genes arbor non-methylated CpG islands which are tightly associated with their promoter regions [22, 23]. As such genes are ubiquitously expresses and also, autosomal CpG islands are non-methylated, housekeeping genes are presumed to be regulated by DNA methylation. It is now established that relation between DNA methylation and gene expression levels of tissue-specific genes is mostly that of inverse correlation. In a recent study, majority of tissue specific genes exhibited a correlation between hypo methylation of promoter region and gene activity. Also, CpG dinucleotides showed weaker correlation throughout the gene body. Notably, de novo methylation of CpG islands in tissue culture cells was observed in a widespread manner [22]. Since CpG islands are non-methylated in normal tissues (in vivo) and associated with non-essential growth genes in tissue culture, it suggests that methylation induced gene silencing is of selective advantage for cell growth.

1.2.6 DNA methylation and Cancer

DNA methylation is often referred to addition of methyl groups to the 5 carbon at cytosine residues that are preceding guanine nucleotides, which are linked together by phosphate bonds (CpG) and by utilizing a methyl donor such as S-adenosylmethionine. Asymmetric arrangement of CpG rich foci are found genome wide. They are clustered in short CpG rich DNA sequences, often referred to as CpG islands and also in regions of large repetitive

sequences such as centromeric repeats, retrotransposons etc [24,25]. CpG islands are specifically targeted by DNA methyl transferases (DNMT) class of enzymes. Four DNMTs are identified, DNMT 1, 2, and 3a and 3b [26]. DNA methylation often involves DNMT1 and DNMT3. DNA methylation affects transcription directly by interfering with transcriptional factor binding with target sites as observed in c-myc and other genes [27]. Alternatively, methylated cytosine residues provides a docking platform for methylated DNA binding proteins (MBD1, MBD2, MBD3, and Mecp2). These proteins are readily identified by histone modifying enzymes like histone deacetylases (HDACs), responsible for repression of genes [28–30]. Generally, it has been observed that a normal cell shows a characteristic pattern of genome wide methylation, except for the CpG (cytosine-phosphate-guanine) islands, which are found to be unmethylated [31]. Numerous triggering events or triggers in cancerous cells leads to hypo methylation genome-wide, except for the CpG island promoters, which undergo hyper methylation [32].

1.2.7 Hypo methylation and its role in Cancer

For tumorigenesis to occur, extensive hypo methylation is required at the repetitive sequences as this increases genomic instability due to chromosomal rearrangement [33]. Such activation is aided by hypo methylation of retrotransposons, further leading to retrotransposon translocation to other genomic regions that can potentially disturb genomic instability [34]. Documented evidences include DNA hypo methylation responsible for activation of Ras (growth promoting genes) and mammary serine protease inhibitor (MAPSIN) for gastric carcinoma, S-100 in case of colonic cancer, melanoma-associated antigen (MAGE) in melanoma [35]. DNA hypo methylation is also observed in loss of imprinting, growth factor 2 (IGF-2) in Wilms' tumor [36] and colorectal cancer [37].

1.2.8 Hyper methylation and its role in Cancer

Hyper methylation of CpG islands induces tumorigenesis by completely shutting down tumor suppressor genes expression. This is in stark contrast to hypo methylation mechanism. Such mechanism is achieved by directly involving tumor suppressor genes and also by silencing of the associated tumor suppressor genes' transcription factors and inhibiting the expression of DNA repair genes. Documented evidence includes Rb promoter gene (retinoblastoma associated tumor suppressor gene) hyper methylation. Here hyper methylation of the CpG promoter island site silences tumor suppressor gene, thereby promoting retinoblastoma malignancy [38]. Other examples includes genes such as p16 and BRCA1 which are silenced by hyper methylation [39]. These genes play a vital role in cellular adhesion, apoptosis, and angiogenesis, involved in the cancer development and progression. Alternatively, hyper methylation of CpG promoter regions induced silencing of transcription factors leads to downstream target inactivation of the tumor suppressor genes. This further leads to cancer cell propagation. Examples include RUNX3, GATA-4, and GATA-5 in esophageal, colorectal, and gastric cancers, respectively [40,41]. Furthermore, MLH1 and BRCA1 (DNA repair genes) upon silencing tend to start accumulating other genetic lesions leading to cancer progression. However, one elusive questions that persists as to how can selective genes targeting by the DNA methylation machinery executed? Possible explanation may include CpG island specific methylation is possibly guided by a nucleotide sequence specific mechanism. This in turn, may direct the DNMTs to their respective genes that have shown previous association with the oncogenic transcription factors. Documented example includes PML-RAR fusion protein led abnormal hyper methylation, plus the specific target promoter genes' silencing observed in acute promyelocytic leukemia [42]. Also, in various cancers, long DNA sequence stretches undergo methylation, leading to CpG island hyper methylation as they fall under genomic regions that have potentially undergone large scale epigenetic reprogramming [43]. A third possibility may involve histone marks that can play a vital role in CpG island specific de novo DNA hyper methylation.

1.3 TET Proteins and DNA methylation

Although hyper- and hypo methylation produces varyingly opposite results, they seem to coexist in a single tumor. Also, they afflict different genomic regions by varying mechanisms. It is highly likely that, hyper- and hypo methylation mechanisms can cross-talk or interact at different levels and can possibly give rise to numerous cancer sub- phenotypes. Additionally, DNA methylation is a reversible epigenetic process adding to the already complex cancer genome. This opens up a plethora of modifications that can occur in cellular environment, suggesting that DNA methylation might not be a stable but rather a non-stable and susceptible chromatin modification. DNA methylation mapping in high resolution (in both differentiated and pluripotent cells) has further increased the flexibility and complexity of DNA methylation. Such a flexible and highly complex mechanism has to be supported by a highly efficient enzymatic system. This system might have the capability to completely abolish or alter epigenetic modifications [44]. However, such a hypothesis was proven to be wrong by the discovery or identification of ten-eleven translocation (TET13) group of proteins. The origin of TET terminology is associated with a recurrent chromosomal translocation (10; 11) (q22; q23). This is placed closely to mixed-lineage leukemia or myeloid-lymphoid leukemia (MLL) gene with TET1 protein in a few acute myelocytic leukemia (AML) patients. TET protein family are basically DNA hydroxylase that can convert 5-methyl cytosine (5mC) to 5-hydroxymethylcytosine (5hmC). This, upon further oxidation, yields various oxidation products, like 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), believed to be the vital DNA methylation intermediates and in either active or passive form. It is also speculated that they might either prevent or enhance attachment of methyl CpG binding domain (MBD) proteins. They might even regulate recruitment of chromatin regulators. Furthermore, the genome wide distribution of 5hmC indicated that 5hmC and TET proteins can possibly influence both transcriptional activation and silencing [45].

Table 1.1: DNMT inhibitors in cancer

Drug	Therapeutic use	Developmental stage
Nucleoside analogue inhibitors		
(1) 5-azacytidine	Myelodysplastic syndrome	Approved [United States-Food and Drug Administration (US-FDA)]
	Acute myeloid leukemia	Phase 2
	Solid tumors	Phase 2
(2) Decitabine	Myelodysplastic syndrome	Approved (US-FDA)
	Acute myeloid leukemia	Approved [European Commission (EC)]
(3) Zebularine	Solid tumors like breast, urinary bladder, hepatocellular cancer	Preclinical
(4) SGI-110	Myelodysplastic syndrome	Phase 1
	Acute myeloid leukemia	Phase 1
	Solid tumors like bladder	Preclinical
Nonnucleoside analogue inhibitors		
(1) Procainamide	Solid tumors like bladder, breast,prostate, cervix	Preclinical
(2) Procaine		
(3) Epigallocatechin -3-gallate		
(4) SGI-1027	Leukemia	Preclinical
(5) Hydralazine	Breast cancer	Phase 2
	Ovarian cancer	Phase 3

1.4 DNA methylation for therapeutic use

Hypo methylating agents have fast acquired the status of being the first epigenetic therapeutic agent, approved by the Food and Drug Administration (FDA).

Hypomethylating agents have proved to be effective against hematological malignancies. They are highly effective Myelodysplastic syndrome (MDS) [46–50]. However, these hypo methylating agents/DNMT inhibitors (DNMTi) are not effective against solid malignancies [51, 52]. A possible reason for their failure could be the complex nature of solid tumors as compared to hematological neoplasms [53]. Yet another reason for their inefficacy maybe due to their slow rate of replication dependent incorporation of DNMTi inhibitors in solid tumor cells. Also, these inhibitors are inactivated by cytidine deaminase enzyme. Also, toxicity is an issue as DNMTi inhibitors are effective against hematological malignancies at a higher dosage alone. However, it is now established that azacytidine, in phase 2 trial has proven to be effective even in low dosage form [54]. This has paved a new path for DNMTi inhibitor application against solid tumors and possibly new treatment regime. Yet another new strategy to achieve gene demethylation is the use of small nonnucleoside DNMT inhibitor molecules as indicated in Table 2. These molecules partially and competitively

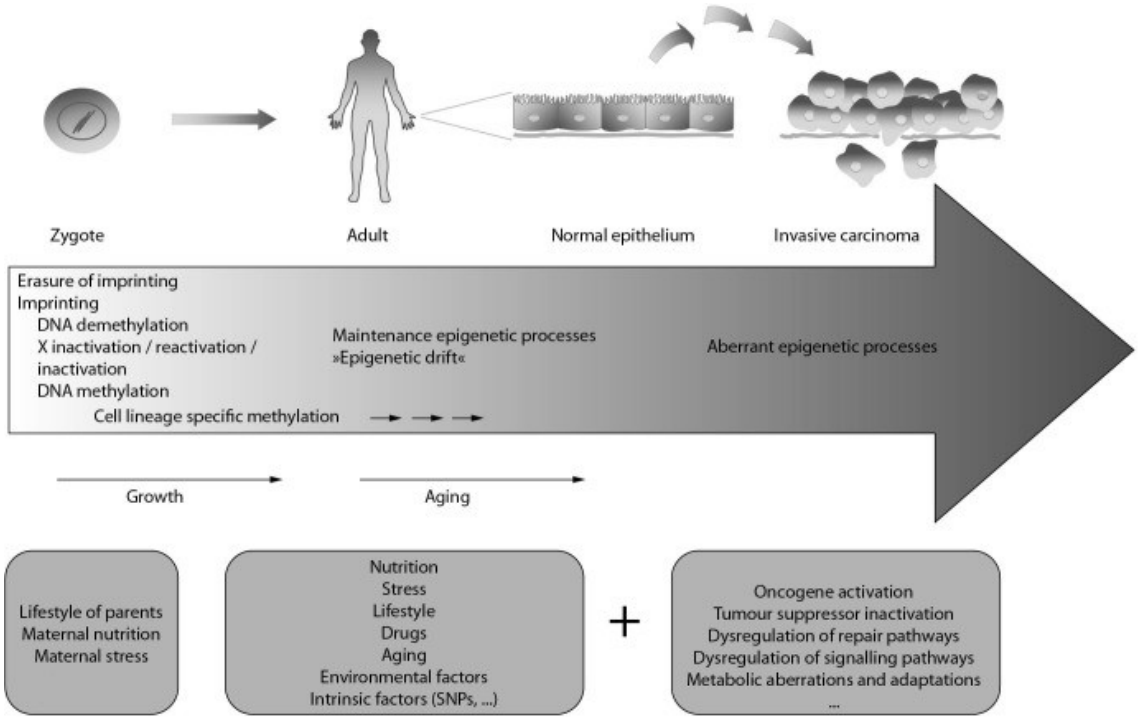


Figure 1.2: DNA methylation and complex diseases

inhibit DNMT1 and also decrease DNMTs affinity towards their substrate. This leads to DNMT1 and hemi-methylated DNA to dissociate. In an another therapy approach using azanucleosides in combination with standard nucleoside analogues like 5-fluorouracil have exhibited excellent efficacy compared to DNMT inhibitors as these can reignite the dormant or silenced pro-apoptic genes [55, 56]. Also, HDAC inhibitors and DNMTi upon synergistic usage, may yield superior results, thereby providing new treatment avenue [57–59]

1.5 Clinical perspective of aberrant methylation patterns in cancer

Feinberg and Vogelstein, first reported the possible association between differences in DNA methylation status to cancer [60] (Figure 1.2). Research since then, has accumulated a wealth of information documenting aberrant DNA methylation in complex diseases.

Research focusing on DNA methylation and DNA methylation mapping techniques have

provided a platform for translation of basic research to therapy/treatment regimes. Aberrant gene methylation has been observed in diseases such as colorectal cancer, Prader-Willi, Angelman, Beckwith-Wiedmann syndromes and now part of routine detection diagnostics. DNA methylation works in close harmony with histone modifications and chromatin structure, either building transcriptionally active or repressed chromatin [61–63]. The complex cross-talk/ dynamics is currently the biggest challenge to be deciphered and aberrant DNA methylation definitely impacts histone modifications and chromatin structure. Also, the reversible effect of histone modifications and dysregulation of histone modifying proteins can also influence DNA methylation patterns.

1.6 Discovery and detection of DNA methylation

A new platform in cancer diagnostics has arouse based on DNA methylation. This is due to biomarker discovery for both diagnostic and prognostic use [64]. DNA methylation research started with locus specific approach to a now genome-wide determination of methylome data at a fine base pair resolution [65]. Methods or detection techniques for determining DNA methylation are numerous, and selection of an appropriate technique depends largely on the nature and number of samples, the purpose of investigation and expenses involved. The three broad methodology or approaches that exists are: 1) methylation-specific restriction enzyme digestion 2) Affinity purification of methylated DNA and 3) Bisulfite conversion of DNA [65]. DNA subjected to investigation follow one of the two detection regimes: A molecular genetic approach where in a single locus is analyzed using a PCR based analysis or a genome-wide investigation based on microarray technology, mass spectroscopy or next generation sequencing analysis. DNA methylation detection started initially with methylation-sensitive restriction enzymes and southern blotting, whereas today numerous detection techniques are based on bisulfite conversion of DNA and subsequent PCR-based method [65]. Bisulfite treatment technique usually demands a good pair of research hands for conducting protocols, as it may lead to DNA degradation and unnecessary conversion

Table 1.2: Commonly used techniques for locus specific DNA methylation determination based on bisulfite sequencing with potential for translation into clinical practice.

Method	Advantages	Disadvantages
Methylation specific PCR (MSP-PCR)	Very sensitive. Cost-effective.	Need for two different pairs of primers, one for methylated DNA and one for non-methylated. Risk for false positive results if primer design is not appropriate. Only qualitative.
SMART-MSP	Low rate of false positive results. High sensitivity. Closed tube technique low risk for sample contamination.	Determination of methylated DNA only. Not suitable for detection of heterogeneous methylation.
MethyLight	Very high analytical sensitivity. Low false positive rates. Closed tube technique low risk for sample contamination.	Only for detection of methylated DNA. When samples with heterogeneous DNA methylation are analyzed it is only semi-quantitative.
Methylation-sensitive high resolution melting (MS-HRM)	Useful for screening purposes high throughput, inexpensive, fast. Real - time tracking of methylation status. Applicable also for small amounts of DNA. Closed tube technique low risk for sample contamination.	Information on methylation degree based on standard curve analysis semi-quantitative. No information on specific sites of methylation patterns are hard to recognize.
Sanger sequencing of bisulfite treated DNA	Data on complete sequence composition. Relatively long sequence reads possible.	Only semi - quantitative. Low quality results at the beginning of the reads.
Pyrosequencing	Quantitative analysis of individual CpG islands with real - time monitoring. Appropriate for degraded formalin-fixed, paraffin - embedded (FFPE) samples.	Relatively short sequences ($\sim$ 50 nucleotides) can be reliably analyzed.
Next generation sequencing	High throughput. Data on complete sequence reads genetic and epigenetic data. Quantitative.	Need for high-quality DNA. Relatively labor demanding. Still associated with high costs. Currently used applicable for research use only. Purchase of an expensive instrument is required.
MassARRAY EpiTYPER	Quantitative analysis, high throughput, applicable for heterogeneous DNA methylation patterns.	Investment into expensive instruments is required.

of methylated cytosines to thymines based on sensitive incubation time and protocols involved. Hence, commercially available kits for bisulfite conversion of DNA isolated from various sample types were developed and now in extensive usage [66].

DNA methylation detection analysis on specific locus requires the investigation region to be ideally unmethylated in normal tissue and methylated in cancerous tissue or vice versa. Also, yet another requirement requires the differentiation between the methylation levels between the two statuses of the samples [67]. Pyrosequencing has emerged to be a popular technique for locus specific methylation biomarkers method. This is most appropriate for degraded formalin-fixed, paraffin-embedded (FFPE) samples that forms an integral and important part of tissue bio-banks. Pyrosequencing yields quantitative analysis of each CpG position [67]. High resolution melting (HRM) curve analysis is another fast emerging

and robust method in DNA methylation detection. Methylation-sensitive HRM (MS-HRM) curve analysis and quantitative real time PCR, such as SMART-MSP are widely used now (Kristensen LS, et al., 2009). These techniques are sensitive and inexpensive, accompanied by a good throughput and quantification, and are closed tubes techniques. Quantitative real time PCR (SMART-MSP) has an advantage of minimizing the risk involved in sample confusion and cross-contamination which is a very critical factor in clinical laboratory [67, 68].MS-HRM method has already been employed on small DNA samples and has proved to be a sensitive and reliable for screening investigations [69]. The above two techniques are found to be successful on FFPE tissues also [70]. However, to overcome the obstacle of false positive results, well designed primers and very stringent annealing temperatures are required. Also, these techniques are not well-suited for analyzing heterogeneous DNA methylation patterns.

Matrix-assisted laser desorption ionization - time of flight (MALDI-TOF) is yet another technique being considered for single locus analysis. Sequenom Inc. has recently developed a very sensitive and high throughput assay MassARRAY EpiTYPER, enabling a quantitative screening and differential methylation analysis in cancer samples [71]. Two other commercial ventures, Roche 454 Genome Sequencer and Illumina Genome analyser are now very popular for their usage of next generation sequencing platforms in research and most likely to be validated and approved for clinical use [72]. They are now key players in cancer genome-wide methylome determination that could result in determination of an array of biomarkers of practical application. The arrays developed by these commercial ventures, are now being subjected to testing thoroughly on larger sample cohorts by using a more cost effective methodology. Currently, next-generation sequencing costs are too high for large sample testing studies. Such studies have led to providing good fingerprints of cancer methylomes that are highly helpful for classifying cancer subtypes. However, establishing a safe cancer specific methylation signatures is nowhere near, as limited knowledge of functional consequences of methylation aberrations, enormous number of discovered changes and overlapping changes between different cancers, still pose an enormous challenge that

needs to be overcome. Roche 454 system was a pioneer platform enabling a comprehensive multi-sample, multi-gene, and ultra-deep sequencing of cancer DNA generating specific methylation patterns. Adding to the high number of reads, and therefore detailed sequence coverage, a significant advantage of this technology was simultaneous exploration of genetic and epigenetic data at a genome-wide level [73]. The Infinium HumanMethylation BeadChip microarray platform developed by Illumina is yet another platform allowing for genome-wide methylome studies which has attained popularity. One of their array platforms allowed for detection and analysis of 27,578 highly informative CpG islands located within the proximal promoter regions [74]. A disadvantage to this analysis platform is the requirement of high-quality DNA, which is not the most optimal for clinical setting as the samples are mostly stored as FFPE. Furthermore, studies involving comparison of fresh-frozen samples with FFPE showed their correlation of results between them was not optimal [75]. Although, DNA methylation detection and analysis methods are numerous, their applications for clinical diagnostic purposes are yet to overcome significant obstacles like standardization of methods between laboratories, determination of reference standards, and associated expenses involving the training of personnel and obtaining expensive new equipment [67]. Key aspects that needs to be developed for DNA methylation techniques for clinical setting are the ease of use, high throughput, preferably automation, applicability on degraded DNA, cost-effectiveness, and should be able to provide quantitative methylation data [67]. To add value to such development and in its favor is the fact that DNA methylation is a stable covalent modification, present at single or multiple CpG sites, and can be easily translated into highly robust and high throughput routine laboratory diagnostic tests [76]. However, biomarker discovery and evaluation should be possible and readily accessible diagnostic specimens, such as blood, urine, faeces or saliva for early stage detection of the disease.

Table 1.3: Overview of bladder cancer biomarkers.

Biomarker	Sample	Type	Diagnosis, treatment, or prognosis
RUNX3	Tissue	DNA methylation	Diagnosis
RSPH9	Urine	DNA methylation	Diagnosis, prognosis
PCDH10, PCDH17	Urine	DNA hypermethylation	Treatment, prognosis
PCDH17, POU4F2	Urine	DNA methylation	All
TWIST1, NID2	Urine	DNA methylation	Diagnosis
CDH1, CDH13, RASSF1A, APC	Urine	DNA methylation	Prognosis
RASSF1A, CDH1, TNFSR25, EDNRB, APC	Urine	DNA methylation	Prognosis
H4K20	Tissue	Histone modification	Prognosis
KLF4	Urine	Histone modification	Treatment
H4K20me3	Tissue	Expression level	Prognosis
miR-422a-3p			
miR-486-3p			
miR-103a-3p	Tissue (serum)	Overexpression	Prognosis
miR-27a-3p			
miRNA-146a-5p	Urine	Overexpression	Prognosis
miRNA-145	Urine	Overexpression	Prognosis

1.7 Current knowledge, advances and applications of DNA methylation biomarkers in various cancers

1.7.1 DNA Methylation biomarkers in Urological Cancer

Urological cancer comprises of prostate, testis, kidney and bladder cancers. These cancers are usually silenced in early stages and hence there is loss of early diagnosis and treatment. Clinical biomarkers are scarce and existing ones are not specific or sensitive for applications. However, detection of epigenetic conditions is easily accessed through urine samples.

1.7.2 Epigenetic biomarkers in bladder cancer

Current trend in bladder cancer diagnosis is mainly invasive. This is highly discomforting to patients and only provides a generalized outcome for the subject. Noninvasive screening and diagnosis is the need of the hour. Discovery of epigenetic biomarker will ease the use or entirely erase the use of invasive methods and can also provide diagnostic value at early stages for an effective treatment regime. RUNX3 gene, a tumor suppressor gene shows a high level of methylation increase in bladder cancer in an analysis involving 124 tumor tissue samples, indicating a potential valued role for RUNX3 gene (Peng Wu., et

Table 1.4: Overview of kidney cancer biomarkers

Biomarker	Sample	Type	Diagnosis, treatment, or prognosis
Wnt family genes	Tissue (Serum)	DNA methylation	Diagnosis, prognosis
VHL, RASSF1A	Tissue	DNA methylation	Diagnosis
SMPD3, FBXW10	Tissue	Hyper methylation	Diagnosis
DAB2IP	Tissue	Methylation	Prognosis
H3K4me2, H3K18Ac	Tissue	Histone modification	Prognosis
hMOF	Tissue	Histone modification	Diagnosis
HDAC	Tissue	Histone modification	Treatment
miRNA-126	Tissue	Downregulated	Treatment
miR-146a-5p			
miR-128a-3p	Tissue	Downregulated	Prognosis
miR-17-5p			

al., 2016). Recently, Yoon et al., discovered a prognostic indicator in patients with non-muscle-invasive bladder cancer (NMIBC). Quantitative Pyrosequencing has revealed the clinical significance of RSPH9 using 136 human bladder specimens (8 normal controls and 128 NMIBCs). From this study, it was concluded that RSPH9 methylation showed clinical value for the assessment of disease recurrence and can be used as an independent prognostic indicator in NMIBC patients. Furthermore, Lin and Luo et al., reported from their study, that the hyper methylation of PCDH10 (50%,) and PCDH17 (52%,) was closely related to the bladder cancer development and was an independent predictor with regards to the cancer-specific survival time [77, 78].

1.7.3 Epigenetic biomarkers in kidney cancer

Kidney cancer is reported to be the third most commonly occurring urological malignancy in China. At present, there does not exist any tumor markers for clinical diagnosis of renal cell carcinoma and to add to the complexity, clinical diagnosis of which depends on imaging examination and precise diagnostic confirmation can be obtained after pathological examination alone. Hauser et al., reported and demonstrated that, using tumor and serum DNA, Wnt antagonist family genes could possibly be used as a biomarkers for diagnosis, staging, and prognosis in kidney cancer. In this particular study, Hauser et al., adopted methylation-specific PCR method to identify level of genes panels. This gene panel comprised of sFRP-1, sFRP-2, sFRP-4, sFRP-5, Wif-1, and Dkk-3 in 62 RCC samples and

their corresponding normal renal tissue. Results showed that Wnt antagonist family genes detection showed sensitivity of 79.0% and specificity of 75.8% Also, serum DNA significantly correlated with tumor grade and stage [79]. It is also reported that certain genes are highly specific for RCC patients in the level of DNA hyper methylation that includes VHL (91%) and RASSF1A (93%) [80]. Similar to the above two studies, genes SMPD3 and FBXW10, showed hyper methylation in ccRCC tissue samples as compared to paired normal tissues. However, upon 5-aza-2-deoxycytidine treatment, mRNA expression of SMPD3 and FBXW10 showed high levels of upregulation. Hence SMPD3 and FBXW10 genes can be utilized as target for treatment or prognostic value [81]. Furthermore, it has been recently reported that, DAB2IP, tumor suppressive gene, its CpG1 methylation is a practical and repeatable biomarker for ccRCC that provides prognostic value and also complements the present staging system. Also, they showed that there exists a relationship between CpG methylation biomarker (DAB2IP CpG1) and poor overall survival in TCGA by pyrosequencing quantitative methylation assay [82].

1.7.4 Epigenetic biomarkers in prostate cancer

Currently, the PSA test is a subject of increasing criticism, primarily due to potential overtreatment and less comprehensive evaluation [83]. For prostate cancer, candidate biomarkers is classified in few groups such as molecular class, soluble proteins DNA methylation, mRNA and microRNA [84–86].

PCDH17 and TCF21 gene methylation quantification studies involving a total of 12 cancer cell lines and 318 clinical samples provided data revealing a sensitivity rate of 96% for prostate cancer. High methylation exposure in prostate cancer cell lines was significantly different from that of primary tumor tissues. Additionally, methylation levels showed significantly lower levels in bladder and prostate non-tumorous tissues, providing a possible evidence for potential cancer biomarkers [87, 88]. Also, diagnostic platform may be extended and covered by using gene panels including GSTP1/ARF/CDNK2A/MGMT and GSTP1/APC/RARB2/RASSF1A for urine and GSTP1/PTGS2/RPRM/TIG1 for serum

Table 1.5: Overview of prostate cancer biomarkers

Biomarker	**Sample**	**Type**	**Diagnosis, treatment, or prognosis**
PCDh17, TCF21	Tissue	DNA methylation	Diagnosis
GSTP1, ARF, CDNK2A, MGMT	Urine	DNA methylation	Diagnosis
GSTP1, APC, RARB2, RASSF1A	Urine	DNA methylation	Diagnosis
GSTP1, PTGS2, RPRM, TIG1	Tissue (serum)	DNA methylation	Diagnosis
HOXB13	Tissue	Overexpression	Prognosis
ADAM19	Tissue	Overexpression	Treatment
SFRP1	Tissue	Decreased expression	Diagnosis, prognosis
PSF1	Tissue	Overexpression	Diagnosis, prognosis
EN2	Tissue, urine	overexpression	Diagnosis
SLC18A2	Tissue	Downregulated	Diagnosis
TRPM4	Tissue	Overexpression	Prognosis
SUX2	Tissue	Downregulated	Prognosis
XPO6	Tissue	Overexpression	Prognosis

samples.

In another study, HOXB13 showed overexpression during malignant progression of the prostatic tissue. The study also revealed an important role in the pathogenesis of the prostate gland and that it can be used as a novel biomarker for the prognosis of prostate cancer [89]. ADAM19 (a disintegrin and metalloproteinase 19) is a transmembrane and soluble protein which is linked to cell phenotype through cell adhesion and proteolysis. A study involving special immune histochemical approach showed that ADAM19 protein levels showed increased expression compared to normal prostate tissue during prostate cancer biopsies [90]. It is also reported that expression of SFRP1 shows inverse correlation with the Gleason score, survival rate and response for endocrine therapy expression, thus substantiating it as a favorable predictive and prognostic biomarker [91]. Other study groups have reported PSF1 expression in high-grade prostate cancer could be a potential biomarker to identify patients for diagnosis [92]. Engrailed-2 (EN2) protein, a homeodomain-containing transcription factor showed expression in prostate cancer. This protein is secreted in urine and shows a high specificity and sensitivity values, adding value as a novel biomarker for prostate cancer [93]. Downregulated protein like SLC18A2 and unregulated protein like TRPM4 in prostate cancer, also show similar functions as EN2 proteins.

1.7.5 Epigenetic biomarkers in testicular cancer

Recent reports have identified risk SNPs in testicular germ cell tumors (TGCT). High levels or increased PDE11A, SPRY4, and BAK1 promoter methylation and decreased KITLG promoter methylation in familial TGCT cases versus healthy male family controls was used to diagnose TGCT in the early time [94, 95]. Other groups have reported that Long Interspersed Nuclear Elements (LINE-1, retrotransposons) methylation may be gender-specific, with a strong correlation between LINE-1 methylation levels associated with disease risk (L. Mirabello, S.., et al., 2010). A knock down miR-199a-3p study, in a normal human testicular cell line (HT) showed a marked elevation of DNMT3A2 (DNMT3A gene isoform 2) mRNA and protein levels. In clinical studies, DNMT3A2 was significantly overexpressed in malignant testicular tumor and showed inverse correlation with miR-199a-3p expression [96]. Methylation profiles of oncogenes in testicular cancer shows correlation with histological types and cancer-specific genes. Furthermore, methylation analysis in a larger cohort is necessary for deciphering the complexity of gene roles in testicular cancer development and can shed light on its therapy, early detection, and disease monitoring [97].

1.7.6 Epigenetic biomarkers in gastric cancer

Gastric cancer (GC) and colorectal cancer (CRC), are the two most frequently occurring gastrointestinal tract cancer. Genetic and Epigenetic factors control initiation and progression of GastroIntestinal Cancer (GIC). DNA methylation, specific histone modifications, chromatin remodeling and noncoding RNA-mediated gene silencing, together comprise epigenetic changes and are potentially reversible and heritable [98].

Numerous gene show altered DNA methylation levels across the CRC genome. These include the genes associated with the Wnt signal transduction pathway (APC, AXIN2, DKK1, SFRP1, SFRP2 and WNT5A), DNA repair genes (MGMT, MLH1 and MLH2), Cell-cycle related genes (CDKN2A) and RAS signaling genes (RASSF1A and RASSF1B) [63, 99]. It is determined that highest CGI hyper methylation frequency takes place in GC,

using DNA methylation mapping [44, 100]. It is reported that HOP homeobox methylation can be used as a potential biomarker as it exhibited 84% of hyper methylated samples versus 10% of matched adjacent normal tissues [101]. It was also observed that, the expression of ADAMTS9 (a disintegrin and metalloproteinase with thrombospondin motifs 9 and belonging to ADAMTS family), was silenced in 75% of GC cell lines and inhibited the expression of AKT/mTOR pathway genes. This is found to be due to promoter hyper methylation [102].

Genes that are differentially methylated and can be detected in various body fluids, using can be of clinical relevance. They are useful, easily available and noninvasive biomarkers for GIC. Methyl-BEAMing for the absolute quantification of methylated molecules in DNA from plasma or fecal samples is one such method developed for identification of clinically relevant DNA methylation biomarker [103, 104]. A pioneer study reported using blood-based PCR tests to detect the presence of the methylated septin 9 gene in CRC patients had a sensitivity and a specificity of nearly 90% [105]. Such an approach wherein, CRC screening test via blood-based, using the methylated SEPT9 biomarker (septin 9), (encoding a GTPase involved in dysfunctional cytoskeletal organization) specifically detects the majority of CRCs at all stages and locations in the colorectal region. The test showed an overall sensitivity of 90% and a specificity of 88% [106].

Likewise, stool-based test for detecting gene methylation that codes for vimentin, when conducted with colonoscopy exhibits a degree of sensitivity for CRC that ranges from 40 to 80% [106]. Other stool-based tests developed for CRC diagnosis and to detect clinically relevant hyper methylated genes are targeted towards those that encode for fibrillin-1, APC, CDKN2A, MLH1, MGMT, SFRP1, SFRP2 and NDRG4. Their levels of sensitivity that range from 60 to 80% [107, 108]. Recent reports have indicated that TFPI2 is expressed in almost all colorectal adenomas (97%, n = 56) and stage I to IV CRCs (99%, n = 115). Also, DNA-based stool assays have been used from I-III CRC stages and showed a sensitivity of 76-89% and a specificity of 79-93%. This suggests that TFPI2 methylation levels in stool DNA samples can be a potential noninvasive biomarker for the early screening of CRC [109].

Table 1.6: Selected genes with promotor hyper methylation and their clinical correlations in ovarian carcinomas

Genes	Clinical correlations	Ref(s).
RASSF1A	Detection of OC	81, 82
BRCA1	Detection of OC; poor prognosis; improved chemotherapy response	81, 83, 84
APC	Serum/ascites diagnosis of OC	81, 83
MGMT	Detection of OC; improved chemotheraphy response	83, 85
hMLH1	Poor prognosis; improved chemotheraphy response	86-88
HOXA9	Detection of OC	89
OPCML	Detection of OC	90
SFRP-1, -2, -4, -5	Detection of OC; Cancer recurrence; Poor prognosis	91
FZD4, DVL1, NFATC3, ROCK1, LRP5, AXIN1, and NKD1	Poor prognosis	92
FBXo32	Poor prognosis	93
HOXA11	Poor clinical outcome	94
FANCF	Cisplatin resistance	95

Another research group have suggested that, rather than detecting a single methylated gene, sensitivity of stool DNA testing when combined with a panel of different biomarkers for the detection of CRCs showed an increase up to 92.3%. This combined screening approach including the panel of methylated genes is under evaluation for improving sensitivity and also specificity [110].

1.7.7 Epigenetic biomarkers in Ovarian Carcinoma

Ovarian carcinoma (OC) is reported to be the most lethal gynecological malignancy worldwide. Most OCs fall under a category of high grade serous ovarian carcinomas (HGSOC). Common diagnosis occur in advanced stages involving peritoneal dissemination and massive ascites. Advanced OC patients survival rate is ∼30%, even after administered with standard combined therapy of debulking surgery and neoadjuvant chemotherapy of paclitaxel and carboplatin [111]. Epigenetic biomarkers, particularly DNA methylations, have proven to be highly beneficial in terms of clinical utility for detection/diagnosis, chemotherapy response and prognosis in OC (Table 7)[111, 112].

It has been reported that, epigenetic regulation of Wnt and Akt/mTOR pathways may be utilized as biomarkers for prognosis and/or treatment response in OC [113]. In another study, examination of promoter methylation at 302 loci in a panel of 137 Wnt pathway genes in 111 screening cases and 61 validation cases showed that methylations at 7 loci

(FZD4, DVL1, NFATC3, ROCK1, LRP5, AXIN1, and NKD1) were associated with poor progression-free survival. Also, hypermethylations of DVL1 and NFATC3 responded very poorly to platinum chemotherapy [114]. Also, hypermethylations of DVL1 and NFATC3 showed similar poor response to platinum chemotherapy. Additionally, subjects with progressive or stable disease had increased methylation levels compared to those with partial or complete response. The same research group reported that, promoter methylations of VEGFB, VEGFA, HDAC11, FANCA, E2F1, GPX4, PRDX2, RAD54L and RECQL4 were associated with increased hazard of disease progression. This was independent from conventional clinical prognostic factors both in the screening cohort (n=150) and the TCGA validation cohort (n=311). Furthermore, methylations at VEGFB and GPX4 showed poor response to chemotherapy [114]. Next, a diagnostic model was built using methylation profile in the previous Wnt pathway and the methylations of NKD1, VEGFB and PRDX2, from which, methylation index was calculated to identify two distinct prognostic groups. Subjects with increased methylation index exhibited a very poor response to chemotherapy.

Studies have also been conducted using genome-wide identification of methylated biomarkers in OCs. A approach known as methylated DNA immunoprecipitation microarray (MeDIP-chip) was able to identify 367 CpG islands specifically methylated in OC, compared to normal ovaries [115]. 168 genes are reported to be epigenetically silenced (Nature report. 2011; 474:609615). Three genes AMT, CCL21, and SPARCL1 exhibits promoter hyper methylation in most cancers, including OC and may serve as biomarkers for the presence of OC. Four subtypes was generated upon consensus clustering of methylations across tumors, with valid prognostic differences. Genome wide associated studies (GWAS) studies in OC have yielded methylation signatures associated with progression-free survival [116, 117]. Methylation analysis across genome-wide can potentially identify biomarkers of good prognostic value. 220 differentially methylated regions were identified, in tumor tissue of patients with short vs. long progression-free survival (106 hypo- and 114 hyper methylated regions) using genome-wide array analysis approach [118]. This was validated when subjects harboring methylation at the CpG island of RUNX3/CAMK2N1 had a significantly

lower progression free survival. Such identified biomarkers using genome screening needs to be further investigated in large cohorts supplemented by good clinical documentation.

NOTE: A comprehensive coverage on the current status, discovery and development of all cancer DNA methylation biomarker is beyond the scope of this dissertation work. An attempt is made above to highlight the past, current and on-going discovery and development of the same.

1.8 Need for DNA methylation biomarker discovery

DNA methylation biomarkers have currently advanced to diagnostic laboratories, particularly those which are being used for early stage cancer detections. Lack of standardized methodologies and inconsistent reference standards for detection of valuable biomarkers are the biggest challenges that needs to be overcome today. Inappropriate methodologies involving inappropriate controls are leading to non-replicating results which is hampering biomarker discovery and development. Quantitative DNA methylation detection is the need of the hour and is critical in cases, where only small differences in methylation values determine a diseased or disease-free state. Also, DNA methylation biomarkers in non-cancer related disorders will greatly benefit from the valuable knowledge and results obtained from cancer related studies.

1.9 Future prospects

Standardization of appropriate methods intended towards DNA methylation detection and building reliable reference standards will accelerate the discovery as well as the development of DNA methylation biomarkers for cancer and other disorders. Next generation sequencing has added immense value in this direction, allowing for routine testing of DNA methylation biomarker panels rather than the selective choice of individual biomarkers. This is greatly helpful in cases where disease phenotype exists in quite heterogeneous state. Additionally, genetic disease components will be revealed allowing the validation and strengthening of

biomarker panels by combining genetic and DNA methylation biomarker panels [119]. Future research will not only focus on detection of appropriate epi (genetic) biomarker panels available for diseases or risk stratification but also to translate them into clinical actionable information with substantial validation. Translational approach is utmost important in this context as there is a risk of adverse psychological impacts among patients. There also exists risk of those patients being disadvantaged by healthcare providers. However, those denied for healthcare or affected patients can avail the knowledge to their benefit. This will allow them to actively prevent or delay the early onset of certain diseases, upon early detection.

Chapter 2: MATERIALS AND METHODS

2.1 The Cancer Genome Atlas (TCGA) overview

Overview on TCGA: The Cancer Genome Atlas (TCGA) is a public funded project. The primary objective of this initiative is to discover and catalogue genomic alterations. These catalogued data are used to create a comprehensive Atlas of cancer genomic profiles. To this day, TCGA initiative has analyzed over 30 large cohorts of human tumors using large-scale genome sequencing and integrated multi-dimensional analyses. Also, specific cancer type studies and comprehensive pan-cancer analyses have been enriched from TCGA cancer research initiative. The main goal of this TCGA cancer initiative is to provide publicly available datasets in order to help improve diagnostic methods, treatment standards, and finally to prevent cancer.

In 2005, The Cancer Genome Atlas (TCGA) and in 2008 the International Cancer Genome Consortium (ICGC) were launched. These two main projects aims at accelerating the comprehensive understanding of cancer genetics. This would be achieved through the use of innovative genome analysis technologies and thus would help to generate new cancer therapies, diagnostic methods, and preventive strategies. TCGA was set up in phases. Aim of Phase 1 was to develop and test the research infrastructure which was based on characterization of tumors with poor prognosis. This included brain, lung and ovarian cancers. This was a 3-year pilot study. Phase 2 study started in 2009. The study expanded to additional cancer types and covered 30 tumor types. The analysis was completed in 2014.

Table 2.1: The Cancer Genome Atlas (TCGA) organization centers

Centre Name	Centre Description	Localization
Tissue Source Sites (TSSs)	Collection of the samples (blood and tissue from tumor and normal controls) and clinical metadata from patients (donors) Shipment of the annotated bio specimens to Bio specimen Core Resources (BCR) https://wiki.nci.nih.gov/display/TCGA/Tissue+Source+Site	https://tcgadata.nci.nih.gov/datareports/codeTablesReport.htm?codeTable=tissue%20source%20site
Bio specimen Core Resource (BCR)	Coordination of sample delivery and data collection, cataloguing, processing, and verifying the quality and quantity Isolation and distribution of RNA and DNA from bio specimens to other institutions for genomic characterization and high-throughput sequencing http://cancergenome.nih.gov/abouttcga/overview/howitworks/bcr http://www.nationwidechildrens.org/biospecimen-core-resource-about-us	Research Institute at Nationwide Children's Hospital in Columbus, Ohio
Genome Sequencing Centers (GSCS)	High-throughput sequencing (data are available in TCGA Data Portal or at NIH's database of Genotype and Phenotype) Identification of the DNA alterations http://cancergenome.nih.gov/abouttcga/overview/howitworks/sequencingcenters	Broad Institute Sequencing Platform in Cambridge Human Genome Sequencing Center, Baylor College of Medicine in Houston The Genome Institute at Washington University

2.1.1 The Cancer Genome Atlas (TCGA) Data collection and Research Network

TCGA is well-structured and is supported by cooperating centers which are responsible for collection and sample processing. This is then followed by high-throughput sequencing and sophisticated bioinformatics data analyses (Table 8) [120] (Figure 2.1).

The Cancer Genome Atlas (TCGA) organization centers

At first, various Tissue Source Sites (TSSs) collects the bio-specimen/samples (blood, tissue etc.) from eligible cancer patients. It is then delivered to the Bio-specimen Core Resource (BCR). The BCR then catalogues, processes and verifies the sample (quality and quantity). It then submits clinical data and metadata to the Data Coordinating Center (DCC). It also provides molecular analytes for the Genome Characterization Centers (GCCs) and Genome Sequencing Centers (GSCs) for further genomic characterization and high-throughput sequencing. At this point, sequence-related data are deposited with DCC. The GCC also submits trace files, sequences and alignment mappings to NCIs Cancer Genomics Hub (CGHub) secure repository. Such compiled data source is made publicly available to the research community and Genomic Data Analysis Centers (GDACs). The role of GDACs is to process new information, its analysis and provide visualization tools for

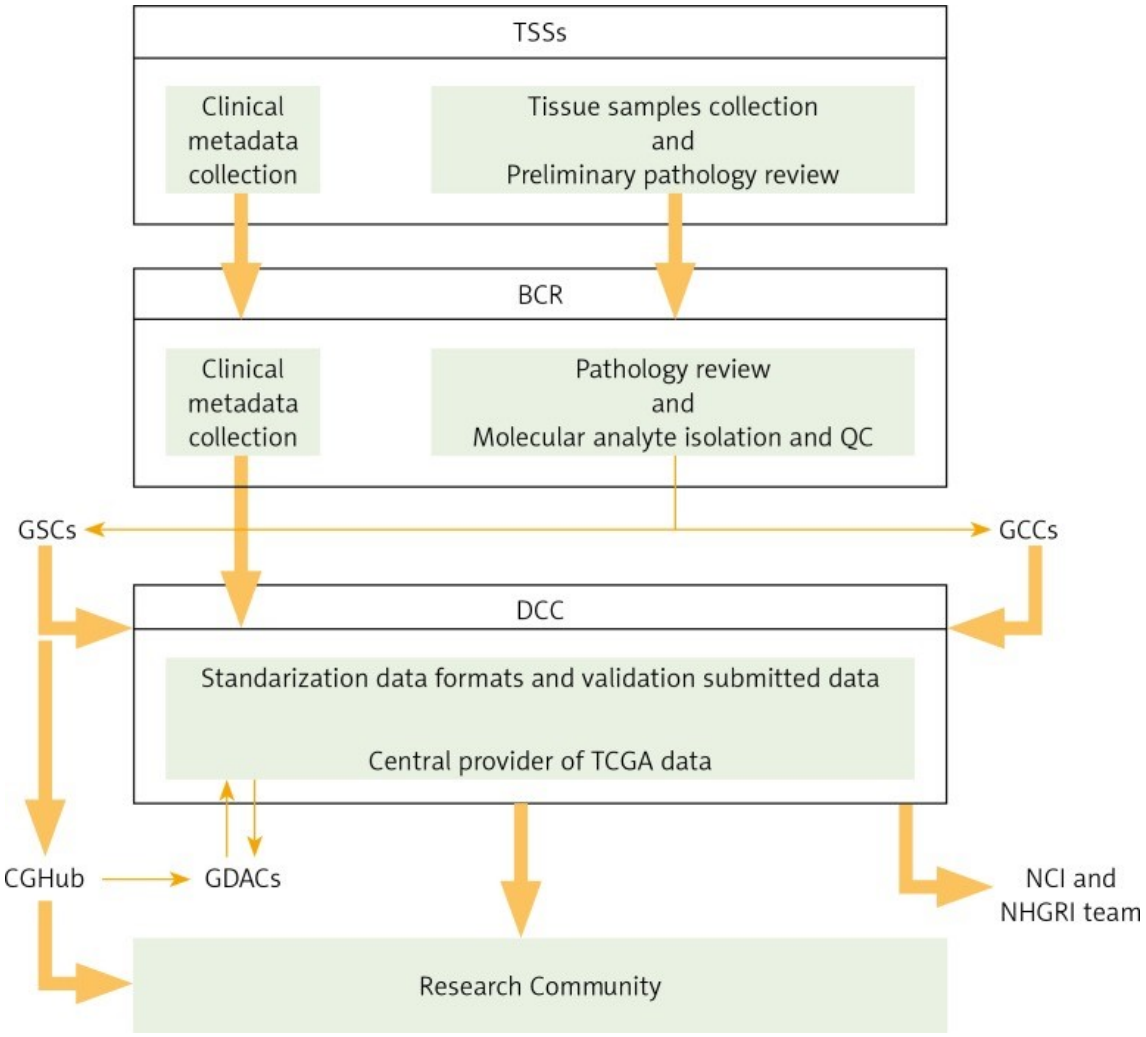


Figure 2.1: The Cancer Genome Atlas (TCGA) Research Network Centers flowchart.

a wider audience. DCC also is a central management center for the entire data generated by TCGA. DCC feeds the data into public free-access databases[1] (TCGA Portal, NCBI's Trace Archive, CGHub).

[1] http://cancergenome.nih.gov/abouttcga/overview

2.1.2 TCGA platform and data types.

TCGA has extensively used high-throughput technologies based on microarrays (to test nucleic acids and proteins) and next-generation sequencing methods (for genome-wide analysis of nucleic acids). Also, the TCGA research network structure is supported by centers that utilizes different platforms to provide a comprehensive cancer genomics data. Some of the applied methods includes: RNA sequencing (RNAseq), MicroRNA sequencing (miRNAseq), DNA sequencing (DNAseq), SNP-based platforms, Array-based DNA methylation sequencing and Reverse-phase protein array (RPPA).

RNA sequencing (RNAseq): This is a high-throughput technology for transcriptome (total RNA) profiling. This can obtain strand information with excellent precision. RNAseq can quickly and efficiently identify and quantify novel transcripts, isoforms, common transcripts, gene fusions and non-coding RNAs from numerous samples, even if the samples are of low quality grade [121]. TCGA utilizes Illumina system for transcriptome analysis. Submitted data contains information pertaining to nucleotide sequence and gene expression. RNA sequence alignment provides a comprehensive information including RNA sequence coverage, sequence variants (like fusion genes), gene expression, exon and/or junction. dbGaP database[2] from NCBI is the repository database for the actual sequence data
.

MicroRNA sequencing (miRNAseq): This is a RNA-seq method that utilizes materials enriched in small RNAs and will thus allow the identification and detection of specific sets of short, noncoding RNAs (miRNAs). These miRNAs can regulate numerous genes within and across diverse signaling pathways. Furthermore, miRNA-sequencing is useful for defining tissue-specific miRNA expression profiles, their isoforms, relatedness to disease and discovery of novel miRNAs [122–124].

DNA sequencing (DNAseq): This is a high-throughput method for identifying or determining nucleotides in a DNA molecule. It provides valuable information about DNA

[2]https://wiki.nci.nih.gov/display/TCGA/RNASeq

alterations, like insertions, deletions, polymorphism, copy number variations, mutation frequencies or cellular events like viral infections. Genomic diversity across many cancer types is catalogued by TCBGA using Sanger Sequencing technique at the TCGA Genome Sequencing Centers [125, 126, 126].

SNP-based platforms: These platforms are utilized for analysis of genome-wide structural variations across numerous cancer genomes. A wide pool of powerful genotypic tool sets are made use of for this purpose. Single Nucleotide Polymorphisms (SNPs) detected using an array-based detection includes platforms that can define SNP, CNV and loss of LOH across multiple samples[3] [127].

Array-based DNA methylation sequencing: This is a high-throughput, genome-wide analysis of DNA methylation profiling. This provides changes in epigenetics in the genome. The most common and the earliest alterations in cancer is abnormal profiles of DNA methylation of CpG sites[4] [128] and Illumina is the main platform utilized by TCGA for DNA methylation assay. This platform ensures single-base-pair resolution, high accuracy, easy workflows and low DNA input requirements. Methylation profiling approaches or methodologies are based on highly multiplexed genotyping of bisulphite-converted genomic DNA. TCGA provides the DNA methylation data files which includes signal intensities (both raw and normalized), confidence of detection, and calculated beta values for methylated (M) and unmethylated probes[5] (U).

Reverse-phase protein array (RPPA): This is a high-sensitivity, reproducible, high-throughput, functional and quantitative proteomic method. This method can detect nanograms of proteins. This is used for large-scale protein expression profiling, biomarker discovery and also for cancer diagnostics. RPPA is based on the antibody-based technique and allows for analysis of more than 1000 samples at any given instance. It also includes 500 different antibodies at the same time. Protein arrays contains information pertaining to both expression and concentration. TCGA submits protein array data to DCC. Such data also

[3] http://www.broadinstitute.org/collaboration/gcc/methods/technology
[4] http://res.illumina.com/documents/products/datasheets/datasheet_dna_methylation_analysis.pdf
[5] https://wiki.nci.nih.gov/display/TCGA/DNA+methylation

includes original images of protein arrays, its raw signals, relative protein concentrations and normalized protein signals[6].

Many data types (kinds or variety in data) can be obtained from each platform. These include gene expression, exon expression, miRNA expression, copy number variation (CNV), single nucleotide polymorphism (SNP), loss of heterozygosity (LOH), mutations, DNA methylation, and also protein expression. Such data obtained are categorized not only by data type but also by data levels. Raw, unnormalized data (Level 1), processed data (Level 2) and segmented or interpreted data (Level 3) from each individual samples and summarized data (level 4) is the data that refers to analysis across sample sets. Level 3 and 4 data is publicly available data, whereas Level 1 and 2 will need special permission for accession[7].

2.1.3 Analysis and visualization of TCGA data

Next-generation sequencing (NGS) and Array-based profiling yields vast amounts of diverse data types. This provides a good platform for cancer genome analysis. Data interpretation and visualization that involves integration and multi-dimensional data is utmost essential. Hence, the need for advanced visualization tools has emerged quite drastically. Various useful imaging tools and databases are now employed for cancer genome analysis[8]. This includes:

The Cancer Imaging Archive, TCIA[9]: This was created by NCI to collect and share numerous medical images of cancer (radiological imaging data), from TCGA cases for public use. In short, it supports the imaging phenotype-genotype research [128].

Berkeley Morphometric Visualization and Quantification from H & E sections[10]: This is a repository for data pertaining to histology-based images of various tumor

[6] http://www.mdanderson.org/education-and-research/resources-for-professionals/scientific-resources/core-facilities-and-services/functional-proteomics-rppa-core/index.html

[7] https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp

[8] https://tcga-data.nci.nih.gov/tcga/tcgaAnalyticalTools.jsp

[9] http://www.cancerimagingarchive.net

[10] http://tcga.lbl.gov/biosig/tcgadownload.do

samples of TCGA cases. This is supported by the Lawrence Berkeley National Laboratory (Chang H., et al., 2013).

The Cancer Digital Slide Archive, CDSA[11]**:** This is an online tool built for the purpose for viewing and annotating diagnostic and tissue slide images of various tumor types from TCGA project. Its creators are Dr. David Gutman and Dr. Lee Cooper of Emory University who have facilitated a broader access to TCGA data [129].

The Broad GDAC Firehose[12]**:** This was created by Broad Institute. It coordinates the smooth flow of datasets in the order of terabyte-scale, thus providing a large amount of different quantitative algorithms including GISTIC, MutSig, Clustering and Correlation [13].

The MD Anderson GDAC's MBatch[14]**:** This platform is useful in identification and quantification of the batch effects accompanying the TCGA data sets. This is in accordance to hierarchical clustering and enriched PCA plots[15].

Cancer Genome Workbench, CGWB[16]**:** This was developed by NCI, to provide an integrative platform and also for displaying sample-level genomic and transcription alterations in various cancers. Major views on this platform are Integrated tracks view, Heat map view and Bambino (Alignment viewer) [130].

UCSC Cancer Genomics Browser[17]**:** An important platform wherein users can find an open-access, web-based tools developed and supported by USCS Cancer genomics Group. This is used to visualize and analyze cancer genome combined with clinical data by using genomic coordinate heat maps. The site provides interactive visual outputs of genomic regions. This is supplemented with annotated cellular pathways and also allows for quantitative analysis for all datasets and integrates with statistical tools also [131].

Integrative Genomics Viewer, IGV[18]**:** Freely available high-performance visualization tool from Broad Institute. Its purpose is to provide interactive exploration of large,

[11] http://cancer.digitalslidearchive.net/
[12] https://confluence.broadinstitute.org/display/GDAC/Home
[13] http://www.broadinstitute.org/cancer/cga/Firehose
[14] http://bioinformatics.mdanderson.org/tcgabatcheffects
[15] https://wiki.nci.nih.gov/display/TCGA/MD+Anderson+GDAC+MBatch
[16] https://cgwb.nci.nih.gov/
[17] https://genome-cancer.soe.ucsc.edu/
[18] http://www.broadinstitute.org/igv

heterogeneous, integrated data sets. IGV provides a platform for easy analysis of user-friendly data or for those from IGV server and TCGA data too. It has coordinate-type data that provides genome annotations with specific labels for viewing genomes.

The cBioPortal for Cancer Genomics[19]: This is offered by the Memorial Sloan-Kettering Cancer Centre (MSKCC). It provides for the visualization, analysis, and download of large-scale cancer genomics data sets. Also, it allows for interactive exploration of customs datasets. This is done by direct accession to OncoPrinter or MutationMapper web tools. This site presently holds data from 69 cancer genome studies including data such as DNA copy-number data, mRNA and miRNA expression data, mutations, RPPA data, DNA methylation data, and limited clinical data related to survival. Visualization interface involves networks, matrices as well as heat maps. This site highly compliments existing tools from TCGA and ICGC data portals, IGV, USCS genome browser and also IntOGen [132, 133].

Regulome Explorer[20]: This portal allows for integrative exploration of relations or associations between molecular features and clinical aspects of TCGA data. This allows users to search and visualize data by applying suitable filters. The visualized data may include either circular or linear genomic coordinates or networks. This explorer is supported by the Center for Systems Analysis of the Cancer Regulome (CSACR), associated with the TCGA initiative and also with the Institute for Systems Biology and The University of Texas MD Anderson Cancer Center [134].

2.1.4 Data mining the vast TCGA resource.

Cancer types with data available via The Cancer Genome Atlas

All TCGA data is made publicly available and centralized at the TCGA data portal. TCGA data has been utilized for various analysis, including a study to characterize the genomic and molecular landscape of various cancer types and their respective analysis. One such analysis includes that of exome sequencing, RNAseq and MiRNAseq across 12

[19] http://cbioportal.org
[20] http://explorer.cancerregulome.org/

Table 2.2: Cancer types with data available via The Cancer Genome Atlas

Available Cancer Types	**# Cases Shipped by BCR***
Acute Myeloid Leukemia [LAML]	200
Adrenocortical carcinoma [ACC]	80
Bladder Urothelial Carcinoma [BLCA]	412
Brain Lower Grade Glioma [LGG]	516
Breast invasive carcinoma [BRCA]	1100
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	308
Cholangiocarcinoma [CHOL]	36
Colon adenocarcinoma [COAD]	461
Esophageal carcinoma [ESCA]	185
FFPE Pilot Phase II [FPPP]	38
Glioblastoma multiforme [GBM]	529
Head and Neck squamous cell carcinoma [HNSC]	528
Kidney Chromophobe [KICH]	66
Kidney renal clear cell carcinoma [KIRC]	536
Kidney renal papillary cell carcinoma [KIRP]	291
Liver hepatocellular carcinoma [LIHC]	377
Lung adenocarcinoma [LUAD]	521
Lung squamous cell carcinoma [LUSC]	510
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC]	48
Mesothelioma [MESO]	87
Ovarian serous cystadenocarcinoma [OV]	586
Pancreatic adenocarcinoma [PAAD]	185
Pheochromocytoma and Paraganglioma [PCPG]	179
Prostate adenocarcinoma [PRAD]	498
Rectum adenocarcinoma [READ]	172
Sarcoma [SARC]	261
Skin Cutaneous Melanoma [SKCM]	470
Stomach adenocarcinoma [STAD]	445
Testicular Germ Cell Tumors [TGCT]	150
Thymoma [THYM]	124
Thyroid carcinoma [THCA]	507
Uterine Carcinosarcoma [UCS]	57
Uterine Corpus Endometrial Carcinoma [UCEC]	548
Uveal Melanoma [UVM]	80

*Excludes non-canonical cases

cancer types revealed 11 major subtypes and redefined three cancer types into one molecular subgroup (Hoadley KA., et al., 2014). In another analysis, exome sequencing and RNAseq data for six cancer types was used to discover neo-antigen expression and to predict patient survival rates [135].

Data types for each cancer types include somatic mutation, copy number, gene expression, miRNA expression, DNA methylation, reverse protein phase array (RPPA) and clinical information. As mentioned before, each data type has raw and processes data. Exceptions to this rule though are for sequencing files from the exome sequencing, RNA sequencing (RNAseq), microRNA sequencing (miRNAseq) and copy number, which require authorization from the Cancer Genomics Hub (CGHub). Also, pipeline for analysis for each data

type are provided in a text file. These contain a standard method for raw processing of data and annotation. This allows for reproducibility in downstream analysis.

2.1.5 Analysis of TCGA data using publicly available web tools.

Web tools such as cBioportal, GDAC firehose Websites such as cBioportal [132], GDAC firehose[21], canEvolve [136], PROGgeneV2 [137], and the UCSC cancer browser [131] all provide their own analysis and visualization tools for TCGA datasets canEvolve , PROGgeneV2 , and the UCSC cancer browser all provide their own analysis and visualization tools for TCGA datasets (Table 10).

Web tools for TCGA analysis

cBioportal site contains >20 000 tumor samples from 89 cancer studies. Users can select datasets and enter a gene list. This site is invaluable as it offers unique analysis pipeline such as OncoPrint diagrams, MEMo (Mutual Exclusivity Modules) analysis, customizable correlation plots, Kaplan-Meier plots, network analysis and integrative genomics viewer integration. Oncoprint diagrams represent genomic alterations such as somatic mutations and copy number alterations across sample sets. Users are able to detect visually the visually co-occurrence or mutual exclusivity of genomic alternations within a cohort. MEMo analysis helps in identifies gene mutations, that share a common pathway and that exhibits a mutually exclusive mutations pattern across a cohort [138]. Using cBioportal analysis of RNAseq and RPPA data types can be done by setting z-score thresholds for identifying significant genes and proteins, respectively. Cytoscape, a tool for network analysis integration allows for viewing gene networks and their corresponding interactions for the gene/s of interest. Integrative genomics viewer (IGV), can be used for visualizing copy number alterations (CNA), gene expression and mutations across all chromosomes genome-wide.

Annotated data can be preprocessed using GDAC firehose and provides correlations and differential gene analysis in all data types. The firehose platform periodically updates new TCGA cases and automates pipelines every four months. GDAC firehose includes unique

[21] http://gdac.broadinstitute.org

analysis pipelines that include GISTIC2 (analysis of copy number data) [139] MutSig2 (analysis of mutation data) [140] and PARADIGM (analysis of copy number and RNAseq data) [141]. GDAC site also correlates clinical data with miRNA, mRNA, RPPA, copy number and DNA methylation datasets. Clustering analysis can be performed for data types and molecular subtypes can be defined. This can then be correlated to clinical, mutation and copy number data.

canEvolve site contains >10 000 tumor samples from 90 cancer studies, including 15 TCGA data sets. Users can select a database and can conduct multiple downstream analysis including differential gene expression, miRNA expression, copy number analysis, regulatory network analysis using ARACNE, co-expression network analysis using WGCNA, gene set enrichment analysis using the MSigDB 3.0 gene sets (Subramanian A., et al., 2005), integrative gene expression and miRNA expression analysis using GemiNI, integrative gene expression and copy number analysis using DR-Integrator, integrative genomic and gene expression analysis, integrative genomic and protein expression analysis and survival analysis. canEvolve can also be used to query genes across multiple datasets. Users can select a predefined gene list from KEGG or Biocarta pathway or a user-end gene list for interrogating gene expression patterns within any given dataset.

PROGgene V2, second version of PROgene, contains >19 000 samples from 134 cohorts in 21 cancer types. This tool provides for analysis on survival rates based on one gene or ration between two genes. Survival plots are generated using gene signatures from KEGG, Biocarta, GO, and Reactome databases. Covariate data like cancer stages can be adjusted for survival plots. Also, unique feature of this site is that, users can upload their own data here. Omics data is now gaining popularity as it does not require programming experience and hence this above mentioned feature will become increasingly important to help comprehend as to how individual patient data compares with larger cohorts. UCSC cancer browser like other tools provides for visualization and analysis for TCGA data. However, it offers a unique interactive analysis of multiple datatypes for a cancer dataset. Cancer data set can be selected to visualize gene expression or DNA methylation, stratified according

to clinical parameters or another dataset. For instance, users can select a mutated genes dataset and stratify according to clustering of miRNA and DNA methylation signatures, allowing users to define cancer specific subgroups and also perform survival analysis. User can also specify genes or gene signatures to visualize within a dataset. End user annotations can be uploaded to clinical heat map for specific clustering analysis. New platforms like Xena platform for visualization and integration with Galaxy is being introduced [142, 143].

2.1.6 Future promise/perspective from TCGA.

TCGA has provided new insights into the molecular biology of cancer and into cancer genomics. Advances in bioinformatics tools and high-throughput technologies has highlighted the intricate similarities and differences in the genomic architecture of cancer and its relevant subtypes, which is publicly available. Immeasurable and invaluable data is now made publicly available with regards to genetic and epigenetic profiles, highlighting candidate cancer biomarkers and drug targets. Also, personalized medicine can benefit immensely from translation of cancer genomics into therapeutic prospects. On the bioinformatics front, it is essential that the tools eliminate potential noise and improve upon resolution of analysis, and identify or discover biomarkers or therapeutic targets from those refined data sets. Such novel discoveries will aid in the medicine community in diagnosis, treatment and cancer prevention. Progress is being made analysis and disease knowledge resulting in advances in medicine. Recent medical advances include a machine learning approach being taught to an artificially intelligent computer WATSON in order to support doctors in diagnosing patients[22] [23].

[22] http://www3.mdanderson.org/streams/FullVideoPlayer.cfm?xml=cfg%2FMoon-Shots-IBM-Watson-2013

[23] http://www.ibm.com/smarterplanet/us/en/ibmwatson/index.html

2.2 TCGA Data: Genomic Data Commons (GDC)

The TCGA Data Portal for data downloading or public access is no longer operational and all TCGA data now resides at the Genomic Data Commons. GDC Data Model Components can be represented as a graph containing nodes and edges. This is the data store for the GDC. Critical relationship between projects, cases, clinical and molecular subdata is maintained and linked precisely to the actual data file using unique identifiers. It is based on the property graph model wherein nodes represent entities, edges between nodes represent relationships between entities and finally, properties on both nodes and edges represent additional data that describes entities and their relationships. Further, relationships are encoded as edges of a given type which associates exactly two nodes. Properties of relationships or nodes are actually sets of key-value pairs. Metadata are submitted by external users and is extracted and loaded into the graph. Data representation provided by other GDC components are derived from authoritative graph model. Files and archive objects are not stored in the graph. They are stored in an external object store. Structure of node/edge of the graph is depicted in (Figure 2.2)(Figure 2.3) GDC Data Model is a centralized method of organization, wherein all data artifacts are ingested by the GDC. Such a data model is designed to maintain data and metadata consistency, integrity, and availability while accommodating the following:

- Bio-specimen , clinical, and cancer genomic data and metadata
- Multiple, disparate NCI ongoing projects
- Completely new, as yet unthought of projects
- Ongoing changes and technological progress
- Frequent and complex queries from both external users and internal administrators

To meet such stringent requirements, the design and implementation of the data model leverages:

- Flexible but robust graph-oriented data stores
- Indexed document stores for API and front end performance
- Ontology-based concept and data element definition
- Schema-based entity and relationship validation on loading

2.2.1 GDC: Data Types and Format.

Submitted Data:

DNA and RNA sequencing data is being accepted by GDC in both FASTQ (link is external) and BAM (link is external) formats. Sequencing data is to be submitted with accompanying metadata in either simple tab-separated values (TSV) or the JavaScript Object Notation JSON format (link is external), or the latest version (currently 1.5) of the SRA XML format. Clinical and bio-specimen data is to be submitted in either TSV or JSON format, or as XML. This should be validated with respect to the latest version of NCI Bio-specimen Core Resource XML Schema documents.

GDC: Data Types and Format: Submitted Data

Generated Data:

For every submitted sequence data (also BAM alignment files), the GDC generates new alignments in BAM format using the latest human reference genome GRCh38 with standard alignment pipelines. Using these standard alignments, the GDC generates high level derived data which includes normal and tumor variant and mutation calls in VCF and MAF formats, and gene and miRNA expression and splice junction quantification data in TSV formats.

GDC: Data Types and Format: Generated Data

Imported Data: GDC also hosts and distributes previously generated data from The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective

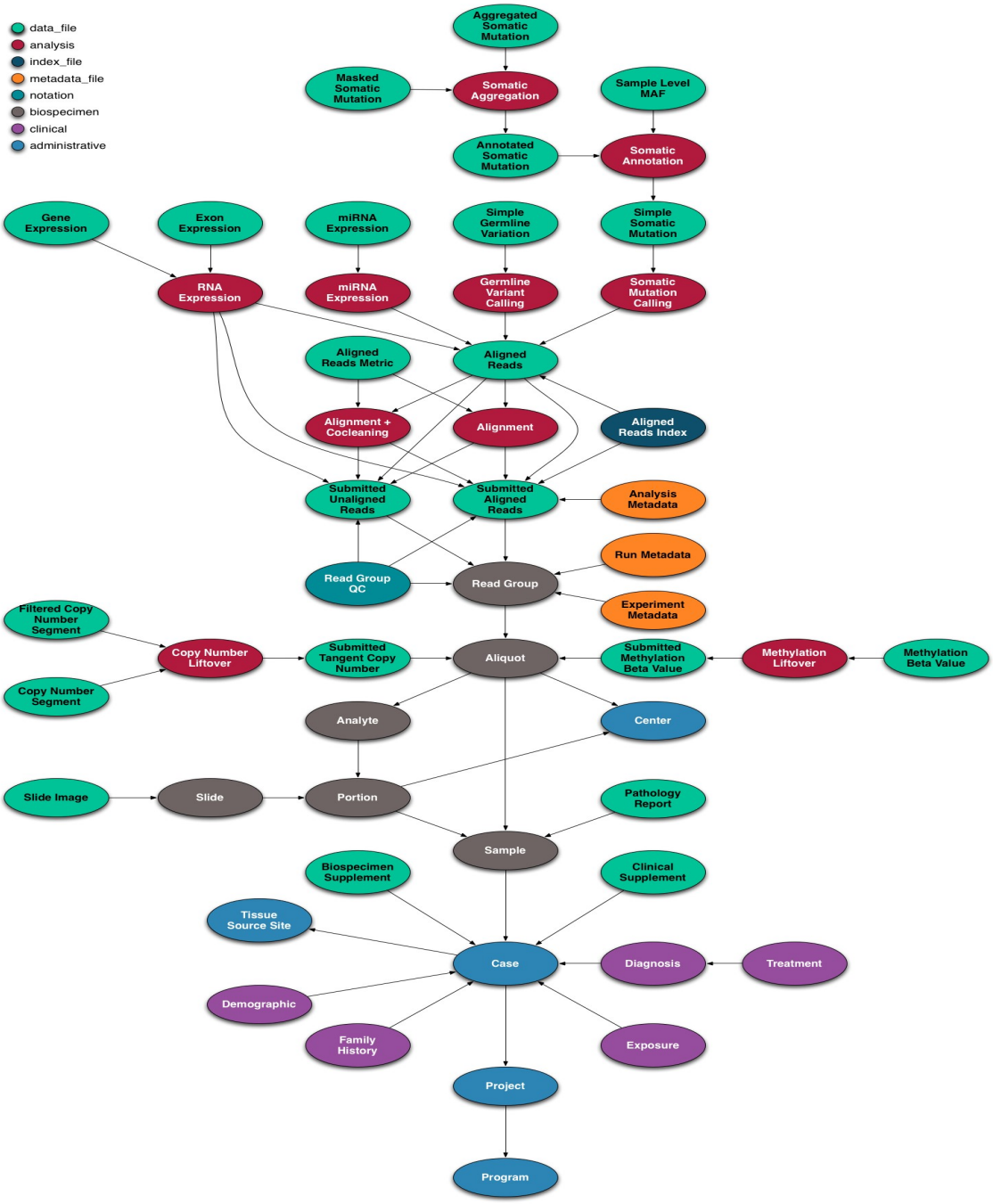


Figure 2.2: Graph Representation of the GDC Data Model

Table 2.3: GDC: Data Types and Format: Generated Data

Entity Category	Entity Name	File Format	File Metadata Template
Administrative	Case	–	TSV, JSON
Biospecimen	Sample	–	TSV, JSON
	Portion	–	TSV, JSON
	Analyte	–	TSV, JSON
	Aliquot	–	TSV, JSON
	Read Group	–	TSV, JSON
	Slide	–	TSV, JSON
Clinical	Demographic	–	TSV, JSON
	Diagnosis	–	TSV, JSON
	Exposure	–	TSV, JSON
	Family History	–	TSV, JSON
	Treatment	–	TSV, JSON
Data File	Analysis Metadata	SRA XML, MAGE-TAB (SDRF, IDF)	TSV, JSON
	Biospecimen Supplement	BCR XML, GDC-approved spreadsheet	TSV, JSON
	Clinical Supplement	BCR XML, GDC-approved spreadsheet	TSV, JSON
	Experiment Metadata	SRA XML	TSV, JSON
	Pathology Report	PDF	TSV, JSON
	Run Metadata	SRA XML	TSV, JSON
	Slide Image	SVS	TSV, JSON
	Submitted Unaligned Reads	FASTQ, BAM(link is external)	TSV, JSON
	Submitted Aligned Reads	BAM(link is external)	TSV, JSON

Treatments (TARGET), and other cancer initiative programs. Original sequence alignments are stored in BAM format, and derived data files are stored and provided in their original formats.

2.3 Methylation analysis and MExpress tool

DNA methylation is now established to be an integral aspect of cancer genomics. This is also reported to have important associations with gene expression, sequence and copy number variations [144]. Large datasets from TCGA is a validation platform with regards to identifying novel biomarkers. It is becoming a standard tool for biomarker research. Also, a significant feature in TCGA platform is the ability to correlate different data types. Recent research has indicated that promoter DNA methylation can influence gene expression and aberrant methylation is found in almost every cancer [145]. This ability for comparing data types is extremely important for identifying novel DNA methylation biomarkers. In view of such a valid, invaluable and vast platform of huge cancer datasets being available for analysis, interactive data visualization tools are critical to understand, especially when

Table 2.4: GDC: Data Types and Format: Imported Data

Data Type	Data Subtype	Format
Raw Sequencing data	Aligned Reads	BAM(link is external)
	Unaligned Reads	FASTQ(link is external)
	Coverage WIG	WIGGLE(link is external)
Simple Nucleotide Variation	Genotypes	TSV
	Simple Germline Variation	MAF, VCF
	Simple Somatic Mutation	
	Simple Nucleotide Variation	
Raw Microarray Data	Raw Intensities	TSV
	CGH Array QC	
	Intensities Log2Ratio	
	Expression Control	
	Intensities	
	Normalized Intensities <	
	Probeset Summary	
	Methylation Array QC Metrics	
Gene Expression	Gene Expression Quantification	TSV
	miRNA Quantification	
	Isoform Expression Quantification	
	Exon Junction Quantification	
	Exon Quantification	
	Gene Expression Summary	
Structural Rearrangement	Structural Germline Variation	VCF, FASTA
	Structural Variation	
DNA Methylation	Bisulfite Sequence Alignment	BAM(link is external)
	Methylation Beta Value	TSV
	Methylation Percentage	
Clinical	Clinical Data	XML
	Biospecimen Data	
	Tissue Slide Image	SVS
	Diagnostic Image	
	Pathology Report	PDF
Copy Number Variation	Copy Number Segmentation	TSV
	Copy Number Estimate	
	Copy Number Germline Variation <	
	LOH	
	Copy Number QC Metrics	
	Copy Number Variation	
	Normalized Copy Numbers	
	Copy Number Summary	
	Probeset Call	
Protein Expression	Protein Expression Quantification	TSV
	Protein Expression Control	
Other	Microsattelite Instability	FSA
	ABI Sequence Trace	TR
	Auxiliary Test	
About the Data Data Types and File Formats Generated Data Types and File Formats Imported Data Types and File Formats Submitted Data Types and File Formats Data Dictionary Data Harmonization and Generation Data Standards Data Availability Matrix Data Download Statistics		

Table 2.5: TCGA Data portal last status and updates

Available Cancer Types	**# Cases Shipped by BCR***	**# Cases with Data***	**Date Last Updated (mm/dd/yy)**
Acute Myeloid Leukemia [LAML]	200	200	5/31/2016
Adrenocortical carcinoma [ACC]	80	80	5/31/2016
Bladder Urothelial Carcinoma [BLCA]	412	412	5/27/2016
Brain Lower Grade Glioma [LGG]	516	516	5/2/2016
Breast invasive carcinoma [BRCA]	1100	1097	5/31/2016
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	308	307	5/26/2016
Cholangiocarcinoma [CHOL]	36	36	5/31/2016
Colon adenocarcinoma [COAD]	461	461	5/27/2016
Esophageal carcinoma [ESCA]	185	185	5/31/2016
FFPE Pilot Phase II [FPPP]	38	38	4/28/2016
Glioblastoma multiforme [GBM]	529	528	5/27/2016
Head and Neck squamous cell carcinoma [HNSC]	528	528	5/3/2016
Kidney Chromophobe [KICH]	66	66	6/1/2016
Kidney renal clear cell carcinoma [KIRC]	536	536	5/27/2016
Kidney renal papillary cell carcinoma [KIRP]	291	291	5/31/2016
Liver hepatocellular carcinoma [LIHC]	377	377	6/2/2016
Lung adenocarcinoma [LUAD]	521	521	6/1/2016
Lung squamous cell carcinoma [LUSC]	510	504	5/26/2016
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC]	48	48	5/31/2016
Mesothelioma [MESO]	87	87	4/8/2016
Ovarian serous cystadenocarcinoma [OV]	586	586	5/31/2016
Pancreatic adenocarcinoma [PAAD]	185	185	5/6/2016
Pheochromocytoma and Paraganglioma [PCPG]	179	179	5/3/2016
Prostate adenocarcinoma [PRAD]	498	498	5/31/2016
Rectum adenocarcinoma [READ]	172	171	6/1/2016
Sarcoma [SARC]	261	261	6/1/2016
Skin Cutaneous Melanoma [SKCM]	470	470	4/8/2016
Stomach adenocarcinoma [STAD]	445	443	5/26/2016
Testicular Germ Cell Tumors [TGCT]	150	150	6/2/2016
Thymoma [THYM]	124	124	5/31/2016
Thyroid carcinoma [THCA]	507	507	5/5/2016
Uterine Carcinosarcoma [UCS]	57	57	4/29/2016
Uterine Corpus Endometrial Carcinoma [UCEC]	548	548	6/2/2016
Uveal Melanoma [UVM]	80	80	4/29/2016

*Excludes non-canonical cases

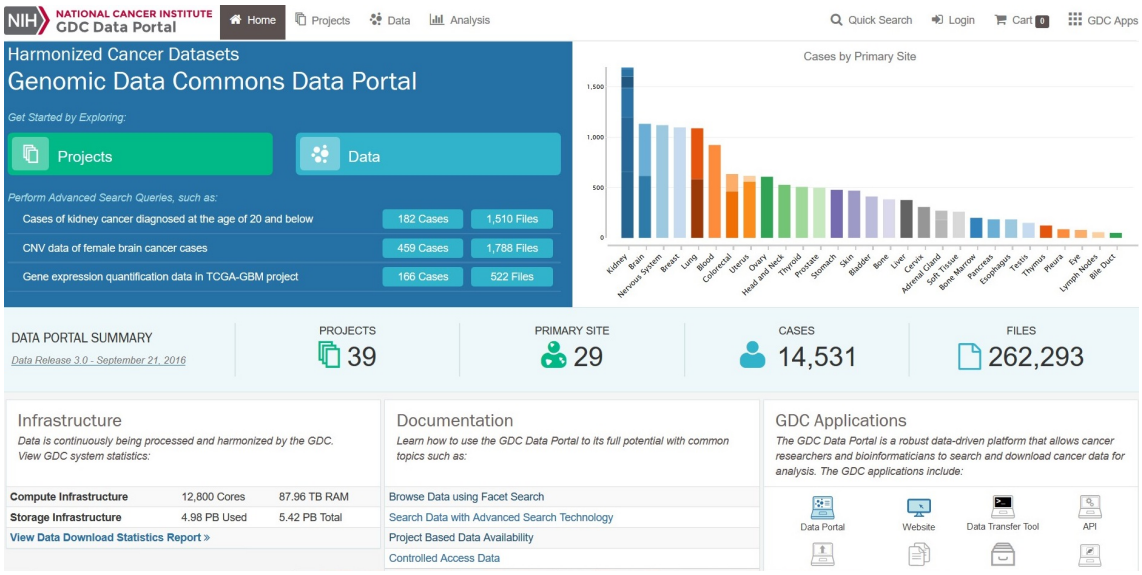


Figure 2.3: Genomic Data Commons Data portal Webpage

multiple samples and data types are to be compared and analyzed. Each visualization tools that has been developed for TCGA data analysis, has focused on one or a more specific research question and offers a wide variety of visualization output and analysis pipeline [132, 132, 146, 147]. Although a wide variety of visualization output and analysis tools are available, none of these tools are easy, fast and straightforward for usage and analysis. MEXPRESS, a novel tool has been developed that is used in our study for TCGA data analysis. This tool combines clinical, methylation and expression data. MEXPRESS, is a powerful tool since users do not need any programming or bioinformatics expertise to use the tool or in analyzing and identifying genes of interest or novel biomarkers in the TCGA data. MEXPRESS is mainly utilized for simple, but quick querying and visualization of clinical, expression and methylation data and also to determine relationship between the TCGA datasets on a single-gene level. MEXPRESS has been designed along the lines of graphical excellence described by Edward Tufte [148]. MEXPRESS tool designed in these lines has demonstrated that such complex and multidimensional TCGA data is presented in a clear, precise and efficient way for the end-user. Also, the user benefits from the fact that, analysis and visualization from MEXPRESS is very easy to use and does not require computational or bioinformatics expertise in any way. Thus, MEXPRESS, virtually does

not have any learning curve or requires any formal training. Such ease of use have facilitated researchers, particularly clinicians to get their results quickly, easily and effectively.

2.3.1 MEXPRESS: Implementation and Output visualization

MEXPRESS carries a key feature, that being simplicity. A visualization output plot is created upon selecting a gene of interest and a cancer type and querying it. An example of a visualization output is demonstrated below with its transcripts and any CpG islands that are involved (Figure 2.4)(Figure 2.5).

A. In the default MEXPRESS output of the visualization plot, the samples are ordered by their gene expression value. Here in this visualization plot, the Pearson Correlation coefficient value clearly demonstrates the negative correlation between GSTP1 expression and promoter methylation. Tumor samples are observed to have lower GSTP1expression when compared to normal samples. B. The visualization plot can also be ordered by another data type, the Sample type. This output shows a clear difference in expression and methylation between normal and tumor samples.

The visual output shows samples are ordered by breast cancer subtype. Results indicate significant differences in expression and methylation. Also, HER2, estrogen and progesterone receptor status indicates clear differences, between the different subtypes.

Gene expression data, methylation data and clinical data can be visualized and analyzed at the same time using MEXPRESS. Each probe generates a methylation data indicated in blue line plot (Infinium HumanMethylation 450 Microarray data) and is present next to the gene (vertical downward arrow line). RNA-seq derived expression data is depicted as a yellow line plot, while the grey line plot depicts the clinical data of the patients. Significance of relation (P value or correlation coefficient) depending on the data types that are compared (methylation, expression or clinical data) between each row is indicated on the

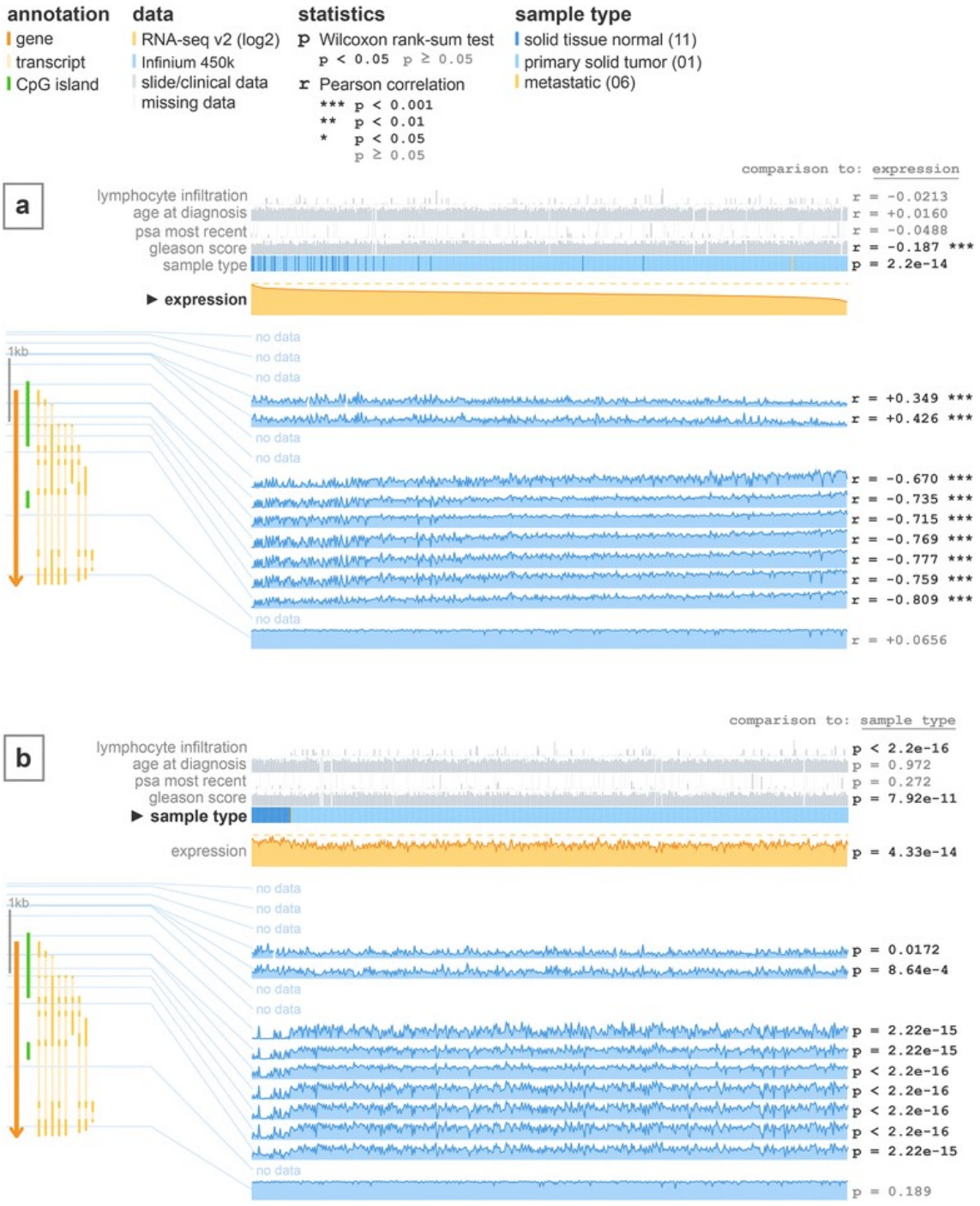


Figure 2.4: Visualization of the TCGA data for GSTP1 in prostate adenocarcinoma using MEXPRESS

Figure 2.5: MEXPRESS view of the TCGA data for MLPH in breast invasive carcinoma

far right of the visual output. The selected sorter is also indicated in the plot. MEXPRESS tool has the default sorter parameter set for expression. This means that samples or data for samples are ordered by their expression values. MEXPRESS carries the flexibility of changing the order by which the samples can be ordered. It can be called up or sorted using clinical or methylation data types too. The tool will then query based upon the selected sorter and samples are reordered with the relevant recalculated relationships or significance values. The visual output can be saved and downloaded in PNG or SVG extension format.

2.4 MEXPRESS and TCGA Data

MEXPRESS, directly downloads TCGA data from its ftp (file transfer protocol). RNA-seq v2 expression data from IlluminaHiSeq_RNASeqV2 from Level 3 of TCGA, HumanMethylation450 derived DNA methylation data from Level 3 and Biotab format derived clinical patient and tumor sample data. MEXPRESS tool, which runs on Bash scripts on the back-end Linux servers automatically checks the TCGA ftp site on a monthly basis. Such updates are then identified and automatically updated to the MEXPRESS database. Also, TCGA makes it publically available about cancer types which is also automatically updated by MEXPRESS scripts. The tool is facilitated initially by the R scripts (R version 3.0.2). These scripts are responsible for significant data processing and address issues like missing values. It also facilitates the combination of different files into one upon identifying the requirement, reformats data into relevant accessible types and to generate SQL scripts for uploading the processed and also new data. MEXPRESS tool, does a log-transformation on the RNA-seq data before such data is utilized for visualization plot. Also, only the most relevant clinical parameters is utilized by the MEXPRESS plot to minimize data clutter and for efficient data analysis.

2.4.1 MEXPRESS and other data sources

MEXPRESS, greatly facilitates incorporation of novel data types (mutation, proteomic or other omics data). MEXPRESS, can access UCSC cancer genome for various cancer subtypes (normal, basal, luminal A, luminal B and Her2, in case of invasive carcinoma sample) [149]. The tool can specifically access CpG data from the USCS table browser using the following path: clade: Mammal, genome: Human, assembly: Feb. 2009 (GRCh37/hg19), group: Regulation, track: CpG Islands, table: cpgIslandsExt [150]. MEXPRESS, can access and obtain annotation data (exon or transcript) using Ensembl with the aid of BioMart tool.

2.4.2 MEXPRESS and statistical analyses

Two main statistical tests that is incorporated in the tool are: Pearson correlation and the non-parametric Wilcoxons rank-sum test. These tests are created using JavaScript. Pearson correlation test mainly compares two data types which are at different levels such as comparison of methylation and expression data. Non-parametric Wilcoxons rank-sum test calculates the variable between two groups for which comparison is undertaken (for example, difference in methylation with respect to gender). A false discovery rate correction step is also incorporated in the tool [151].

2.4.3 Methods in Statistical Analysis

Pearson correlation Test

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a statistical measure of the strength of a linear association between two variables and is denoted by r. It should be noted that, the symbol for Pearson's correlation is "" when it is measured in the population and "r" when it is measured in a sample. The difference between Pearson product-moment correlation and the Pearson correlation coefficient can be explained as follows. Basically, a Pearson product-moment correlation attempts to draw

Table 2.6: Guidelines proposed to interpret Pearson's correlation coefficient

	Coefficient,r	
Strength of Association	Positive	Negative
Small	.1 to .3	-.1 to .3
Medium	.3 to .5	-.3 to .5
Large	.5 to .10	-.5 to 1.0

a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

Assigned values and its interpretation: The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

Determination of strength association: The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit there are no data points that show any variation away from this line. Values for r between +1 and -1 (for example, r = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of r to 0 the greater the variation around the line of best fit. (Table 2.6)

Variables used in this test: When using this statistical test, the two variables have to be measured on either an interval or ratio scale. However, both variables do not need to be measured on the same scale (e.g., one variable can be ratio and one can be interval).

Measuring the variables: Also, the two variables can be measured in entirely different units. For example, you could correlate a person's age with their blood sugar levels. Here, the units are completely different; age is measured in years and blood sugar level measured in mmol/L (a measure of concentration). Indeed, the calculations for Pearson's correlation

coefficient were designed such that the units of measurement do not affect the calculation. This allows the correlation coefficient to be comparable and not influenced by the units of the variables used.

Units of measurement for variables: The two variables can be measured in entirely different units.

Dependent and independent variables: The Pearson product-moment correlation does not take into consideration whether a variable has been classified as a dependent or independent variable. It treats all variables equally.

Slope of the line: It is important to realize that the Pearson correlation coefficient, r, does not represent the slope of the line of best fit. Therefore, if you get a Pearson correlation coefficient of +1 this does not mean that for every unit increase in one variable there is a unit increase in another. It simply means that there is no variation between the data points and the line of best fit.

5 assumptions made in this test:

- The variables must be either interval or ratio measurements.
- The variables must be approximately normally distributed.
- There is a linear relationship between the two variables.
- Outliers are either kept to a minimum or are removed entirely.
- There is homoscedasticity of the data.

To detect a linear relationship: To test to see whether your two variables form a linear relationship, the user needs to simply need to plot them on a graph (a scatterplot, for example) and visually inspect the graph's shape.

Pearson's correlation determines the degree to which a relationship is linear. Put another way, it determines whether there is a linear component of association between two continuous variables. As such, linearity is not actually an assumption of Pearson's correlation. However, you would not normally want to pursue a Pearson's correlation to determine

the strength and direction of a linear relationship when you already know the relationship between your two variables is not linear. Instead, the relationship between your two variables might be better described by another statistical measure. For this reason, it is not uncommon to view the relationship between your two variables in a scatterplot to see if running a Pearson's correlation is the best choice as a measure of association or whether another measure would be better.

Wilcoxons rank-sum test

The Wilcoxon rank-sum test is a nonparametric alternative to the two-sample t-test which is based solely on the order in which the observations from the two samples fall. The Wilcoxon rank-sum test tests the null hypothesis that two sets of measurements are drawn from the same distribution. The alternative hypothesis is that values in one sample are more likely to be larger than the values in the other sample. This test should be used to compare two samples from continuous distributions. It does not handle ties between measurements in x and y.

An alternative explanation would be as follows: A popular nonparametric test to compare outcomes between two independent groups is the Mann Whitney U test. The Mann Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape). Some investigators interpret this test as comparing the medians between the two populations. Recall that the parametric test compares the means (H0: 1=2) between independent groups.

In contrast, the null and two-sided research hypotheses for the nonparametric test are stated as follows:

H0: The two populations are equal versus

H1: The two populations are not equal.

This test is often performed as a two-sided test and, thus, the research hypothesis indicates that the populations are not equal as opposed to specifying directionality. A one-sided research hypothesis is used if interest lies in detecting a positive or negative shift

in one population as compared to the other. The procedure for the test involves pooling the observations from the two samples into one combined sample, keeping track of which sample each observation comes from, and then ranking lowest to highest from 1 to n1+n2, respectively.

A common experiment design is to have a test and control conditions. A two sample t-test would have been a good choice if the test and control groups are independent and follow Normal distribution. If conditions are not met, nonparametric test methods are needed. This section covers one such test, called Wilcoxon rank-sum test (equivalent to the Mann-Whiney U-test) for two samples. The test is preferred when:

Comparing two samples.

- The two groups of data are independent
- The type of variable could be continuous or ordinal
- The data might not be normally distributed

Wilcoxon Rank Sum Test for Independent Samples:

When the requirements for the t-test for two independent samples are not satisfied, the Wilcoxon Rank-Sum non-parametric test can often be used provided the two independent samples are drawn from populations with an ordinal distribution.

For this test we use the following null hypothesis:

H0: the observations come from the same population

From a practical point of view, this implies:

H0: if one observation is made at random from each population (call them x0 and y0), then the probability that x0 > y0 is the same as the probability that x0 < y0, and so the populations for each sample have the same medians.

2.5 MEXPRESS as a visualization tool

MEXPRESS tool/site runs on Apache server. The back end database is accessed using PHP. Interactive plots are created and statistical analysis is done by employing JavaScript, the

jQuery JavaScript library (version 1.11.0), Ajax autocomplete for jQuery(version 1.2.10, [24]) and the d3.js JavaScript library (version 3.0.6, [25]). The visualization output from SVG format can be converted into PNG format with the aid of Inkscape which is a freely available vector graphics editor ([26]). MEXPRESS database is conceptually created using MySQL database which contains the TCGA data for visualization and analysis. This forms the main backbone of the tool. The way this tool functions is that PHP scripts handles all the queries from the user which is then directed to the database, results then packaged in JSON and sent back to the user. The entire code of MEXPRESS, is highly validated (back-end, front-end and data processing) as it can be cloned or downloaded from this GitHub repository[27].

2.6 Gene query against BioMuta and BioXpress databases

BIOMUTA DATABASE

URL: *http://hive.biochemistry.gwu.edu/tools/biomuta/index.php*

CSR: *http://hive.biochemistry.gwu.edu/dna.cgi?cmd=csr*

HIVE: *http://hive.biochemistry.gwu.edu*

BioMuta, a database created with integrated sequence features, provides a framework for both automated and manual curation and integration of cancer-related sequence features for NGS analysis pipelines, was utilized (40- 42). Sequence feature information in BioMuta is integrated from a variety of source such as Catalogue of Somatic Mutations in Cancer (COSMIC), ClinVar, UniProtKB and biocuration of published data. BioMuta also contains non-synonymous single-nucleotide variations (nsSNVs) identified from NGS data. The High-performance Integrated Virtual Environment (HIVE) was created for handling petabytes of data for storage, analysis, computing and curating NGS data and related metadata support BioMuta too. Different algorithms were used to identify and tackle variations in

[24]https://github.com/devbridge/jQuery-Autocomplete
[25]http://d3js.org/
[26]http://www.inkscape.org/
[27]https://github.com/akoch8/mexpress

cancer data. We queried the five selected genes BLCAP, GDF15, PIWIL4, DMRT1 and ITPKA against BioMuta for validating or supplementing the above mentioned MEXPRESS study/results that identifies epigenetic alterations (methylation) affecting gene expression in various cancers.

BIOXPRESS DATABASE

URL: *http://hive.biochemistry.gwu.edu/tools/bioxpress*

CSR: *http://hive.biochemistry.gwu.edu/dna.cgi?cmd=csr*

HIVE: *http://hive.biochemistry.gwu.edu*

BioXpress is a gene expression and cancer association database wherein expression levels are mapped to genes using RNA-seq data obtained from TCGA, International Cancer Genome Consortium, Expression Atlas and literature reviews. BioXpress encompasses expression data from 64 cancer types, 6361 patients and 17469 genes, of which 9513 genes exhibit differential expression between tumor and normal samples. Data from RNA-seq data repositories is supplemented with manual curation of cancer data from literature reviews. Pan-cancer analysis is also facilitated by mapping cancer types to Disease Ontology terms. BioXpress can be queried using HUGO Gene Nomenclature Committee gene symbol, UniProtKB/RefSeq accession or by cancer types with specialized filters. This database is invaluable in identifying cancer-related genes using a pre-computed downloadable file containing differentially expressed genes in multiple cancers (43). Again, we queried the five selected genes BLCAP, GDF15, PIWIL4, DMRT1 and ITPKA against BioXpress for validating or supplementing the above mentioned MEXPRESS study/results that identifies epigenetic alterations (methylation) affecting gene expression in various cancers .

Chapter 3: RESULTS

3.1 MEXPRESS plot details

Mxpress, for our gene/s of interest when selected with a particular cancer type generates the Mxpress data/plot. Here, the height of the orange line (expression data) represents the logarithm of the level 3 RNA-sequencing data in TCGA (normalized RNASeqV2 values per gene). The expression data forms the basis of the whole plot. This is because the samples are ranked based on their expression value for the ITPKA gene selected. Samples with the highest expression appear on the left side and the lowest on the right. In the panel below the expression data, on the left hand side, the gene is designated by a solid orange line, the CpG islands in green and the different transcripts in broken or dotted orange lines. The arrow on the gene indicates its direction. If the arrow points down, the gene is located on the + strand. If it points up, the gene lies on the - strand. Also, in this panel, to the right, the Infinium 450k probes are linked to our gene of interest. The height of the blue lines indicates the beta value for a probe. When there is no data available for a certain probe, no line is plotted and instead it simply says "no data". Gaps in the line will indicate that there was no methylation data for one or more samples. Similar to the expression data, the samples are also ranked along the x axis (they are ordered based off our gene of interest expression value). Thin blue lines connect the probes to their respective genomic locations. When a user hovers over a methylation data, the plot will highlight the corresponding probe on the left hand side and the name of the probe will also be shown. The user can fix the highlighting of a probe by clicking on its data plot. Also, by clicking the same data plot a second time will clear the highlighting. Hence, Mexpress is an invaluable tool to detect clinical, methylation and expression data simultaneously and to detect the significance that exist between these data sets.

Common features of MEXPRESS result plot analysis

ITPKA gene when queried for BRCA in MEXPRESS, generates the above plot. On the left hand top corner, ITPKA as a gene entry in entered. Here both HGNC symbols and Ensembl gene IDs are recognized as valid entries. The plot demonstrates that the samples are arranged from left to right, while the different data types (clinical, expression and methylation) are arranged from the top to the bottom of the plot. On the resulting plot, if users can hover over one of the methylation line plots will exhibit the ID of the corresponding probe. Here users can click on a methylation line plot to fix the probe ID on the figure. By clicking it again will enable users to remove the probe ID (or click another methylation line plot). Users can also highlight the promoter probes by clicking the button right above the legend. User can download the figure by simply clicking on the png or svg button in the upper right corner. Users can emphasize the probes that are located in a gene's promoter region by clicking on the highlight promoter probes button. This will turn the highlighting of the promoter probes on.

The expression data is represented by the yellow/orange line plot. The height of the orange line represents the logarithm of the level 3 RNA-sequencing data in TCGA (normalized RNASeqV2 values per gene). The expression data forms the basis of the whole plot, because the samples are ranked based on their expression value for the gene that is selected for query. Here, the resulting plot shows the highest expression on the left side and the lowest on the right.

The methylation data is indicated by the blue line plot. On the left hand side, solid orange vertical line indicates the gene, CpG islands are indicated using the solid vertical green line and the different gene transcripts are indicated by the dotted/ broken orange lines. The arrow on the gene indicates its direction. If the arrow points down, the gene is located on the + strand. If it points up, the gene lies on the - strand.

On the right hand side, the Infinium 450k probes that are linked to the gene can be observed. The height of the blue lines corresponds with the beta value for a probe. If data is unavailable for a certain probe, no line is plotted and instead it simply says "no data".

Gaps in the line indicate that there is no methylation data available for one or more amples.

Like the expression data, the samples are ranked along the x axis (they are ordered based on their gene expression value). Thin blue lines on the plot, connect the probes to their respective genomic locations. Users can hover over a methylation data plot to highlight the corresponding probe on the left hand side and the name of the probe will be also be showed. Users can fix the highlighting of a probe by clicking on its data plot. Clicking the same data plot a second time will clear the highlighting. Once users have fixed a probe's highlighting by clicking on the data plot, users can click on the probe's name to reveal the probe's genomic location and annotation.

The values on the far right of the plot represents the Pearson product-moment correlation coefficient between the methylation values for a probe and the expression values. If probes exhibit a strong negative correlation between methylation and expression, it indicates that gene expression might be controlled through DNA methylation. The asterisks gives an indication of the significance of the correlations.

The clinical data is represented using the grey line plot. For every cancer type, the most appropriate relevant clinical parameters is extracted from TCGA. In order to represent all the data as bar plots, some clinical parameters have been converted to numeric values. One example is the pathologic stage where values such as Stage IIA and Stage IV were converted to the values 2 and 4 respectively.

Labels/ names of different clinical parameters are listed on the left and the Pearson product-moment correlation values or the p values for Wilcoxon rank-sum test can be found on the right. If a clinical parameter contains only two levels (e.g. male or female) a p value is calculated instead of a correlation coefficient. This p value indicates the difference in expression between the two groups for this parameter. For the sample type parameter, the expression is always compared between the normal and tumor samples.

As indicated before, the samples are sorted based on their expression value by default. By clicking on the name of the annotation parameter that users are interested in, they can rearrange the samples by the annotation that they selected. So if users, for example, like

Figure 3.1: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer

to compare age to the expression and methylation of a certain gene or to the other clinical parameters, they have to click on "age at diagnosis" and the samples will be reordered.

3.1.1 BLCAP (bladder cancer associated protein) as a DNA methylation biomarker gene

(Figure 3.1)(Figure 3.2)(Figure 3.3)

Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer:

MEXPRESS plot for BLCAP gene expression for BLCA cancer reveals the following details: A) there are quite a few probes with a strong negative correlation between methylation and expression, indicating that BLCAP expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear

Figure 3.2: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer

Figure 3.3: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer

Figure 3.4: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

that the normal samples tend to have a lower BLCAP expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for BLCAP gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence BLCAP gene expression. When samples are ordered by expression, sample type p= 6.02e-4 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 9.17e-4

Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer: (Figure 3.4)(Figure 3.5)(Figure 3.6)

MEXPRESS plot for BLCAP gene expression for BRCA cancer reveals the following details: A) there are numerous probes with a strong negative correlation between methylation and expression, indicating that BLCAP expression might be controlled through DNA

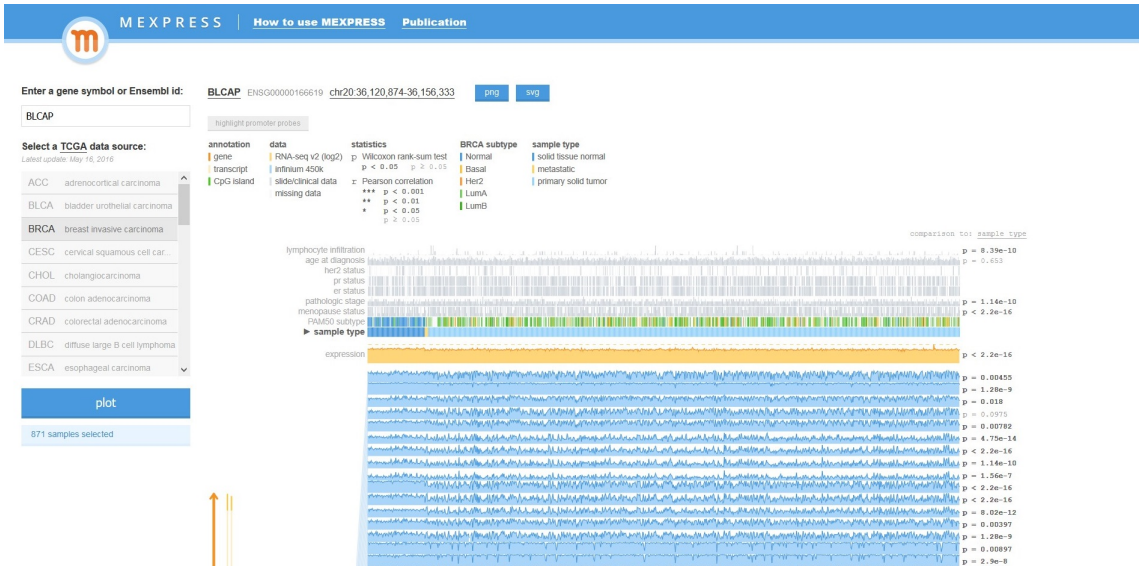


Figure 3.5: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

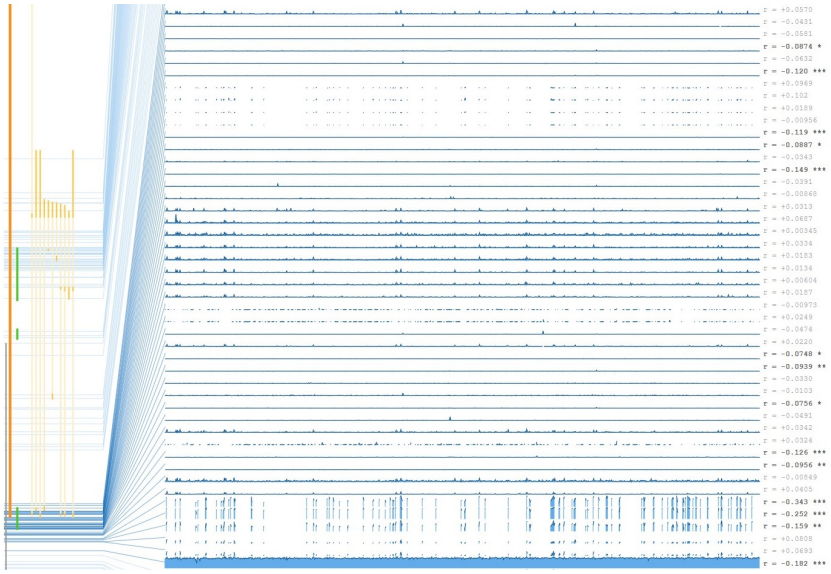


Figure 3.6: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

Figure 3.7: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma) cancer

methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have a higher BLCAP expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for BLCAP gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence BLCAP gene expression. When samples are ordered by expression, sample type p $<$ 2.2e-16 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p $<$ 2.2e-16

Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma) cancer: (Figure 3.7)(Figure3.8)(Figure3.9)

MEXPRESS plot for BLCAP gene expression for COAD cancer reveals the following

Figure 3.8: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma) cancer

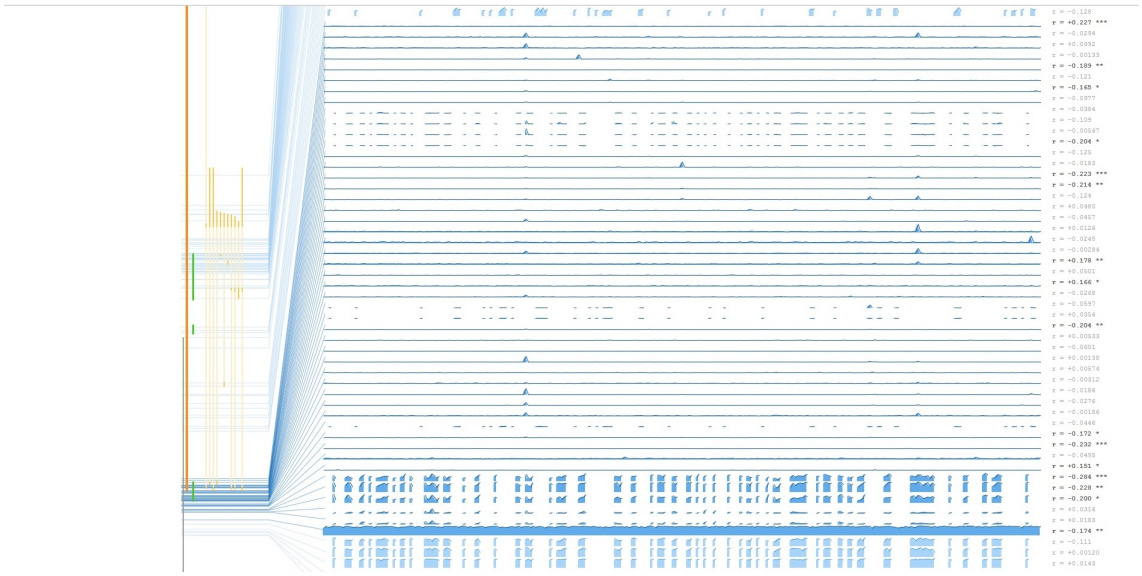


Figure 3.9: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma) cancer

details: A) there are more probes with a strong negative correlation as compared to strong positive correlation probes between methylation and expression, indicating that BLCAP expression might be slightly controlled or influenced through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have a slightly lower BLCAP expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for BLCAP gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence BLCAP gene expression. When samples are ordered by expression, sample type p= 0.0562 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.0687

Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for CRAD (Colorectal Adeno Carcinoma) cancer: (Figure 3.10)(Figure3.11)(Figure3.12)

MEXPRESS plot for BLCAP gene expression for CRAD cancer reveals the following details: A) there are more probes with a strong negative correlation as compared to the ones with strong positive correlation probes between methylation and expression, indicating that BLCAP expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower BLCAP expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for BLCAP

Figure 3.10: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for CRAD (Colorectal Adeno Carcinoma)) cancer

Figure 3.11: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for CRAD (Colorectal Adeno Carcinoma)) cancer

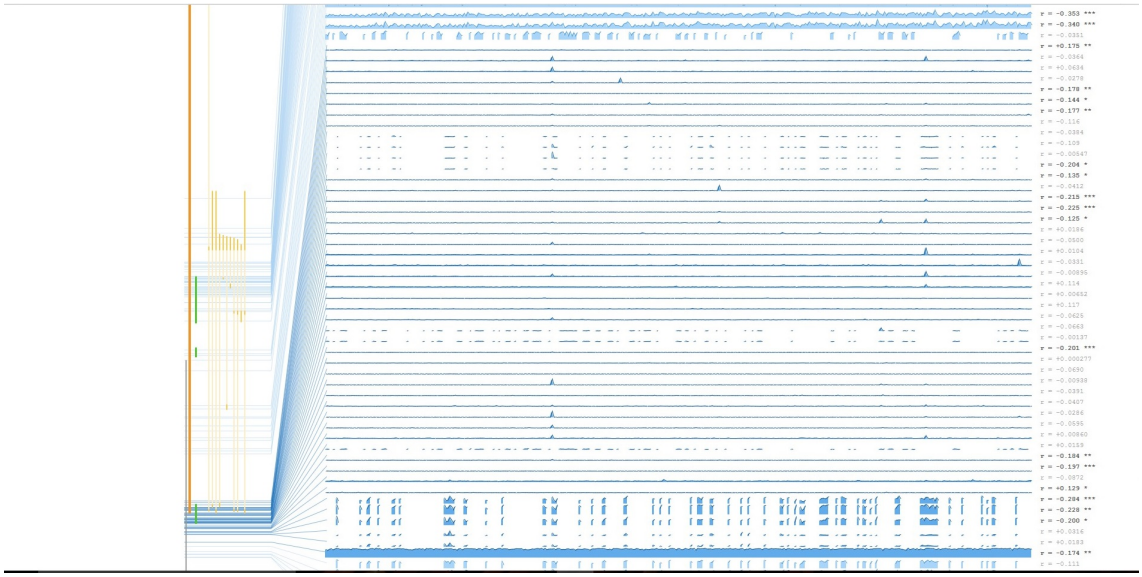


Figure 3.12: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for CRAD (Colorectal Adeno Carcinoma) cancer

gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence BLCAP gene expression. When samples are ordered by expression, sample type p= 0.023 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.0294

Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer: (Figure 3.13)(Figure3.14)(Figure3.15)

MEXPRESS plot for BLCAP gene expression for KIRC cancer reveals the following details: A) there are numerous probes with a strong negative correlation between methylation and expression, indicating that BLCAP expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have higher BLCAP expression than the tumor samples.

Figure 3.13: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer

Figure 3.14: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer

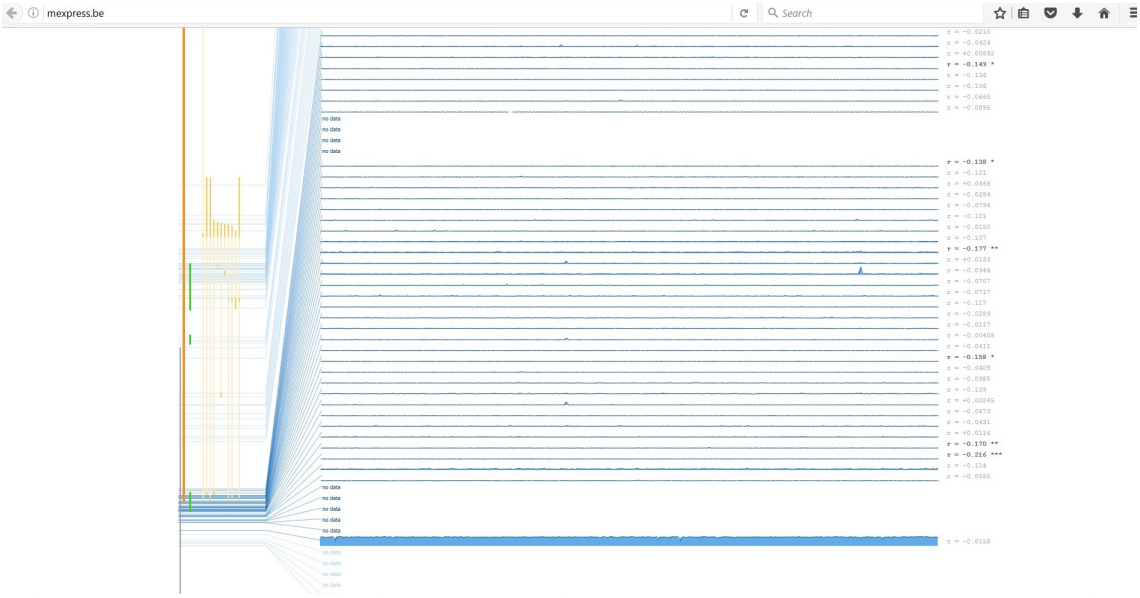


Figure 3.15: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer

C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for BLCAP gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence BLCAP gene expression. When samples are ordered by expression, sample type p= 2.77e-10 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 1.36e-10

Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma) cancer: (Figure 3.16)(Figure3.17)(Figure3.18)

MEXPRESS plot for BLCAP gene expression for KIRP cancer reveals the following details: A) there are numerous probes with a strong negative correlation between methylation and expression, indicating that BLCAP expression might be controlled through DNA

Figure 3.16: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma) cancer

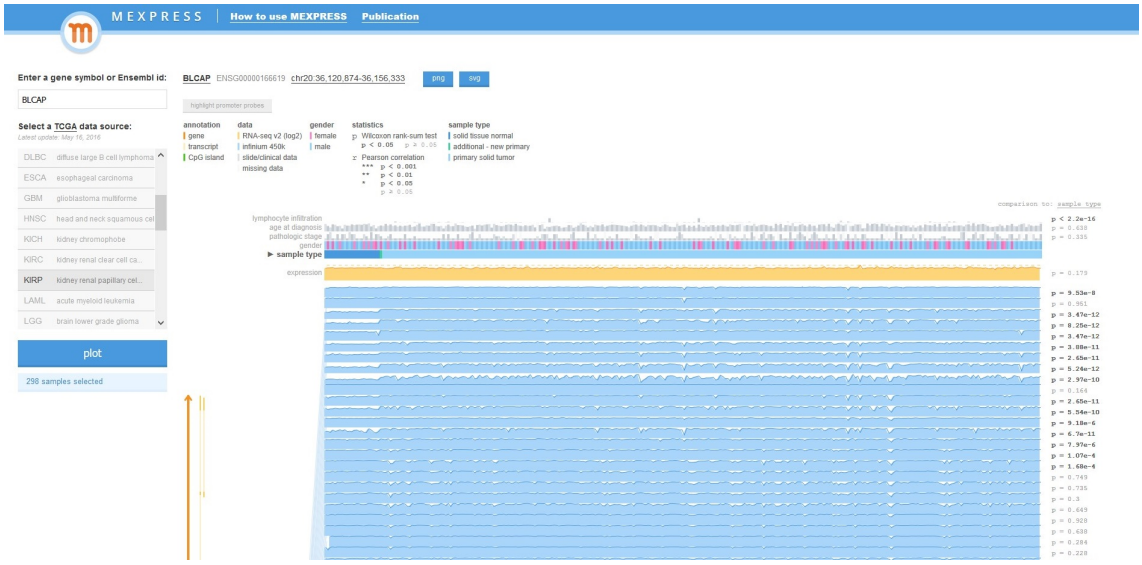


Figure 3.17: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma) cancer

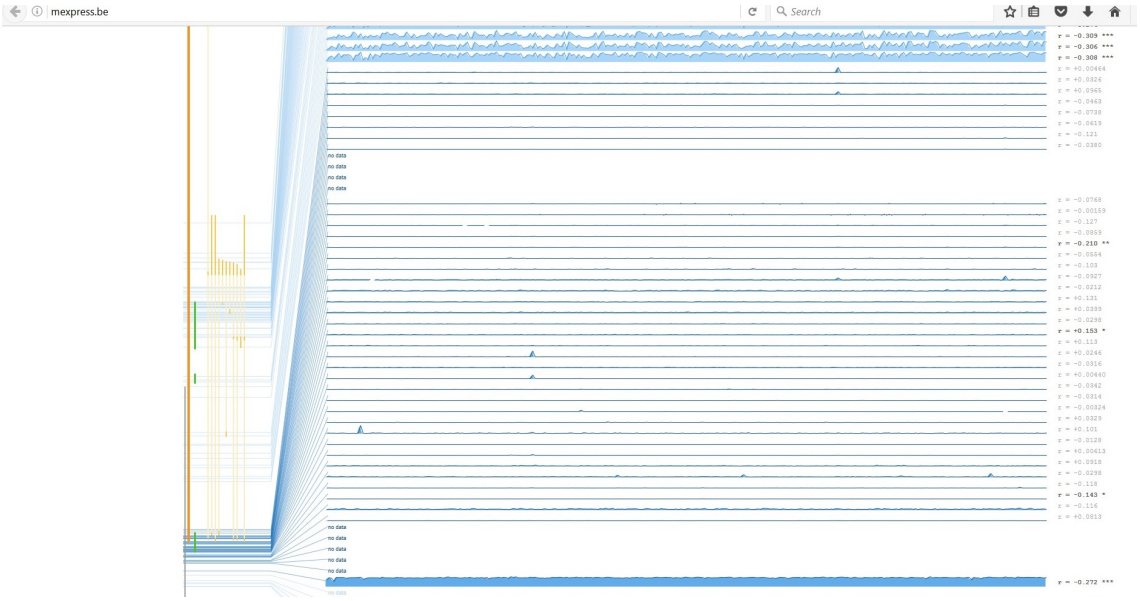


Figure 3.18: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma) cancer

methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower BLCAP expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for BLCAP gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence BLCAP gene expression. When samples are ordered by expression, sample type p= 0.246 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.179

Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer: Figures (Figure 3.19)(Figure 3.20)(Figure 3.21)

Figure 3.19: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer

Figure 3.20: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer

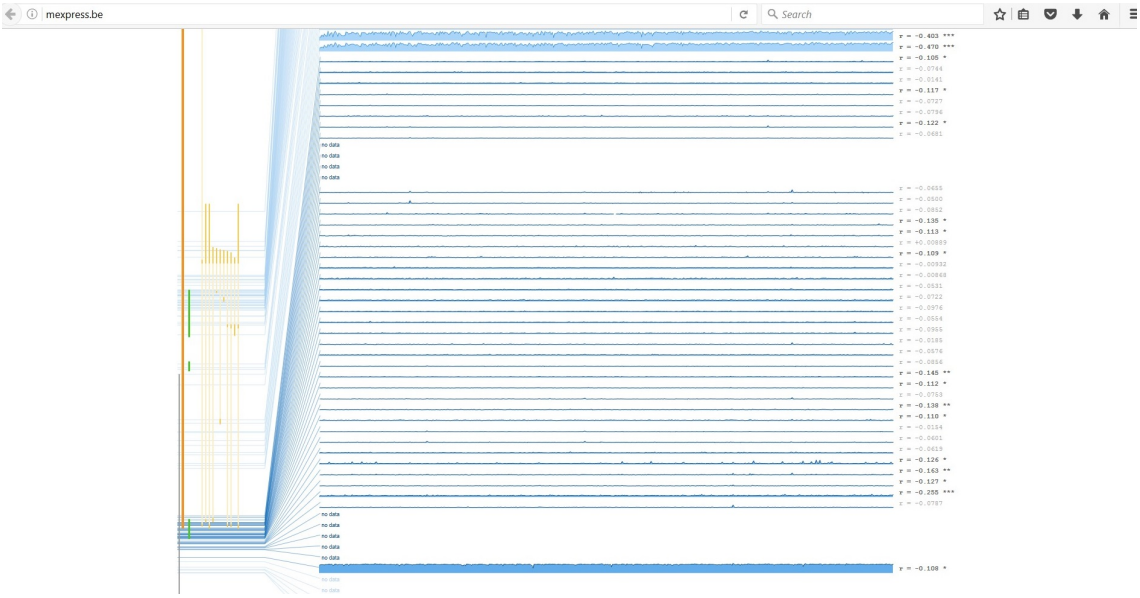


Figure 3.21: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer

MEXPRESS plot for BLCAP gene expression for LUAD cancer reveals the following details: A) there are numerous probes with a strong negative correlation between methylation and expression, indicating that BLCAP expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower BLCAP expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for BLCAP gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence BLCAP gene expression. When samples are ordered by expression, sample type p= 0.578 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.623 **Analysis of BLCAP (Bladder Cancer Associated**

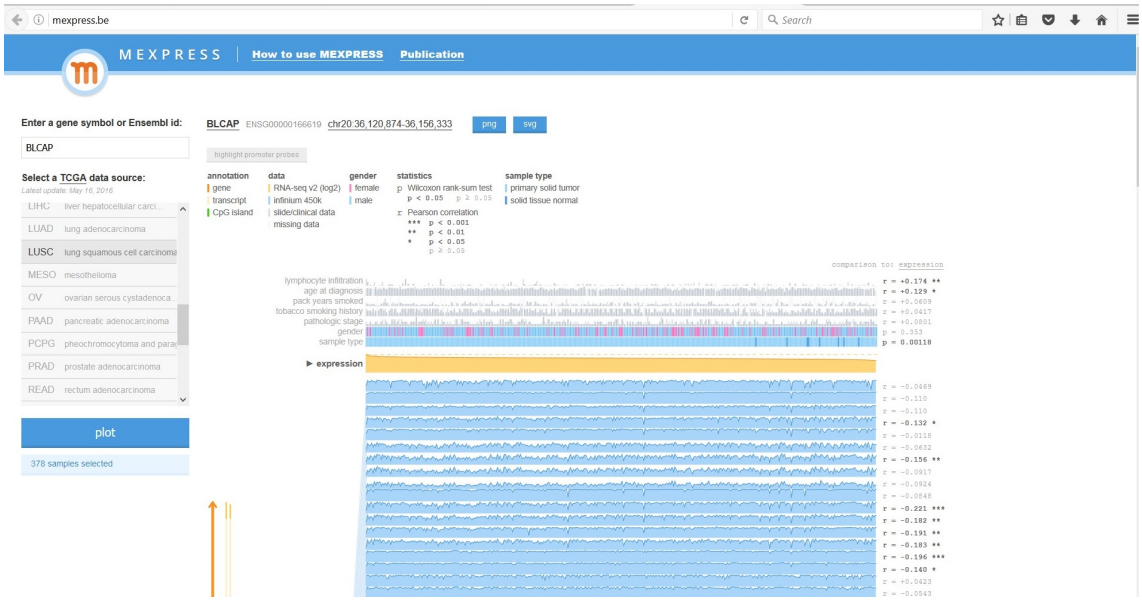


Figure 3.22: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer

Protein) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer:(Figure 3.22)(Figure3.23)(Figure3.24)

MEXPRESS plot for BLCAP gene expression for LUSC cancer reveals the following details: A) there are numerous probes with a strong negative correlation between methylation and expression, indicating that BLCAP expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have lower BLCAP expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for BLCAP gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence BLCAP gene expression. When samples are ordered by expression, sample type

Figure 3.23: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma)) cancer

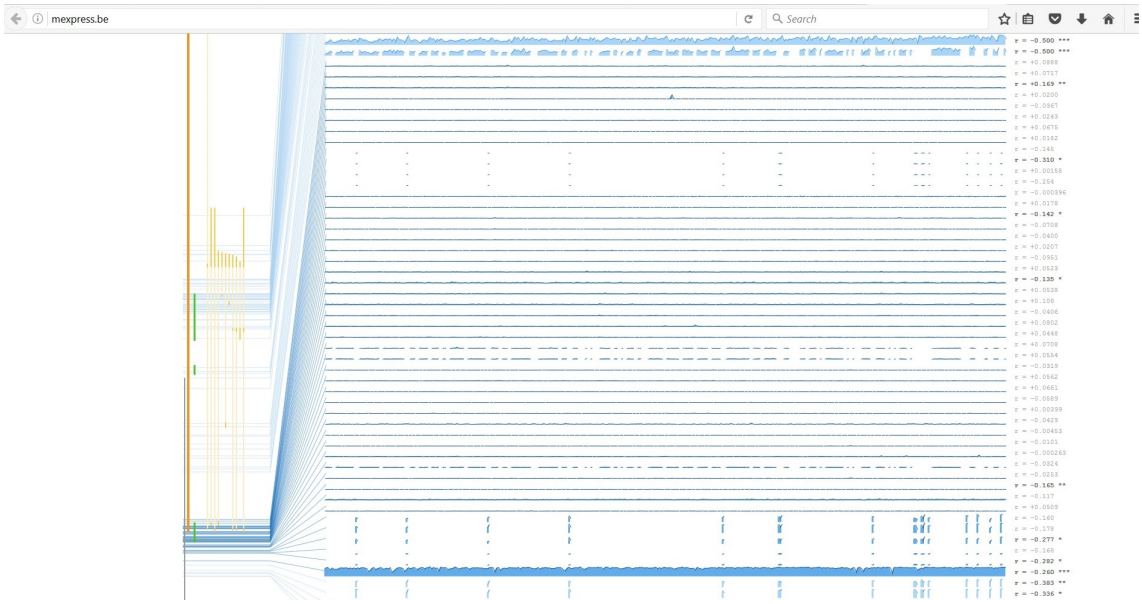


Figure 3.24: Analysis of BLCAP (Bladder Cancer Associated Protein) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer

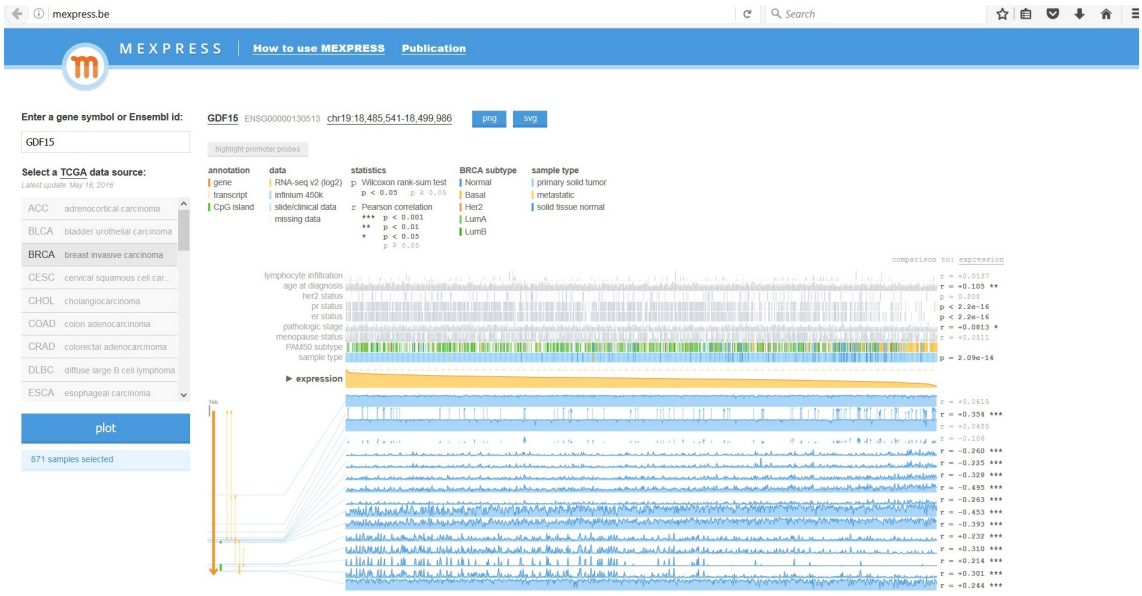


Figure 3.25: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

p= 0.0118 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.00128

3.2 GDF15 (Growth Differentiation Factor 15) as a DNA methylation biomarker gene

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer: (Figure 3.25)(Figure3.26)(Figure3.27)

MEXPRESS plot for GDF15 gene expression for BRCA cancer reveals the following details: A) there are more probes with a strong negative correlation as compared to the ones with a strong positive correlation between methylation and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to

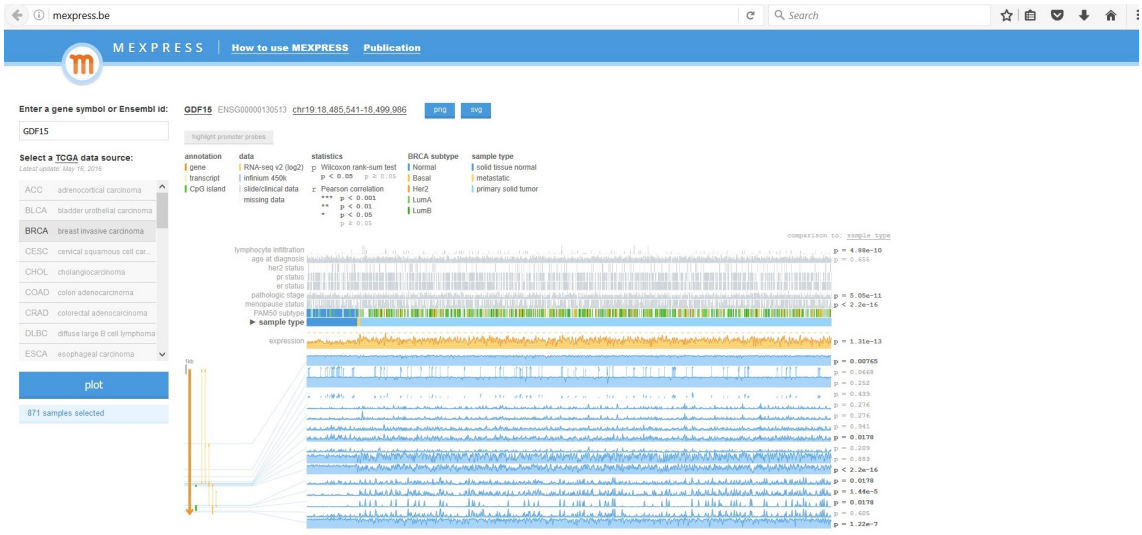


Figure 3.26: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

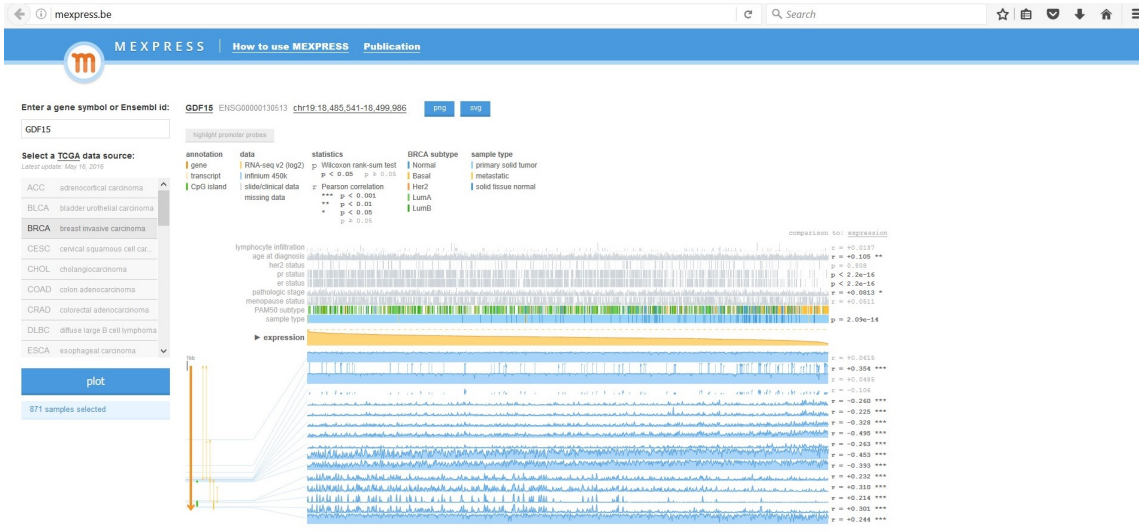


Figure 3.27: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

Figure 3.28: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma) cancer

have very lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 2.09e-14 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 1.31e-13

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma) cancer: (Figure 3.28)(Figure3.29)(Figure3.30)

MEXPRESS plot for GDF15 gene expression for COAD cancer reveals the following details: A) there are couple probes with a strong negative correlation between methylation and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that

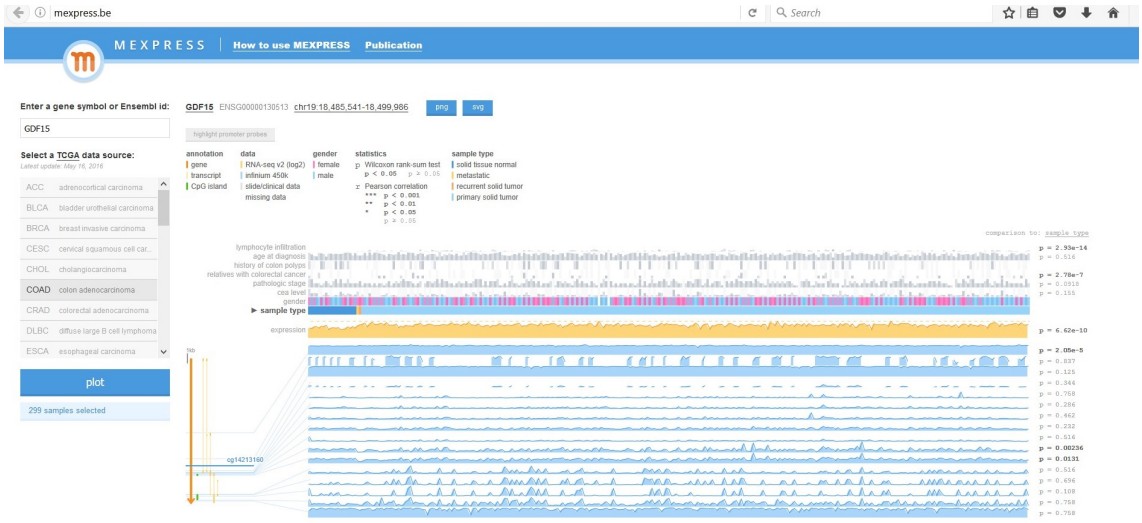


Figure 3.29: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma) cancer

Figure 3.30: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma)) cancer

Figure 3.31: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for CRAD (Colo Rectal Adeno Carcinoma) cancer

the normal samples tend to have very lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 1.37e-9 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 6.62e-10

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for CRAD (Colo Rectal Adeno Carcinoma) cancer: (Figure 3.31)(Figure3.32)(Figure3.33)

MEXPRESS plot for GDF15 gene expression for CRAD cancer reveals the following details: A) there are couple probes with a strong negative correlation between methylation and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance

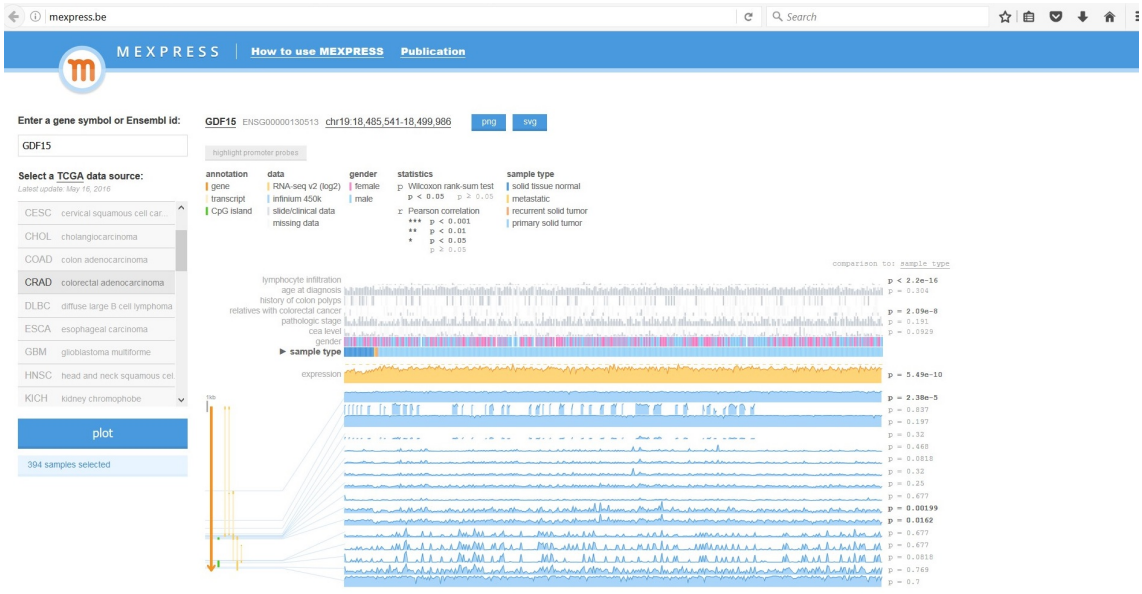


Figure 3.32: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for CRAD (Colo Rectal Adeno Carcinoma) cancer

Figure 3.33: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for CRAD (Colo Rectal Adeno Carcinoma) cancer

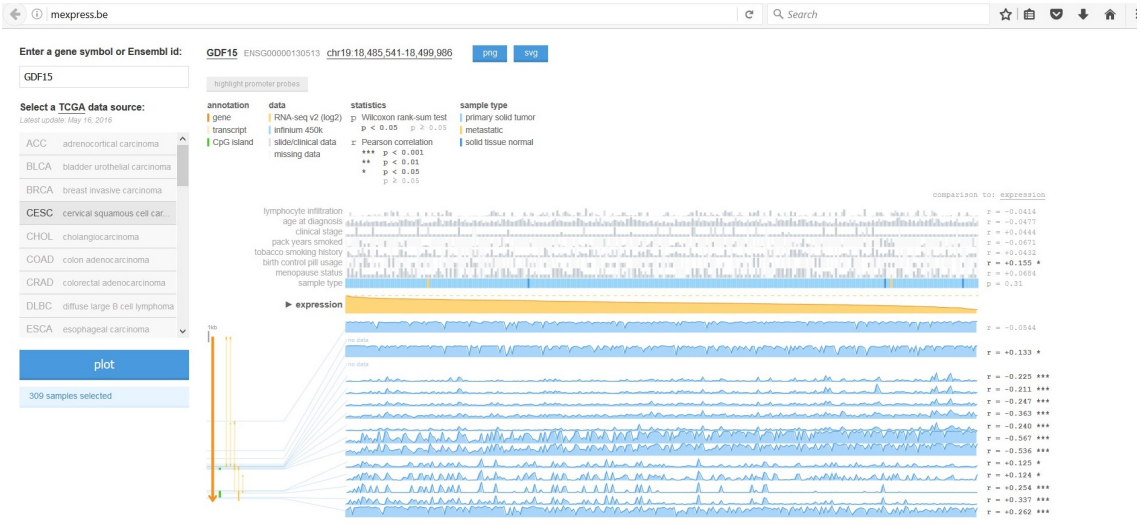


Figure 3.34: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for CESC (Cervical Squamous Cell Carcinoma) cancer

of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have very lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 5.38e-10 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 5.49e-10

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for CESC (Cervical Squamous Cell Carcinoma) cancer: (Figure 3.34)(Figure3.35)(Figure3.36)

MEXPRESS plot for GDF15 gene expression for CESC cancer reveals the following details: A) there are couple probes with a strong negative correlation between methylation

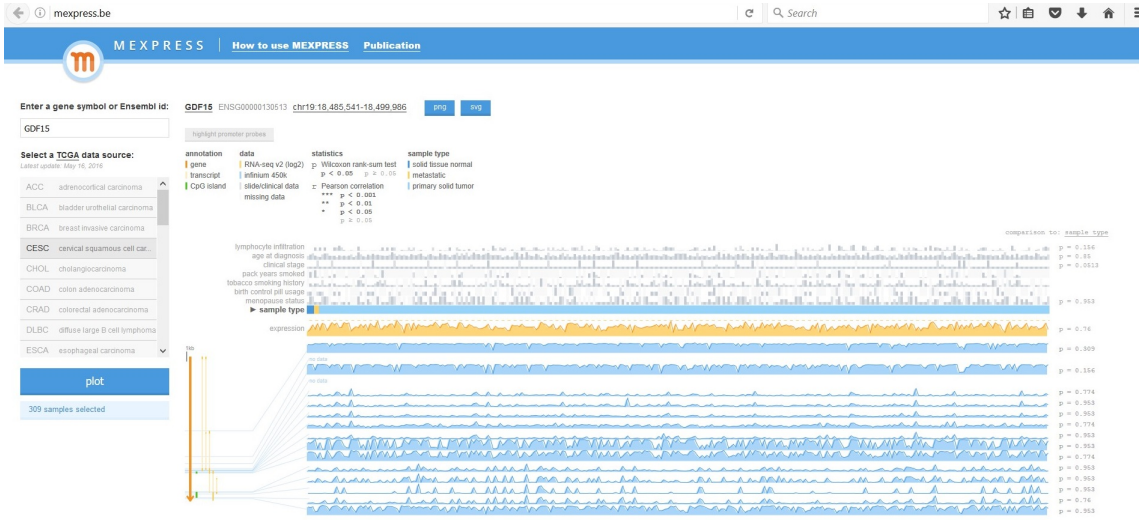


Figure 3.35: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for CESC (Cervical Squamous Cell Carcinoma) cancer

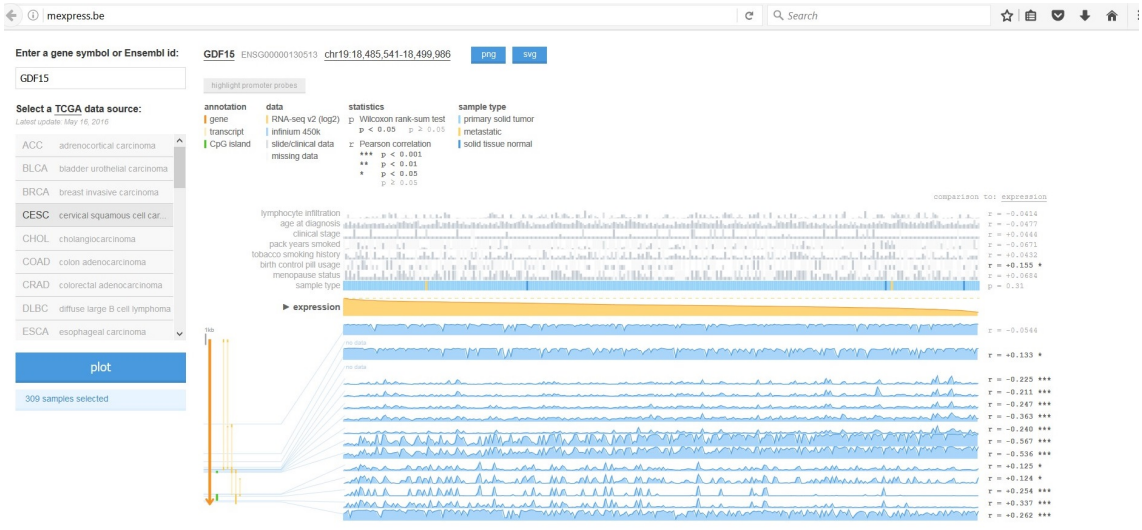


Figure 3.36: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for CESC (Cervical Squamous Cell Carcinoma) cancer

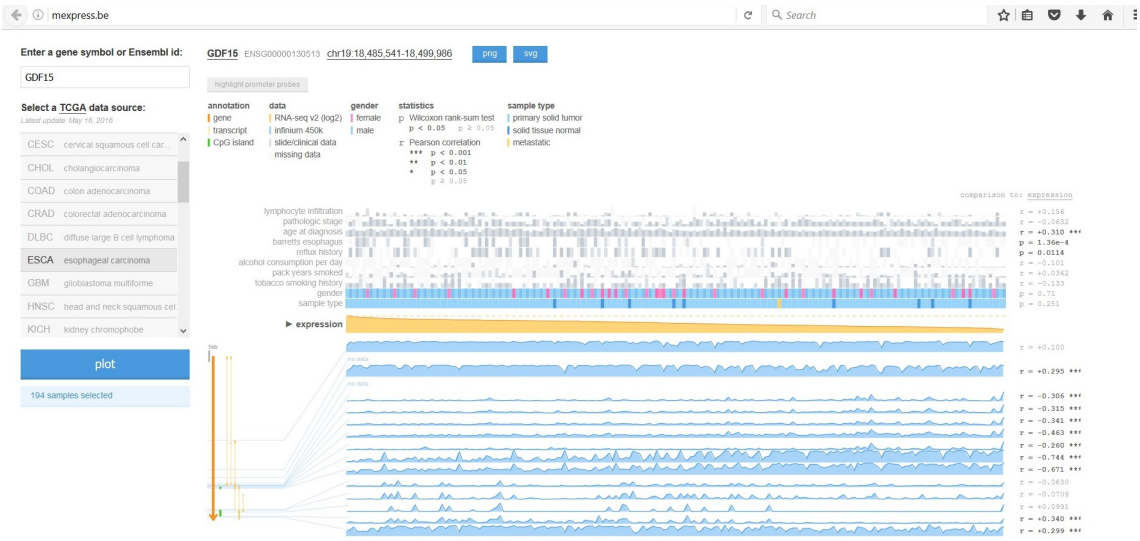


Figure 3.37: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for ESCA (Esophageal Carcinoma) cancer

and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have very lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 0.31 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.76

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for ESCA (Esophageal Carcinoma) cancer: Figures 3.37 to 3.39(Figure 3.37)(Figure3.38)(Figure3.39)

Figure 3.38: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for ESCA (Esophageal Carcinoma) cancer

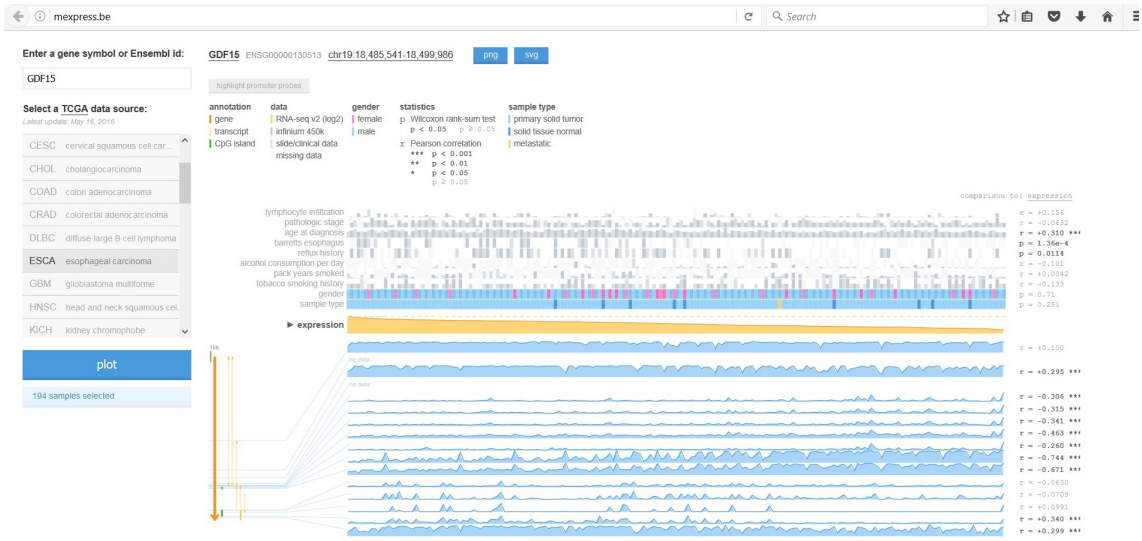


Figure 3.39: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for ESCA (Esophageal Carcinoma) cancer

MEXPRESS plot for GDF15 gene expression for ESCA cancer reveals the following details: A) there are more probes with a strong negative correlation as compared to the ones with a strong positive correlations between methylation and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 0.251 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.695

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer:(Figure 3.40)(Figure3.41)(Figure3.42)

MEXPRESS plot for GDF15 gene expression for HNSC cancer reveals the following details: A) there are more probes with a strong negative correlation as compared to the ones with a strong positive correlations between methylation and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene

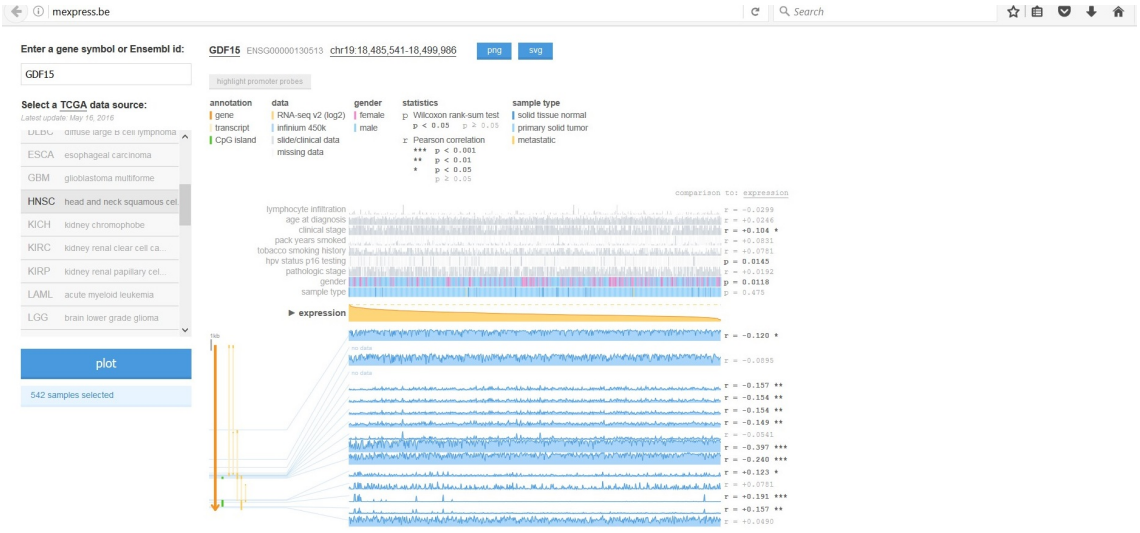


Figure 3.40: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer

Figure 3.41: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer

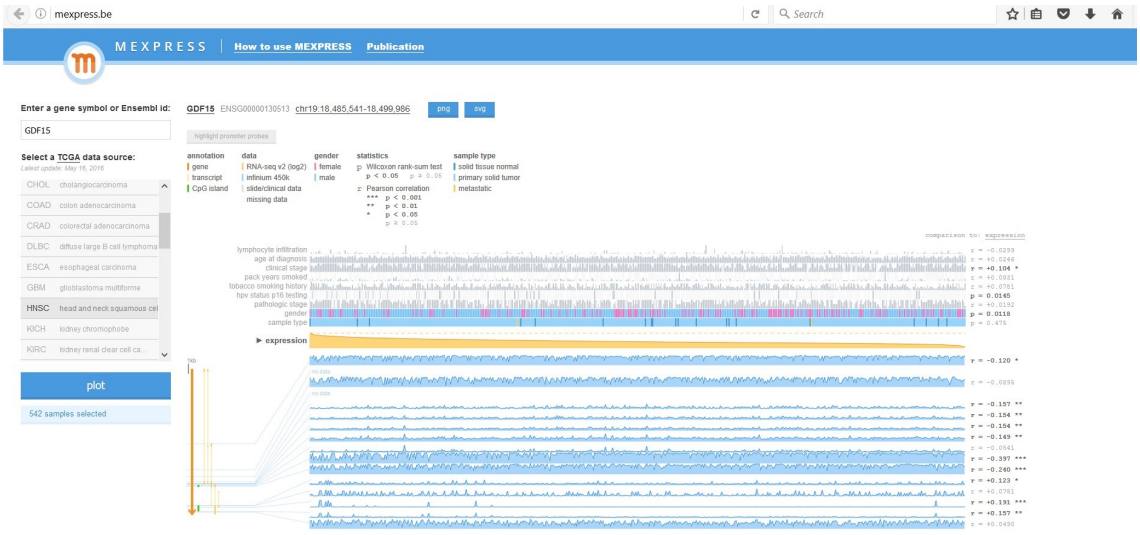


Figure 3.42: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer

expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 0.475 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.482

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma) cancer: (Figure 3.43)(Figure3.44)(Figure3.45)

MEXPRESS plot for GDF15 gene expression for KIRP cancer reveals the following details: A) there are more probes with a strong negative correlation as compared to the ones with a strong positive correlations between methylation and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower GDF15 expression than the tumor samples. C) Highlighted promoter

Figure 3.43: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma) cancer

Figure 3.44: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma) cancer

Figure 3.45: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma) cancer

probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 0.461 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.478

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LIHC (Liver Hepato Cellular Carcinoma) cancer: (Figure 3.46)(Figure3.47)(Figure3.48)

MEXPRESS plot for GDF15 gene expression for LIHC cancer reveals the following details: A) there are numerous strong negative correlation between methylation and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the

Figure 3.46: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LIHC (Liver Hepato Cellular Carcinoma) cancer

Figure 3.47: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LIHC (Liver Hepato Cellular Carcinoma) cancer

Figure 3.48: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LIHC (Liver Hepato Cellular Carcinoma) cancer

normal samples tend to have slightly lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 0.874 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.731

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer: (Figure 3.49)(Figure3.50)(Figure3.51)

MEXPRESS plot for GDF15 gene expression for LUAD cancer reveals the following details: A) there are numerous strong negative correlation between methylation and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the

Figure 3.49: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer

Figure 3.50: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer

Figure 3.51: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer

correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 0.00196 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.00436

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer: (Figure 3.52)(Figure3.53)(Figure3.54)

MEXPRESS plot for GDF15 gene expression for LUSC cancer reveals the following details: A) there are numerous strong positive correlation as compared to the ones with strong negative correlations between methylation and expression, indicating that GDF15

Figure 3.52: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer

Figure 3.53: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer

Figure 3.54: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer

expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 0.00692 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.00843

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for PRAD (Prostate Adeno Carcinoma) cancer: Figures 3.55 to 3.57(Figure 3.55)(Figure3.56)(Figure3.57)

MEXPRESS plot for GDF15 gene expression for PRAD cancer reveals the following

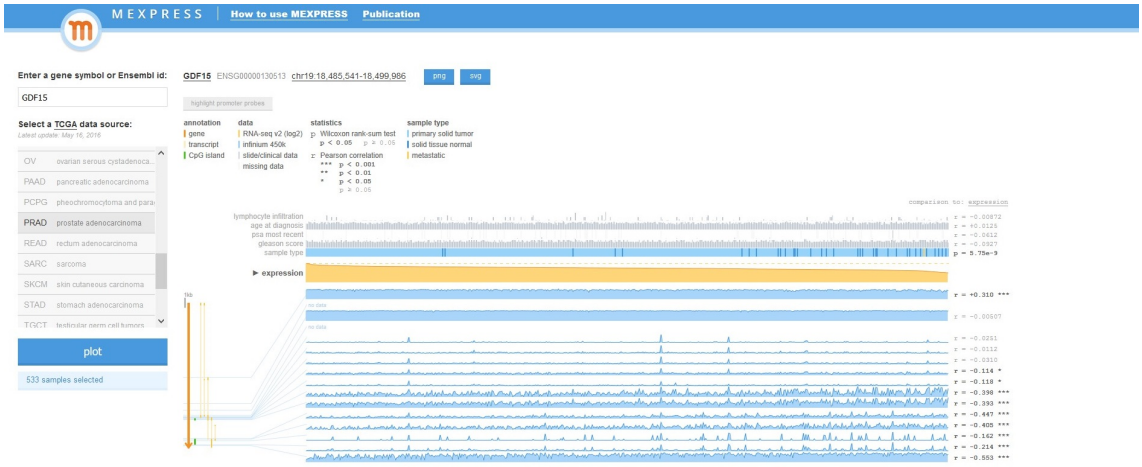


Figure 3.55: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for PRAD (Prostate Adeno Carcinoma) cancer

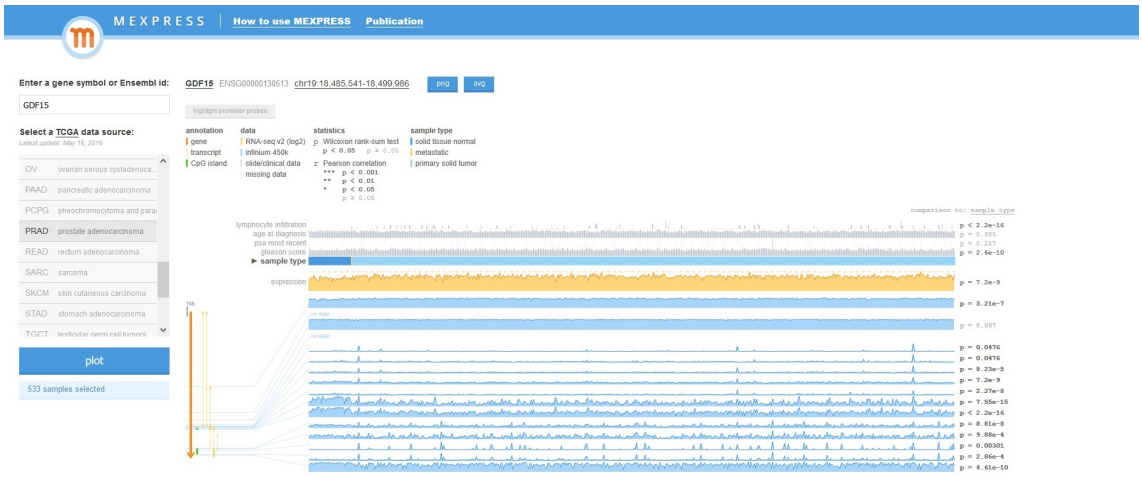


Figure 3.56: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for PRAD (Prostate Adeno Carcinoma) cancer

Figure 3.57: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for PRAD (Prostate Adeno Carcinoma) cancer

details: A) there are numerous strong negative correlation between methylation and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 5.75e-9 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 7.2e-9

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer: (Figure 3.58)(Figure3.59)(Figure3.60)

Figure 3.58: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer

Figure 3.59: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer

Figure 3.60: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer

MEXPRESS plot for GDF15 gene expression for THCA cancer reveals the following details: A) there are more number of strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type $p < 2.2e\text{-}16$ When samples are ordered by sample type i.e., difference in expression between normal and tumor type $p < 2.2e\text{-}16$

Figure 3.61: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer

Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer:(Figure 3.61)(Figure3.62)(Figure3.63)

MEXPRESS plot for GDF15 gene expression for UCEC cancer reveals the following details: A) there are slightly more number of strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that GDF15 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower GDF15 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes being highlighted indicating that the promoter region for GDF15 gene might NOT be involved in the regulation of GDF15 gene expression through DNA methylation. Also, the promoter region is NOT involved in influencing the methylation of CpG islands or its subsequent effect on GDF15 gene expression. When samples are ordered by expression, sample type p= 4.82e-8 When

Figure 3.62: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer

Figure 3.63: Analysis of GDF15 (Growth Differentiation Factor 15) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer

Figure 3.64: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer

samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 2.49e-6

3.3 PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) as a DNA methylation biomarker gene

(Figure 3.64)(Figure 3.65)(Figure 3.66)

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer:

MEXPRESS plot for PIWIL4 gene expression for BLCA cancer reveals the following details: A) there are numerous strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that PIWIL4 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to

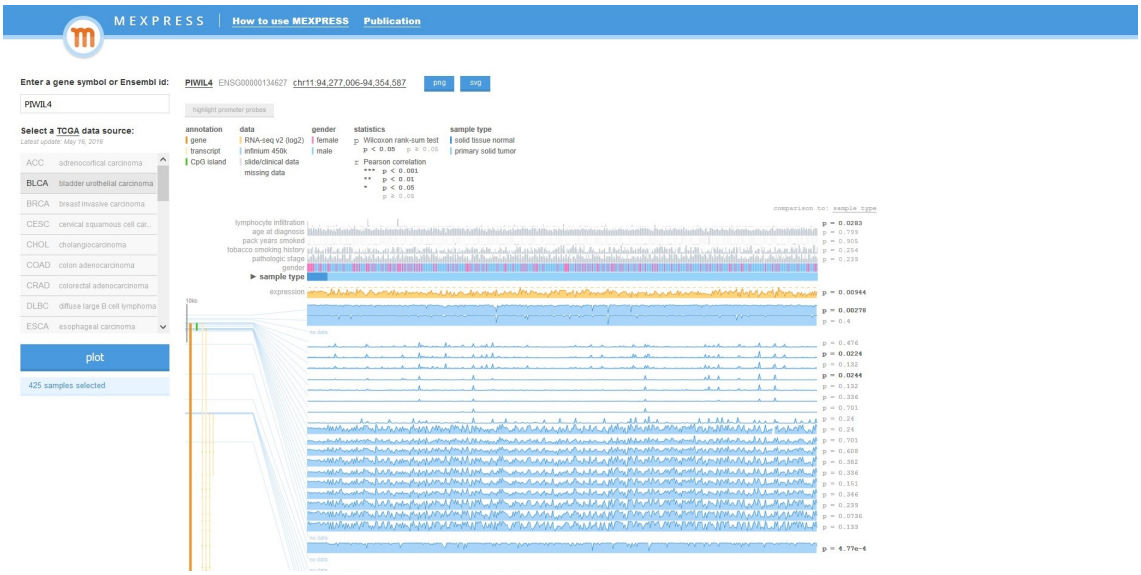


Figure 3.65: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer

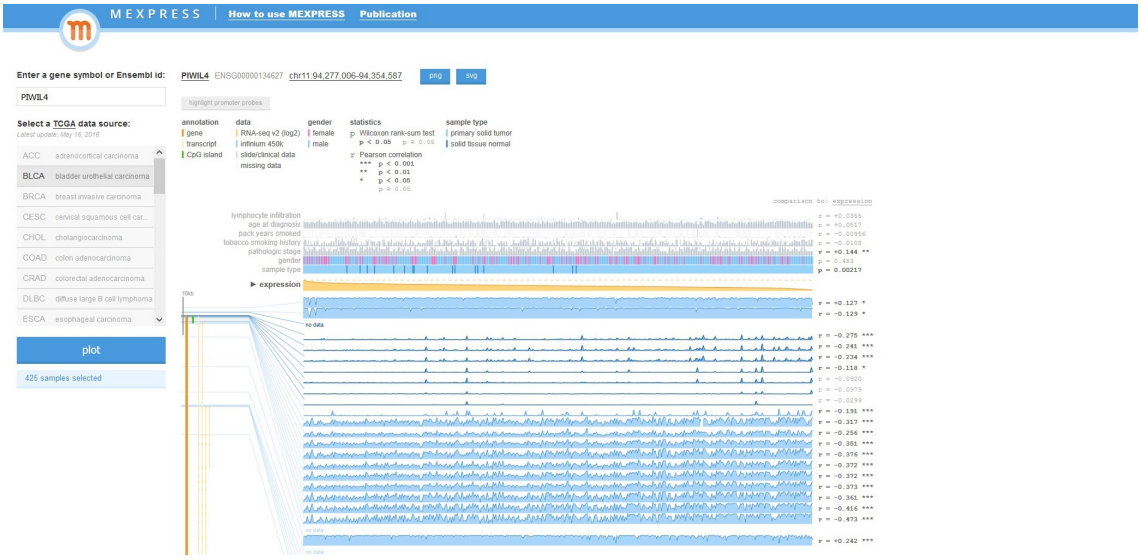


Figure 3.66: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer

Figure 3.67: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

have lower PIWIL4 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are few, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression. When samples are ordered by expression, sample type p= 0.00217 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.00944

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer: (Figure 3.67)(Figure 3.68)(Figure 3.69)

MEXPRESS plot for PIWIL4 gene expression for BRCA cancer reveals the following details: A) there are numerous strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that PIWIL4 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in

Figure 3.68: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

Figure 3.69: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

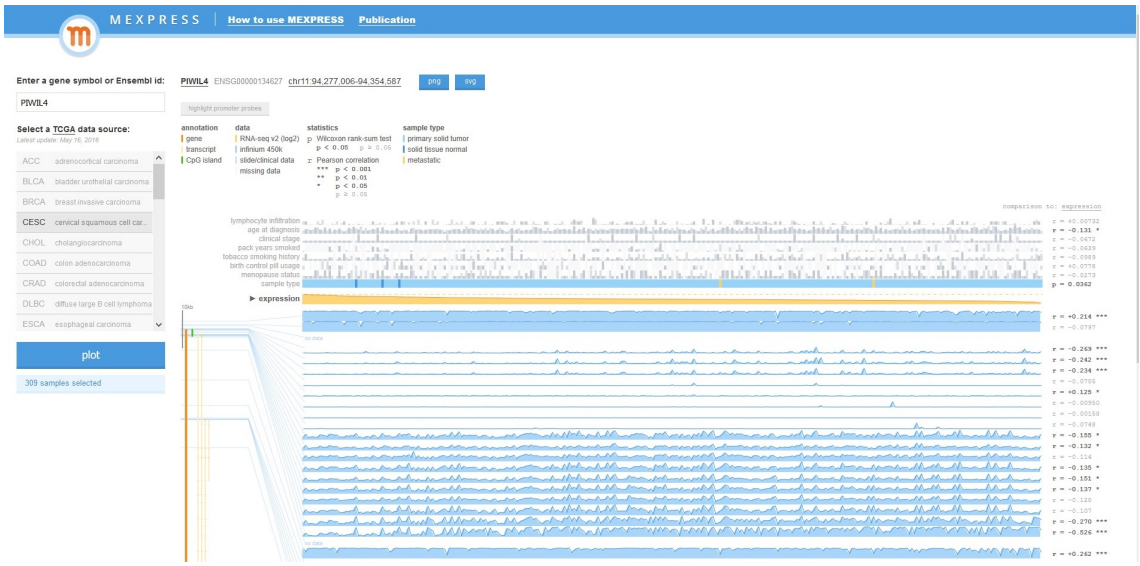


Figure 3.70: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CESC (Cervical Squamous Cell Carcinoma) cancer

all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have higher PIWIL4 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression. When samples are ordered by expression, sample type $p < 2.2e\text{-}16$ When samples are ordered by sample type i.e., difference in expression between normal and tumor type $p < 2.2e\text{-}16$

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CESC (Cervical Squamous Cell Carcinoma) cancer:(Figure 3.70)(Figure 3.71)(Figure 3.72)

MEXPRESS plot for PIWIL4 gene expression for CESC cancer reveals the following details: A) there are more number of strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that

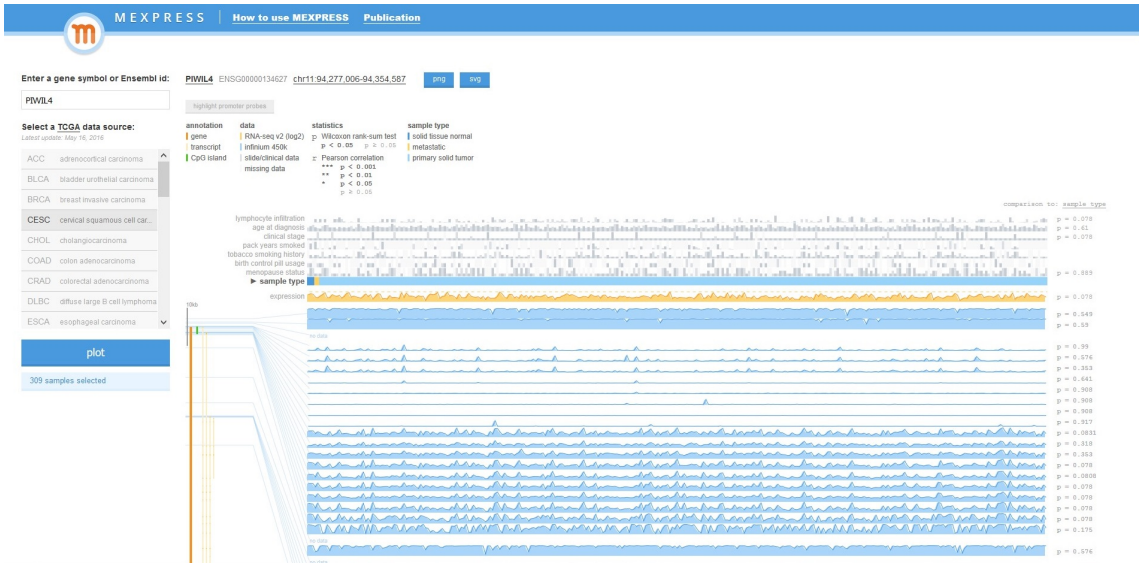


Figure 3.71: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CESC (Cervical Squamous Cell Carcinoma) cancer

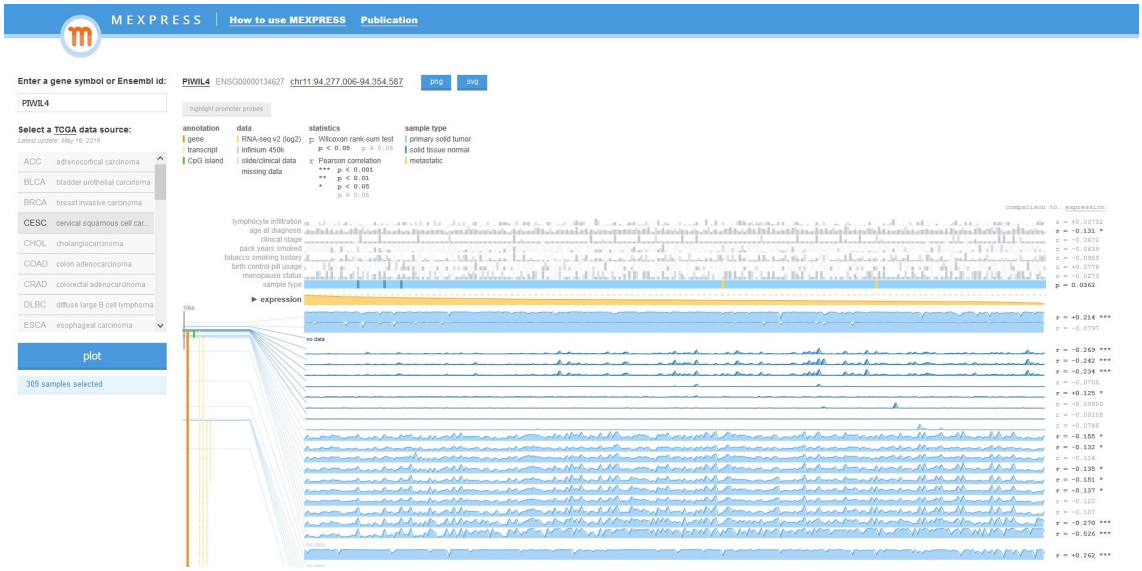


Figure 3.72: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CESC (Cervical Squamous Cell Carcinoma) cancer

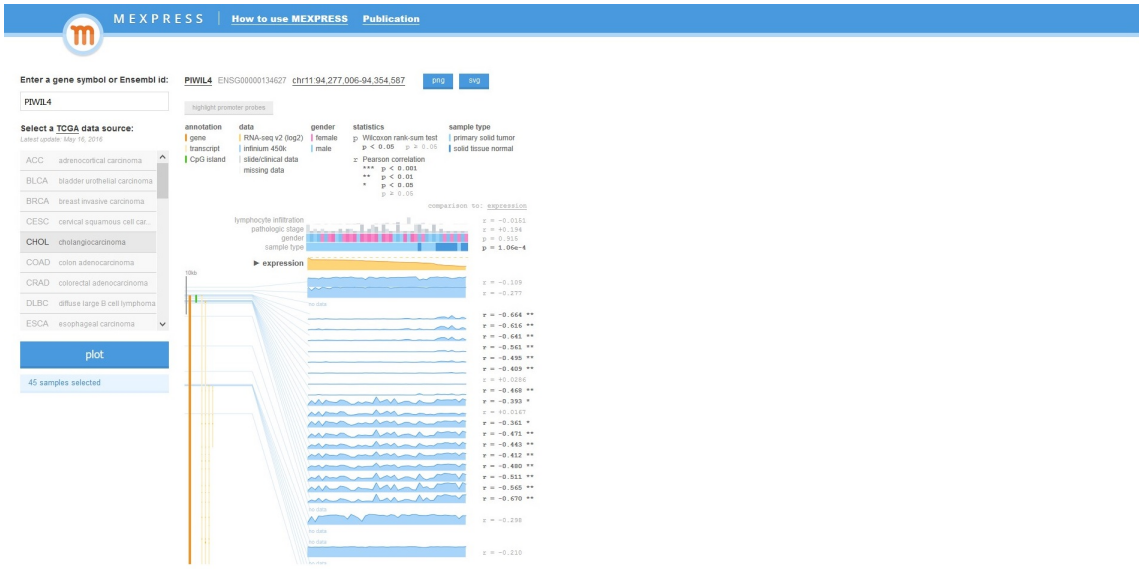


Figure 3.73: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CHOL (Cholangio Carcinoma) cancer

PIWIL4 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly higher PIWIL4 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are few, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression. When samples are ordered by expression, sample type p= 0.0362 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.078

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CHOL (Cholangio Carcinoma) cancer:(Figure 3.73)(Figure 3.74)(Figure 3.75)

MEXPRESS plot for PIWIL4 gene expression for CHOL cancer reveals the following

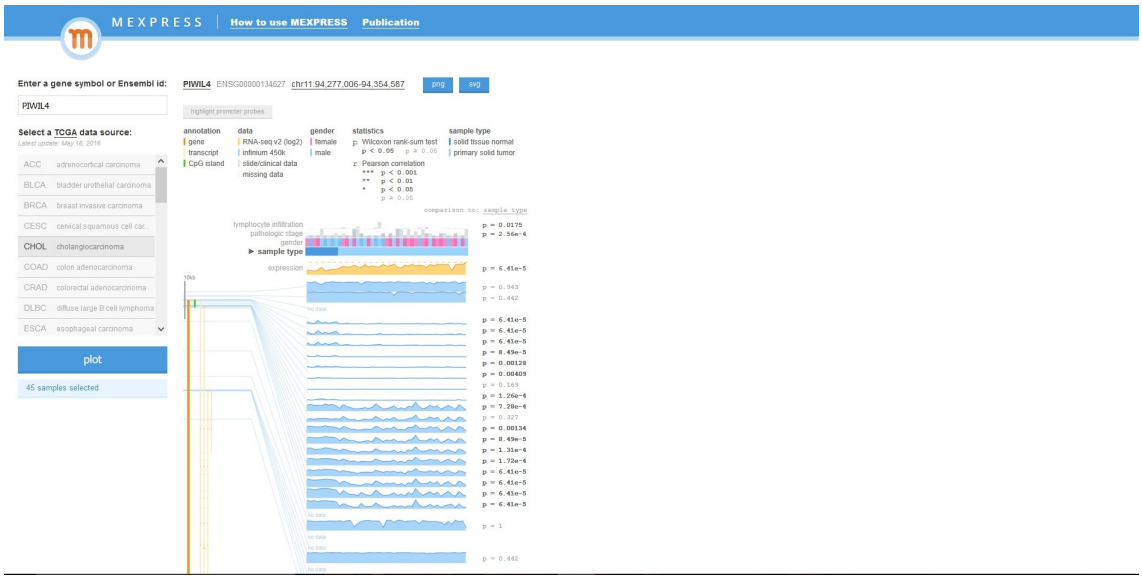


Figure 3.74: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CHOL (Cholangio Carcinoma) cancer

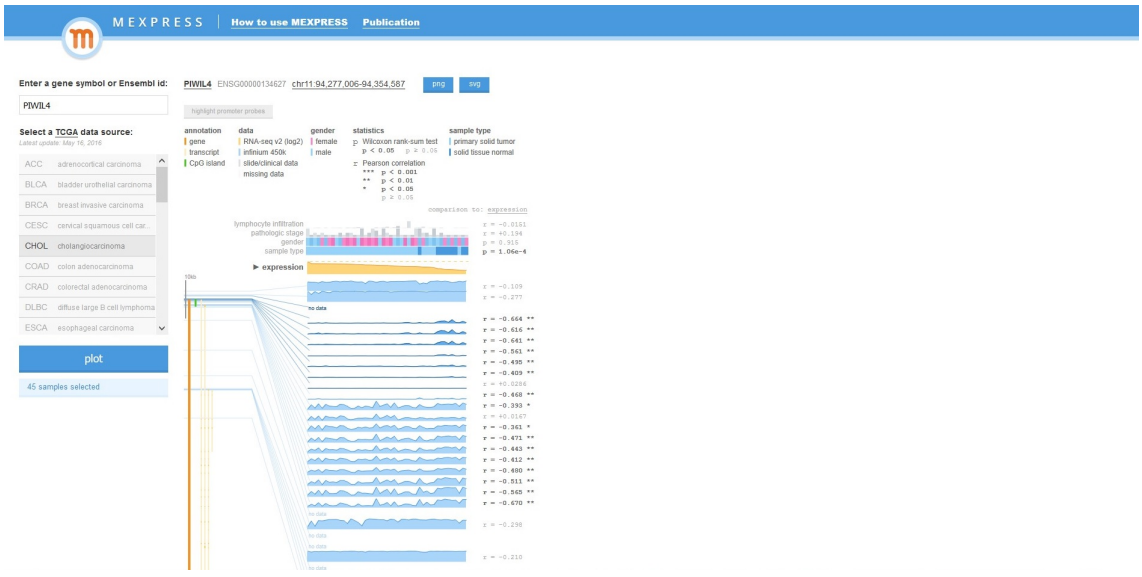


Figure 3.75: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CHOL (Cholangio Carcinoma) cancer

details: A) there are numerous strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that PIWIL4 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower PIWIL4 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are few, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression. When samples are ordered by expression, sample type p= 1.06e-4 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 6.41e-5

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma) cancer:(Figure 3.76)(Figure 3.77)(Figure 3.78)

MEXPRESS plot for PIWIL4 gene expression for COAD cancer reveals the following details: A) there are numerous strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that PIWIL4 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower PIWIL4 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4

Figure 3.76: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma) cancer

Figure 3.77: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma) cancer

Figure 3.78: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for COAD (Colon Adeno Carcinoma) cancer

gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression. When samples are ordered by expression, sample type p= 0.0363 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.0334

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CRAD (Colo Rectal Adeno Carcinoma) cancer:(Figure 3.79)(Figure 3.80)(Figure 3.81)

MEXPRESS plot for PIWIL4 gene expression for CRAD cancer reveals the following details: A) there are numerous strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that PIWIL4 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower PIWIL4 expression than the tumor samples. C) Highlighted promoter

Figure 3.79: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CRAD (Colo Rectal Adeno Carcinoma) cancer

Figure 3.80: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CRAD (Colo Rectal Adeno Carcinoma) cancer

Figure 3.81: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for CRAD (Colo Rectal Adeno Carcinoma) cancer

probes plot data reveals that there are few, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression. When samples are ordered by expression, sample type p= 0.0219 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.0206

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for ESCA (Esophageal Carcinoma) cancer:(Figure 3.82)(Figure 3.83)(Figure 3.84)

MEXPRESS plot for PIWIL4 gene expression for ESCA cancer reveals the following details: A) there are numerous strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that PIWIL4 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared

Figure 3.82: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for ESCA (Esophageal Carcinoma) cancer

Figure 3.83: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for ESCA (Esophageal Carcinoma) cancer

Figure 3.84: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for ESCA (Esophageal Carcinoma) cancer

between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly higher PIWIL4 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are few, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression. When samples are ordered by expression, sample type p= 0.491 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.887

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer:(Figure 3.85)(Figure3.86)(Figure3.87)

MEXPRESS plot for PIWIL4 gene expression for HNSC cancer reveals the following details: A) there are numerous strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that PIWIL4 expression might be controlled through DNA methylation. As the plot's legend

Figure 3.85: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer

Figure 3.86: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer

Figure 3.87: Analysis ofPIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer

explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower PIWIL4 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are few, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression. When samples are ordered by expression, sample type p= 0.895 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.900

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer:(Figure 3.88)(Figure3.89)(Figure3.90)

MEXPRESS plot for PIWIL4 gene expression for KIRC cancer reveals the following details: A) there are numerous strong negative correlation values of probes as compared to

Figure 3.88: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer

Figure 3.89: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer

Figure 3.90: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer

those of strong positive correlations, between methylation and expression, indicating that PIWIL4 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower PIWIL4 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are few, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression. When samples are ordered by expression, sample type p= 1.68e-10 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 4.7e-10

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma)

Figure 3.91: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma) cancer

cancer: (Figure 3.91)(Figure3.92)(Figure3.93)

MEXPRESS plot for PIWIL4 gene expression for KIRP cancer reveals the following details: A) there are numerous strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that PIWIL4 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower PIWIL4 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression. When samples are ordered by expression, sample type p = 0.283 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.327

Figure 3.92: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma) cancer

Figure 3.93: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Cell Carcinoma) cancer

Figure 3.94: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for LIHC (Liver Hepatocellular Carcinoma) cancer

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for LIHC (Liver Hepatocellular Carcinoma) cancer:(Figure 3.94)(Figure3.95)(Figure3.96)

MEXPRESS plot for PIWIL4 gene expression for LIHC cancer reveals the following details: A) there are numerous strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that PIWIL4 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly lower PIWIL4 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are numerous, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression.

Figure 3.95: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for LIHC (Liver Hepatocellular Carcinoma) cancer

Figure 3.96: Analysis ofPIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for LIHC (Liver Hepatocellular Carcinoma) cancer

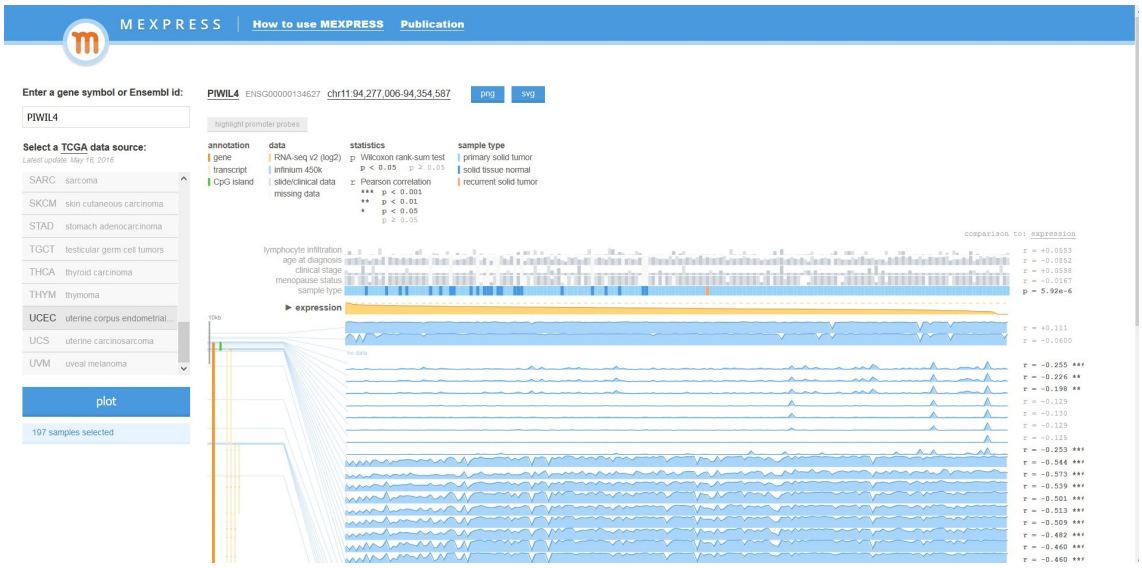


Figure 3.97: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer

When samples are ordered by expression, sample type p= 0.126 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.118

Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer:(Figure 3.97)(Figure3.98)(Figure3.99)

MEXPRESS plot for PIWIL4 gene expression for UCEC cancer reveals the following details: A) there are numerous strong negative correlation values of probes as compared to those of strong positive correlations, between methylation and expression, indicating that PIWIL4 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have slightly higher PIWIL4 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are few, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for PIWIL4 gene might be regulated through DNA methylation. Such promoter probes are also found on the

Figure 3.98: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer

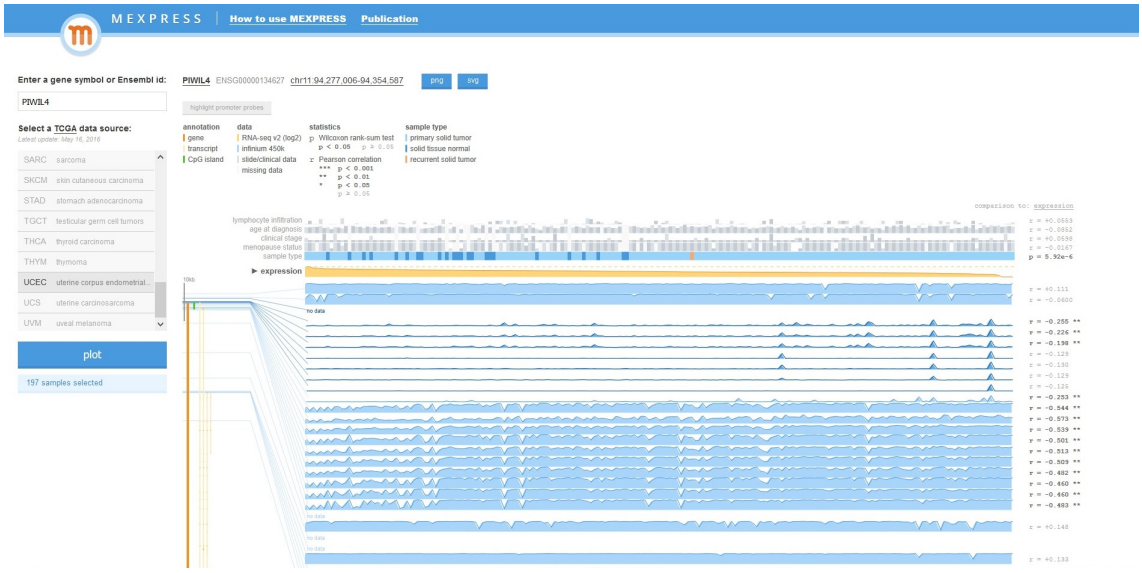


Figure 3.99: Analysis of PIWIL4 (Piwi Like RNA-Mediated Gene Silencing 4) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer

Figure 3.100: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence PIWIL4 gene expression. When samples are ordered by expression, sample type p= 5.92e-6 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 9.49e-4

3.4 DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) as a DNA methylation biomarker gene

(Figure 3.100)(Figure3.101)(Figure3.102)

Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer:

MEXPRESS plot for DMRT1 gene expression for BRCA cancer reveals the following details: A) there are numerous strong positive correlation values of probes as compared to those of strong negative correlations, between methylation and expression, indicating that DMRT1 expression might be controlled through DNA methylation. As the plot's

Figure 3.101: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

Figure 3.102: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

Figure 3.103: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer

legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have lower DMRT1 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are few, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for DMRT1 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence DMRT1 gene expression. When samples are ordered by expression, sample type p= 0.119 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.105

Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer:(Figure 3.103)(Figure3.104)(Figure3.105)

MEXPRESS plot for DMRT1 gene expression for LUSC cancer reveals the following details: A) there are numerous strong positive correlation values of probes as compared to

Figure 3.104: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer

Figure 3.105: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer

those of strong negative correlations, between methylation and expression, indicating that DMRT1 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower DMRT1 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are few, yet highly significant negative correlation values between methylation and expression indicating that the promoter region for DMRT1 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence DMRT1 gene expression. When samples are ordered by expression, sample type p= 0.00639 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.00553

Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer: (Figure 3.106)(Figure3.107)(Figure3.108)

MEXPRESS plot for DMRT1 gene expression for THCA cancer reveals the following details: A) there are numerous strong positive correlation values of probes as compared to those of strong negative correlations, between methylation and expression, indicating that DMRT1 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower DMRT1 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are few, yet highly significant positive correlation values between methylation and expression indicating that the promoter region for DMRT1 gene might be regulated through DNA methylation. Such promoter probes are also found

Figure 3.106: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer

Figure 3.107: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer

Figure 3.108: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer

on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence DMRT1 gene expression. When samples are ordered by expression, sample type p= 0.168 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.178

Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer:(Figure 3.109)(Figure3.110)(Figure3.111)

MEXPRESS plot for DMRT1 gene expression for UCEC cancer reveals the following details: A) there are numerous strong positive correlation values of probes as compared to those of strong negative correlations, between methylation and expression, indicating that DMRT1 expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower DMRT1 expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are very few, yet highly significant positive

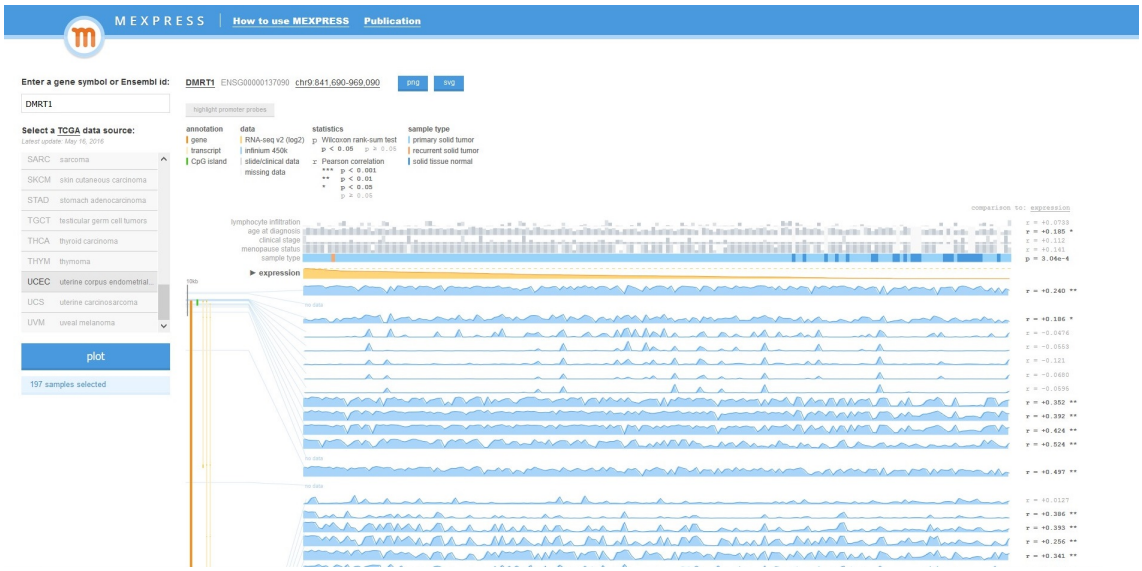


Figure 3.109: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer

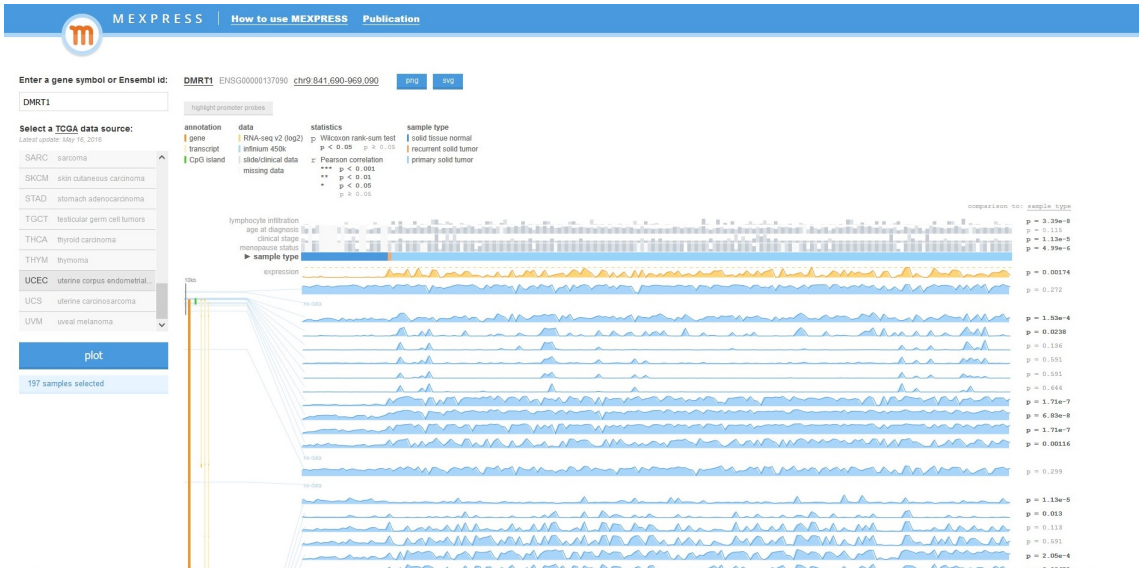


Figure 3.110: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer

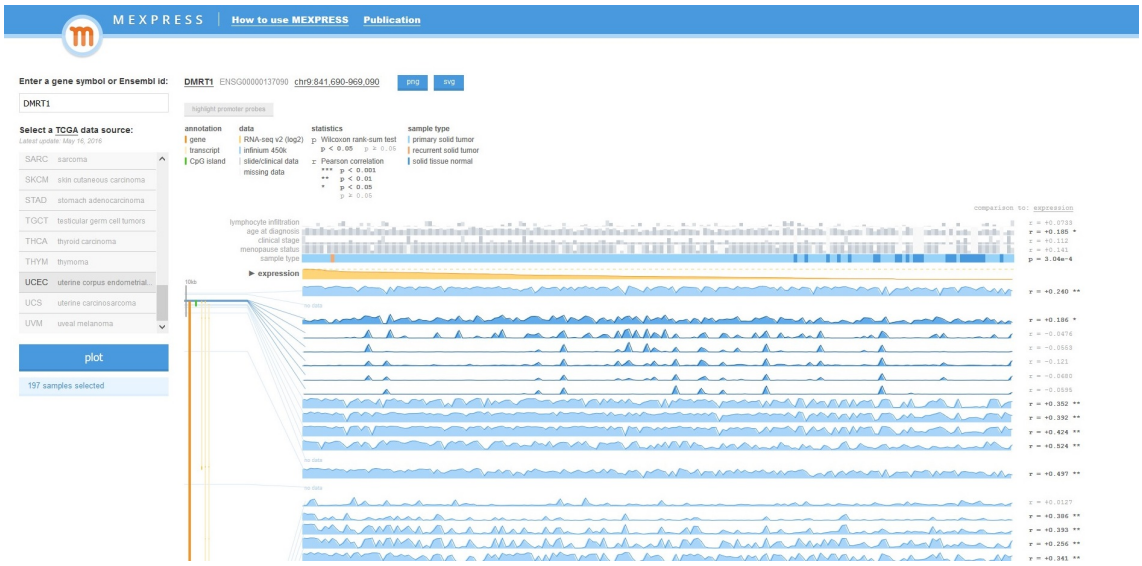


Figure 3.111: Analysis of DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) gene expression using MEXPRESS for UCEC (Uterine Corpus Endometrial Carcinoma) cancer

correlation values between methylation and expression indicating that the promoter region for DMRT1 gene might be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence DMRT1 gene expression. When samples are ordered by expression, sample type p= 3.04e-4 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.00174

3.5 ITPKA (inositol-trisphosphate 3-kinase A) as a DNA methylation biomarker gene

Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer:(Figure 3.112)(Figure3.113)(Figure3.114)

MEXPRESS plot for ITPKA gene expression for BLCA cancer reveals the following details: A) there are numerous strong positive correlation values of probes as compared to

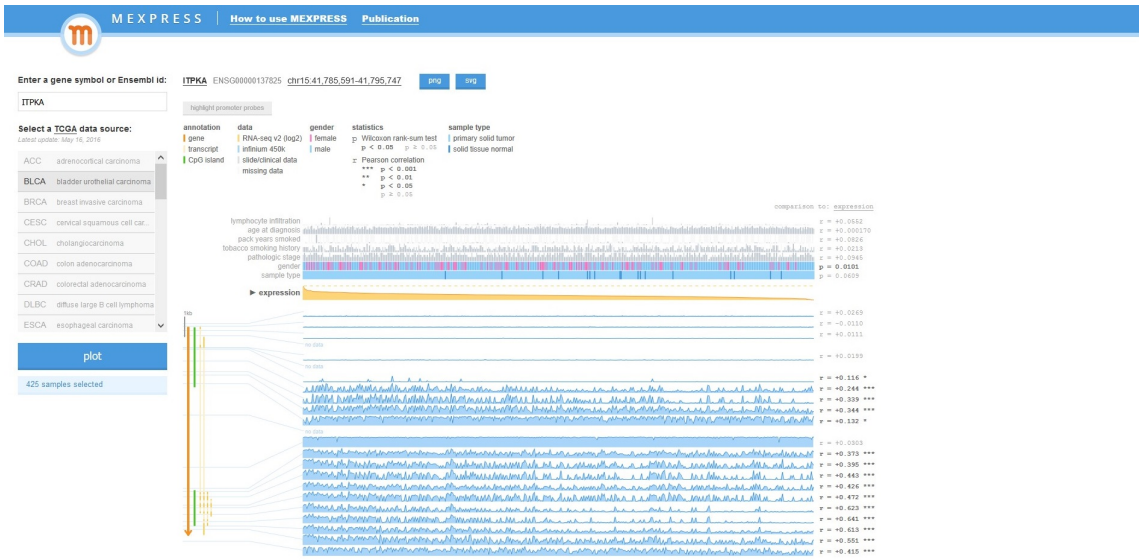


Figure 3.112: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer

Figure 3.113: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer

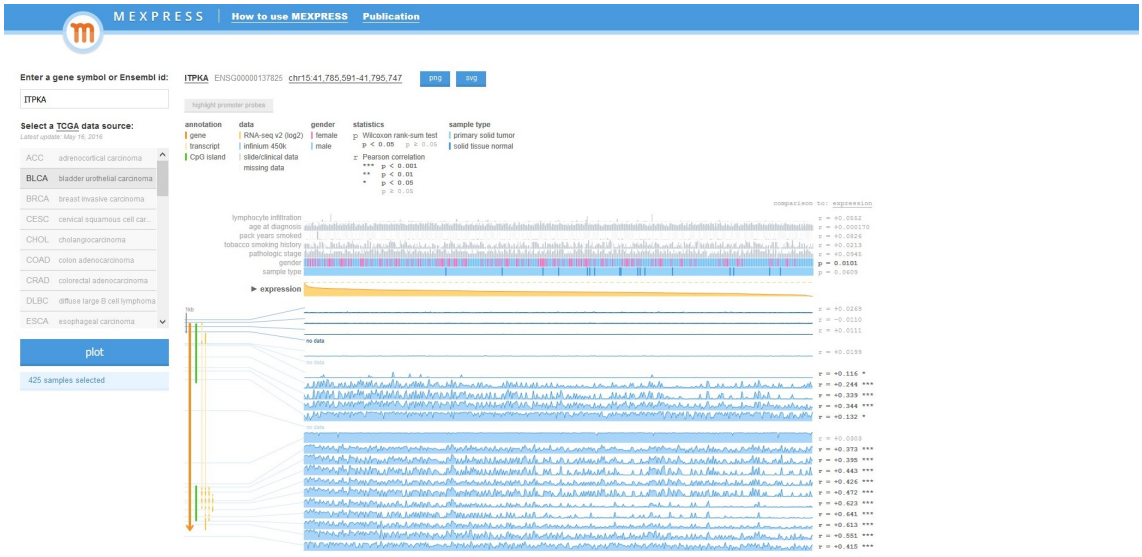


Figure 3.114: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for BLCA (Bladder Urothelial Carcinoma) cancer

those of strong negative correlations, between methylation and expression, indicating that ITPKA expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower ITPKA expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are very few and slightly positive correlation values between methylation and expression indicating that the promoter region for ITPKA gene might or might not be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence ITPKA gene expression. When samples are ordered by expression, sample type p= 0.0609 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p= 0.0996

Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using

Figure 3.115: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer:(Figure 3.115)(Figure3.116)(Figure3.117)

MEXPRESS plot for ITPKA gene expression for BRCA cancer reveals the following details: A) there are numerous strong positive correlation values of probes as compared to those of strong negative correlations, between methylation and expression, indicating that ITPKA expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower ITPKA expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are just a couple of probes and slightly negative correlation values between methylation and expression indicating that the promoter region for ITPKA gene might or might not be regulated through DNA methylation. Such promoter probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence ITPKA gene expression. When samples are ordered by expression, sample type $p < 2.2e\text{-}16$

Figure 3.116: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

Figure 3.117: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for BRCA (Breast Invasive Carcinoma) cancer

Figure 3.118: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer

When samples are ordered by sample type i.e., difference in expression between normal and tumor type p < 2.2e-16

Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer: (Figure 3.118)(Figure3.119)(Figure3.120)

MEXPRESS plot for ITPKA gene expression for HNSC cancer reveals the following details: A) there are numerous strong positive correlation values of probes as compared to those of strong negative correlations, between methylation and expression, indicating that ITPKA expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower ITPKA expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are just a couple of probes and low negative correlation values between methylation and expression indicating that the promoter region for ITPKA gene might or might not be regulated through DNA methylation. Such promoter

Figure 3.119: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer

Figure 3.120: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for HNSC (Head and Neck Squamous Cell Carcinoma) cancer

Figure 3.121: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer

probes are also found on the CpG island region (indicated in green color) indicating that DNA methylation has an effect on the CpG island region which can subsequently influence ITPKA gene expression. When samples are ordered by expression, sample type p = 6.46e-4 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p = 8.42e-4

Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer:(Figure 3.121)(Figure3.122)(Figure3.123)

MEXPRESS plot for ITPKA gene expression for KIRC cancer reveals the following details: A) there are numerous strong positive correlation values of probes as compared to those of strong negative correlations, between methylation and expression, indicating that ITPKA expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower ITPKA expression than the tumor samples. C) Highlighted

Figure 3.122: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer

Figure 3.123: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for KIRC (Kidney Renal Clear Cell Carcinoma) cancer

Figure 3.124: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Carcinoma) cancer

promoter probes plot data reveals that there is just one probe with a low negative correlation values between methylation and expression indicating that the promoter region for ITPKA gene might or might not be regulated through DNA methylation. Such a promoter probe is also found very close to CpG island region (indicated in green color) indicating that DNA methylation might have an effect on the CpG island region which can subsequently influence ITPKA gene expression. When samples are ordered by expression, sample type p = 9.47e-9 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p = 2.03e-8

Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Carcinoma) cancer:(Figure 3.124)(Figure3.125)(Figure3.126)

MEXPRESS plot for ITPKA gene expression for KIRP cancer reveals the following details: A) there are numerous strong positive correlation values of probes as compared to those of strong negative correlations, between methylation and expression, indicating that ITPKA expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all

Figure 3.125: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Carcinoma) cancer

Figure 3.126: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for KIRP (Kidney Renal Papillary Carcinoma) cancer

Figure 3.127: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LIHC (Liver Hepatocellular Carcinoma) cancer

MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower ITPKA expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are few probes with very low negative correlation values between methylation and expression indicating that the promoter region for ITPKA gene might or might not be regulated through DNA methylation. Such a promoter probe is also found in the CpG island region (indicated in green color) indicating that DNA methylation might have an effect on the CpG island region which can subsequently influence ITPKA gene expression. When samples are ordered by expression, sample type p = 3.84e-7 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p = 1.94e-6

Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LIHC (Liver Hepatocellular Carcinoma) cancer:(Figure 3.127)(Figure3.128)(Figure3.129)

MEXPRESS plot for ITPKA gene expression for LIHC cancer reveals the following details: A) there are a couple of strong negative correlation values of probes as compared to

Figure 3.128: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LIHC (Liver Hepatocellular Carcinoma) cancer

Figure 3.129: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LIHC (Liver Hepatocellular Carcinoma) cancer

those of strong positive correlations, between methylation and expression, indicating that ITPKA expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower ITPKA expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are couple probes with very low negative correlation values between methylation and expression indicating that the promoter region for ITPKA gene might or might not be regulated through DNA methylation. Such a promoter probe is also found in the CpG island region (indicated in green color) indicating that DNA methylation might have an effect on the CpG island region which can subsequently influence ITPKA gene expression. When samples are ordered by expression, sample type p = 2.53e-11 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p = 2.44e-11

Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer:(Figure 3.130)(Figure3.131)(Figure3.132)

MEXPRESS plot for ITPKA gene expression for LUAD cancer reveals the following details: A) there are a number of strong positive correlation values of probes as compared to those of strong negative correlations, between methylation and expression, indicating that ITPKA expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower ITPKA expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are couple probes with very low negative correlation values between methylation and expression indicating that the promoter region for ITPKA gene might or might not be regulated through DNA methylation. Such a promoter

Figure 3.130: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer

Figure 3.131: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer

Figure 3.132: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LUAD (Lung Adeno Carcinoma) cancer

probe is also found in the CpG island region (indicated in green color) indicating that DNA methylation might have an effect on the CpG island region which can subsequently influence ITPKA gene expression. When samples are ordered by expression, sample type p = 1.05e-10 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p = 2.52e-10

Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer:(Figure 3.133)(Figure 3.134)(Figure 3.135)

MEXPRESS plot for ITPKA gene expression for LUSC cancer reveals the following details: A) there are a number of strong positive correlation values of probes as compared to those of strong negative correlations, between methylation and expression, indicating that ITPKA expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower ITPKA expression than the tumor samples. C) Highlighted

Figure 3.133: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer

Figure 3.134: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer

Figure 3.135: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for LUSC (Lung Squamous Cell Carcinoma) cancer

promoter probes plot data reveals that there are couple probes with very low negative correlation values between methylation and expression indicating that the promoter region for ITPKA gene might or might not be regulated through DNA methylation. Such a promoter probe is also found in the CpG island region (indicated in green color) indicating that DNA methylation might have an effect on the CpG island region which can subsequently influence ITPKA gene expression. When samples are ordered by expression, sample type p = 2.23e-5 When samples are ordered by sample type i.e., difference in expression between normal and tumor type p = 7.77e-5

Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer:(Figure 3.136)(Figure 3.137)(Figure 3.138)

MEXPRESS plot for ITPKA gene expression for THCA cancer reveals the following details: A) there are a number of strong positive correlation values of probes as compared to those of strong negative correlations, between methylation and expression, indicating that ITPKA expression might be controlled through DNA methylation. As the plot's legend explains, the asterisks gives an indication of the significance of the correlations. B) As in

Figure 3.136: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer

Figure 3.137: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer

Figure 3.138: Analysis of ITPKA (inositol-trisphosphate 3-kinase A) gene expression using MEXPRESS for THCA (Thyroid Carcinoma) cancer

all MEXPRESS plotting, for the sample type parameter, the expression is always compared between the normal and tumor samples. Here, it is clear that the normal samples tend to have significantly lower ITPKA expression than the tumor samples. C) Highlighted promoter probes plot data reveals that there are no probes with low correlation values between methylation and expression indicating that the promoter region for ITPKA gene might not be regulated through DNA methylation. Such a promoter probe is also not found in the CpG island region (indicated in green color) indicating that DNA methylation might not have an effect on the CpG island region which can subsequently influence ITPKA gene expression. When samples are ordered by expression, sample type $p < 2.2e\text{-}16$ When samples are ordered by sample type i.e., difference in expression between normal and tumor type $p = 5.33e\text{-}15$

The below table is a comprehensive listing for the five gene analyzed using MEXPRESS tool. The significance of the relation as determined by the p-value is listed in the above table for each of the gene analyzed against 34 cancer types. The p-value so obtained is using the default setting in MEXPRESS tool. These p-values are obtained when samples are ordered by their expression values. Samples with the highest expression values are placed on the left

	SAMPLE SIZE	BLCAP	GDF15	PIWIL4	DMRT1	ITPKA
ACC	79					
BLCA	425	6.02E-04				
BRCA	871	< 2.2e-16	2.09E-14	< 2.2e-16		< 2.2e-16
CESC	309					
CHOL	45			1.06E-04		
COAD	299		1.37E-09			1.12E-09
CRAD	394		5.38E-10			6.21E-11
DLBC	48					
ESCA	194					
GBM	65					
HNSC	542					6.46E-04
KICH	66					
KIRC	343	2.77E-10	1.65E-07	1.68E-10		9.47E-09
KIRP	298					3.84E-07
LAML	170					
LGG	530					
LIHC	414					2.53E-11
LUAD	477					1.05E-10
LUSC	378					2.23E-05
MESO	87					
OV	299	NIL	NIL	NIL	NIL	NIL
PAAD	183					
PCPG	187					
PRAD	533		5.75E-09			4.22E-07
READ	95					
SARC	263					
SKCM	472					
STAD	372					
TGCT	156				6.70E-04	
THCA	563		< 2.2e-16	5.18E-12		< 2.2e-16
THYM	122					
UCEC	197		4.82E-08	5.92E-06	3.04E-04	
UCS	57		-0.16		NIL	
UVM	80					

Figure 3.139: Comprehensive Result Table of Gene analysis using MEXPRESS and their p-or significance values (When samples are ordered by value of their expression i.e., by using MEXPRESS default setting

and those with the lowest expression values are placed at the right of the line plot (yellow line). These expression values are the logarithm of the level3 RNASeqV2 values. These RNASeq values are normalized values for a gene. It must be noted that the expression data forms the basis of the whole plot, because the samples are ranked based on their expression value for the gene we selected with the highest expression on the left side and the lowest on the right. Sample size in the table indicates the number of samples or patients from whom the samples were obtained. Most significant p-values are indicated in red, meaning in these cancer types, the gene expression is highly influenced by the corresponding DNA methylation either in their promoter or regulatory region. Samples can also be re-ordered by sample type, which is always a measure of or which indicates the difference between the expression values of normal vs tumor samples. The above table is an indication of several hits or leads obtained in terms of the gene of our interest being considered as a DNA methylation biomarker gene. (Figure 3.139)

Overall results of the five genes analyzed (Figure 3.140)(Figure 3.141)(Figure 3.142)(Figure 3.143)(Figure 3.144)(Figure 3.145)(Figure 3.146)(Figure 3.147)(Figure 3.148)(Figure 3.149)

Representative results of querying ITPKA gene Vs BioMuta and BioXpress

When ITPKA is queried against BioMuta, the results are indicated as follows: 66 possible singlenucleotide variations (SNVs) are identified for the ITPKA gene with the highest number found for Urinary bladder cancer and Lung cancer. The tabular result indicates important information such as the chromosome number and position at which the SNV is found and its possible phenotypic effect. Results also show that five of the 66 SNVs are nsSNVs that affect functional sites (three gain of phosphorylation and two gain of glycosylation). Fig 2B: Pie-chart representing different cancer types and the number of positions affected by SNVs in them. Utility: BioMuta is a curated single-nucleotide variation (SNV) and disease association database. It is an important source of variations, particularly because the variations are mapped to the genome/protein/gene. Such query helps to identify variations and since the database is compiled from various sources through biocuration, it paves ways

Cancer Type	p-values: Samples ordered by expression values	p-values: Samples ordered by Sample type	Majority of correlation values of methylation probes	Significance of relationship	CpG island present in promoter region
BLCA	6.02e-4	9.17e-4	Negative	* * *	Yes
BRCA	<2.2e-16	<2.2e-16	Negative	* * *	Yes
COAD	0.0562	0.0687	Negative	* * *	Yes
CRAD	0.023	0.0249	Negative	* * *	Yes
KIRC	2.77e-10	1.36e-10	Negative	* * *	Yes
KIRP	0.246	0.179	Negative	* * *	Yes
LUAD	0.578	0.623	Negative	* * *	Yes
LUSC	0.00118	0.00128	Negative	* * *	Yes

Figure 3.140: Overall analysis of BLCAP gene as a biomarker using MEXPRESS tool

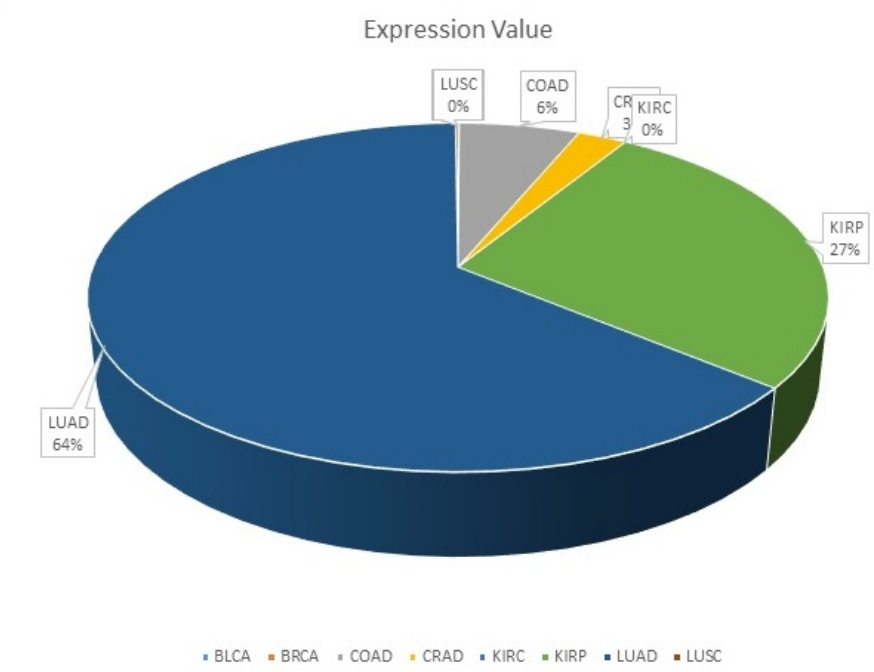


Figure 3.141: Overall analysis of BLCAP gene as a biomarker using MEXPRESS tool

Cancer Type	p-values: Samples ordered by expression values	p-values: Samples ordered by Sample type	Majority of correlation values of methylation probes	Significance of relationship	CpG island present in promoter region
BRCA	2.09e-14	1.31e-13	Negative	* * *	No
COAD	1.37e-9	6.62e-10	Negative	* * *	No
CRAD	5.38e-10	5.49e-10	Negative	* *	No
CESC	0.31	0.76	Negative	* *	No
ESCA	0.251	0.695	Negative	* *	No
HNSC	0.475	0.482	Negative	* * *	No
KIRP	0.461	0.478	Negative	* * *	No
LIHC	0.874	0.731	Negative	* * *	No
LUAD	0.00196	0.00436	Negative	* * *	No
LUSC	0.00692	0.00843	Positive	* * *	No
PRAD	5.75e-9	7.2e-9	Negative	* * *	No
THCA	<2.2e-16	<2.2e-16	Negative	* * *	No
UCEC	4.82e-8	2.49e-6	Negative	* *	No

Figure 3.142: Overall analysis of GDF15 gene as a biomarker using MEXPRESS tool

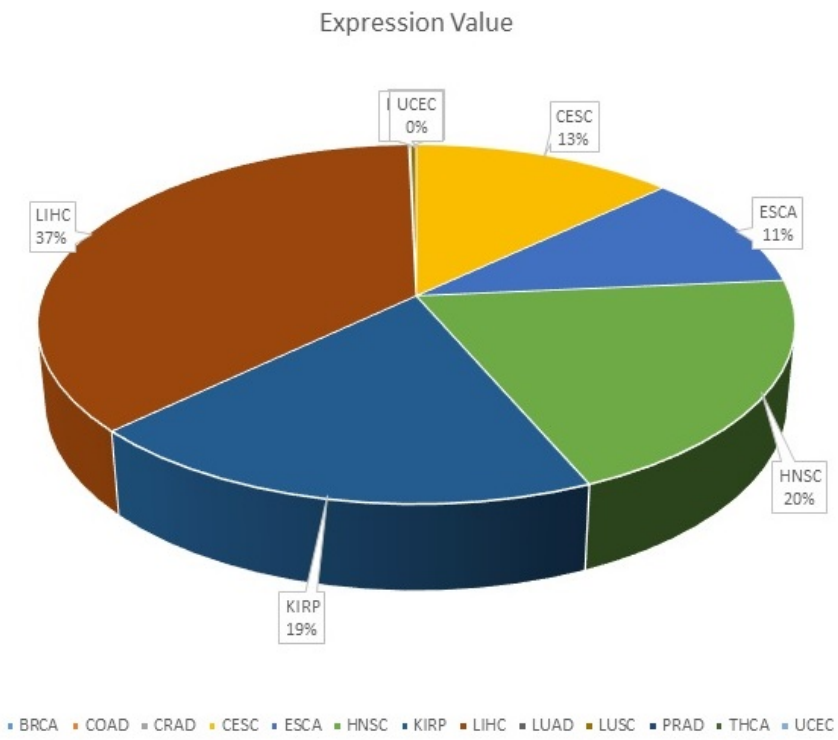


Figure 3.143: Overall analysis of GDF15 gene as a biomarker using MEXPRESS tool

for prioritizing variations for further experimental validations (Figure 3.150)(Figure 3.151)

When ITPKA is queried against BioXpress, the results obtained are as follows: the ITPKA gene is shown to be over-expressed in Thyroid Carcinoma (THCA), which validates and confirms our findings from the MEXPRESS study. Our MEXPRESS study also reveals that ITPKA exhibits epigenetic aberrations in other cancer types such as BRCA, COAD, CRAD, HNSC, KIRC, KIRP, LIHC, LUAD, LUSC and PRAD, which is also reproduced here (clear differential expression observed) when queried against BioXpress. Fig3B: Pie-chart representing different cancer types and the over-expression of ITPKA gene in percent. Utility: BioXpress is a curated gene expression and disease association database where the expression levels are mapped to genes. BioXpress is useful in identifying differences between expression levels in disease and normal pairs and to discover differential expression for a gene. It also helps in identification of potential biomarkers or pathways that lead to tumor formation or to explore the overall expression of specific genes across multiple cancer types. BioXpress can be queried using HGNC-approved gene symbols (HUGO Gene

Cancer Type	p-values: Samples ordered by expression values	p-values: Samples ordered by Sample type	Majority of correlation values of methylation probes	Significance of relationship	CpG island present in promoter region
BLCA	0.00217	0.00944	Negative	* * *	Yes
BRCA	<2.2e-16	<2.2e-16	Negative	* * *	Yes
CESC	0.0362	0.078	Negative	* * *	Yes
CHOL	1.06e-4	6.41e-5	Negative	* *	Yes
COAD	0.0363	0.0334	Negative	* * *	Yes
CRAD	0.0219	0.0206	Negative	* * *	Yes
ESCA	0.491	0.887	Negative	* *	Yes
HNSC	0.895	0.900	Negative	* * *	Yes
KIRC	1.68e-10	4.7e-10	Negative	* * *	Yes
KIRP	0.283	0.327	Negative	* * *	Yes
LIHC	0.126	0.118	Negative	* * *	Yes
UCEC	5.92e-6	9.49e-4	Negative	* *	Yes

Figure 3.144: Overall analysis of PIWIL4 gene as a biomarker using MEXPRESS tool

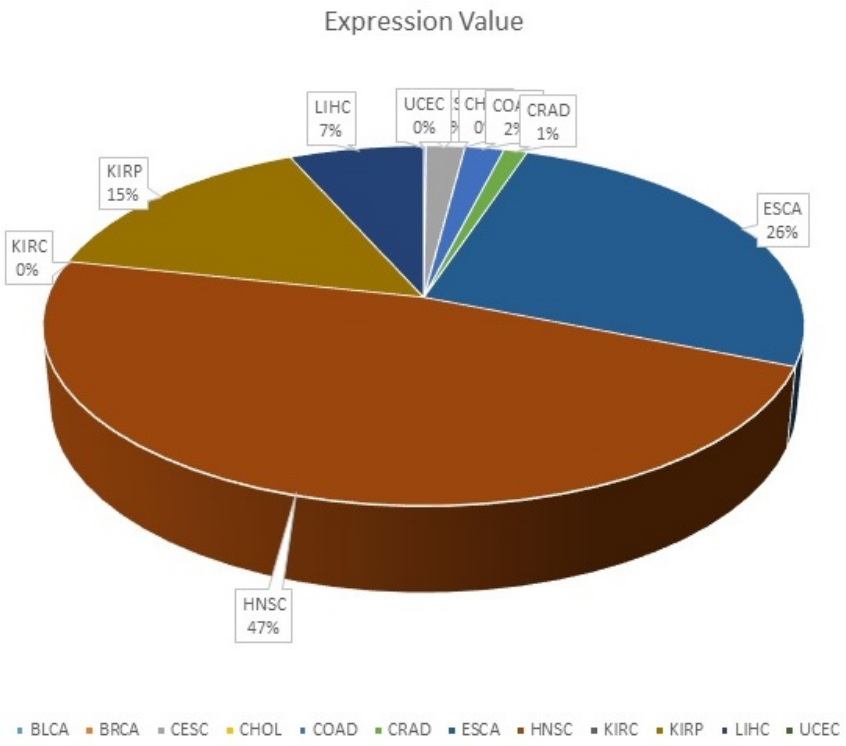


Figure 3.145: Overall analysis of PIWIL4 gene as a biomarker using MEXPRESS tool

Cancer Type	p-values: Samples ordered by expression values	p-values: Samples ordered by Sample type	Majority of correlation values of methylation probes	Significance of relationship	CpG island present in promoter region
BRCA	**0.119**	**0.105**	**Positive**	*** * ***	**Yes**
LUSC	**0.00639**	**0.00553**	**Positive**	*** * ***	**Yes**
THCA	**0.168**	**0.178**	**Positive**	*** * ***	**Yes**
UCEC	**3.04e-4**	**0.00174**	**Positive**	*** ***	**Yes**

Figure 3.146: Overall analysis of DMRT1 gene as a biomarker using MEXPRESS tool

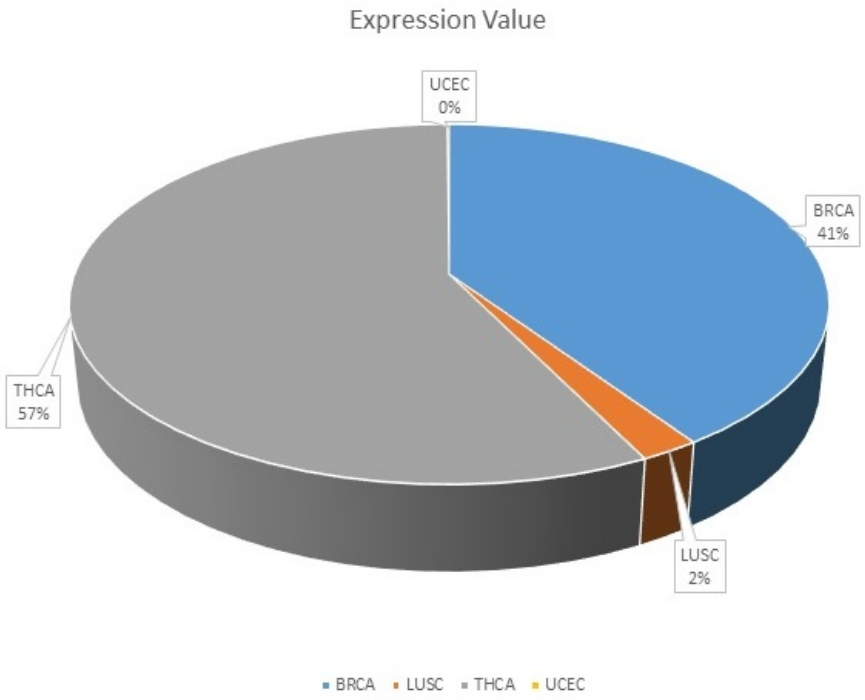


Figure 3.147: Overall analysis of DMRT1 gene as a biomarker using MEXPRESS tool

Cancer Type	p-values: Samples ordered by expression values	p-values: Samples ordered by Sample type	Majority of correlation values of methylation probes	Significance of relationship	CpG island present in promoter region
BLCA	0.0609	0.0996	Positive	* * *	Yes
BRCA	<2.2e-16	<2.2e-16	Positive	* * *	Yes
HNSC	6.46e-4	8.42e-4	Positive	* * *	Yes
KIRC	9.47e-9	2.03e-8	Positive	* * *	Yes
KIRP	3.84e-7	1.94e-6	Positive	* * *	Yes
LIHC	2.53e-11	2.44e-11	Negative	* * *	Yes
LUAD	1.05e-10	2.52e-10	Positive	* * *	Yes
LUSC	2.23e-5	7.77e-5	Positive	* * *	Yes

Figure 3.148: Overall analysis of ITPKA gene as a biomarker using MEXPRESS tool

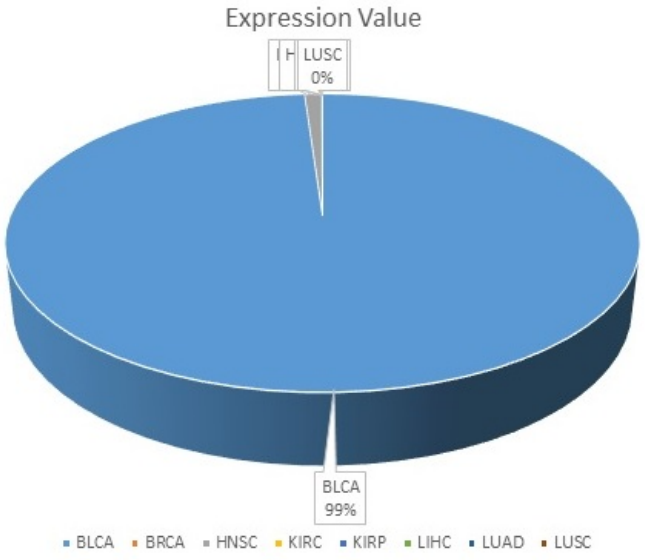


Figure 3.149: Overall analysis of ITPKA gene as a biomarker using MEXPRESS tool

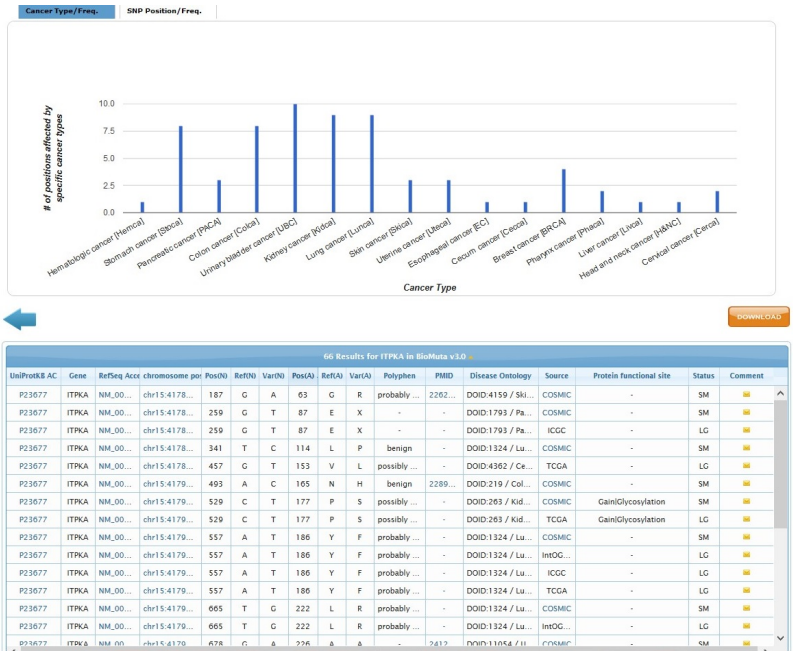


Figure 3.150: Overall analysis of ITPKA gene as a biomarker using MEXPRESS tool

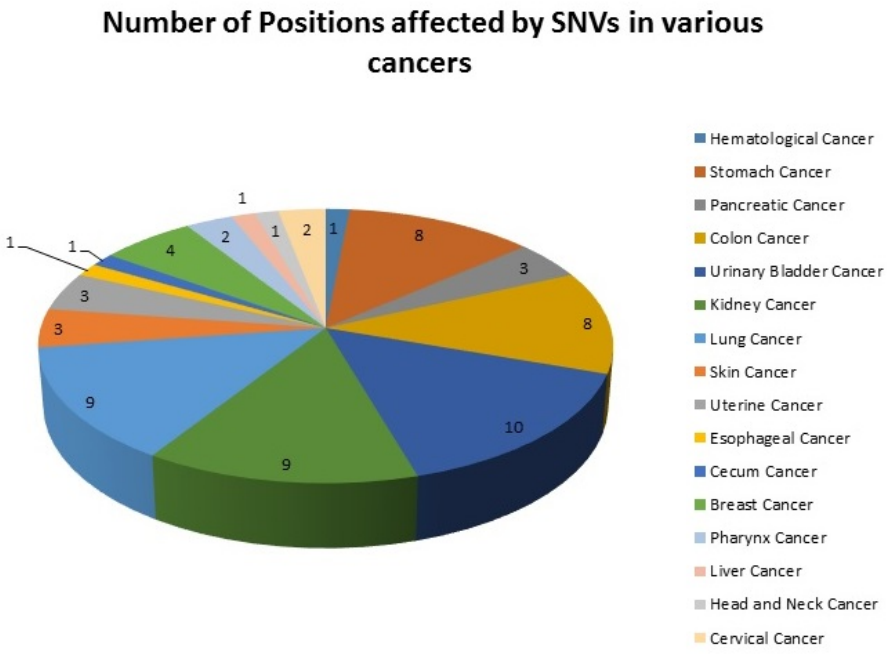
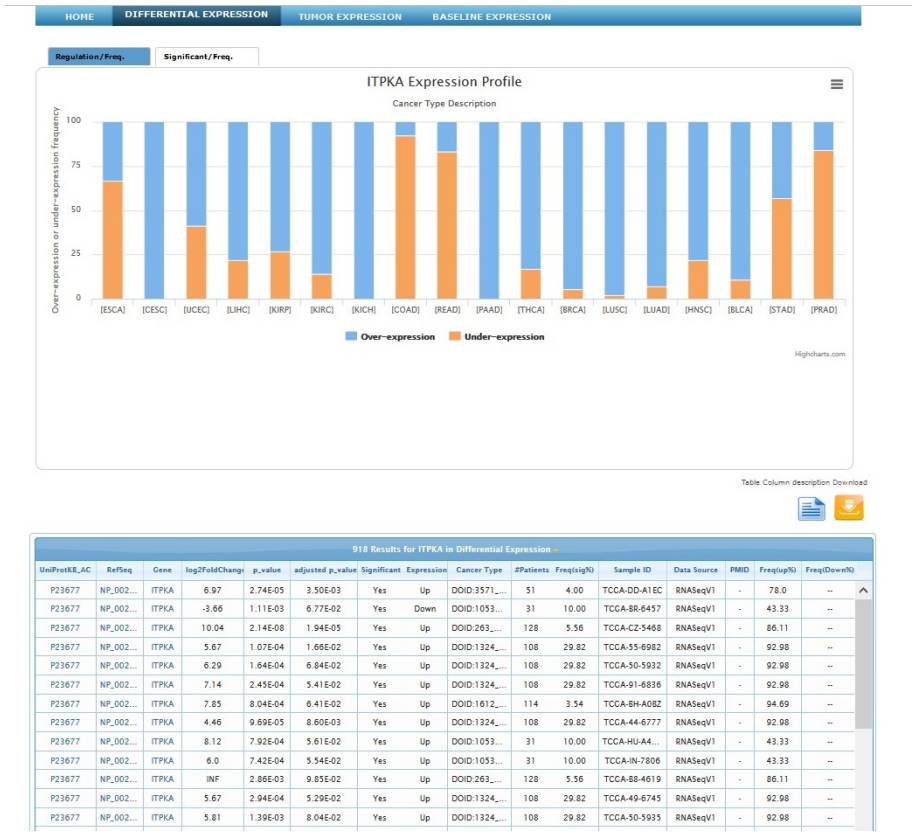


Figure 3.151: Overall analysis of ITPKA gene as a biomarker using MEXPRESS tool

Nomenclature Committee), UniProtKB/Swiss-Prot accessions or RefSeq accessions. Genes that are differentially expressed for a specific cancer type can also be retrieved. Also, all data in BioXpress, including lists of genes that are significantly differentially expressed in two or more cancer types, can be downloaded (Figure 3.152)(Figure 3.153)



UniProtKB_AC	RefSeq	Gene	log2FoldChange	p_value	adjusted p_value	Significant	Expression	Cancer Type	#Patients	Freq(sig%)	Sample ID	Data Source	PMID	Freq(up%)	Freq(Down%)
P23677	NP_002...	ITPKA	6.97	2.74E-05	3.50E-03	Yes	Up	DOID:3571_...	51	4.00	TCGA-DD-A1EC	RNASeqV1	-	78.0	--
P23677	NP_002...	ITPKA	-3.66	1.11E-03	6.77E-02	Yes	Down	DOID:1053...	31	10.00	TCGA-BR-6457	RNASeqV1	-	43.33	--
P23677	NP_002...	ITPKA	10.04	2.14E-08	1.94E-05	Yes	Up	DOID:263_...	128	5.56	TCGA-CZ-5468	RNASeqV1	-	86.11	--
P23677	NP_002...	ITPKA	5.67	1.07E-04	1.66E-02	Yes	Up	DOID:1324_...	108	29.82	TCGA-55-6982	RNASeqV1	-	92.98	--
P23677	NP_002...	ITPKA	6.29	1.64E-04	6.84E-02	Yes	Up	DOID:1324_...	108	29.82	TCGA-50-5932	RNASeqV1	-	92.98	--
P23677	NP_002...	ITPKA	7.14	2.45E-04	5.41E-02	Yes	Up	DOID:1324_...	108	29.82	TCGA-91-6836	RNASeqV1	-	92.98	--
P23677	NP_002...	ITPKA	7.85	8.04E-04	6.41E-02	Yes	Up	DOID:1612_...	114	3.54	TCGA-BH-A0BZ	RNASeqV1	-	94.69	--
P23677	NP_002...	ITPKA	4.46	9.69E-05	8.60E-03	Yes	Up	DOID:1324_...	108	29.82	TCGA-44-6777	RNASeqV1	-	92.98	--
P23677	NP_002...	ITPKA	8.12	7.92E-04	5.61E-02	Yes	Up	DOID:1053...	31	10.00	TCGA-HU-A4...	RNASeqV1	-	43.33	--
P23677	NP_002...	ITPKA	6.0	7.42E-04	5.54E-02	Yes	Up	DOID:1053...	31	10.00	TCGA-IN-7806	RNASeqV1	-	43.33	--
P23677	NP_002...	ITPKA	INF	2.86E-03	9.85E-02	Yes	Up	DOID:263_...	128	5.56	TCGA-B8-4619	RNASeqV1	-	86.11	--
P23677	NP_002...	ITPKA	5.67	2.94E-04	5.29E-02	Yes	Up	DOID:1324_...	108	29.82	TCGA-49-6745	RNASeqV1	-	92.98	--
P23677	NP_002...	ITPKA	5.81	1.39E-03	8.04E-02	Yes	Up	DOID:1324_...	108	29.82	TCGA-50-5935	RNASeqV1	-	92.98	--

Figure 3.152: Overall analysis of ITPKA gene as a biomarker using MEXPRESS tool

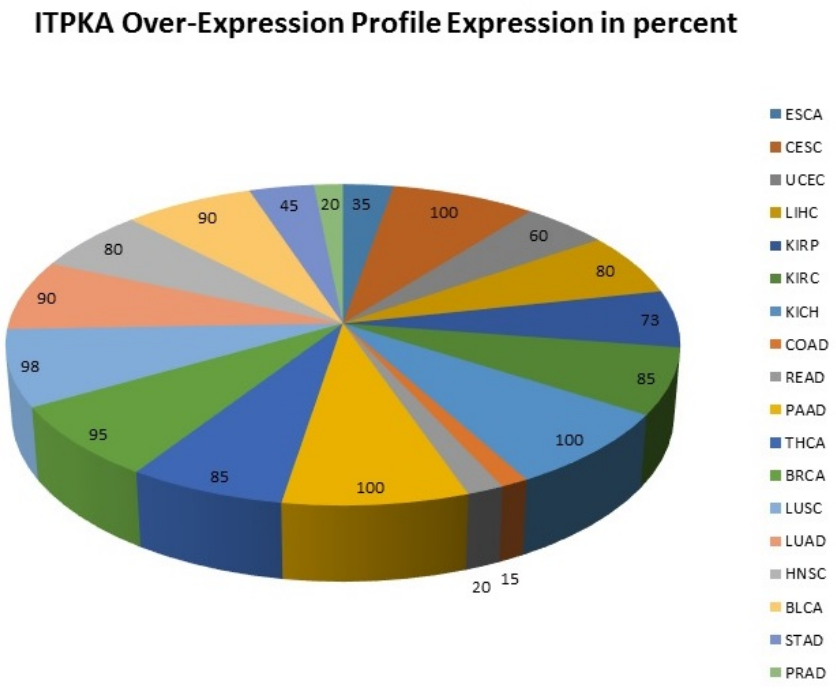


Figure 3.153: Overall analysis of ITPKA gene as a biomarker using MEXPRESS tool

Chapter 4: DISCUSSION

ITPKA (inositol-trisphosphate 3-kinase A) as a novel DNA methylation biomarker gene

ITPKA gene is known to regulate inositol phosphate metabolism by phosphorylation of second messenger inositol 1,4,5-trisphosphate to Ins(1,3,4,5)P4. The activity of the inositol 1,4,5-trisphosphate 3-kinase is responsible for regulating the levels of a large number of inositol polyphosphates that play a key role in cellular signaling. Both calcium/calmodulin and protein phosphorylation mechanisms control its activity. It is also a substrate for the cyclic AMP-dependent protein kinase, calcium/calmodulin- dependent protein kinase II, and protein kinase C in vitro

Recent research findings have pointed out to ITPKA (inositol-trisphosphate 3-kinase A), possibly being a novel DNA methylation biomarker for few cancer types. Yi-Wei Wang.,et al.,2016 in a recent studies demonstrated that Inositol-trisphosphate 3-kinase A gene (ITPKA) was identified as a potential oncogene and its distribution was found limited in certain tissue. They also showed that ITPKA is up-regulated in its gene expression in many cancers. Such an over-expressed ITPKA contributes to tumorigenesis in few cancers like lung and breast cancers. ITPKA expression was also demonstrated to be regulated by epigenetic DNA methylation. This was due to modulation of the SP1 transcription factor binding to ITPKA promoter region. Methylation levels were significantly different in normal versus cancer conditions. Low methylation levels were found in normal tissue but showed high methylation levels in malignant tumors. They finally demonstrated that, particular to lung cancer, ITPKA gene methylation appears foremost in situ carcinoma stage and increases progressively after invasion. To summarize their findings, they have demonstrated that ITPKA gene expression is upregulated in lung, breast and other cancer types. Such overexpression ITPKA is shown to promote malignant transformation in vitro and in

vivo. This is due to ITPKA expression is highly regulated by its gene body methylation. It has been shown that numerous tumor suppressor genes have been silenced by epigenetic modifications mainly through DNA methylation of gene promoter regions [152–156]. It is speculated that promoter hyper methylation mainly acts as a repressor and that this epigenetic change down-regulates gene expression. On the contrary, methylation of a gene body is more prominent as compared to promoter hyper methylation and is seen to be observed or responsible for increased gene expression [157]. However, the influence of methylation of the gene body on its expression is very poorly understood. It has been speculated that gene body methylation may possibly repress false intragenic transcription and therefore might permit or ease the process of efficient transcriptional elongation[158]. However, majority of the gene body methylation is observed to be associated with non-transcription initiation sites. Gene body methylation may possibly influence gene expression by modulating or interfering with transcription factor binding. This can directly alleviate gene expression [159]. It is has been reported from various sources that hyper methylation of the SP1-DNA can directly inhibit the binding of SP1 [160–162]. This group has also given a possible explanation saying that two SP1-DNA binding motifs are found in the ITPKA gene body CpG island 2 region and hypothesized that these two SP1-DNA binding sites in the gene body may serve as decoys to recruit and sequester SP1 from binding to the promoter. DNMT3B, (DNA (Cytosine-5-)-Methyltransferase 3 Beta) upon its action, the fully methylated body region turns refractory to SP1 binding. It thereby releases SP1 for promoter binding to drive gene expression. Substantiative or validating experiment involving bisulfate sequencing analysis demonstrated that SP1 binding motifs in the gene body (13-16 CpG sites and 81-84 CpG sites within the 99 CpG sites in the CpG island 2 region) were seen to be hyper methylated in high-ITPKA-expressing cells and hypo methylated in low ITPKA-expressing cells (Supplemental Figure 1).

This suggests that methylation levels within SP1-DNA binding site in the ITPKA gene body is highly correlated with its gene expression. SP1 binding motif 1 displayed a higher and much significant difference in methylation levels between the high and low ITPKA

expressing cell lines. This suggests that SP1 binding motif has an important role in methylation regulated expression of ITPKA. Further studies on such mechanisms might reveal clear insights into the regulation of gene expression by gene body methylation and oncogenes identification, based on similar regulation mechanisms as observed involving ITPKA. Transcription factors SP1 and RE1 which are silencing transcription factor (REST)/NRSF have been investigated and are reported to bind to ITPKA promoter. Sp1 is positively and REST/NRSF negatively regulate the gene expression of ITPKA [163].

After the demonstrated evidence showing that SP-1 mediated ITPKA expression is modulated by its gene body methylation, this group further questioned whether SP1 and REST levels may have a role in the contribution of deregulated expression of ITPKA upon malignant transformation. This can be achieved by assessing the correlation between ITPKA expression and the SP1 and REST expression using microarray analysis of lung cancer cell lines, including 113 NSCLCs, 29 small cell lung cancers, and 59 HRECs. It was found that expression level of SP1 and REST was not significant and that it played a very minor role in regulating the ITPKA gene expression. The Pearson correlation coefficients of ITPKA with SP1 and REST was 0.01 and -0.20 respectively.

In supplemental figure 2, investigation of the correlation of ITPKA methylation and its gene expression was examined. Results showed that Spearman correlation coefficients in SCC and ADC were not satisfactory (r 0.52 and 0.6). However, a trend of positive correlation was observed between gene body methylation and expression. Investigation on whether normal cells can infiltrate into tumor tissues and whether any other factors can dilute the correlation or significance. Using data from genome wide analysis of DNA methylation patterns, it has been demonstrated that the human secretin gene (SCT) promoter is frequently hyper methylated in lung cancer [152]. It is seen that SCT is expressed at undetectable levels in normal and malignant cells, irrespective of its promoter methylation status. Their study validated SCT to be a lung cancer biomarker, although functional implications or biological significance of SCT promoter methylation is far from being understood [152]. An important finding from this study is that CpG Island 2 in ITPKA gene body is observed

to be highly methylated in lung cancer and such ITPKA gene body methylation can be utilized as an early biomarker for detection of lung cancer. Also, it has been demonstrated that ITPKA gene body methylation promotes its expression, facilitating the development of malignant phenotypes. Contrastingly to SCT, ITPKA methylation is associated with gene expression and facilitates the malignant phenotype development. Also, since ITPKA is overexpressed in multiple cancer types and drives tumorigenesis, this ITPKA gene may serve as a potential therapeutic target agent.

In normal physiological conditions, ITPKA is highly expressed in neurons during brain development and also in testis [164]. During normal brain development, brain cells consistently display ITPKA gene body methylation at very high levels. Interestingly, placenta demonstrates very high levels of ITPKA body methylation. Placental tissues like cytotrophoblast and syncytiotrophoblast, and the extravillous trophoblast cells carry the ability to migrate, invade and remodel the maternal decidua and can develop a vascular supply similar to cancer progression [165].

Numerous tumor suppressor genes and oncogenes play an important role in normal placental development and the epigenetic program of the placenta exhibits similarities to those of cancer cells [166,167]. These evidences points out to the fact that placenta is a self-limited malignancy, further consolidating that ITPKA body methylation significantly higher in malignant tumors and results in tumorigenesis. To summarize the above findings, it can be said that deregulation of ITPKA plays an important role in pathogenesis of cancer. This is due to the fact that highly specific and sensitive patterns of ITPKA expression and gene body methylation is observed. ITPKA body methylation is not observed in nonmalignant or normal lung cells. This appears at premalignant stages and will progressively increase with cancer development. This clearly suggests that ITPKA can be utilized as a DNA methylation biomarker for early lung and other cancer type detection.

Our study has complemented, supplemented and also validated ITPKA methylation and expression correlation with respect to its being considered as a DNA methylation biomarker. Our study has implicated ITPKA gene in eight cancer types (BLCA, BRCA, HNSC, KIRC,

KIRP, LIHC, LUAD and LUSC). Of the eight cancer types mentioned above, ITPKA genes potential use as a DNA methylation biomarker is shown as a novel hit by us in all these eight cancer types. Detailed explanation regarding the molecular aspects or methylation/expression correlation for each cancer type is beyond the scope of this dissertation research. However, our results clearly establishes the direct or indirect relationship between the two and this is complemented by the statistical data analysis and interpretation of the same using MEXPRESS. Supplemental Table 1 shows a comparison of different tools for the visualization of TCGA data.

GDF15 (growth differentiation factor 15) as a DNA methylation biomarker gene

GDF15 gene is known to encode a secreted ligand of the TGF-beta (transforming growth factor-beta) superfamily of proteins. Also, ligands of this superfamily bind to various TGF-beta receptors. This leads to recruitment and activation of SMAD family transcription factors that can regulate gene expression. The encoded preproprotein is proteolytically processed to generate each subunit of the disulfide-linked homodimer. Such processed protein is expressed in a wide variety of cell types. This protein acts as a pleiotropic cytokine and is involved in the stress response program of cells after cellular injury. Increased protein levels are implicated in disease states such as tissue hypoxia, inflammation, acute injury and oxidative stress.

In a study by Vera L. Costa., et al., 2010, an attempt to identify a list of novel epigenetic methylation candidates for BLCA (bladder cancer) was undertaken using urine samples. Gene expression microarray was used and analyzed with BLCA cell lines upon treatment with 5-aza-2-deoxycytidine and trichostatin A as well as 26 tissue samples were also part of this study design. Candidate genes methylation level were quantified in 4 BLCA cell lines, 50 BLCA tissues and 20 normal bladder mucosas (NBM) and urine sediments from BLCA patients and 20 healthy donors, 19 renal cancer patients, and 20 prostate cancer patients. Receiver operator characteristic (ROC) curve analysis was used to assess the diagnostic performance of the gene panel. Results indicated that GDF15, HSPA2, TMEFF2, and VIM

were identified as epigenetic biomarkers for BLCA. It was observed that methylation levels of BLCA tissues were far higher than those of NBM ($P < 0.001$) and cancer specificity was found to be ($P < 0.001$) in urine samples. GDF15, TMEFF2, and VIM was able to identify BLCA tissues from a methylation panel list with 100% specificity and sensitivity. Using the urine samples, methylation panel achieved a sensitivity of 94% and a 100% specificity and an area under the curve of 0.975. Also, the compiled methylation panel could easily differentiate BLCA between normal and renal or prostate cancer patients (sensitivity, 94%; specificity, 90%). Therefore, Vera L. Costa., et al. 2010, showed that by using a genome-wide approach, they were able to identify a novel epigenetic biomarker panel that can be utilized for early and accurate detection of BLCA in urine samples with an additional advantage of it being noninvasive.

Results pertaining to the methylation status of novel candidate genes in vitro and in vivo showed the following: 21 of the DNA methylation candidate genes were analyzed by MSP in BLCA cell lines. Among the candidate list the top 4 biomarkers which exhibited hyper methylation in a minimum of 3 cell lines were selected for further validation. These were the GDF15, HSPA2, TMEFF2, and VIM (Supplementary Table 2). Three of these biomarkers were methylated in BLCA cells as compared to kidney and prostate cancer cell lines, except for TMEFF2. GDF15 was found to be methylated at 64% in bladder tumors. Also, quantitative analysis in methylation levels were significantly different in normal vs cancer patients for all the above mentioned genes (MannWhitney, $P < 0.001$).

Our study has complemented, supplemented and also validated GDF15 methylation and expression correlation with respect to its being considered as a DNA methylation biomarker. Our study has implicated GDF15 gene in thirteen cancer types (BRCA, COAD, CRAD, CESC, ESCA, HNSC, KIRP, LIHC, LUAD, LUSC, PRAD, THCA and UCEC). Of the thirteen cancer types mentioned above, GDF15 genes potential use as a DNA methylation biomarker is shown as a novel hit by us in nine cancer types (CRAD, CESC, ESCA, KIRP, LIHC, LUAD, LUSC, THCA and UCEC). Detailed explanation regarding the molecular aspects or methylation/expression correlation for each cancer type is beyond the scope of

this dissertation research. However, our results clearly establishes the direct or inverse relationship between the two and this is complemented by the statistical data analysis and interpretation of the same using MEXPRESS.

BLCAP (bladder cancer associated protein) as a DNA methylation biomarker gene

BLCAP gene is known to encode a protein that reduces cell growth by stimulating apoptosis. Multiple transcript variants encoding the same protein are identified which may be the result of mechanisms like alternative splicing and the use of alternative promoters. This gene is imprinted in brain. It is known that different transcript variants are expressed from each parental allele. Also, transcript variants initiating from the upstream promoter are expressed preferentially from the maternal allele, while transcript variants initiating downstream of the interspersed NNAT gene are expressed from the paternal allele. Transcripts at this locus is known to undergo A to I editing, resulting in amino acid changes at three positions in the N-terminus of the protein.

Jos M. A. Moreira., et al., 2009 generated and characterized antibodies that are able to specifically recognize BLCAP. Also, they demonstrated that BLCAP localizes predominantly to the epithelial lining of the urinary bladder. BLCAP IHC staining pattern types B and D are observed to be associated with benign/low grade and high grade invasive lesions respectively. They can be utilized as a diagnostic indicators irrespective of the fact that type A and C staining patterns are not good classifiers. This is because they appear ubiquitously in all grade and stages of cancer. Staining type A was prominently associated with poor disease-specific survival. 2D Western blot analysis of samples classified by IHC as type A or B (Supplemental Figure 3, e and f respectively), showed increased immunoreactivity for BLCAp antigen observed by IHC. This corresponds to elevated protein and both polypeptides expression levels (unmodified and modified forms; Fig. 3, e and f, black and white arrows, respectively) is increased. These demonstrated data indicates the loss of BLCAP expression is directly associated tumor progression. However, an increased percentage of cells with high nuclear levels of BLCAP confers to poor prognosis. BLCAP is

seen to be overexpressed in ∼20% of the cases examined and it is liked with poor survival. This indicates that BLCAP expression does not carry good prognostic value. In cases of invasive tumors, BLCAP expression offers an adverse patient outcome, especially with those bearing tumors that have lost expression of BLCAP are better performers as compared to those with tumors expressing BLCAP at any expression levels. (Supplemental Figure 7c, pT2-4 tumors).

Loss of BLCAP expression was observed in both epithelial and vascular endothelial cells indicates that this mechanism brings about a change in cellular microenvironment as compared to being a process related to epithelial carcinogenesis. Multiple cancer types such as cervical, renal, human tongue carcinoma and osteosarcoma exhibit differential expression of BLCAP suggesting that micro environmental changes corresponding to differential expression of BLCAP, triggers this cellular response and is of general nature rather than being tissue-specific [168–171, 171, 172].

In another study, it was examined the expression pattern of BLCA- 1 in tissues and urine samples from bladder cancer patients and also from normal controls. This was done by utilizing BLCA-1 sequence data to produce antibodies to this protein, which was further used in immunoblot and ELISA. Their results indicated that BLCA-1 was detectable in tissues from patients with bladder cancer but not detectable in normal adjacent areas of bladder or in normal donor bladder tissue. This protein was also found in urine of patients with bladder cancer using immunoblot and immunoassay. The cutoff optical density units (absorbance value) was assigned as 0.025, BLCA-1 was detected in 20 of 25 urine samples from patients with bladder cancer but in just 6 of 46 normal, high risk, prostate or renal cancer samples tested. This results in a test with 80% sensitivity and 87% specificity. BLCA-1 expression did not correlate with the tumor grade. This suggested that BLCA-1 is a urine based marker of bladder cancer and could be utilized as an early stage detection for this disease.

Our study has complemented, supplemented and also validated BLCAP methylation and expression correlation with respect to its being considered as a DNA methylation biomarker.

Of particular importance is our finding that BLCAP gene expression is influenced by DNA methylation and is detectable in seven cancer types (BRCA, COAD, CRAD, KIRC, KIRP, LUAD AND LUSC). Thus, BLCAP gene can be utilized as an early stage DNA methylation biomarker for these cancer types. Detailed explanation regarding the molecular aspects or methylation/expression correlation for each cancer type is beyond the scope of this dissertation research. However, our results clearly establishes the direct or inverse relationship between the two and this is complemented by the statistical data analysis and interpretation of the same using MEXPRESS.

PIWIL4 (piwi like RNA-mediated gene silencing 4) as a DNA methylation biomarker gene

PIWIL4 gene is known to play a central role during spermatogenesis. It achieves this by repressing transposable elements and preventing their mobilization, which is essential for the germline integrity. It acts via the piRNA metabolic process, which mediates the repression of transposable elements during meiosis by forming complexes composed of piRNAs and Piwi proteins and governs the methylation and subsequent repression of transposons. It also binds to piRNAs directly (class of 24 to 30 nucleotide RNAs that are generated by a Dicer-independent mechanism and are primarily derived from transposons) and other repeated sequence elements. It is also known to associate with secondary piRNAs antisense and PIWIL2/MILI is required for such association. The piRNA process acts upstream of known mediators of DNA methylation. It participates in a piRNA amplification loop. In addition to their role in transposable elements repression, piRNAs are probably involved in other processes during meiosis such as translation regulation. They may be involved in the chromatin-modifying pathway by inducing Lys-9 methylation of histone H3 at some loci.

Preethi Krishnan., et al., 2016, in their study were able to identify 8 non-redundant piRNAs as a novel prognostic biomarkers for breast cancer. They also identified PIWI genes as potential prognostic markers for breast cancer. PIWI genes are of 4 homologues and PIWIL3 and PIWIL4 are observed to be associated with OS, and PIWIL3 alone is seen to be associated with RFS (Supplemental figure 5). Not much information is available with

regards to the clinical significance of PIWIL3 and PIWIL4. This study was in fact the first, to report these genes to breast cancer prognosis. Further studies are required to validate their prognostic role. This study group used a cohort with complete clinical annotation and follow-up for long term, thereby validating piRNAs and PIWI genes to be a novel prognostic markers for breast cancer.

In another study, investigation of the expression of PIWI genes was conducted in order to determine the activity and potential prognostic role of the PIWI/piRNA pathway in NSCLC. It was reported that PIWIL1 participates in the primary pathway and PIWIL2 and PIWIL4 in the secondary pathway, both of which are active in NSCLC. The re-expression of the PIWIL1 gene, which can be confirmed by immunohistochemistry, is related to poor prognosis and is associated with a stem-cell signature. Furthermore, the downregulation of PIWIL4 is also related to poor prognosis and is associated with lower methylation. Further investigation in a larger cohort of patients is warranted to validate these findings and to examine potential diagnostic and therapeutic approaches.

Our study has complemented, supplemented and also validated PIWIL4 methylation and expression correlation with respect to its being considered as a DNA methylation biomarker. PIWIL4 gene is observed to be involved in at-least twelve types of cancer (BLCA, BRCA, CESC, CHOL, COAD, CRAD, ESCA, HNSC, KIRC, KIRP, LIHC and UCEC), wherein the methylation probes correspond to negative values which are highly significant. Our study has shown that PIWIL4 gene can be utilized as an early stage DNA methylation biomarker for eleven of the above mentioned cancer types with HNSC being the exception. Any efforts in attempting a detailed explanation regarding the molecular aspects or methylation/expression correlation for each cancer type is beyond the scope of this dissertation research. However, our results clearly establishes the direct or indirect relationship between the two and this is complemented by the statistical data analysis and interpretation of the same using MEXPRESS.

DMRT1 (Doublesex and Mab-3 Related Transcription Factor 1) as a DNA methylation biomarker gene

Transcription factors play a very critical role in the early development process. It is known that transcription factor plays a key role in male sex determination and differentiation by controlling testis development and male germ cell proliferation. It also plays a central role in spermatogonia by inhibiting meiosis in undifferentiated spermatogonia and promoting mitosis, leading to spermatogonial development and allowing abundant and continuous production of sperm. It acts both as a transcription repressor and activator: prevents meiosis by restricting retinoic acid (RA)-dependent transcription and repressing STRA8 expression and promotes spermatogonial development by activating spermatogonial differentiation genes, such as SOHLH1. Also plays a key role in postnatal sex maintenance by maintaining testis determination and preventing feminization: represses transcription of female promoting genes such as FOXL2 and activates male-specific genes. They may act as a tumor suppressor and also play a minor role in oogenesis

Spermatogonial stem cells (SSCs) are capable of acquiring pluripotency under specific culture conditions. The frequency of pluripotent cell derivation, is however, very low. Also, the mechanism of SSC reprogramming remains unknown. Seiji Takashima, et al., 2013, reported the induction of global DNA hypo methylation in germline stem cells (GS) (cultured SSCs) induces pluripotent cell derivation. GS cells seems to undergo apoptosis, when DNA demethylation was triggered by Dnmt1 depletion. However, GS cells converted to embryonic stem (ES)-like cells accompanying the double knockdown of Dnmt1 and p53. DMRT1 is downregulated by this treatment. DMRT1 is a gene involved in sexual differentiation, meiosis and pluripotency. DMRT1 depletion results in apoptosis of GS cells, however, a combination of DMRT1 and p53 depletion can also induce pluripotency. Putative DMRT1 target genes upon undergoing functional screening and undergoing depletion will upregulate SoX2. SoX2 transfection up-regulates Oct4 and can produce pluripotent cells. This conversion is enhanced by Oct1 depletion which suggests balance of Oct proteins maintains SSC identity. These results suggest that SSC reprogramming on a spontaneous basis is caused by unstable DNA methylation and that a DMRT1-SoX2 cascade is very important for regulating pluripotency in SSCs.

Our study has complemented, supplemented and also validated DMRT1 methylation and expression correlation with respect to its being considered as a DNA methylation biomarker. DMRT1 gene is observed to be involved in at-least four types of cancer (BRCA, LUSC, THCA and UCEC), wherein the methylation probes correspond to negative values which are highly significant. Our study has shown that DMRT1 gene can be utilized as an early stage DNA methylation biomarker for all the four cancer types (BRCA, LUSC, THCA and UCEC). Detailed explanation regarding the molecular aspects or methylation/expression correlation for each cancer type is beyond the scope of this dissertation research. However, our results clearly establishes the direct or indirect relationship between the two and this is complemented by the statistical data analysis and interpretation of the same using MEXPRESS [173].

Query of ITPKA gene Vs BioMuta and BioXpress - A representative result

We queried the ITPKA gene against the BioMuta database to identify and evaluate variations (both synonymous and non-synonymous) for any possible functional impact on protein structure and functions. 66 SNVs are found when ITPKA gene is queried against BioMuta. Of these 66 SNVs, five are nsSNVs that affect functional sites (three gain of phosphorylation and two gain of glycosylation). The five nsSNVs identified were mapped to functional sites that are obtained from UniProtKB sequence feature annotation. Precise nucleotide positions at which the post-translational modifications (PTMs) and active and binding sites are affected by nsSNVs were identified. In order to investigate whether certain types of PTM or other functional sites are resistant to variations, P-values were calculated in BioMuta, to estimate the significance between observed and expected numbers. Of the 66 SNVs, the majority of functional sites analyzed are protected from mutation (significantly less observed variations than expected). Further studies need to be conducted as to why, in certain cancer types, some of the functional sites appear to be less protected. The identified variations were integrated into SNVDis. Effects of identified variations can be analyzed as they are coupled with PolyPhen-based predictions and are included in the BioMuta table. SNVDis is a useful application as it evaluates the distribution of nsSNVs on protein

functional sites, domains and pathways at the entire proteome level. Such proteome-wide analysis complements the functional impact analysis using methods such as PolyPhen and SIFT, and similar algorithms. It should be noted that BioMuta is supported on the High-performance Integrated Virtual Environment (HIVE). HIVE is a bio-computing operating system serving as an ideal backbone to integrate modular software into a data analytics backbone. In short, HIVE (High-performance Integrated Virtual Environment) is a bio-computing environment for storing, analyzing, computing and curating huge genomic data and associated metadata. Once again, such identified nsSNVs (from our computational approach) and their relevant effects at the proteomic and cellular level need to be validated by subsequent in-vitro studies.

We also queried ITPKA against the BioXpress database. When ITPKA gene is queried using the HGNC-approved gene symbol or UniProt/RefSeq accession, BioXpress retrieves three types of information: differential expression information (cancer vs. normal), tumor-only expression data (where normal samples are not available) and baseline expression information from normal human tissues (Illumina Human Body Map Project). Our research primarily focuses on differential expression in normal vs cancer types. For the ITPKA gene in THCA, over-expression is clearly observed from BioXpress results, there by validating and reproducing the MEXPRESS study. In the default view, BioXpress provides expression frequency (over- or under-expression) in the patients. Additionally, the number of patients for a particular cancer type, P value and a variety of other information is available in the table below which can be downloaded. Complete cancer names can be retrieved on clicking the cancer abbreviations in the figure and additional details can be retrieved by clicking the Table column description link. Data collected in BioXpress can be used to sort, filter and further analyze the gene expression and to compare and contrast expression of genes across many patients and cancer types. Such a computational approach wherein querying datasets proves to be very useful in identifying expression levels between disease and normal pairs leads to analysis of differential expression for a gene. Also, potential leads on biomarkers or pathways involved in tumor formation can be identified and overall expression of specific

genes across multiple cancer types can be conveniently studied.

Chapter 5: CONCLUSION

The field of Oncogenomics (sub-field of genomics which characterizes cancer associated genes) has three broad applications. To improve diagnosis (use molecular markers of gene mutations for early cancer detection), prognosis (use markers of gene mutations to classify cancers and predict their outcomes) and therapeutics (use gene mutations found in cancer as targets of drug therapy). Oncogenomics is growing at an exponential rate with the help of databases and datasets being created and available publically. Their value and significance will continue to gain prominence. Such a rapid progress and implementation creates a demand for developing intuitive and straightforward tools that enable researchers to quickly analyze and visualize the data of interest. The Cancer Genome Atlas (TCGA) is one such invaluable database. We have selected and used the MEXPRESS tool from a list of methylation analysis tools based on its ease of use and the integrated visualization of different data types over hundreds of samples from TCGA. Not only does this tool help identify novel DNA methylation biomarker gene but also help in testing hypotheses that concern the discovery of DNA methylation or expression-based biomarkers. We have undertaken a set of five genes of interest based on literature search. BLCAP gene is observed to be involved in at-least eight types of cancer (BLCA, BRCA, COAD, CRAD, KIRC, KIRP, LUAD AND LUSC), wherein the methylation probes correspond to negative values which are highly significant. In all these eight genes a strong negative correlation (between methylation and expression) exists, indicating that the corresponding gene expression might be controlled through DNA methylation. Of the eight cancer types mentioned above, six of these cancer types (BRCA, COAD, CRAD, KIRP, LUAD AND LUSC) are novel hits and have previously not found in literature or any research findings. GDF15 gene is observed to be involved in at-least thirteen cancer types wherein the methylation probes correspond

to negative values which are highly significant. These are (BRCA, COAD, CRAD, CESC, ESCA, HNSC, KIRP, LIHC, LUAD, LUSC, PRAD, THCA and UCEC). We have shown that, GDF15 gene can be potentially use as a DNA methylation biomarker in nine cancer types as they showed either direct or inverse relation between DNA methylation and gene expression. These nine cancer types are CRAD, CESC, ESCA, KIRP, LIHC, LUAD, LUSC, THCA and UCEC. PIWIL4 gene is observed to be involved in at-least twelve types of cancer (BLCA, BRCA, CESC, CHOL, COAD, CRAD, ESCA, HNSC, KIRC, KIRP, LIHC and UCEC), wherein the methylation probes correspond to negative values which are highly significant. Our study has shown that PIWIL4 gene can be utilized as an early stage DNA methylation biomarker for eleven of the above mentioned cancer types with HNSC being the exception. Our study has implicated ITPKA gene in eight cancer types (BLCA, BRCA, HNSC, KIRC, KIRP, LIHC, LUAD and LUSC). Of the eight cancer types mentioned above, ITPKA genes potential use as a DNA methylation biomarker is shown as a novel hit by us in all these eight cancer types. Our study has implicated DMRT1 gene to be involved in at-least four types of cancer (BRCA, LUSC, THCA and UCEC), wherein the methylation probes correspond to negative values which are highly significant. We have also shown that DMRT1 gene can be utilized as an early stage DNA methylation biomarker for all the four cancer types (BRCA, LUSC, THCA and UCEC). Although our research has multiple novel hits in terms of identifying novel DNA methylation biomarker genes, a greater challenge still remains with regards to its clinical implementation and development. Never the less our research is a first step in identification of novel DNA methylation biomarker genes using a methylation tool that is easy to use with an added advantage of TCGA data visualization involving clinical, gene expression and methylation data simultaneously for comparison.

	UCSC genome browser	cBioPortal	CGW	IGV	MEXPRESS
All TCGA cancer and data types available	yes	yes	no	no	no
Integration of expression, DNA methylation and clinical data	no	no	no	no	yes
Statistical interpretation of the relationships	no	yes	no	no	yes
Registration and download required	no	no	no	yes	no

Figure 5.1: A comparison of different tools for the visualization of TCGA data

LIST OF SUPPLEMENTAL TABLES

Supplemental Table 1

A comparison of different tools for the visualization of TCGA data. As illustrated by the Additional file 1: Figures S1, S2, S3 and S4, there are obvious differences between existing tools and MEXPRESS in both the data and the features these tools offer. This table lists the most relevant of these differences, thereby highlighting some of the strengths and weaknesses of each tool. (CGW Cancer Genomics Workbench, IGV Integrative Genomics Viewer) (Figure 5.1)

CGW Cancer Genomics Workbench, IGV Integrative Genomics Viewer

Supplemental Table 2(Figure 5.2)

Supplementary Table S5 – Gene promoter methylation status in bladder (BlCa), renal (RCT) and prostate (PCa) cancer cell lines analyzed by methylation-specific PCR (MSP).

	GDF15	*HSPA2*	*TMEFF2*	*VIM*
BlCa cell lines				
5637	M	M	U/M	U
J82	M	M	U/**M**	U/**M**
SCaBER	U/**M**	M	M	M
TCCSUP	**U**/M	U/**M**	U/M	M
RCT cell lines				
786-O	U	U	U/**M**	U
ACHN	M	U	U/M	U
Caki-1	U	U	U/M	U
Caki-2	U	U	U/M	U
PCa cell lines				
22Rv1	U	U	U/M	**U**/M
DU145	M	U	M	U
LNCaP	**U**/M	M	U/**M**	U
PC-3	U	**U**/M	U/**M**	U

U, unmethylated; **M**, methylated; **U/M**, partial methylated

Figure 5.2: Gene promoter methylation status analyzed using PCR

Supplemental Table 3, 4 and 5

Supplemental Table 3,4 and 5(Figure 5.3) (Figure 5.4) (Figure 5.5) COHCAP quality control metrics. (A) Dendrogram: the sample ID for each sample is shown in the dendrogram representing the hierarchical clustering of the genome-wide beta values for each sample. Sample IDs are colored based on the sample grouping (in this case, the parental HCT116 strain is shown in blue and the mutant strain is shown in red). Notice that the samples in each group cluster together. (B) Sample histogram: density distribution for all the samples in a COHCAP project is shown in the histogram. Again, the color for

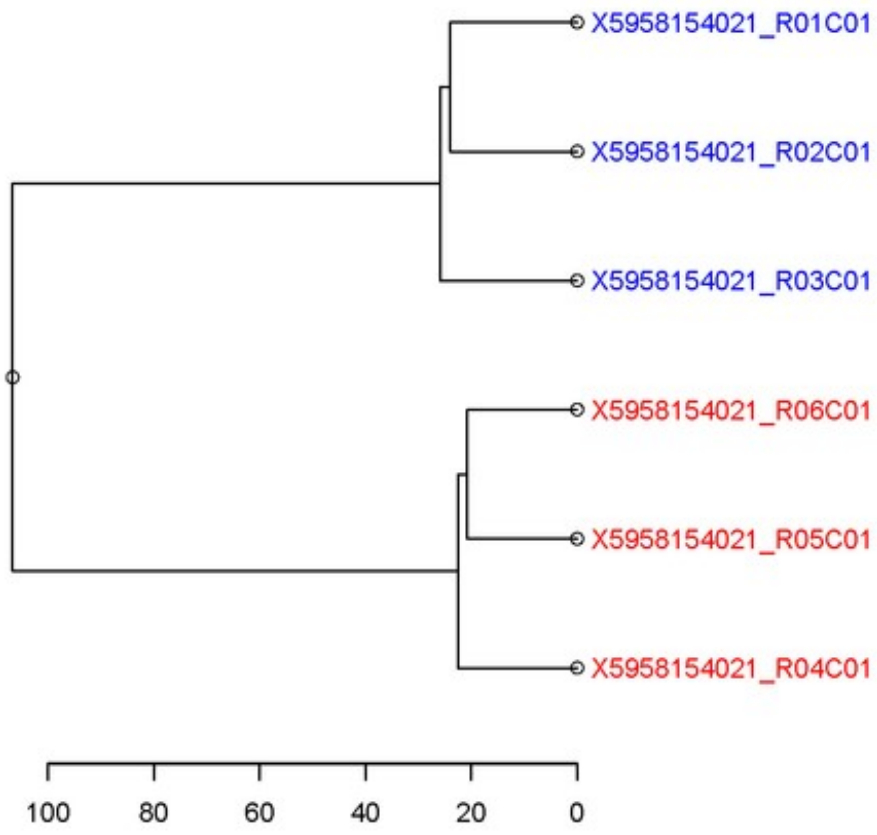


Figure 5.3: COHCAP quality control metrics: Dendrogram

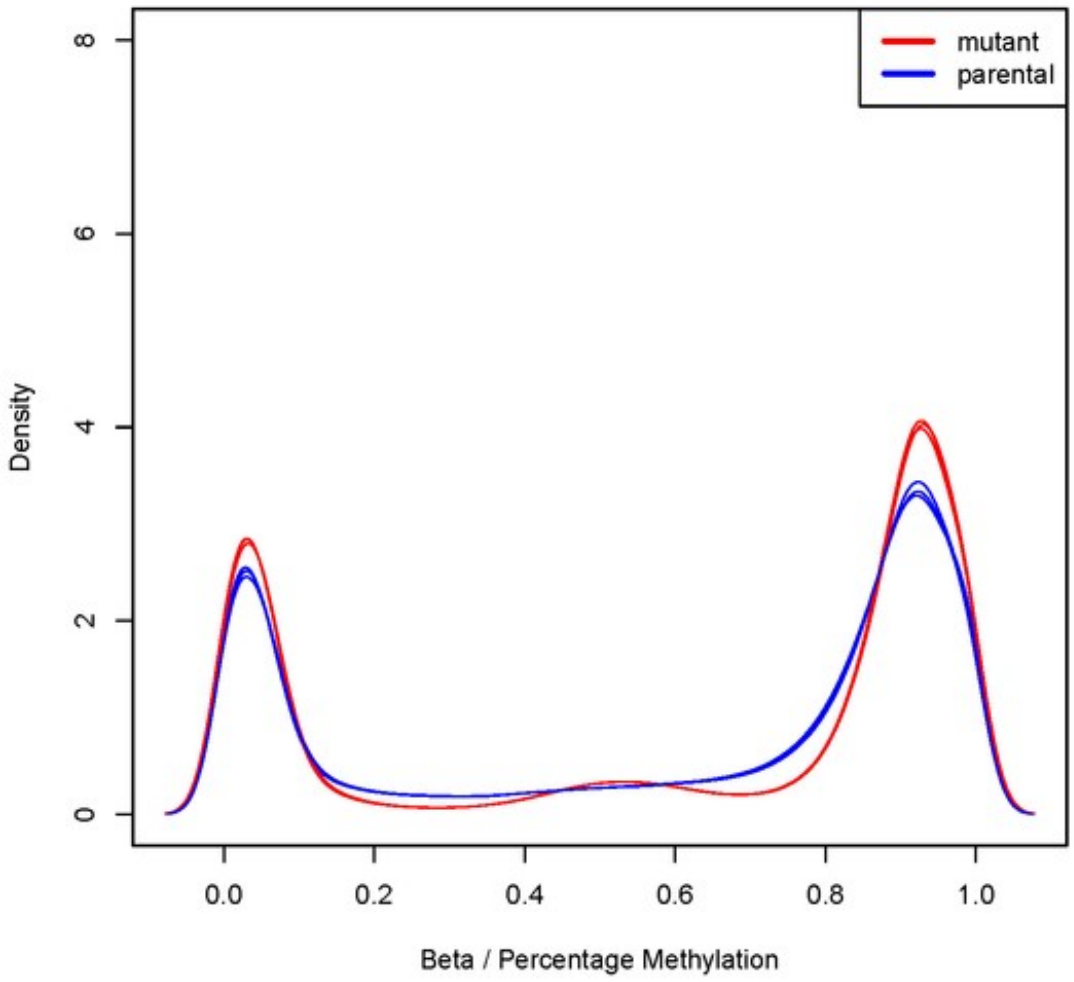


Figure 5.4: COHCAP quality control metrics: Histogram

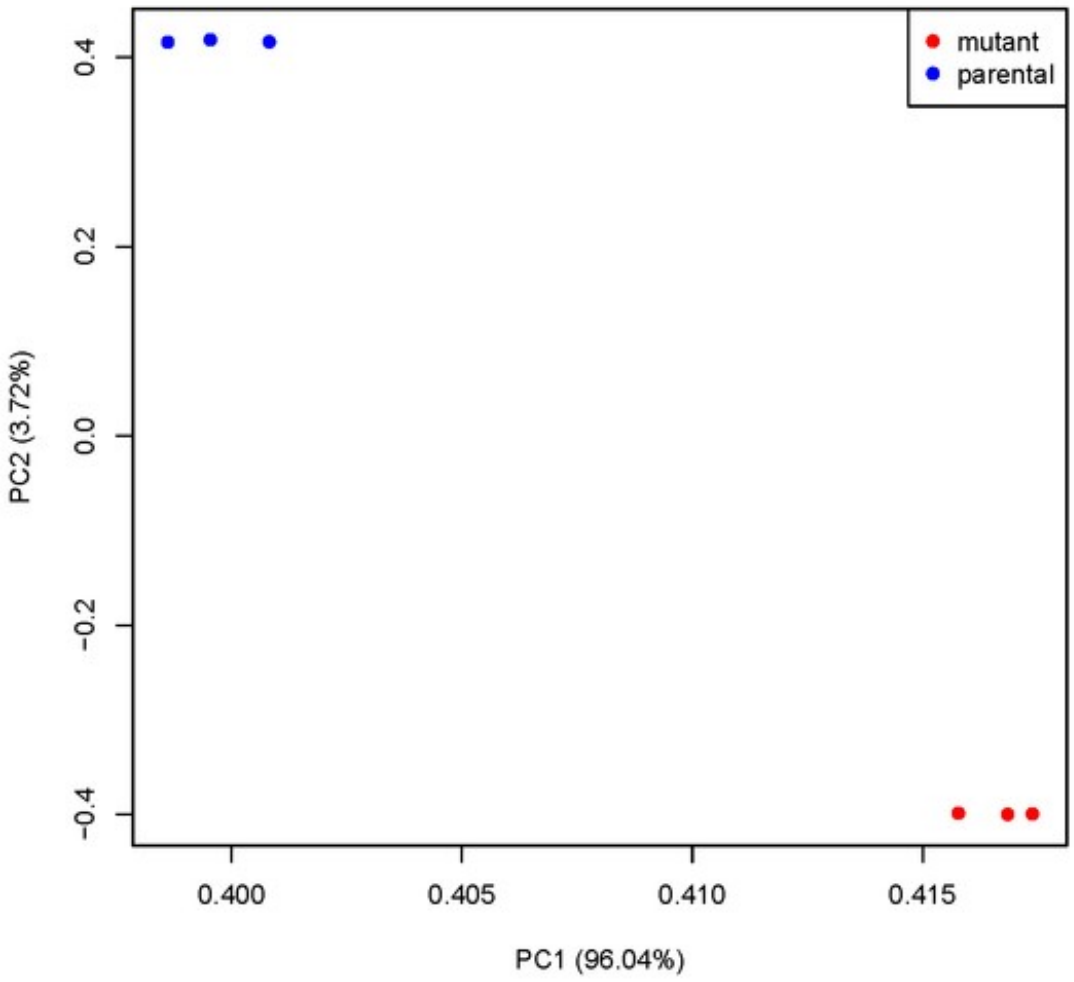


Figure 5.5: COHCAP quality control metrics: PCA plot

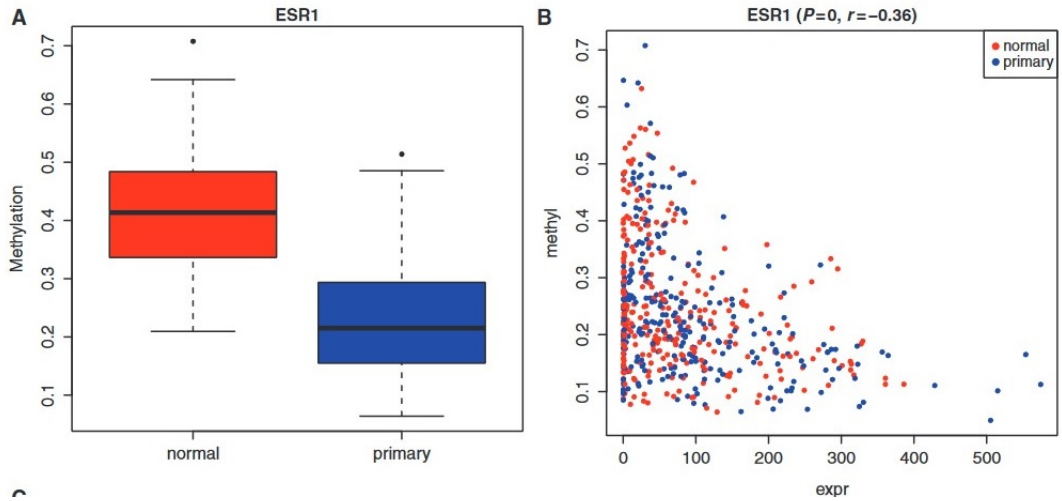


Figure 5.6: Box plot

each sample is determined by the sample grouping. Notice the strong bimodal distribution, corresponding to methylated and unmethylated CpG sites. Sample statistics (median, top quartile, bottom quartile, minimum and maximum) are provided in a text file. (C) Principal component analysis (PCA) plot: samples are plotted based on their coordinates defined by the first two principal components. All the principal component values can be found in a text file. Samples are colored based on sample grouping. Notice that the groups show clear clustering from one another in the PCA plot.

Supplemental Table 6 Box plot and Scatter plot

Box plot: the average beta value for a normal sample is higher than the primary sample, indicating that this CpG island (mapped to ESR1) shows decreased methylation in breast tumors. The box plot shows the median, minimum, maximum and quartiles for beta values for each group. This figure was produced using the Average by Island workflow. (Figure 5.6)

Scatter plot: methylation levels of ESR1 are negatively correlated with RNA-Seq expression levels. Individual samples are colored based on their sample grouping. This

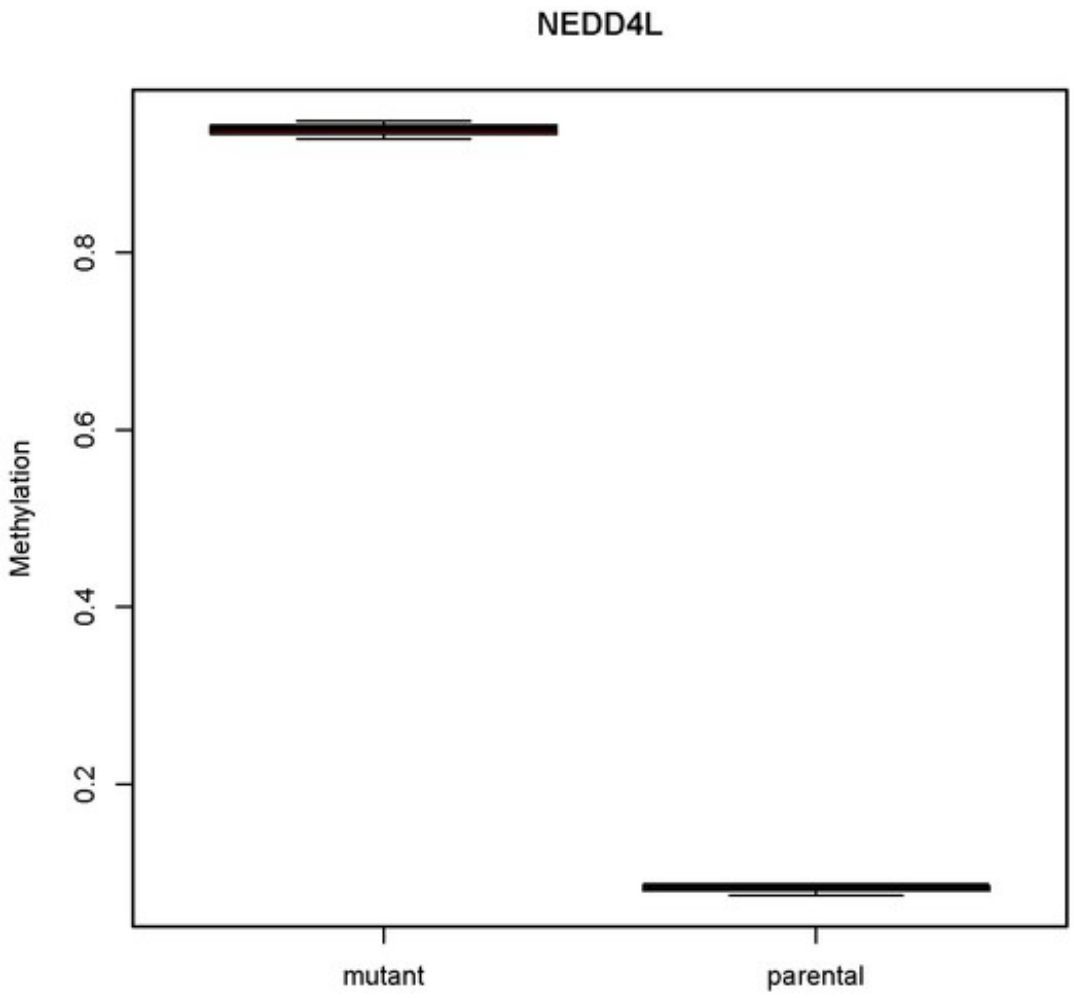


Figure 5.7: Scatter plot

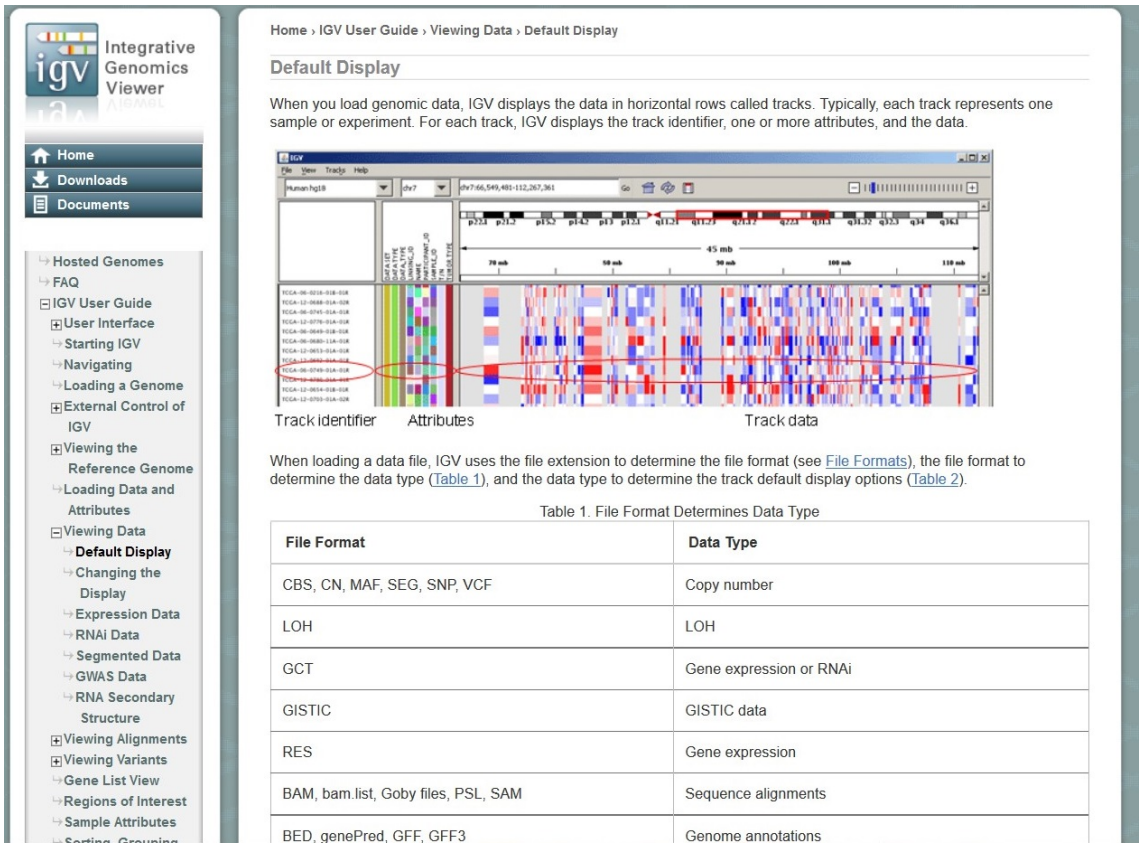


Figure 5.8: Integrative Genomics Viewer: Home page

figure was produced using the Average by Island workflow.(Figure 5.7)

Supplemental Table 7

Integrative Genomics Viewer

When you load genomic data, IGV displays the data in horizontal rows called tracks. Typically, each track represents one sample or experiment. For each track, IGV displays the track identifier, one or more attributes, and the data. When loading a data file, IGV uses the file extension to determine the file format (see File Formats), the file format to determine the data type (Table 1), and the data type to determine the track default display options (Figure 5.8) (Figure 5.9) .

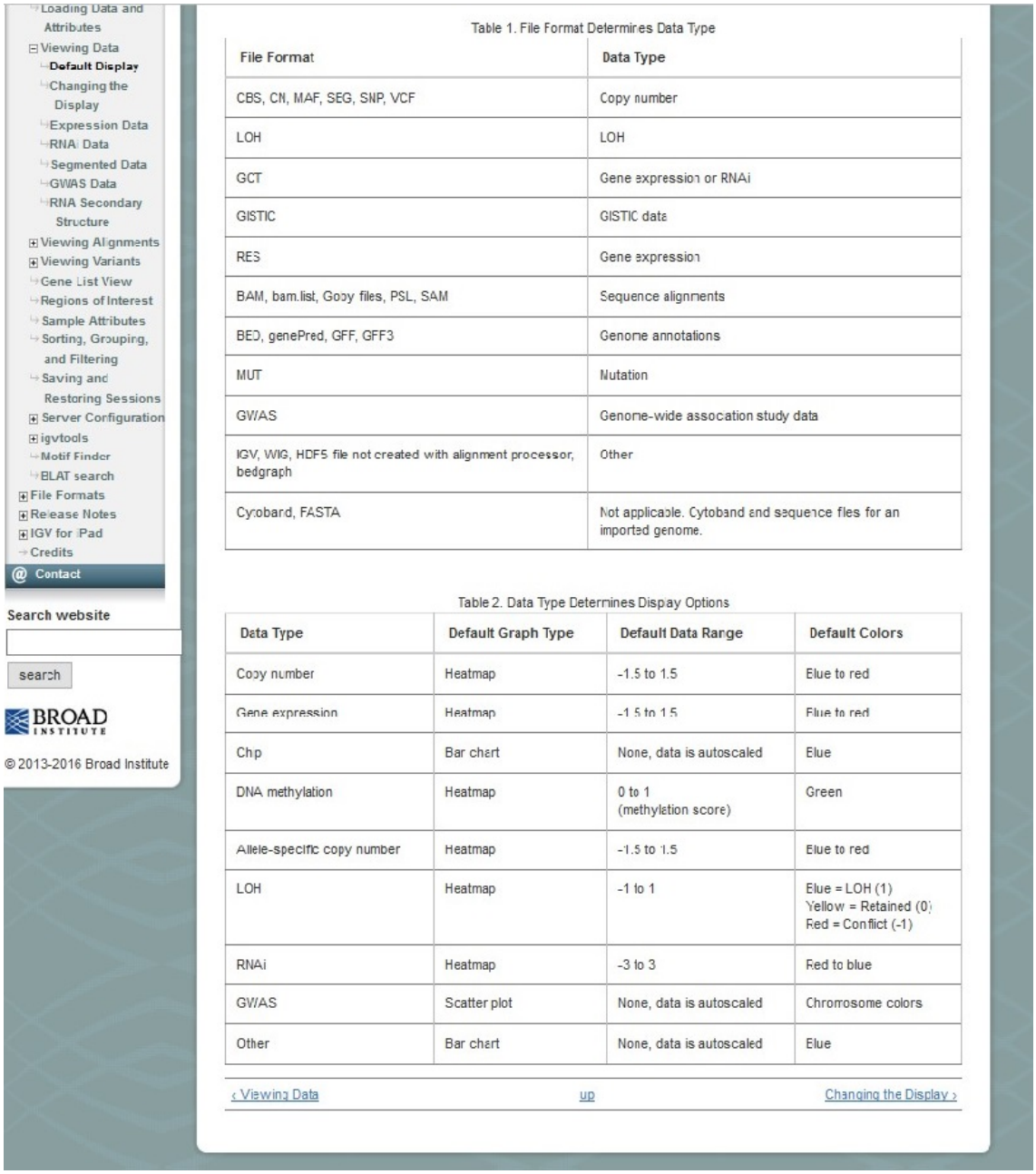


Table 1. File Format Determines Data Type

File Format	Data Type
CBS, CN, MAF, SEG, SNP, VCF	Copy number
LOH	LOH
GCT	Gene expression or RNAi
GISTIC	GISTIC data
RES	Gene expression
BAM, bam.list, Copy files, PSL, SAM	Sequence alignments
BED, genePred, GFF, GFF3	Genome annotations
MUT	Mutation
GWAS	Genome-wide association study data
IGV, WIG, HDF5 file not created with alignment processor, bedgraph	Other
Cytoband, FASTA	Not applicable. Cytoband and sequence files for an imported genome.

Table 2. Data Type Determines Display Options

Data Type	Default Graph Type	Default Data Range	Default Colors
Copy number	Heatmap	-1.5 to 1.5	Blue to red
Gene expression	Heatmap	-1.5 to 1.5	Blue to red
Chip	Bar chart	None, data is autoscaled	Blue
DNA methylation	Heatmap	0 to 1 (methylation score)	Green
Allele-specific copy number	Heatmap	-1.5 to 1.5	Blue to red
LOH	Heatmap	-1 to 1	Blue = LOH (1) Yellow = Retained (0) Red = Conflict (-1)
RNAi	Heatmap	-3 to 3	Red to blue
GWAS	Scatter plot	None, data is autoscaled	Chromosome colors
Other	Bar chart	None, data is autoscaled	Blue

Figure 5.9: Integrative Genomics Viewer: File format determination data type

Bibliography

Bibliography

[1] D. E. Handy, R. Castro, and J. Loscalzo, "Epigenetic modifications basic mechanisms and role in cardiovascular disease," *Circulation*, vol. 123, no. 19, pp. 2145–2156, 2011.

[2] E. I. Campos and D. Reinberg, "Histones: annotating chromatin," *Annual review of genetics*, vol. 43, pp. 559–599, 2009.

[3] A. P. Bird, "Cpg-rich islands and the function of dna methylation." *Nature*, vol. 321, no. 6067, pp. 209–213, 1985.

[4] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo *et al.*, "Human dna methylomes at base resolution show widespread epigenomic differences," *nature*, vol. 462, no. 7271, pp. 315–322, 2009.

[5] W. Reik, "Stability and flexibility of epigenetic gene regulation in mammalian development," *Nature*, vol. 447, no. 7143, pp. 425–432, 2007.

[6] R. S. Illingworth and A. P. Bird, "Cpg islands–a rough guide," *FEBS letters*, vol. 583, no. 11, pp. 1713–1720, 2009.

[7] L. Shen, Y. Kondo, Y. Guo, J. Zhang, L. Zhang, S. Ahmed, J. Shu, X. Chen, R. A. Waterland, and J.-P. J. Issa, "Genome-wide profiling of dna methylation reveals a class of normally methylated cpg island promoters," *PLoS Genet*, vol. 3, no. 10, p. e181, 2007.

[8] M. Weber, I. Hellmann, M. B. Stadler, L. Ramos, S. Pääbo, M. Rebhan, and D. Schübeler, "Distribution, silencing potential and evolutionary impact of promoter dna methylation in the human genome," *Nature genetics*, vol. 39, no. 4, pp. 457–466, 2007.

[9] J. Nomura, A. Hisatsune, T. Miyata, and Y. Isohama, "The role of cpg methylation in cell type-specific expression of the aquaporin-5 gene," *Biochemical and biophysical research communications*, vol. 353, no. 4, pp. 1017–1022, 2007.

[10] T. Aoyama, T. Okamoto, S. Nagayama, K. Nishijo, T. Ishibe, K. Yasura, T. Nakayama, T. Nakamura, and J. Toguchida, "Methylation in the core-promoter region of the chondromodulin-i gene determines the cell-specific expression by regulating the binding of transcriptional activator sp3," *Journal of Biological Chemistry*, vol. 279, no. 27, pp. 28 789–28 797, 2004.

[11] J. Rössler, I. Stolze, S. Frede, P. Freitag, L. Schweigerer, W. Havers, and J. Fandrey, "Hypoxia-induced erythropoietin expression in human neuroblastoma requires a methylation free hif-1 binding site," *Journal of cellular biochemistry*, vol. 93, no. 1, pp. 153–161, 2004.

[12] B. Hendrich and A. Bird, "Identification and characterization of a family of mammalian methyl-cpg binding proteins," *Molecular and cellular biology*, vol. 18, no. 11, pp. 6538–6547, 1998.

[13] X. Nan, H.-H. Ng, C. A. Johnson, C. D. Laherty, B. M. Turner, R. N. Eisenman, and A. Bird, "Transcriptional repression by the methyl-cpg-binding protein mecp2 involves a histone deacetylase complex," *Nature*, vol. 393, no. 6683, pp. 386–389, 1998.

[14] P. A. Wade, A. Gegonne, P. L. Jones, E. Ballestar, F. Aubry, and A. P. Wolffe, "Mi-2 complex couples dna methylation to chromatin remodelling and histone deacetylation," *Nature genetics*, vol. 23, no. 1, pp. 62–66, 1999.

[15] M. Okano, D. W. Bell, D. A. Haber, and E. Li, "Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development," *Cell*, vol. 99, no. 3, pp. 247–257, 1999.

[16] T. Chen, Y. Ueda, J. E. Dodge, Z. Wang, and E. Li, "Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by dnmt3a and dnmt3b," *Molecular and cellular biology*, vol. 23, no. 16, pp. 5594–5605, 2003.

[17] G. Liang, M. F. Chan, Y. Tomigahara, Y. C. Tsai, F. A. Gonzales, E. Li, P. W. Laird, and P. A. Jones, "Cooperativity between dna methyltransferases in the maintenance methylation of repetitive elements," *Molecular and cellular biology*, vol. 22, no. 2, pp. 480–491, 2002.

[18] M. C. Cirio, J. Martel, M. Mann, M. Toppings, M. Bartolomei, J. Trasler, and J. R. Chaillet, "Dna methyltransferase 1o functions during preimplantation development to preclude a profound level of epigenetic variation," *Developmental biology*, vol. 324, no. 1, pp. 139–150, 2008.

[19] H. Leonhardt, A. W. Page, H.-U. Weier, and T. H. Bestor, "A targeting sequence directs dna methyltransferase to sites of dna replication in mammalian nuclei," *Cell*, vol. 71, no. 5, pp. 865–873, 1992.

[20] C. Ling, P. Poulsen, S. Simonsson, T. Rönn, J. Holmkvist, P. Almgren, P. Hagert, E. Nilsson, A. G. Mabey, P. Nilsson *et al.*, "Genetic and epigenetic factors are associated with expression of respiratory chain component ndufb6 in human skeletal muscle," *The Journal of clinical investigation*, vol. 117, no. 11, pp. 3427–3435, 2007.

[21] A. H. Reis, F. R. Vargas, and B. Lemos, "Biomarkers of genome instability and cancer epigenetics," *Tumor Biology*, vol. 37, no. 10, pp. 13 029–13 038, 2016.

[22] J. Jost and H. Saluz, *DNA methylation: molecular biology and biological significance*. Birkhäuser, 2013, vol. 64.

[23] F. Larsen, G. Gundersen, R. Lopez, and H. Prydz, "Cpg islands as gene markers in the human genome," *Genomics*, vol. 13, no. 4, pp. 1095–1107, 1992.

[24] A. Bird, "Dna methylation patterns and epigenetic memory," *Genes & development*, vol. 16, no. 1, pp. 6–21, 2002.

[25] D. Takai and P. A. Jones, "Comprehensive analysis of cpg islands in human chromosomes 21 and 22," *Proceedings of the national academy of sciences*, vol. 99, no. 6, pp. 3740–3745, 2002.

[26] B. Brueckner and F. Lyko, "Dna methyltransferase inhibitors: old and new drugs for an epigenetic cancer therapy," *Trends in pharmacological sciences*, vol. 25, no. 11, pp. 551–554, 2004.

[27] G. C. Prendergast and E. B. Ziff, "Methylation-sensitive sequence-specific dna binding by the c-myc basic region," *Science*, vol. 251, no. 4990, pp. 186–189, 1991.

[28] M. Curradi, A. Izzo, G. Badaracco, and N. Landsberger, "Molecular mechanisms of gene silencing mediated by dna methylation," *Molecular and cellular biology*, vol. 22, no. 9, pp. 3157–3173, 2002.

[29] M. W. Łuczak and P. P. Jagodziński, "The role of dna methylation in cancer development," *Folia Histochem Cytobiol*, vol. 44, pp. 143–154, 2006.

[30] M. Szyf, "Targeting dna methylation in cancer," *Ageing research reviews*, vol. 2, no. 3, pp. 299–328, 2003.

[31] M. M. Suzuki and A. Bird, "Dna methylation landscapes: provocative insights from epigenomics," *Nature Reviews Genetics*, vol. 9, no. 6, pp. 465–476, 2008.

[32] P. A. Jones and S. B. Baylin, "The fundamental role of epigenetic events in cancer," *Nature reviews genetics*, vol. 3, no. 6, pp. 415–428, 2002.

[33] A. Eden, F. Gaudet, A. Waghmare, R. Jaenisch *et al.*, "Chromosomal instability and tumors promoted by dna hypomethylation," *Science*, vol. 300, no. 5618, pp. 455–455, 2003.

[34] G. Howard, R. Eiges, F. Gaudet, R. Jaenisch, and A. Eden, "Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice," *Oncogene*, vol. 27, no. 3, pp. 404–408, 2008.

[35] A. S. Wilson, B. E. Power, and P. L. Molloy, "Dna hypomethylation and human diseases," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, vol. 1775, no. 1, pp. 138–162, 2007.

[36] O. Ogawa, M. R. Eccles, J. Szeto, L. A. McNoe, K. Yun, M. A. Maw, P. J. Smith, and A. E. Reeve, "Relaxation of insulin-like growth factor ii gene imprinting implicated in wilms' tumour." *Nature*, vol. 362, no. 6422, pp. 749–751, 1993.

[37] H. Cui, M. Cruz-Correa, F. M. Giardiello, D. F. Hutcheon, D. R. Kafonek, S. Brandenburg, Y. Wu, X. He, N. R. Powe, and A. P. Feinberg, "Loss of igf2 imprinting: a potential marker of colorectal cancer risk," *Science*, vol. 299, no. 5613, pp. 1753–1755, 2003.

[38] V. Greger, E. Passarge, W. Höpping, E. Messmer, and B. Horsthemke, "Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma," *Human genetics*, vol. 83, no. 2, pp. 155–158, 1989.

[39] S. B. Baylin, "Dna methylation and gene silencing in cancer," *Nature clinical practice Oncology*, vol. 2, pp. S4–S11, 2005.

[40] C. Long, B. Yin, Q. Lu, X. Zhou, J. Hu, Y. Yang, F. Yu, and Y. Yuan, "Promoter hypermethylation of the runx3 gene in esophageal squamous cell carcinoma," *Cancer investigation*, vol. 25, no. 8, pp. 685–690, 2007.

[41] Y. Akiyama, N. Watkins, H. Suzuki, K.-W. Jair, M. van Engeland, M. Esteller, H. Sakai, C.-Y. Ren, Y. Yuasa, J. G. Herman *et al.*, "Gata-4 and gata-5 transcription factor genes and potential downstream antitumor target genes are epigenetically silenced in colorectal and gastric cancer," *Molecular and cellular biology*, vol. 23, no. 23, pp. 8429–8439, 2003.

[42] L. Di Croce, V. A. Raker, M. Corsaro, F. Fazi, M. Fanelli, M. Faretta, F. Fuks, F. L. Coco, T. Kouzarides, C. Nervi *et al.*, "Methyltransferase recruitment and dna hypermethylation of target promoters by an oncogenic transcription factor," *Science*, vol. 295, no. 5557, pp. 1079–1082, 2002.

[43] J. Frigola, J. Song, C. Stirzaker, R. A. Hinshelwood, M. A. Peinado, and S. J. Clark, "Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band," *Nature genetics*, vol. 38, no. 5, pp. 540–549, 2006.

[44] S. B. Baylin and P. A. Jones, "A decade of exploring the cancer epigenomebiological and translational implications," *Nature Reviews Cancer*, vol. 11, no. 10, pp. 726–734, 2011.

[45] H. Wu and Y. Zhang, "Mechanisms and functions of tet protein-mediated 5-methylcytosine oxidation," *Genes & development*, vol. 25, no. 23, pp. 2436–2452, 2011.

[46] P. Fenaux, G. J. Mufti, E. Hellstrom-Lindberg, V. Santini, C. Finelli, A. Giagounidis, R. Schoch, N. Gattermann, G. Sanz, A. List *et al.*, "Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase iii study," *The lancet oncology*, vol. 10, no. 3, pp. 223–232, 2009.

[47] H. Kantarjian, J.-P. J. Issa, C. S. Rosenfeld, J. M. Bennett, M. Albitar, J. DiPersio, V. Klimek, J. Slack, C. De Castro, F. Ravandi *et al.*, "Decitabine improves patient outcomes in myelodysplastic syndromes," *Cancer*, vol. 106, no. 8, pp. 1794–1803, 2006.

[48] H. I. Saba, M. Lübbert, and P. Wijermans, "Response rates of phase 2 and phase 3 trials of decitabine (dac) in patients with myelodysplastic syndromes (mds)." *Blood*, vol. 106, no. 11, pp. 2515–2515, 2005.

[49] P. Wijermans, M. Lübbert, G. Verhoef, A. Bosly, C. Ravoet, M. Andre, and A. Ferrant, "Low-dose 5-aza-2-deoxycytidine, a dna hypomethylating agent, for the treatment of high-risk myelodysplastic syndrome: a multicenter phase ii study in elderly patients," *Journal of Clinical Oncology*, vol. 18, no. 5, pp. 956–956, 2000.

[50] L. R. Silverman, E. P. Demakos, B. L. Peterson, A. B. Kornblith, J. C. Holland, R. Odchimar-Reissig, R. M. Stone, D. Nelson, B. L. Powell, C. M. DeCastro *et al.*, "Randomized controlled trial of azacitidine in patients with the myelodysplastic syndrome: a study of the cancer and leukemia group b," *Journal of Clinical oncology*, vol. 20, no. 10, pp. 2429–2440, 2002.

[51] J. A. Gollob and C. J. Sciambi, "Decitabine up-regulates s100a2 expression and synergizes with ifn-γ to kill uveal melanoma cells," *Clinical Cancer Research*, vol. 13, no. 17, pp. 5219–5225, 2007.

[52] K. Appleton, H. J. Mackay, I. Judson, J. A. Plumb, C. McCormick, G. Strathdee, C. Lee, S. Barrett, S. Reade, D. Jadayel *et al.*, "Phase i and pharmacodynamic trial of the dna methyltransferase inhibitor decitabine and carboplatin in solid tumors," *Journal of Clinical Oncology*, vol. 25, no. 29, pp. 4603–4609, 2007.

[53] V. Valdespino and P. M. Valdespino, "Potential of epigenetic therapies in the management of solid tumors," *Cancer management and research*, vol. 7, p. 241, 2015.

[54] D. Matei, F. Fang, C. Shen, J. Schilder, A. Arnold, Y. Zeng, W. A. Berry, T. Huang, and K. P. Nephew, "Epigenetic resensitization to platinum in ovarian cancer," *Cancer research*, vol. 72, no. 9, pp. 2197–2205, 2012.

[55] T. Kanda, M. Tada, F. Imazeki, O. Yokosuka, K. Nagao, and H. Saisho, "5-aza-2'-deoxycytidine sensitizes hepatoma and pancreatic cancer cell lines," *Oncology reports*, vol. 14, no. 4, pp. 975–979, 2005.

[56] S. Morita, S. Iida, K. Kato, Y. Takagi, H. Uetake, and K. Sugihara, "The synergistic effect of 5-aza-2-deoxycytidine and 5-fluorouracil on drug-resistant tumors," *Oncology*, vol. 71, no. 5-6, pp. 437–445, 2007.

[57] K. Schmelz, M. Wagner, B. Dörken, and I. Tamm, "5-aza-2-deoxycytidine induces p21waf expression by demethylation of p73 leading to p53-independent apoptosis in myeloid leukemia," *International journal of cancer*, vol. 114, no. 5, pp. 683–695, 2005.

[58] X. Tang, W. Wu, S.-Y. Sun, I. I. Wistuba, W. Hong, and L. Mao, "Hypermethylation of the death-associated protein kinase promoter attenuates the sensitivity to trail-induced apoptosis in human non-small cell lung cancer cells." *Molecular cancer research: MCR*, vol. 2, no. 12, pp. 685–691, 2004.

[59] T. Walton, G. Li, R. Seth, S. McArdle, M. Bishop, and R. Rees, "Dna demethylation and histone deacetylation inhibition co-operate to re-express estrogen receptor beta and induce apoptosis in prostate cancer cell-lines," *The Prostate*, vol. 68, no. 2, pp. 210–222, 2008.

[60] A. P. Feinberg, B. Vogelstein *et al.*, "Hypomethylation distinguishes genes of some human cancers from their normal counterparts," *Nature*, vol. 301, no. 5895, pp. 89–92, 1983.

[61] A. Korkmaz, L. C. Manchester, T. Topal, S. Ma, D.-X. Tan, and R. J. Reiter, "Epigenetic mechanisms in human physiology and diseases," *J Exp Integr Med*, vol. 1, no. 3, pp. 139–47, 2011.

[62] A. E. Handel, G. C. Ebers, and S. V. Ramagopalan, "Epigenetics: molecular mechanisms and implications for disease," *Trends in molecular medicine*, vol. 16, no. 1, pp. 7–16, 2010.

[63] V. V. Lao and W. M. Grady, "Epigenetics and colorectal cancer," *Nature Reviews Gastroenterology and Hepatology*, vol. 8, no. 12, pp. 686–700, 2011.

[64] T. Mikeska and J. M. Craig, "Dna methylation biomarkers: cancer and beyond," *Genes*, vol. 5, no. 3, pp. 821–864, 2014.

[65] P. W. Laird, "Principles and challenges of genome-wide dna methylation analysis," *Nature Reviews Genetics*, vol. 11, no. 3, pp. 191–203, 2010.

[66] E. E. Holmes, M. Jung, S. Meller, A. Leisse, V. Sailer, J. Zech, M. Mengdehl, L.-A. Garbe, B. Uhl, G. Kristiansen *et al.*, "Performance evaluation of kits for bisulfite-conversion of dna from tissues, cell lines, ffpe tissues, aspirates, lavages, effusions, plasma, serum, and urine," *PloS one*, vol. 9, no. 4, p. e93933, 2014.

[67] T. Mikeska, C. Bock, H. Do, and A. Dobrovic, "Dna methylation biomarkers in cancer: progress towards clinical implementation," *Expert review of molecular diagnostics*, vol. 12, no. 5, pp. 473–487, 2012.

[68] T. K. Wojdacz, "Methylation-sensitive high-resolution melting in the context of legislative requirements for validation of analytical procedures for diagnostic applications," *Expert review of molecular diagnostics*, vol. 12, no. 1, pp. 39–47, 2012.

[69] Z. Xiao, B. Li, G. Wang, W. Zhu, Z. Wang, J. Lin, A. Xu, and X. Wang, "Validation of methylation-sensitive high-resolution melting (ms-hrm) for the detection of stool dna methylation in colorectal neoplasms," *Clinica Chimica Acta*, vol. 431, pp. 154–163, 2014.

[70] L. S. Kristensen and L. L. Hansen, "Pcr-based methods for detecting single-locus dna methylation biomarkers in cancer diagnostics, prognostics, and response to treatment," *Clinical chemistry*, vol. 55, no. 8, pp. 1471–1483, 2009.

[71] M. W. Coolen, A. L. Statham, M. Gardiner-Garden, and S. J. Clark, "Genomic profiling of cpg methylation and allelic specificity using quantitative high-throughput mass spectrometry: critical evaluation and improvements," *Nucleic acids research*, vol. 35, no. 18, p. e119, 2007.

[72] J. Garcia-Gimenez, F. Sanchis-Gomar, G. Lippi, S. Mena, D. Ivars, M. Gomez-Cabrera, J. Viña, and F. Pallardó, "Epigenetic biomarkers: A new perspective in laboratory diagnostics," *Clinica Chimica Acta*, vol. 413, no. 19, pp. 1576–1582, 2012.

[73] K. H. Taylor, R. S. Kramer, J. W. Davis, J. Guo, D. J. Duff, D. Xu, C. W. Caldwell, and H. Shi, "Ultradeep bisulfite sequencing analysis of dna methylation patterns in multiple gene promoters by 454 sequencing," *Cancer research*, vol. 67, no. 18, pp. 8511–8518, 2007.

[74] J. Sandoval, H. Heyn, S. Moran, J. Serra-Musach, M. A. Pujana, M. Bibikova, and M. Esteller, "Validation of a dna methylation microarray for 450,000 cpg sites in the human genome," *Epigenetics*, vol. 6, no. 6, pp. 692–702, 2011.

[75] F. Jasmine, R. Rahaman, S. Roy, M. Raza, R. Paul, M. Rakibuz-Zaman, R. Paul-Brutus, C. Dodsworth, M. Kamal, H. Ahsan *et al.*, "Interpretation of genome-wide infinium methylation data from ligated dna in formalin-fixed, paraffin-embedded paired tumor and normal tissue," *BMC research notes*, vol. 5, no. 1, p. 1, 2012.

[76] K. A. Heichman and J. D. Warren, "Dna methylation biomarkers and their utility for solid cancer diagnostics," *Clinical chemistry and laboratory medicine*, vol. 50, no. 10, pp. 1707–1721, 2012.

[77] P. Wu, Z. Cao, and S. Wu, "New progress of epigenetic biomarkers in urological cancer," *Disease Markers*, vol. 2016, 2016.

[78] S. Guil and M. Esteller, "Dna methylomes, histone codes and mirnas: tying it all together," *The international journal of biochemistry & cell biology*, vol. 41, no. 1, pp. 87–95, 2009.

[79] S. Hauser, T. Zahalka, G. Fechner, S. C. Mueller, and J. Ellinger, "Serum dna hypermethylation in patients with kidney cancer: results of a prospective study," *Anticancer research*, vol. 33, no. 10, pp. 4651–4656, 2013.

[80] M. De Martino, T. Klatte, A. Haitel, and M. Marberger, "268 serum cell-free dna in renal cell carcinoma: A diagnostic and prognostic marker," *European Urology Supplements*, vol. 10, no. 2, p. 105, 2011.

[81] J. Wang, J. Li, J. Gu, J. Yu, S. Guo, Y. Zhu, and D. Ye, "Abnormal methylation status of fbxw10 and smpd3, and associations with clinical characteristics in clear cell renal cell carcinoma," *Oncology letters*, vol. 10, no. 5, pp. 3073–3080, 2015.

[82] Z. Wang, J. Wei, J. Zhou, A. Haddad, L. Zhao, P. Kapur, K. Wu, B. Wang, Y. Yu, B. Liao *et al.*, "Validation of dab2ip methylation and its relative significance in predicting outcome in renal cell carcinoma." *Oncotarget*, 2016.

[83] A. Bettin, I. Reyes, and N. Reyes, "Gene expression profiling of prostate cancer-associated genes identifies fibromodulin as potential novel biomarker for prostate cancer." *The International journal of biological markers*, vol. 31, no. 2, pp. e153–62, 2015.

[84] C. Jerónimo and R. Henrique, "Epigenetic biomarkers in urological tumors: A systematic review," *Cancer letters*, vol. 342, no. 2, pp. 264–274, 2014.

[85] S. Dijkstra, I. L. Birker, F. P. Smit, G. H. Leyten, T. M. de Reijke, I. M. van Oort, P. F. Mulders, S. A. Jannink, and J. A. Schalken, "Prostate cancer biomarker profiles in urinary sediments and exosomes," *The Journal of urology*, vol. 191, no. 4, pp. 1132–1138, 2014.

[86] V. Urquidi, C. J Rosser, and S. Goodison, "Molecular diagnostic trends in urological cancer: biomarkers for non-invasive diagnosis," *Current medicinal chemistry*, vol. 19, no. 22, pp. 3653–3663, 2012.

[87] V. L. Costa, R. Henrique, S. A. Danielsen, M. Eknaes, P. Patrício, A. Morais, J. Oliveira, R. A. Lothe, M. R. Teixeira, G. E. Lind *et al.*, "Tcf21 and pcdh17 methylation: An innovative panel of biomarkers for a simultaneous detection of urological cancers," *Epigenetics*, vol. 6, no. 9, pp. 1120–1130, 2011.

[88] L. Van Neste, R. J. Hendriks, S. Dijkstra, G. Trooskens, E. B. Cornel, S. A. Jannink, H. de Jong, D. Hessels, F. P. Smit, W. J. Melchers *et al.*, "Detection of high-grade prostate cancer using a urinary molecular biomarker–based risk score," *European urology*, 2016.

[89] A. Ouhtit, M. Al-Kindi, P. Kumar, I. Gupta, S. Shanmuganathan, and Y. Tamimi, "Hoxb13, a potential prognostic biomarker for prostate cancer." *Frontiers in bioscience (Elite edition)*, vol. 8, pp. 40–45, 2015.

[90] G. Hoyne, C. Rudnicka, Q.-X. Sang, M. Roycik, S. Howarth, P. Leedman, M. Schlaich, P. Candy, and V. Matthews, "Genetic and cellular studies highlight that a disintegrin and metalloproteinase 19 is a protective biomarker in human prostate cancer," *BMC cancer*, vol. 16, no. 1, p. 1, 2016.

[91] L. Zheng, D. Sun, W. Fan, Z. Zhang, Q. Li, and T. Jiang, "Diagnostic value of sfrp1 as a favorable predictive and prognostic biomarker in patients with prostate cancer," *PloS one*, vol. 10, no. 2, p. e0118276, 2015.

[92] H. Tahara, H. Naito, K. Kise, T. Wakabayashi, K. Kamoi, K. Okihara, A. Yanagisawa, Y. Nakai, N. Nonomura, E. Morii *et al.*, "Evaluation of psf1 as a prognostic biomarker for prostate cancer," *Prostate cancer and prostatic diseases*, vol. 18, no. 1, pp. 56–62, 2015.

[93] R. Morgan, A. Boxall, A. Bhatt, M. Bailey, R. Hindley, S. Langley, H. C. Whitaker, D. E. Neal, M. Ismail, H. Whitaker *et al.*, "Engrailed-2 (en2): a tumor specific urinary biomarker for the early diagnosis of prostate cancer," *Clinical Cancer Research*, vol. 17, no. 5, pp. 1090–1098, 2011.

[94] W. Onstenk, W. de Klaver, R. de Wit, M. Lolkema, J. Foekens, and S. Sleijfer, "The use of circulating tumor cells in guiding treatment decisions for patients with metastatic castration-resistant prostate cancer," *Cancer treatment reviews*, vol. 46, pp. 42–50, 2016.

[95] L. Mirabello, S. A. Savage, L. Korde, S. M. Gadalla, and M. H. Greene, "Line-1 methylation is inherited in familial testicular cancer kindreds," *BMC medical genetics*, vol. 11, no. 1, p. 1, 2010.

[96] B.-F. Chen, S. Gu, Y.-K. Suen, L. Li, and W.-Y. Chan, "microrna-199a-3p, dnmt3a, and aberrant dna methylation in testicular cancer," *Epigenetics*, vol. 9, no. 1, pp. 119–128, 2014.

[97] M. Brait, L. Maldonado, S. Begum, M. Loyo, D. Wehle, F. Tavora, L. Looijenga, J. Kowalski, Z. Zhang, E. Rosenbaum *et al.*, "Dna methylation profiles delineate epigenetic heterogeneity in seminoma and non-seminoma," *British journal of cancer*, vol. 106, no. 2, pp. 414–423, 2012.

[98] H.-M. Chen and J.-Y. Fang, "Epigenetic biomarkers for the early detection of gastrointestinal cancer," *Gastrointestinal Tumors*, vol. 1, no. 4, pp. 201–208, 2015.

[99] L. Migliore, F. Migheli, R. Spisni, and F. Coppedè, "Genetics, cytogenetics, and epigenetics of colorectal cancer," *BioMed Research International*, vol. 2011, 2011.

[100] K. Yamashita, S. Sakuramoto, and M. Watanabe, "Genomic and epigenetic profiles of gastric cancer: potential diagnostic and therapeutic applications," *Surgery today*, vol. 41, no. 1, pp. 24–38, 2011.

[101] A. Ooki, K. Yamashita, S. Kikuchi, S. Sakuramoto, N. Katada, K. Kokubo, H. Kobayashi, M. Kim, D. Sidransky, and M. Watanabe, "Potential utility of hop homeobox gene promoter methylation as a marker of tumor aggressiveness in gastric cancer," *Oncogene*, vol. 29, no. 22, pp. 3263–3275, 2010.

[102] W. Du, S. Wang, Q. Zhou, X. Li, J. Chu, Z. Chang, Q. Tao, E. Ng, J. Fang, J. Sung *et al.*, "Adamts9 is a functional tumor suppressor through inhibiting akt/mtor pathway and associated with poor survival in gastric cancer," *Oncogene*, vol. 32, no. 28, pp. 3319–3328, 2013.

[103] F. Coppedè, A. Lopomo, R. Spisni, and L. Migliore, "Genetic and epigenetic biomarkers for diagnosis, prognosis and treatment of colorectal cancer," *World J Gastroenterol*, vol. 20, no. 4, pp. 943–956, 2014.

[104] M. Li, W.-d. Chen, N. Papadopoulos, S. N. Goodman, N. C. Bjerregaard, S. Laurberg, B. Levin, H. Juhl, N. Arber, H. Moinova *et al.*, "Sensitive digital quantification of dna methylation in clinical samples," *Nature biotechnology*, vol. 27, no. 9, pp. 858–863, 2009.

[105] R. Grützmann, B. Molnar, C. Pilarsky, J. K. Habermann, P. M. Schlag, H. D. Saeger, S. Miehlke, T. Stolz, F. Model, U. J. Roblick *et al.*, "Sensitive detection of colorectal cancer in peripheral blood by septin 9 dna methylation assay," *PloS one*, vol. 3, no. 11, p. e3759, 2008.

[106] J. D. Warren, W. Xiong, A. M. Bunker, C. P. Vaughn, L. V. Furtado, W. L. Roberts, J. C. Fang, W. S. Samowitz, and K. A. Heichman, "Septin 9 methylated dna is a sensitive and specific blood test for colorectal cancer," *BMC medicine*, vol. 9, no. 1, p. 1, 2011.

[107] H. Yang, B.-Q. Xia, B. Jiang, G. Wang, Y.-P. Yang, H. Chen, B.-S. Li, A.-G. Xu, Y.-B. Huang, and X.-Y. Wang, "Diagnostic value of stool dna testing for multiple markers

of colorectal cancer and advanced adenoma: a meta-analysis," *Canadian Journal of Gastroenterology and Hepatology*, vol. 27, no. 8, pp. 467–475, 2013.

[108] Q. Guo, Y. Song, H. Zhang, X. Wu, P. Xia, and C. Dang, "Detection of hypermethylated fibrillin-1 in the stool samples of colorectal cancer patients," *Medical Oncology*, vol. 30, no. 4, pp. 1–5, 2013.

[109] S. C. Glöckner, M. Dhir, J. M. Yi, K. E. McGarvey, L. Van Neste, J. Louwagie, T. A. Chan, W. Kleeberger, A. P. de Bruïne, K. M. Smits *et al.*, "Methylation of tfpi2 in stool dna: a potential novel biomarker for the detection of colorectal cancer," *Cancer research*, vol. 69, no. 11, pp. 4691–4699, 2009.

[110] T. F. Imperiale, D. F. Ransohoff, S. H. Itzkowitz, T. R. Levin, P. Lavin, G. P. Lidgard, D. A. Ahlquist, and B. M. Berger, "Multitarget stool dna testing for colorectal-cancer screening," *N Engl J Med*, vol. 2014, no. 370, pp. 1287–1297, 2014.

[111] R. J. Ginsberg, L. V. Rubinstein, L. C. S. Group *et al.*, "Randomized trial of lobectomy versus limited resection for t1 n0 non-small cell lung cancer," *The Annals of Thoracic Surgery*, vol. 60, no. 3, pp. 615–623, 1995.

[112] T. E. Liggett, A. Melnikov, Q. Yi, C. Replogle, W. Hu, J. Rotmensch, A. Kamat, A. K. Sood, and V. Levenson, "Distinctive dna methylation patterns of cell-free plasma dna in women with malignant ovarian tumors," *Gynecologic oncology*, vol. 120, no. 1, pp. 113–120, 2011.

[113] H.-Y. Su, H.-C. Lai, Y.-W. Lin, Y.-C. Chou, C.-Y. Liu, and M.-H. Yu, "An epigenetic marker panel for screening and prognostic prediction of ovarian cancer," *International Journal of Cancer*, vol. 124, no. 2, pp. 387–393, 2009.

[114] C. G. A. R. Network *et al.*, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–615, 2011.

[115] R. Michaelson-Cohen, I. Keshet, R. Straussman, M. Hecht, H. Cedar, and U. Beller, "Genome-wide de novo methylation in epithelial ovarian cancer," *International Journal of Gynecological Cancer*, vol. 21, no. 2, pp. 269–279, 2011.

[116] S. H. Wei, C.-M. Chen, G. Strathdee, J. Harnsomburana, C.-R. Shyu, F. Rahmatpanah, H. Shi, S.-W. Ng, P. S. Yan, K. P. Nephew *et al.*, "Methylation microarray analysis of late-stage ovarian carcinomas distinguishes progression-free survival in patients and identifies candidate epigenetic markers," *Clinical Cancer Research*, vol. 8, no. 7, pp. 2246–2252, 2002.

[117] D. O. Bauerschlag, O. Ammerpohl, K. Bräutigam, C. Schem, Q. Lin, M. T. Weigel, F. Hilpert, N. Arnold, N. Maass, I. Meinhold-Heerlein *et al.*, "Progression-free survival in ovarian cancer is reflected in epigenetic dna methylation profiles," *Oncology*, vol. 80, no. 1-2, pp. 12–20, 2011.

[118] N. Häfner, D. Steinbach, L. Jansen, H. Diebolder, M. Dürst, and I. B. Runnebaum, "Runx3 and camk2n1 hypermethylation as prognostic marker for epithelial ovarian cancer," *International Journal of Cancer*, vol. 138, no. 1, pp. 217–228, 2016.

[119] D. S. Kim, J. Y. Lee, S. M. Lee, J. E. Choi, S. Cho, and J. Y. Park, "Promoter methylation of the rgc32 gene in nonsmall cell lung cancer," *Cancer*, vol. 117, no. 3, pp. 590–596, 2011.

[120] K. Tomczak, P. Czerwinska, M. Wiznerowicz *et al.*, "The cancer genome atlas (tcga): an immeasurable source of knowledge," *Contemp Oncol (Pozn)*, vol. 19, no. 1A, pp. A68–A77, 2015.

[121] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[122] D. P. Bartel, "Micrornas: target recognition and regulatory functions," *cell*, vol. 136, no. 2, pp. 215–233, 2009.

[123] T. A. Farazi, J. I. Spitzer, P. Morozov, and T. Tuschl, "mirnas in human cancer," *The Journal of pathology*, vol. 223, no. 2, pp. 102–115, 2011.

[124] S. Sandhu and R. Garzon, "Potential applications of micrornas in cancer diagnosis, prognosis, and treatment," in *Seminars in oncology*, vol. 38, no. 6. Elsevier, 2011, pp. 781–787.

[125] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in dna by primed synthesis with dna polymerase," *Journal of molecular biology*, vol. 94, no. 3, pp. 441–448, 1975.

[126] H. Bayley, "Sequencing single molecules of dna," *Current opinion in chemical biology*, vol. 10, no. 6, pp. 628–637, 2006.

[127] S. A. McCarroll, F. G. Kuruvilla, J. M. Korn, S. Cawley, J. Nemesh, A. Wysoker, M. H. Shapero, P. I. de Bakker, J. B. Maller, A. Kirby *et al.*, "Integrated detection and population-genetic analysis of snps and copy number variation," *Nature genetics*, vol. 40, no. 10, pp. 1166–1174, 2008.

[128] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[129] D. A. Gutman, J. Cobb, D. Somanna, Y. Park, F. Wang, T. Kurc, J. H. Saltz, D. J. Brat, L. A. Cooper, and J. Kong, "Cancer digital slide archive: an informatics resource to support integrated in silico analysis of tcga pathology data," *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1091–1098, 2013.

[130] J. Zhang, R. Finney, M. Edmonson, C. Schaefer, W. Rowe, C. Yan, R. Clifford, S. Greenblum, G. Wu, H. Zhang *et al.*, "The cancer genome workbench: identifying and visualizing complex genetic alterations in tumors," *NCI Nature Pathway Interaction Database*, vol. 10, 2010.

[131] J. Z. Sanborn, S. C. Benz, B. Craft, C. Szeto, K. M. Kober, L. Meyer, C. J. Vaske, M. Goldman, K. E. Smith, R. M. Kuhn *et al.*, "The ucsc cancer genomics browser: update 2011," *Nucleic acids research*, p. gkq1113, 2010.

[132] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson *et al.*, "The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," *Cancer discovery*, vol. 2, no. 5, pp. 401–404, 2012.

[133] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal," *Science signaling*, vol. 6, no. 269, p. pl1, 2013.

[134] K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov *et al.*, "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin," *Cell*, vol. 158, no. 4, pp. 929–944, 2014.

[135] S. D. Brown, R. L. Warren, E. A. Gibb, S. D. Martin, J. J. Spinelli, B. H. Nelson, and R. A. Holt, "Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival," *Genome research*, vol. 24, no. 5, pp. 743–750, 2014.

[136] M. K. Samur, Z. Yan, X. Wang, Q. Cao, N. C. Munshi, C. Li, and P. K. Shah, "canevolve: a web portal for integrative oncogenomics," *PLoS One*, vol. 8, no. 2, p. e56228, 2013.

[137] C. P. Goswami and H. Nakshatri, "Proggenev2: enhancements on the existing database," *BMC cancer*, vol. 14, no. 1, p. 1, 2014.

[138] G. Ciriello, E. Cerami, C. Sander, and N. Schultz, "Mutual exclusivity analysis identifies oncogenic network modules," *Genome research*, vol. 22, no. 2, pp. 398–406, 2012.

[139] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz, "Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers," *Genome biology*, vol. 12, no. 4, p. 1, 2011.

[140] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts *et al.*, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, 2013.

[141] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm," *Bioinformatics*, vol. 26, no. 12, pp. i237–i245, 2010.

[142] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor *et al.*, "Galaxy: a platform for interactive large-scale genome analysis," *Genome research*, vol. 15, no. 10, pp. 1451–1455, 2005.

[143] M. Goldman, B. Craft, T. Swatloski, M. Cline, O. Morozova, M. Diekhans, D. Haussler, and J. Zhu, "The ucsc cancer genomics browser: update 2015," *Nucleic acids research*, p. gku1073, 2014.

[144] D. J. Weisenberger, "Characterizing dna methylation alterations from the cancer genome atlas," *The Journal of clinical investigation*, vol. 124, no. 1, pp. 17–23, 2014.

[145] A. Koch, T. De Meyer, J. Jeschke, and W. Van Criekinge, "Mexpress: visualizing expression, dna methylation and clinical tcga data," *BMC genomics*, vol. 16, no. 1, p. 636, 2015.

[146] J. Zhang, R. P. Finney, W. Rowe, M. Edmonson, S. H. Yang, T. Dracheva, J. Jen, J. P. Struewing, and K. H. Buetow, "Systematic analysis of genetic alterations in tumors using cancer genome workbench (cgwb)," *Genome research*, vol. 17, no. 7, pp. 1111–1117, 2007.

[147] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, "Integrative genomics viewer (igv): high-performance genomics data visualization and exploration," *Briefings in bioinformatics*, vol. 14, no. 2, pp. 178–192, 2013.

[148] E. R. Tufte and P. Graves-Morris, *The visual display of quantitative information.* Graphics press Cheshire, CT, 1983, vol. 2, no. 9.

[149] M. Goldman, B. Craft, T. Swatloski, K. Ellrott, M. Cline, M. Diekhans, S. Ma, C. Wilks, J. Stuart, D. Haussler *et al.*, "The ucsc cancer genomics browser: update 2013," *Nucleic acids research*, vol. 41, no. D1, pp. D949–D954, 2013.

[150] W. Kent, C. Sugnet, T. Furey, K. Roskin, T. Pringle, A. Zahler, and D. Haussler, "The human genome browser at ucsc genome res 2002 12."

[151] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[152] Y.-W. Wang, X. Ma, Y.-A. Zhang, M.-J. Wang, Y. Yatabe, S. Lam, L. Girard, J.-Y. Chen, and A. F. Gazdar, "Itpka gene body methylation regulates gene expression and serves as an early diagnostic marker in lung and other cancers," *Journal of Thoracic Oncology*, 2016.

[153] J. G. Herman, A. Merlo, L. Mao, R. G. Lapidus, J.-P. J. Issa, N. E. Davidson, D. Sidransky, and S. B. Baylin, "Inactivation of the cdkn2/p16/mts1 gene is frequently associated with aberrant dna methylation in all common human cancers," *Cancer research*, vol. 55, no. 20, pp. 4525–4530, 1995.

[154] J. G. Herman, J. Jen, A. Merlo, and S. B. Baylin, "Hypermethylation-associated inactivation indicates a tumor suppressor role for p15ink4b," *Cancer research*, vol. 56, no. 4, pp. 722–727, 1996.

[155] M. Sanchez-Cespedes, M. Esteller, L. Wu, H. Nawroz-Danish, G. H. Yoo, W. M. Koch, J. Jen, J. G. Herman, and D. Sidransky, "Gene promoter hypermethylation in tumors and serum of head and neck cancer patients," *Cancer research*, vol. 60, no. 4, pp. 892–895, 2000.

[156] M. F. Kane, M. Loda, G. M. Gaida, J. Lipman, R. Mishra, H. Goldman, J. M. Jessup, and R. Kolodner, "Methylation of the hmlh1 promoter correlates with lack of expression of hmlh1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines," *Cancer research*, vol. 57, no. 5, pp. 808–811, 1997.

[157] P. A. Jones, "Functions of dna methylation: islands, start sites, gene bodies and beyond," *Nature Reviews Genetics*, vol. 13, no. 7, pp. 484–492, 2012.

[158] M. Renner, T. Wolf, H. Meyer, W. Hartmann, R. Penzel, A. Ulrich, B. Lehner, V. Hovestadt, E. Czwan, G. Egerer *et al.*, "Integrative dna methylation and gene expression analysis in high-grade soft tissue sarcomas," *Genome biology*, vol. 14, no. 12, p. 1, 2013.

[159] X. Yang, H. Han, D. D. De Carvalho, F. D. Lay, P. A. Jones, and G. Liang, "Gene body methylation can alter gene expression and is a therapeutic target in cancer," *Cancer cell*, vol. 26, no. 4, pp. 577–590, 2014.

[160] X. Zhang, R. Yang, Y. Jia, D. Cai, B. Zhou, X. Qu, H. Han, L. Xu, L. Wang, Y. Yao *et al.*, "Hypermethylation of sp1 binding site suppresses hypothalamic pomc in neonates and may contribute to metabolic disorders in adults: impact of maternal dietary clas," *Diabetes*, vol. 63, no. 5, pp. 1475–1487, 2014.

[161] W.-G. Zhu, K. Srinivasan, Z. Dai, W. Duan, L. J. Druhan, H. Ding, L. Yee, M. A. Villalona-Calero, C. Plass, and G. A. Otterson, "Methylation of adjacent cpg sites affects sp1/sp3 binding and activity in the p21cip1 promoter," *Molecular and cellular biology*, vol. 23, no. 12, pp. 4056–4065, 2003.

[162] I. N. Zelko, M. R. Mueller, and R. J. Folz, "Cpg methylation attenuates sp1 and sp3 binding to the human extracellular superoxide dismutase promoter and regulates its cell-specific expression," *Free Radical Biology and Medicine*, vol. 48, no. 7, pp. 895–904, 2010.

[163] G. Perini, D. Diolaiti, A. Porro, and G. Della Valle, "In vivo transcriptional regulation of n-myc target genes is controlled by e-box methylation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 34, pp. 12 117–12 122, 2005.

[164] V. Vanweyenberg, D. Communi, C. D'santos, and C. Erneux, "Tissue-and cell-specific expression of ins (1, 4, 5) p3 3-kinase isoenzymes," *Biochemical Journal*, vol. 306, no. 2, pp. 429–435, 1995.

[165] B. Novakovic and R. Saffery, "Placental pseudo-malignancy from a dna methylation perspective: unanswered questions and future directions," *Frontiers in genetics*, vol. 4, p. 285, 2013.

[166] ——, "Dna methylation profiling highlights the unique nature of the human placental epigenome," 2010.

[167] D. I. Schroeder, J. D. Blair, P. Lott, H. O. K. Yu, D. Hong, F. Crary, P. Ashwood, C. Walker, I. Korf, W. P. Robinson *et al.*, "The human placenta methylome," *Proceedings of the national academy of sciences*, vol. 110, no. 15, pp. 6037–6042, 2013.

[168] V. L. Costa, R. Henrique, S. A. Danielsen, S. Duarte-Pereira, M. Eknaes, R. I. Skotheim, Â. Rodrigues, J. S. Magalhães, J. Oliveira, R. A. Lothe *et al.*, "Three epigenetic biomarkers, gdf15, tmeff2, and vim, accurately predict bladder cancer from dna-based analyses of urine samples," *Clinical Cancer Research*, vol. 16, no. 23, pp. 5842–5851, 2010.

[169] Z. Zuo, M. Zhao, J. Liu, Y. Wei, and X. Wu, "Inhibitory effect of bladder cancer related protein gene on hela cell proliferation." *Ai zheng= Aizheng= Chinese journal of cancer*, vol. 25, no. 7, pp. 811–817, 2006.

[170] Z. Zuo, M. Zhao, J. Liu, G. Gao, and X. Wu, "Functional analysis of bladder cancer-related protein gene: a putative cervical cancer tumor suppressor gene in cervical carcinoma," *Tumor Biology*, vol. 27, no. 4, pp. 221–226, 2006.

[171] F. K. Rae, S.-A. Stephenson, D. L. Nicol, and J. A. Clements, "Novel association of a diverse range of genes with renal cell carcinoma as identified by differential display," *International journal of cancer*, vol. 88, no. 5, pp. 726–732, 2000.

[172] H. Su, Y. Zhao, D. Fan, Q. Fan, P. Zhang, Y. Wen, and Y. Liu, "[relationship between expression of blcap protein and malignancy of osteosarcoma]." *Xi bao yu fen zi mian yi xue za zhi= Chinese journal of cellular and molecular immunology*, vol. 19, no. 5, pp. 465–466, 2003.

[173] S. Takashima, M. Hirose, N. Ogonuki, M. Ebisuya, K. Inoue, M. Kanatsu-Shinohara, T. Tanaka, E. Nishida, A. Ogura, and T. Shinohara, "Regulation of pluripotency in male germline stem cells by dmrt1," *Genes & development*, vol. 27, no. 18, pp. 1949–1958, 2013.

Curriculum Vitae

Santosh Mahadevana Goud is currently a graduate student in the School of Systems Biology at the George Mason University (GMU) in the Northern Virginia Area. He worked as a graduate teaching assistant in the Biology department at GMU, Fairfax, VA.
Before joining GMU, he has worked in various capacities from being a Research Assistant to Junior Research Fellow in India and Germany. He completed his Bachelor of Science in Microbiology (1998) and Master in Science in Biotechnology (2001) from Bangalore University, Bangalore, India. He also completed a second Masters degree from University of Hartford (2007), West Hartford, US. His research interest can be broadly categorized as Bioinformatics, Systems Biology, Cell Biology, Molecular Biology, Biostatistics and Computational Biology.