METHYLATION LANDSCAPES ACROSS CANCERS RECAPITULATE CLINICAL FINDINGS AND SUGGEST NEW ASSOCIATIONS

Presented by Rathi Kannan

In partial fulfillment of the requirements for graduation with the Dean's Scholars Honors Degree in Biology.

_____ _____

Dr. Claus O. Wilke
Supervising Professor

Date

_____ _____

Dr. Ruth Buskirk
Honors Advisor in Biology

Date

I grant the Dean's Scholars Program permission to post a copy of my thesis on the University of Texas Digital Repository. For more information, visit http://repositories.lib.utexas.edu/about.

METHYLATION LANDSCAPES ACROSS CANCERS RECAPITULATE CLINICAL FINDINGS AND SUGGEST NEW ASSOCIATIONS

Department: Biology

_____ _____

Student Date

_____

Signature

_____ _____

Supervising Professor Date

_____

Signature

Abstract

Genomic instability caused by epigenetic modifications is implicated in promoting tumor progression. DNA methylation is a mode of epigenetic modification that alters gene expression by hypermethylation, a suppressed state, or hypomethylation, an active state. Comparison between healthy tissues and tumors has shown aberrant methylation across various types of cancers. Here, I have compared methylation profiles across eight diverse cancers to identify global methylomes patterns that are common to a tumorigenic state. Using the Illumina Human Methylation 27K platform, I found an average of 65% of interrogated sites in proximal promoter regions display hypomethylation. In addition, I identified putative methylation sub-types within some cancers. Clustering of these subtypes reveals novel associations between cancers as well as recapitulates associations described through clinical observations. Finally, analyses of 39 tumor suppressor genes that have been shown to be hypermethylated in other studies were shown to be overwhelming hypomethylated in my data.

Introduction

Cancer is caused by aberrant cells that have gained the ability to grow and divide, free of normal regulatory constraints. These hyper-proliferative cells form tumors that invade healthy tissue and aggressively consume metabolic resources. The shift from healthy to neoplastic tissue is not so much an instantaneous switch as it is a multistep progression (Hanahan 2011). Healthy cells progress to a tumorigenic state by accumulating growth advantages over neighboring cells. These growth advantages allow subpopulations of cells to grow and proliferate at a greater rate than their neighbors. These growth advantages are heritable and can be genetic or epigenetic in nature.

Epigenetic modifications are heritable changes to a genome that do not change the actual genetic sequence. DNA methylation is one such epigenetic modification that plays a critical role in several cellular processes including genomic imprinting, embryonic development, X chromosome inactivation, and maintaining chromosomal integrity. Sweeping methylation changes have been observed in almost all cancers characterized by global hypomethylation along with targeted hypermethylation (Suzuki 2008). Before we can understand how these competing processes influence tumorigenesis, we must first understand how and why DNA methylation occurs under normal conditions.

In mammalian genomes, DNA methylation occurs primarily at cytosine bases located immediately 5' of guanosine bases (Bird 2002). In double-stranded DNA, this results in two methylated cytosines oriented diagonally from each other on opposing strands. While methylation can occur at any cytosine base, over 98% of methylation marks occur at this cytosine-guanosine dinucleotide motif, termed a CpG dinucleotide. The DNA methyltransferase (DMNTs) family of enzymes, of which there are 5 known

members, is responsible for establishing and maintaining DNA methylation (Bestor 2000). DMNTs are responsible for de novo methylation wherein the original pattern of methylation is established and for maintenance methylation where methylation is "transferred" from an old DNA strand to a new strand following replication. Aberrant expression or mutations in DNMTs have been shown to have dire consequences, especially during development. DNMT1 has a maintenance methylation function and mice with both copies of DNMT1 deleted do not survive past day E9 of embryonic development (Li 1992). DNMT3a and DNMT3b are de novo methylation enzymes and knockout mice for DNMT3a die at around 4 weeks while DNMT3b knockout mice do not survive E18.5 of embryonic development (Okano 1999).

The mammalian methylome is sparsely and globally distributed, with CpG dinucleotides found in all genomic elements including promoters, intragenic regions, and transposons (Suzuki 2008). CpGs are enriched in regions termed CpG islands (CGIs) (Bird 1986). CGIs are at least 500 bp in length with greater than 55% GC content (Takai 2002). CGIs generally exist in a hypomethylated state in normal cells where as much as 98% of all CpGs are normally methylated. Furthermore, CGIs are present in proximal promoter regions of more than 50% of mammalian genes (Antequera 1993). This methylation landscape is not uniform across all eukarya. For example, several model organisms including *Saccharomyces cerevisiae* and *Caenorhabditis elegans* do not display DNA methylation or carry the enzymes necessary for such modifications. Invertebrates most frequently display a mosaic epigenomic landscape with regions of heavily methylated DNA interspersed with regions of unmethylated DNA (Suzuki 2008).

As discussed previously, DNA methylation is a crucial aspect of proper development in mammals. Deletion of specific DNMTs in mice has shown repression of X inactivation in females and aberrant expression and maintenance of imprinted genes (Dodge 2005). These and other experiments have facilitated the view that DNA methylation serves to maintain the repressed chromatin state and thereby, silence gene expression. However, the true picture of how DNA methylation affects gene expression is far more complex and nebulous. For example, we still do not know how de novo DNMTs are guided to the specific CpGs they will methylate. Equally perplexing is the discovery of intragenic methylation, or gene body methylation. Early on, CGI methylation in promoter regions of X chromosome genes was causally linked to the silencing of gene expression from the inactivated chromosome (Mohandas 1981). However, a recent study has shown that the active X chromosome actually displays more than double the amount of methylation throughout than does the inactive X (Weber 2005). Most of this methylation is found within gene bodies. Whether or not this inversion of how methylation affects expression can be expanded to describe autosomal chromosomes is yet to be seen. Gene body methylation has been observed in mammalian autosomal genes and they conspicuously do not occur on the 5' ends of the genes. While it has been established that gene body methylation does not silence gene expression, it is not clear what function this modification might serve. Perhaps comfortingly, CGI methylation in promoters has been unequivocally linked with gene silencing (Costello 2000, Gitan 2002, Gonzalgo 1997, Toyota 1999). This has been observed both on inactivated X chromosomes and in a variety of different somatic tissue types.

Feinberg and Vogelstein first established the link between aberrant DNA methylation and cancer in 1983 (Feinberg 1983). Using primary tumors and adjacent healthy tissue from patients with colon and lung carcinomas, they found that tumor cells are hypomethylated compared to their normal counterparts. Feinberg and Vogelstein selected three genes for their experiment: human growth hormone, gamma-globin, and alpha-globin. They digested their samples with certain restriction enzymes that specifically target 5' CG 3'. These enzymes cannot digest methylated CpGs so they were able to differentiate methylated sites from unmethylated ones. In the same year, Gama-Sosa and Slagel determined the methylation status across primary, secondary, and benign neoplasms as well as healthy tissue (Gama-Sosa 1983). They too digested their samples with methylated-cytosine senstive enzymes and analyzed the fragments from entire genomes using HPLC. They observed hypomethylation across the neoplasms with benign neoplasms displaying the least amount of hypomethylation and secondary neoplasms displaying the most.

Since these early experiments, global hypomethylation has been established as an aberrant feature of cancer. According to the stem cell theory of cancer, global hypomethylation is an early step in tumorigenesis that is reflective of cells reverting to a pluripotent state. Indeed, while hypomethylation alone does not signify up-regulation in transcriptional activity, there is mounting evidence that global hypomethylation contributes to genomic instability that can result in the gross changes in karyotypes we see in cancers. As mentioned earlier, targeted hypermethylation is also a hallmark of epigenetic modification in cancers. Several studies have linked hypermethylation of prominent tumor suppressors with their subsequent silencing. Interestingly, some of these

tumor suppressors are regulated primarily by epigenetic changes and are rarely found mutated in cancers. *O*-6-methylguanine-DNA methyltransferase (MGMT) is one such tumor suppressor that plays an important role in DNA repair (Esteller 1999,2000,2001).

In this study, we aimed to recapitulate certain findings in individual cancers across a set of eight cancers. In particular, we were interested in establishing similarities between methylome landscapes that might reflect aspects of the cancer state as a whole.

Methods

Analysis Pipeline

All of the experiments conducted in this paper can be recapitulated using the scripts and annotation files found at: https://github.com/rpkannan/hm_27k_profile.

For directions and details, please refer to the README in this repository.

Data Collection

All data was obtained through The Cancer Genome Atlas (TCGA) data matrix: https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm. The following filter settings were used to select the appropriate data for each cancer type:

Data Type: DNA Methylation

Batch Number: All

Data Level: Level 3

Availability: Available

Preservation: Frozen

Center/Platform: JHU_USC (HumanMethylation27)

The following cancers were selected for study because they had the most methylome data available:

- Breast invasive carcinoma (BRCA), n = 315
- Colon adenocarcinoma (COAD), n = 166
- Glioblastoma multiforme (GBM), n = 295
- Kidney renal clear cell carcinoma (KIRC), n = 219
- Lung adenocarcinoma (LUAD), n = 126
- Lung squamous cell carcinoma (LUSC), n = 134

- Ovarian serous cystadenocarcinoma (OV), n = 591
- Uterine corpus endometrial carcinoma (UCEC), n = 117

As per TCGA guidelines, all samples are collected from primary tumors and contain at least 60% tumor nuclei.

Methylation profiles are organized by cancer with a separate text file for each sample profile. Samples are uniquely identified by barcodes and each profile is presented as interrogated loci (probe IDs, chromosomal location, etc.) with associated derived beta values.

Methylation Platform

TCGA tissue processing centers used the Illumina Infinium Human Methylation 27K array platform to generate methylation profiles for each sample. This platform utilizes fluorescent probes to interrogate 27,578 CpG sites that correspond to 14,475 genes. Briefly, bisulfite-treated DNA is incubated with probes linked to fluorophores. Methylated CpG sites elicit one signal and un-methylated CpG sites elicit another signal. The ratio of these raw signal intensities, specifically the ratio of methylated probe intensity to the total probe intensity, is termed the beta value. The beta value is continuously valued from 0 to 1 inclusive and is broadly interpreted as the degree of methylation at the interrogated site.

Probes were selected based on their performance (ability to generate accurate signal), location in a CGI, and distance from a transcriptional start site. Based on supporting studies that show adjacent CpGs have similar methylation statuses, each CGI is interrogated by an average of two probes (Eckhardt 2006). Annotation (genomic location, corresponding gene ID, distance from transcription start site) was obtained from

TCGA and maps to the Human Genome Build 36: https://tcga-data.nci.nih.gov/tcga/tcgaPlatformDesign.jsp.

Data Organization and Pre-Processing

Data from TCGA files was handled primarily in Python using the NumPy, Scipy, and Pandas packages. Methylomes were assembled by cancer into dictionaries with barcode keys and associated lists of beta values indexed by probe ID (script: `from_source_hm27k.py`). Beta values are always stored as floats in NumPy data structures to improve calculation efficiencies. TCGA datasets contain NA values for any individual probes that do not cross a signal intensity level or for probes that map to SNPs. I removed all probes that map to SNPs using a NumPy mask for a new total of 24,981 interrogated probes.

Histograms

I extracted beta values by cancer from the source files and divided them into two lists based on membership in CGIs. I removed NA values using a NumPy mask and plotted histograms using the pyplot.hist() function using the following parameters (script: `plots.py)`: bins=100, normed=1.

Discretize Beta Values

I was less interested in the absolute strength of methylation as I was in broadly categorizing probes as hypo- and hyper-methylated. Thus, based on the distributions of beta values found in the histograms, I chose cutoffs to delineate hypomethylation and hypermethylation. I made new dictionaries for each cancer where beta values were discretized to fall under three categories of regulation (script: `from_source_hm27k.py`):

- Raw Score <= 0.2 : Discretized Score: 0.0
- 0.2 < Raw Score <= 0.6 : Discretized Score: 1.0
- Raw Score > 0.6: Discretized Score: 2.0
- NA values were retained as such

Hierarchical Clustering

I performed hierarchical clustering analysis using the hamming distance measure and Ward's linkage method. The hamming distance is a measure of dissimilarity and accounts for differences in methylation status by probe. Ward's method clustering criterion generates clusters based on which merge produces the smallest increase in the sum of squares of the new cluster formed. I computed pairwise hamming distances in Python to utilize the efficiency of NumPy calculations. I performed hierarchical clustering in R using the hclust() function with the following parameter: method="ward". This generates an hclust class object that can be visualized as a dendrogram using the plot() function. This object can also be passed to the function heatmap.2() to generate a heatmap based on the clustering.

Identification of Sub-Types

I computed pairwise hamming distances between samples (script: `to_cluster_in_R.py`) and performed hierarchical clustering analysis (script: `cluster_by_cancer.R`). The generated hclust object contains a list of the height of each cluster in the dendrogram. I used this list to find the heights of the two-four largest clusters in each dendrogram. Given the heights observed across all cancer dendrograms, I chose to subdivide the cancers into the following subtypes:

- BRCA: height= 21117.400; subtypes= 2

- COAD: height= 15563.879, subtypes= 1
- GBM: height= 21342.033, subtypes= 2
- KIRC: height= 19152.379, subtypes= 2
- LUAD: height= 21050.091, subtypes= 2
- LUSC: height= 16895.093, subtypes= 1
- OV: heights= 22028.307, 34561.637, subtypes= 4
- UCEC: height= 13670.358, subtypes= 1

I used the cuttree() function with the heights indicated above to find which samples comprised each cluster (ie. Subtype).

Clustering by Sub-Type

The subtype memberships found previously were used to partition the cancer data dictionaries into subtype data dictionaries in Python (script: `sub_cancers.py`). Subtype dictionaries maintain the same format used in previous data dictionaries. To calculate the consensus sequence for each sub-type, I summed the occurrences of each score at each probe ID. The consensus score assigned was the value that occurred most frequently. Priority to ties was given as follows: hyper-, hypo-, neutral. The consensus sequences for all 15 sub-types were assigned to a matrix. I computed the pairwise hamming distance matrices by subtype on this matrix and exported these values to use in hierarchical clustering (script: `cluster_sub_cancers.R`).

Clustering by Tumor Suppressors

Using previous studies, I compiled a list of 39 tumor suppressors that are commonly silenced in other cancers. To ascertain their methylation status in my cancer sub-types, I compiled the discretized beta values associated with these genes using a resolution

algorithm described below. I generated consensus sequences for each sub-type and calculated the hamming distance between genes as described previously (script: `hypermeth_44.py`). I performed clustering analysis as described previously and generated a heat map based on the computed dendrogram using the heatmap.2() function (script: `cluster_hypermeth_44.R)`.

Resolution of Probes to Genes

Probe IDs were mapped to gene IDs based on the annotation provided by TCGA. Because two or more probes cover each gene, we needed to generate an algorithm to resolve multiple beta values. We used the following resolution criterion:

- If a gene is covered by two probes:
    - If one probe is not scored (NA), we use the score of the other probe
    - If one probe value is neutral and the other is hyper- or hypo- methylated, we use the hyper- or hypo- methylation score
    - If one probe is hyper-methylated and the other is hypo-methylated, we use the hyper-methylation score
- If the gene is covered by more than two probes:
    - We resolve the score to the value (hypo-, hyper-, or neutral) that occurs most frequently. Ties are given the following priority: hyper-, hypo-, neutral.

Similar to the probe-based dictionaries, I stored these resolved beta values in dictionaries with barcode keys and associated beta values indexed by gene ID (script: `cluster_methods.py)`.

Results

Global Hypomethylation

We wanted to assess the global methylation status of interrogated probes across our selected cancers. Previous studies have shown that the CGIs and non-island CpGs are under different selective pressures. Specifically, CGIs are frequently hypermethylated in cancers. It is yet unclear how non-island CpG methylation status affects gene expression. The Illumina Human Methylation 27K platform interrogates 20,006 CpG sites that are within CGIs and 7,572 that are not within CGIs. Un-discretized beta values were divided by CGI and non-CGI status and used to generate histograms. We found that the distributions of hypomethylated and hypermethylated probes did not vary significantly between CGI and non-CGI probes within each cancer type (Fig. 1 A, B). Furthermore, we did not find a significant difference between the number of probes that were hypomethylated and hypermethylated across cancers. OV displayed the greatest frequency of hypomethylated probes (68% of CGI probes and 69% of non-CGI probes) while the lowest frequency of hypomethylated probes at 62% was observed in several categories (CGI probes in BRCA and COAD and non-CGI probes in LUAD) (Fig. 1 C).

Methylome Landscape By Cancer

Methylome landscapes are highly variable by cell type. Given the heterogeneity of our tissue samples (at most 30% of each sample does not display tumor nuclei) as well as the possibility that multiple aberrant cell subtypes can exist within a tumor, we wanted to see if we could identify methylome subtypes within each cancer.

Given that we did not find a difference in distribution of CGI and non-CGI probes, we ceased to use this distinction to segregate probes. We discretized the probe beta values

and performed hierarchical clustering analysis on each cancer using Ward's method as the clustering criterion and the hamming distance as the distance criterion. Based on the height (magnitude of the sum of squares) of the clusters generated, we delineated subtypes within each cancer (Fig. 2). The height cutoffs were determined based on how closely samples clustered within each cancer and how distinct the sub-type clusters appeared. Though these subtype delineations are rough and not substantiated by other measures (ie. patient outcomes or other metrics attached to each sample), they provided a simple distinction with which to compare methylomes across cancers.

To ascertain how these putative cancer-subtypes might relate to each other, we performed another round of hierarchical clustering. First, we generated a consensus sequence for each cancer sub-type wherein each probe was assigned a single beta value that was observed with the greatest frequency in that sub-type. Clustering using Ward's criterion revealed that sub-types did not cluster purely based on the tissue of origin (Fig. 3). However, the distances between each cluster were not as large as the distances observed when clustering within each cancer. This could imply that cancer methylomes are quite heterogeneous in nature such that they bear as much resemble to any other cancer methylome as they do to one of their own tissue type.

Genes Regulated Similarly Across All Cancers

We wanted to ensure that the apparent similarities in methylomes displayed by the probe values would also be reflected when multiple probe values were resolved and mapped onto their corresponding genes. The methylation platform is designed such that two or more probes map to each gene ID. During the resolution process, when probes carried conflicting methylation statuses, we erred on the side of granting the gene a

hypermethylated status. Furthermore, to apply an extra criterion of stringency, we only included genes that scored with 95% consistency within each cancer. We found that, of 14,475 genes represented in this platform, 5,245 genes were hypomethylated across all cancers and 655 were hypermethylated. These numbers do indeed reflect the landscape of global hypomethylation observed earlier.

Regulation of 39 Known Tumor Suppressors

Silencing of tumor suppressors by hypermethylation has become a hallmark of cancer. As such, we wanted to examine the methylation states of 39 tumor suppressors, selected based on known suppression in other studies. Probe beta values were resolved to the selected gene IDs and we performed hierarchical clustering by cancer sub-types (Fig. 4). Surprisingly, we find approximately 20 of the 39 tumor suppressors are not hypermethylated in our samples. The genes that are consistently hypermethylated across all cancer sub-types are Semaphorin 3b, Runt-related transcription factor 3, checkpoint with forkhead and ring finger domain, and *O*-6-methylguanine-DNA methyltransferase.

Discussion

In this work, I have analyzed patterns of global hypomethylation across different cancers. I observed similar methylation profiles between CGI and non-CGI CpG sites. Based on heterogeneity in methylation profiles, I have identified putative sub-types within different cancers and developed consensus sequences for each sub-type. Clustering of these consensus sequences has further revealed novel similarities between cancers and recapitulated similarities identified through clinical results. I surveyed 39 important tumor suppressors that are common targets for silencing in other cancers and found only four to be globally hypermethylated across my cancer sub-types.

I found that the proportion of hypomethylated genes did not vary significantly between cancers. All of my samples were derived from primary tumors so perhaps this finding is indicative of a genetic state shared by all primary tumors. Previous studies have noted that primary tumors are more hypomethylated than healthy tissue and less hypomethylated than secondary tumors, lending credence to the idea that tumors in similar stages of tumorigenesis share molecular features despite different tissue origins (Gama-Sosa 1983).

Furthermore, I did not find significant differences in the methylation states of CGI and non-CGI CpGs. As it is yet unclear how non-CGI CpGs affect gene expression and how CGI CpGs affect chromatin structure, I reserve judgment on how this finding contributes to the cancer state.

Based on the differences in methylation profiles within cancers, I was able to derive putative cancer sub-types and generate a consensus profile for each sub-type. Clustering of these sub-types recapitulate previous findings of clinical associations

between BRCA and GBM and reveal novel associations between other cancers. There is substantial evidence linking primary breast cancer and primary brain tumors in the clinical setting. As early as 1983, researchers have noted a high incidence of primary meningiomas in patients who also have primary breast cancer tumors (Custer 2002, Wallack 1983). A 2005 study also documents incidence of primary glioblastomas in patients who were previously treated for breast cancer (Piccirilli 2005). Furthermore, up to 15% of breast cancer patients develop brain metastases (Millers 2003). The molecular nature of these associations is poorly understood but my finding demonstrates a potential similarity in methylation profiles between breast cancer and GBM.

Finally, upon investigation of 39 clinically relevant tumor suppressors, I found only four were consistently hypermethylated in our samples: *O*-6-methylguanine-DNA methyltransferase (MGMT), Semaphorin 3b (SEMA3B), Runt-related transcription factor 3 (RUNX3), and checkpoint with forkhead and ring finger domain (CHFR). MGMT notably is primarily silenced by promoter hypermethylation and is rarely found mutated in colon, lung, lymphoid, and other cancers. SEMA3B is an antiangiogenic factor and is commonly mutated in LUSC, LUAD, GBM, and BRCA cancers. Interestingly, researchers have found SEMA3B to be silenced by hypermethylation or chromosomal deletion (Kuroki 2003, Nair 2007). RUNX3 is a transcription factor that suppresses cell proliferation in a variety of cell types (Ito 2004). RUNX3 too is found in a region of the genome that is often deleted in many cancers (Lau 2006). RUNX3 is silenced by a combination of mutation and promoter hypermethylation in gastric and breast cancers (Jiang 2008). CHFR is involved in regulating entry into mitosis, possibly implicated in managing microtubule integrity prior to entry into mitosis (Scolnick 2000). CHFR has

been previously identified as hypermethylated in various cancer cell lines as well as in primary colorectal and head and neck cancers (Toyota 2003).

DNA methylation is a diverse epigenetic modification and modulates gene expression and chromatin structure differently based on where the methylation occurs. While gene silencing induced by promoter hypermethylation is a well-established link, it is yet unclear if and how methylation outside of promoter regions and outside of CGIs are linked to chromatin stability and gene expression. A major caveat of my study is that I only investigated the methylation status of CpGs in proximal promoter sites. Evidence in the field suggests that the global hypomethylation observed in cancers is localized to non-promoter regions, especially within gene bodies, in intergenic regions, and in heterochromatin. In these regions, methylation does not affect gene regulation so much as it influences chromatin structure and integrity. Another shortcoming of my study and similar studies that exclusively investigate methylomes is that gene expression cannot be strictly predicted by methylation status. Especially with CpGs in promoter regions not found in CGIs, it is not well understood how methylation affects gene expression. One must integrate gene expression data with methylomic data to conclusively make statements about methylation-modulated changes in gene expression.

Gene regulation by methylation is distinct from regulation by mutations in more ways than simply mechanism. DNA methylation is a dynamic process that can be altered during the life cycle of a cell. Also, unlike genetic mutations that are passed on to progeny intact, the methylation status of progeny can differ from their parent cells. In other words, DNA methylation modifications need not strictly accumulate with every round of replication. Genes that are widely inactivated by hypermethylation and not by

mutation lend support to the idea that gene silencing by these two mechanisms are different and are differentially selected for.

Cancer sub-typing provides valuable insight into how to effectively identify and treat different cancers. The cancer sub-types I identified revealed interesting and novel associations across different cancers. However, these associations must ultimately be substantiated with additional data to provide clinical relevance. For example, a TCGA study used multiple platforms (genomes, transcriptomes, methylomes) to identify four sub-types within BRCA. Similarly, it would be useful to utilize patient information (prognosis, tumor grade) to further define these sub-types. Integrating patient information with molecular information allows us to both further our understanding of how molecular changes manifest themselves in cancer phenotypes and provide us with predictive power as to how a particular tumor genetic program might progress. Cancer sub-typing using small nucleotide polymorphisms (SNPs) is a prolific field of study and is harnessed primarily for its potential to predict patient outcome and clinical progression (O'Brien 2014). DNA methylation landscape characterization may provide us with similar prognostic capabilities.

In this project, I have explored and established the value of studying methylation profiles across cancers. I found that genome-wide hypomethylation occurred in every cancer type I studied. Furthermore, the amount of hypomethylation was similar across cancers and did not differ between CGI and non-CGI sites. Using methylation statuses alone, I was able to establish cancer sub-types that revealed novel associations between different cancers. Furthermore, contrary to previous studies, I found many commonly suppressed tumor suppressors to not be hypermethylated across my samples. Combined

with other molecular and clinical metrics, the study of methylation landscapes promises to reveal novel aspects of tumorigenesis.

References

Antequera, F. & Bird, A. 1993. Number of CpG islands and genes in human and mouse. Proc. Natl. Acad. Sci. USA 90, 11995–11999.

Bestor TH. 2000. The DNA methyltransferases of mammals. Hum Mol Genet 9, 2395 2402.

Bird AP. 1986. CpG-rich islands and the function of DNA methylation. Nature 321, 209–213.

Bird A. 2002. DNA methylation patterns and epigenetic memory. Genes Dev 16, 6-21.

Costello JF. et al. 2000. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. Nature Genet.24, 132–138.

Custer BS, Koepsell TD, Mueller BA. 2002. The association between breast carcinoma and meningioma in women. Cancer. 94:1626 –35.

Dodge, JE. et al. 2005. Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization. J. Biol. Chem. 280: 17986–17991.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J et al. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet 38: 1378-1385.

Esteller M, Hamilton SR, Burger PC, Baylin SB, Herman JG. 1999. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. Cancer Res.59, 793–797.

Esteller, M. et al. 2000. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is associated with G to A mutations in K-ras in colorectal tumorigenesis. Cancer Res. 60, 2368–2371.

Esteller M. et al. 2001. Promoter hypermethylation of the DNA repair gene O(6) methylguanine-DNA methyltransferase is associated with the presence of G:C to A:T transition mutations in p53 in human colorectal tumorigenesis. CancerRes.61, 4689–4692.

Feinberg AP, Vogelstein B. 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature. 1983 Jan 6;301(5895):89-92.

Gama-Sosa MA, Slagel VA, Trewyn RW, Oxenhandler R, Kuo KC, Gehrke CW, Ehrlich M. 1983. The 5-methylcytosine content of DNA from human tumors. Nucleic Acids Res. 11(19): 6883–6894.

Gitan RS, Shi H, Chen CM, Yan PS, Huang TH. 2002. Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. Genome Res.12, 158–164.

Gonzalgo ML et al. 1997 Identification and characterization of differentially methylated regions of genomic DNA by methylation-sensitive arbitrarily primed PCR. Cancer Res.57, 594–599.

Hanahan D, Weinberg RA 2011. Hallmarks of cancer: the next generation. Cell 144, 646-674.

Huang TH, Perry MR, Laux DE. 1999. Methylation profiling of CpG islands in human breast cancer cells. Hum. Mol.Genet.8, 459–470.

Ito Y. 2004. Oncogenic potential of the RUNX gene family: 'overview'. Oncogene 23: 4198–4208.

Jiang Y, Tong D, Lou G, Zhang Y, Geng J. 2008. Expression of RUNX3 gene, methylation status and clinicopathological significance in breast cancer and breast cancer cell lines. Pathobiology 75: 244–251.

Kuroki T, Trapasso F, Yendamuri S et al. 2003. Allelic loss on chromosome 3p21.3 and promoter hypermethylation of semaphorin 3B in non-small cell lung cancer. Cancer Res. 63 (12): 3352–5.

Lau QC, Raja E, Salto-Tellez M, Liu Q, Ito K, Inoue M *et al*. 2006. RUNX3 is frequently inactivated by dual mechanisms of protein mislocalization and promoter hypermethylation in breast cancer. Cancer Res 66: 6512–6520.

Li E, Bestor TH, Jaenisch R. 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 69(6):915-26.

Miller K, Weathers T, Hanley L et al. 2003. Occult central nervous system involvement in patients with metastatic breast cancer: prevalence, predictive factors and impact on overall survival. Ann Oncol 14: 1072–1077.

Mohandas T, Sparkes RS, Shapiro LJ. 1981. Reactivation of an inactive human X-chromosome: evidence for X-inactivation by DNA methylation. Science 211, 393–396.

Nair PN, McArdle L, Cornell J et al. 2007. High-resolution analysis of 3p deletion in neuroblastoma and differential methylation of the SEMA3B tumor suppressor gene. Cancer Genet. Cytogenet. 174 (2): 100–10.

O'Brien KM, Cole SR, Engel LS, Bensen JT, Poole C, Herring AH, Millikan RC. 2014. Breast cancer subtypes and previously established genetic risk factors: a bayesian approach. Cancer Epidemiol Biomarkers Prev. 23(1):84-97.

Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 99(3):247-57.

Piccirilli M, Salvati M, Bistazonni S, et al. 2005. Glioblastoma multi-forme and breast cancer: report on 11 cases and clinico-pathologic remarks. Tumori. 91:256 – 60.

Scolnick DM, Halazonetis TD 2000. CHFR defines a mitotic stress checkpoint that delays entry into metaphase. Nature 406:430–435.

Suzuki MM, Bird A, 2008. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9, 465-476.

Takai, D, Jones PA. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc. Natl Acad. Sci. USA 99, 3740–3745.

Toyota, M. et al. 1999 Identification of differentially methylated sequences in colorectal cancer by methylated CpG island amplification. Cancer Res.59, 2307–2312.

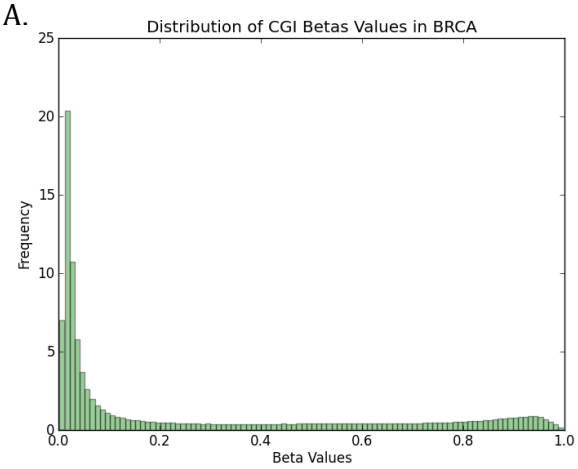
Toyota M, Sasaki Y, Satoh A, Ogi K, Kikuchi T, Suzuki H, Mita H, Tanaka N, Itoh F, Issa JP, Jair KW, Schuebel KE, Imai K, Tokino T. 2003. Epigenetic inactivation of CHFR in human tumors. Proc Natl Acad Sci USA 100:7818–7823

Wallack MK, Wolf JA Jr, Bedwinek J, et al. 1983. Gestational carcinoma of the female breast. Curr Probl Cancer. 7:1– 58.

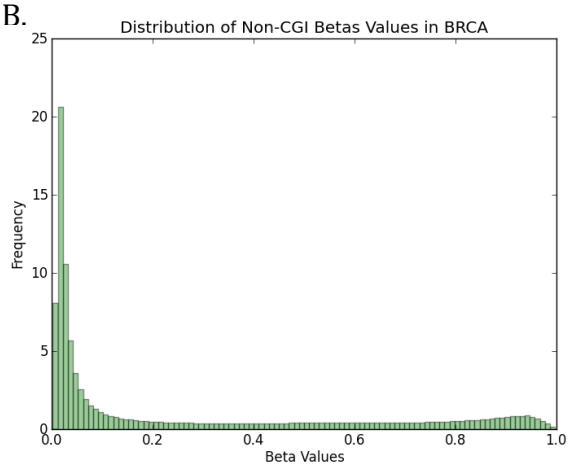
Weber M. et al. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nature Genet.37, 853–862.

Figures

A.



B.



C.

	% CGI Probe Status			% Non-CGI Probe Status		
Cancer	**Hyper-Meth**	**Neutral**	**Hypo-Meth**	**Hyper-Meth**	**Neutral**	**Hypo-Meth**
BRCA	0.22	0.16	0.62	0.21	0.16	0.63
COAD	0.16	0.22	0.62	0.16	0.21	0.63
GBM	0.13	0.23	0.64	0.13	0.22	0.65
KIRC	0.14	0.21	0.65	0.14	0.22	0.64
LUAD	0.17	0.2	0.63	0.17	0.21	0.62
LUSC	0.16	0.19	0.65	0.16	0.18	0.66
OV	0.11	0.21	0.68	0.11	0.2	0.69
UCEC	0.14	0.21	0.65	0.14	0.23	0.63

Fig. 1 Global hypomethylation observed across cancers
A. Representative histogram of the distribution of beta values observed in probes mapped to CGIs. **B.** Representative histogram of the distribution of beta values observed in probes not mapped to CGIs. **C.** The number of probes that were hypermethylated and hypomethylated across cancers by CGI and non-CGI status were significantly variable on either dimension.

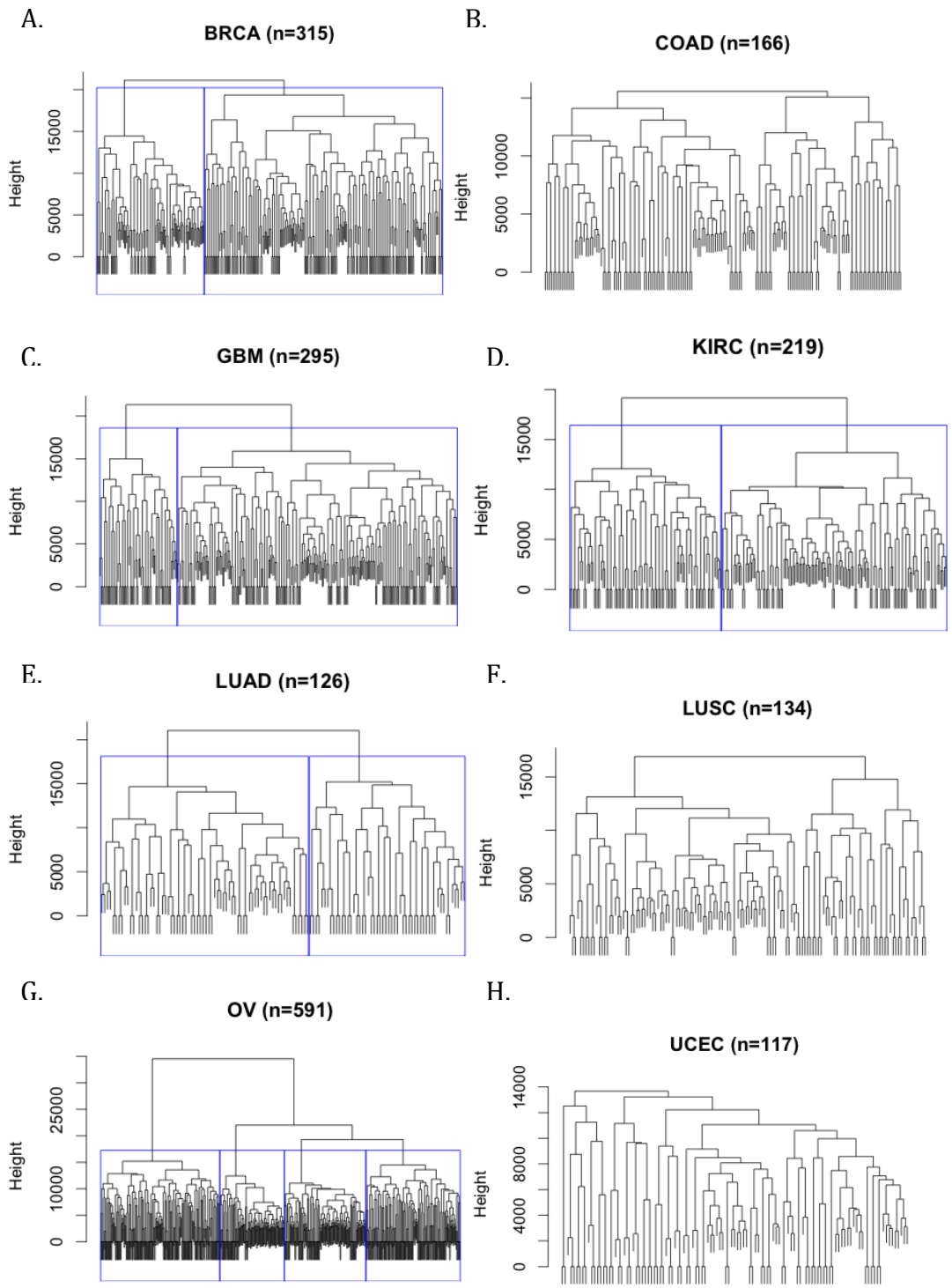


Fig. 2 Identification of cancer sub-types based on clustering height
A-H. Dendrograms of hierarchical clustering using the hamming distance and Ward's criterion show substantial variation in some cancer types.
A,C,D,E,G. Based on the height of the clusters (amount of dissimilarity), these cancers were further divided into sub-types as indicated by the blue boxes.

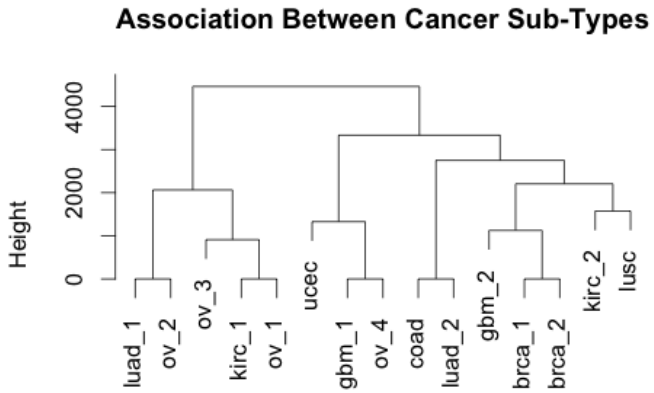


Fig. 3 Clustering of cancer sub-types reveals associations across cancers
Hierarchical clustering using the hamming distance and Ward's criterion shows that some of the consensus sub-types cluster more closely with sub-types from different tissues.

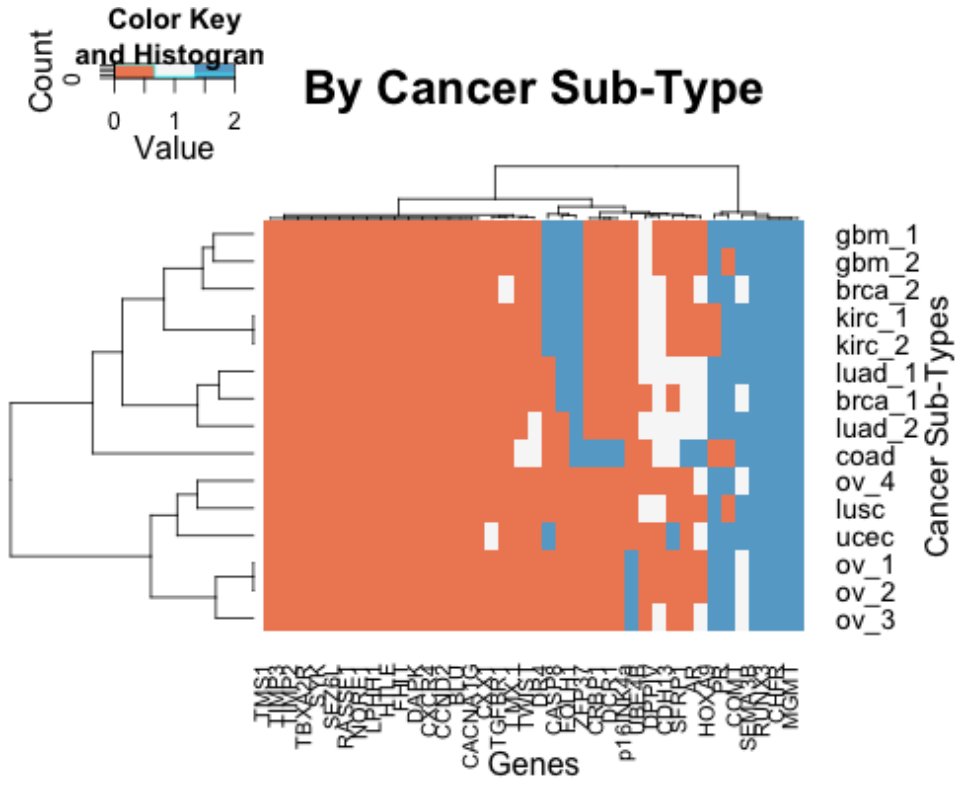


Fig. 4 Analysis of 39 Tumor Suppressor Genes for Putative Hypermethylation
Of the 39 genes interrogated, four genes are hypermethylated across all cancer subtypes: Sema3b, RUNX3, CHFR, and MGMT.