UNIVERSITY OF CALIFORNIA, MERCED

Genomic Interrogation of Melanoma for Identification of Driver Oncogenic Factors

by

Rohit Gupta

A dissertation submitted in partial fulfillment for the degree of Doctor of Philosophy

in the
Program for Quantitative & Systems Biology
School of Natural Sciences

August 2019

Committee in charge:
Professor Miriam Barlow, Advisor
Professor Michael Colvin, Chair
Professor Suzanne Sindi
Professor Mark Sistrom

The Dissertation of Rohit Gupta is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Professor Miriam Barlow, Advisor

Professor Michael Colvin, Chair

Professor Suzanne Sindi

Professor Mark Sistrom

University of California, Merced

2019

"To reach a port we must set sail –
Sail, not tie at anchor
Sail, not drift."

Franklin D. Roosevelt

UNIVERSITY OF CALIFORNIA, MERCED

Abstract

Program for Quantitative & Systems Biology
School of Natural Sciences

Doctor of Philosophy

by Rohit Gupta

Underpinning genomic principles are defining feature in every physiologic and pathologic process. Through advances in metabolomics, systems biologists can now track the dynamic interactions of the metabolome with the epigenome, genome, transcriptome and proteome. Understanding of cross talk between genomic, epigenomic and structural changes at biophysical scale on cellular metabolism is still in its infancy. Using Next-Gen Sequencing data in tandem with biophysical approaches, metabolism was found to play an important role in cellular proliferation, differentiation, metastasis. Mutation, methylation and gene expression level changes at genomic and epigenomic scale orchestrate pathway perturbations crucial for oncogenesis and metastasis. These results show understanding genomic and epigenomic machinery can provide important insights into cellular processes vital for growth and development in tumor cells. Increased attention to how genomic processes support transformational changes necessary for a cancer cell will allow for more precise engineering of biological function and the identification of targeted therapies.. . .

Acknowledgements

This work would not have been possible without the support and help of many throughout my education. I am grateful to my advisor Dr. Fabian Filipp whose guidance has been instrumental. I am also sincerely thankful to my undergraduate advisor Dr. Armindo Salvador for his guidance and support when I needed it the most. You have taught me to strive for excellence in all that I do.

I would like to thank each member of my committee for their continuing support. Dr. Michael Colvin, Dr. Suzanne Sindi and Dr. Mark Sistrom: each of you has served as a positive role model. I extend particular thanks to Dr. Colvin for our timely scientific discussions and working alongside on a long-standing collaborative project.

To my fellow graduate students in the Filipp Lab, your advice, encouragement and assistance have helped me more than I can faithfully express. And a special thanks to my colleague Simar for being a constant source of inspiration and cogent scientific discussions. I have been fortunate to work with an amazing group of undergraduate researchers and greatly appreciate the effort, energy, and enthusiasm they put in.

Finally to my family, your unfaltering support, love and patience have made this work and all else possible. Maa, Papa, Di, Jiju, Avik and Karishma, I am forever grateful to have you in my life. You are my inspiration and I dedicate this milestone to you. In addition, I would also like to thank my friends from VIT who were always there for long talks, much needed excursions, love and support. And finally, Michael Scott (The Office), for being a constant source of joy. I could not have done this without you.

Contents

To Karr...

Chapter 1

Introduction

1.1 Overview

Cancer is a complex genetic disease spanning several molecular events. However, over the last few decades, researchers have begun to recognize the multifaceted complexity which masks the genetic basis of cancer. Advent of high-performance computing and high-precision DNA sequencing technologies confers researchers the capability to obtain an exact readout of a tumor cell environment and factors governing cellular metabolism thereby paving way for an unprecedented expansion of our understanding about what makes cancer cells unique.The advent of genomic technologies have paved way to study regulatory switches, oncogenic mutations, pathway perturbations, chromatin accessibility and gene regulation. What follow in subsequent chapters of this dissertation are a reflection of my published, under-preparation and submitted work applying cancer systems biology strategies and implementing methods to investigate the contribution of genomic processes to diverse tumor factors. Aspects of cancer biology covered in this dissertation include mutation burden[1], hyper mutation driven metastasis[2], epigenomic regulation of tumor suppressors and transcription factors[3], validating structural changes and modeling pathways in cancer metabolism [1,2].

References

1. Guan J, **Gupta R** Filipp F V. Cancer systems biology of TCGA SKCM: efficient detection of genomic drivers in melanoma. Sci Rep. 2015 Jan 20;5:7857. doi: 10.1038/srep07857.

2. Edwards L, **Gupta R**, Filipp FV. Hypermutation of DPYD deregulates pyrimidine metabolism and promotes malignant progression. Mol Cancer Res. 2016 Feb;14(2):196-206. doi: 10.1158/1541-7786.MCR-15-0403.

3. **Gupta R**, Filipp FV. Overexpression of DNMT3a 3b in TCGA SKCM induces targeted regulation led by selective methylation. *To be submitted*

1.2 Background

1.2.1 Overview of Cancer Systems Biology

The genetic composition of a cell is subject to changes from multiple extrinsic and intrinsic factors. Events such as DNA mutations, structural variations, and regulation of epigenetic and transcriptional signatures often manifest into series of transformational changes in a cell that may result in cancer. Even if multiple tumors share similar recurrent genomic events, their relationships with the observed phenotype may vary, and in most cases remains to be studied. Cancer encompasses more than 100 distinct diseases, each with a unique epidemiology and diverse set of risk factors which originate from different cell types and organs of the human body. Furthermore, even within cancers of similar etiology and tissue of origin, heterogeneity at the molecular level and with genetic and genomic factors is prevalent. Generally, tumors arise due to the accumulation of mutations that aid in promoting a cancer phenotype in genes of critical importance with the capability to alter the cell machinery of proliferation, differentiation and death. One approach to the widely expanding understanding of tumors is through the use of Next-Generation Sequencing (NGS) data obtained from cancer cells and tissues from patients.

Genomic techniques are widely regarded as one of the most reliable ways to gain insight into the biology of tumors, with experimental data aimed at integrating multiple molecular events. Cancer Systems Biology refers to the use of relevant genomic and modeling methods for analysis, interpretation and visualization of datasets obtained from cancer cells and cancer tissue sequencing. In cancer, NGS produces a wide array of genomic datasets spanning from gene expression and mutation to epigenetic regulation through methylation. Generally, the sequencing data obtained from NGS spans several molecular events, providing a snapshot of a tumor cell at a specific instance in its growth and development cycle. For melanoma patients, the applicability of genomic data has already produced tangible revolutionary advances. The discovery of cancer-causing genetic and epigenetic factors in tumors has enabled the development of therapies that target such factors and diagnostic tests that aid identification of patients that might benefit from such therapies. For example, genomic analysis of mutation data in Melanoma patients led to the identification of an activating point mutation in *BRAF* kinase (B-Raf proto-oncogene, serine/threonine kinase) that not only established revolutionary treatment options, but also an era of personalized medicine in Melanoma treatment. Personalized therapies with kinase inhibitors of mutated BRAF are often considered for the first line of treatment in Melanoma patients.

In recent years, the use of NGS data, including gene expression, analysis of mutation burden, and a greater understanding of immunoregulatory mechanisms has led to the development of a novel class of treatment options collectively termed immunotherapy. Immunotherapies block immune checkpoints that are otherwise targeted by tumor cells to protect themselves from immune system attacks. Genomic analysis is crucial for identification of the patient subset that responds to immune therapy. Moreover, similar to most therapies and partly attributed to the robustness of a tumor cell, resistance mechanisms quickly evolve, and genomic analysis provides a viable option to understand the underpinning mechanisms that foster resistance to treatment therapies and provide an incredibly detailed snapshot of a tumor cell at a given time.

Most common genomic events studied in cancer include mutations since tumors often arise due to aggregation of mutations and copy number alterations (CNA),

which are DNA mutations on a much larger scale and can vary in size from 1 basepair to an entire chromosome arm. Somatic Copy Number Alterations (SCNAs) serve critical roles in activating oncogenes [1–3]. Regulation of gene expression is often exploited by tumor cells to constitutively activate and repress genes of significance [4, 5]. Quantification of gene expression therefore allows for an understanding of genes and cellular pathways used by tumor cells to remap normal cell machinery in favor of cancer metabolism. Epigenetic mechanisms are required by cells and tissues for maintenance of tissue specific gene expression patterns that serve vital roles in the development, growth and maintenance of a cell [6]. Tumor cells routinely disrupt the epigenetic landscape to alter gene functions [7] and expression [8]. Changes in the methylome are very commonly observed in tumor cells and alterations in gene-wise methylation levels are widely interpreted as one of the prominent mechanisms used to promote the tumor phenotype.

Despite an abundance of rich genomic datasets, comprehensive investigation to discern and analyze interwoven dynamic systems involved in tumor growth, development and metastasis still needs to be identified. Furthermore, most genomic datasets provide a snapshot, while tumor systems are highly complex dynamic systems [9, 10]. Single cell sequencing is an emerging modality that examines sequence information in a cell by optimized NGS technologies. Data produced with single cell sequencing is of higher resolution and helps explore the tumor micro-environment, as clonal differences and heterogeneity is one of the many approaches tumor cells employ to sustain growth and evade immune response [11]. Methodological advances have now enabled clinical applications of large-scale transcriptome profiling and translational therapies [12]. The decreasing cost of NGS coupled with advancements and ease of high performance computing (HPC) has made it possible to analyze and obtain more genomic data than ever before, providing insights into novel mechanisms and therapeutic targets. Cancer Systems Biology facilitates a way to gain insight into the biology of tumors, allowing targeted therapeutics in cancer by providing a host of methods that are used in this dissertation.

1.2.2 Genomic factors in Cancer

Sources of genetic information include biological samples of DNA, information derived from a person's family history of disease, findings from physical examinations, and medical records. DNA-based information captured in cancer patients through NGS technologies can be gathered, stored, and analyzed at any time during an individuals life span, from before conception to after death. All cancers arise as a result of changes that have occurred in the DNA sequence of a cancer cell. In general, cancer cells have more genetic changes than normal cells, attributed mostly due to accumulation of mutations and genetic alterations that promote tumors. But each cancer has a unique combination of genetic alterations that varies from person to person and cancer tissue to cancer tissue. Some of these changes may be the result of cancer, rather than the cause. Even within the same cancer, cancer cells may have different genetic changes attributed to the widespread heterogeneity that occurs in cancer as intra-tumor heterogeneity (ITH) is prevalent in cancers.

Molecular and metabolic profiling includes the evaluation of genomic, proteomic, metabolomic, and epigenomic expression factors, alone or in combination, for cancer diagnosis, prognosis, and therapeutics [13]. A wide array of genomic changes contribute towards genomic and metabolic reprogramming in cancer. The effects and scope of genomic events varies and is often facilitated through multiple factors and signaling pathways. Common genomic variations associated with tumors often manifest through changes to the DNA sequence which alter a genes ability to be transcribed and often profoundly impact its expression; such changes include structural variations to the DNA, mutations, and copy number alterations. However, factors other than DNA sequence variations may also influence gene functioning; most notably, epigenomic regulation has been widely established as a prominent mechanism used by tumor cells and is now considered an emerging hallmark of cancer. Instead of inducing DNA variation, epigenomic regulation includes slightly changing chemical composition of the DNA [14].

Through targeted regulation of the methylome, tumor cells can selectively limit or enhance chromatic accessibility of gene which affects transcription, thereby

altering expression. Much of these data-sets are housed across multiple databases. For example, Catalogue of Somatic Mutations in Cancer (COSMIC) [15] contains a curated list of cancer-associated somatic mutations, whereas NIH Roadmap Epigenomics Mapping Consortium contains epigenomic data associated with normal tissues. Compiling these types of databases across several cancer tissues from multiple patients is the primary objective of consortia such as The Cancer Genome Atlas (TCGA).

The following sections highlight how cells engage the diverse genomic mechanisms listed above to support distinct cellular processes and the methods employed by systems biologists to understand the confluence of genomic factors and metabolism in tumor cells.

1.2.2.1 Mutations and Somatic Variants

Somatic mutation in a cancer cell may encompass several classes of DNA sequence changes, but all cancers carry somatic mutations at varying rates. Mutations are single basepair changes or large-scale changes in the DNA caused by a myriad of intrinsic or extrinsic factors. Changes in DNA sequence, either in germline or somatic mutations, are not limited to substitutions of one base for another, but can also undergo insertion or deletion of small or large segments of DNA, rearrangements, or copy number increase (amplification) or decrease (deletion) of the normal diploid genome. Progressive accumulation of mutations throughout life during cell division or triggered due to an extrinsic event such as prolonged UV exposure can lead to cancer.

Somatic mutations in a cancer genome are classified by their consequence for cancer development. Genomic analysis of somatic mutations in cancer to study alteration in gene expression and pathway dysregulation may serve to identify driver events; therefore, genes harboring aforementioned events may be identified as driver oncogenes. Somatic mutations that confer upon tumor cells a preferential advantage in growth or survival through significantly rewiring cellular machinery to promote disease progression are driver oncogenic mutations. Passenger mutations, on the other hand, provide no distinct phenotypic consequence or selective

growth advantage to the cell harboring them. Passenger mutations are still found in cancer genomes, as they are acquired during cell division. Generally the location of the mutation (intron vs exon), in addition to the kind of substitution (transition vs transversion), as well as its scope (activating vs inactivating), all contribute towards positive selection in a tumor.

Single nucleotide polymorphisms (SNP) are genetic variations resulting from a single nucleotide base (A, C, T, or G) change in the DNA sequence. SNPs account for approximately 90% of the genetic variation in humans. It is estimated that the human genome has more than 30 million SNPs [16, 17]. Subsequent studies have contributed to understanding the role of specific SNPs in genetic predisposition to cancer, heart disease, diabetes, and other chronic diseases. In addition, SNP profiles are useful for identifying cancer genes, risk, prognosis, and comorbidities, as well as drug responses and interactions by using techniques like genome-wide-association studies (GWAS)[18]. Inherited genetic mutations play a significant role in 5 to 10% of cancers, but by and large, somatic structural variations or somatic mutations account for most oncogenic mutations. While SNPs are commonly classified as nucleotide substitutions associated with germ-line cells, mutations of somatic cells are referred as SNVs.

The rate of point mutation varies along the genome, and is typically higher in regions of higher expression levels, repressed chromatin, and late replication times[19, 20]. The rate of mutation varies greatly among tumors and cancer types. A significant increase in the rate of somatic mutations in tumors induced by exogenous mutagens such as UV radiation for Melanoma and tobacco for Lung Cancer is commonly observed. The somatic mutation rate in Skin Cutaneous Melanoma (SKCM) can be up to 17 mutations/mB, followed by Lung Squamous Cell Carcinoma with about 10 mutations/mB [21]. The cause of high mutation rates also varies, and in cancers where loss of repair pathways or integrity check-points are affected, a dramatically high number of mutations can be observed. Thus, both intrinsic and extrinsic events are potential causal factors for high number of somatic mutations. Despite the high number of somatic mutations, only a small fraction show positive selection in cancer, while most mutations are of neutral or mildly deleterious effect to the genome at large. These passenger

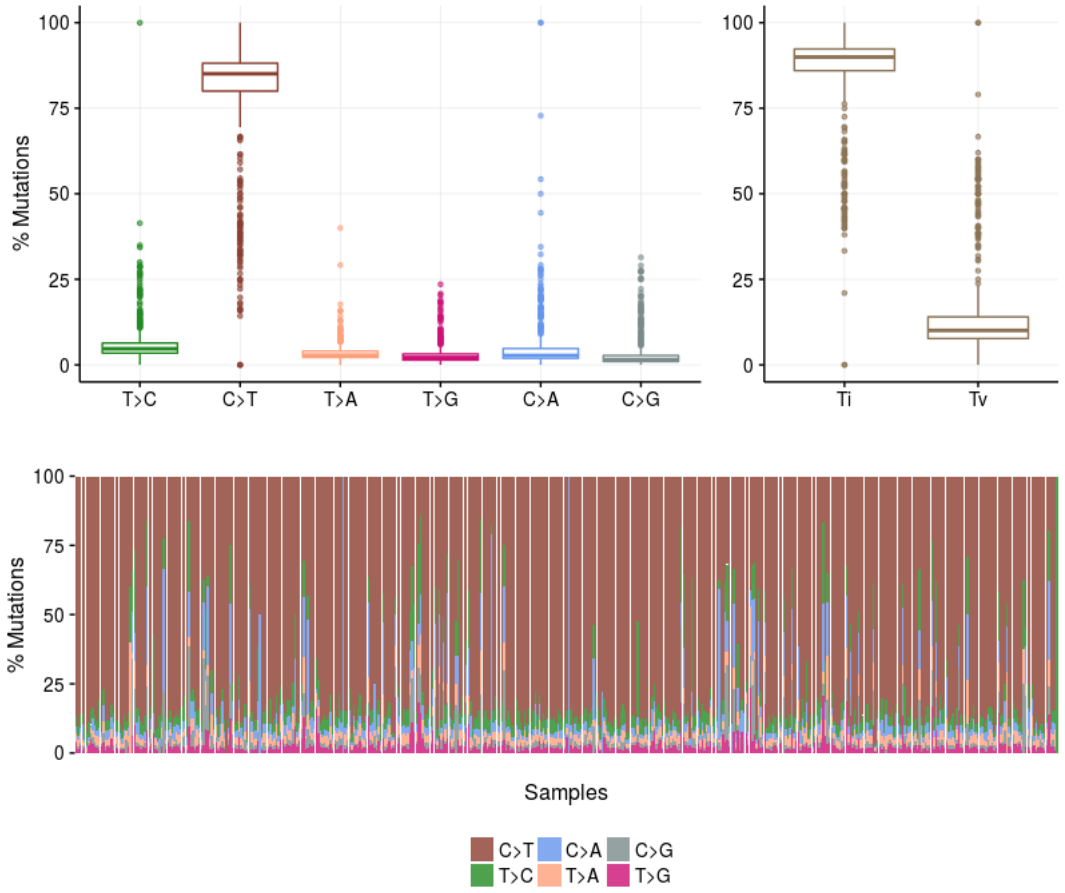


FIGURE 1.1: Distribution of mutation in TCGA SKCM dataset (n=471). The mutation signature of Melanoma is unique and domintated by UV damage induced mutations.

mutation events are found in high numbers due to rapid clonal expansion caused by driver mutations.

The prevailing pattern of mutations is termed the 'mutational signature'. Mutational signatures can be indicative of the processes that promote rewiring of cell metabolism. These signatures are much more profound in Melanoma, as oncogenesis largely occurs due to UV-induced mutation damage. UV light induces pyrimidine dimers, whose erroneous repair leads to C>T mutations at CpC or TpC nucleotides. The majority of mutations in Melanoma can be attributed to UV-induced damage signature. Based on genomic signatures, it is established that UV damage drives genomic instability, mutagenesis, and carcinogenesis. UV damage in particular fuels the malignant transformation of melanocytes into melanoma.

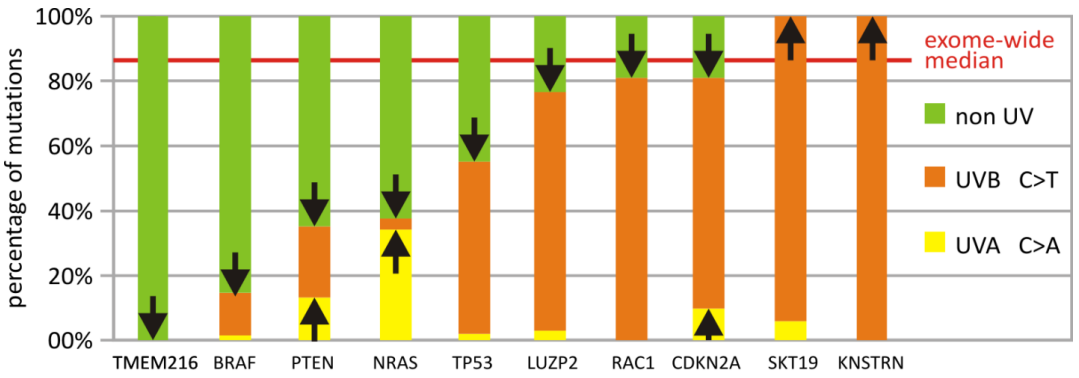


FIGURE 1.2: Signature of UV mutagenesis across driver mutations in melanoma. UV-induced DNA damage in conjunction with DNA mismatch repair are a common cause of somatic mutations in Melanoma, raising the frequency of UVA and UVB associated mutations about four-fold to 83% in comparison to other TCGA cancers [21].

While C>T mutations dominate the landscape of mutations in Melanoma, there exist mutations due to enzymatic damage. BRAF mutations are quite frequent in Melanomas where tumors arise on skin without chronic sun-induced damage [22]. In fact, approximately 50% of Melanomas harbor activating BRAF mutations. BRAF mutations in Melanoma also display a distinct signature with over 90% being at codon 600, and most of these are single nucleotide substitutions from a transversion of T to A at nucleotide 1799 (T1799A) which results in a substitution of valine (V) for glutamic acid (E) (BRAF V600E). The second most commonly mutated gene in Melanomas is NRAS observed in approximately 30% of cases. Mutation signatures are a blunt instrument because most mutagens create mutations similar to those from other mutagens.

Mutations that lead to inhibition/inactivation of a pathway are loss-of-function mutations (inactivating mutations). Loss-of-function mutations induce disruption in the promoter region of a tumor suppressor, which results in a loss of or severely reduced expression. More commonly, loss-of-function mutations can also lead to a functionally inactive gene-encoded protein which results in deletion or inactivation of their functions. Mutations in proto-oncogenes and pathways that transform them into hyperactive forms by a gene amplification event which translates to more copies for an oncogene and increased expression levels are gain-of-function mutations (activating mutations).

Mutations in BRAF are critical to the malignant process in Melanoma and lead to constitutive activation of mitogen-activating protein kinase (MAPK), an important signal transduction pathway involved in cell growth, proliferation and survival[23]. Mutations that lead to enrichment of a pathway (e.g. MAPK in Melanoma), are classified as activating mutations. An enrichment of about 1.34% is observed in the MAPK pathway in BRAF mutated cases[24]. Mutations in *Bad* gene are known to drive inactivation of pro-apoptotic pathways in colon cancer[25]. Similar examples of activating and inactivating mutations are found across most cancer tissues. APOBEC is a widespread mutational signature across several cancer tissues[26]. APOBEC mutation signature is characterized by C>T or C>G substitutions at sites preceded by thymine nucleobase and is caused by off-target modification of DNA by the APOBEC family of proteins. By studying the pattern of mutations and mutational signatures in cancer it is now possible to understand the mechanism of action for many novel mutational processes in cancer.

FIGURE 1.3: Assessing function impact of a variant. Driver mutations would have higher rate of functional manifestation and are more likely to occur in an exon while a passenger mutation would occur in an intron.

Accessing the functional impact of mutation is important to define the scope of a mutation and to infer the degree of positive selection exhibited in cancer. Of the several thousands of mutations typically found in Melanoma only a small subset are positively selected; this subset of mutations represents driver mutations, while passenger mutations do not exhibit strong evidence of positive selection and are passed on due to clonal expansion brought by driver mutations[1]. Typically, driver

mutations in tumors manifest with a higher rate of mutation (several magnitudes higher than background rate) in a gene or region of the genome that is otherwise expected to have neutral/functionally insignificant mutation accumulation. In Melanoma, deviation from exome-wide median of the signature of nucleotide replacement can be indicative for positive selection of cancer genes. As data-sets and exhaustive somatic mutation catalogues across multiple tumor tissues have grown, it has become increasingly important to develop pipelines and models that accurately and efficiently detect background mutation rate and rank mutations with high probability of positive selection. A workflow for discovery of enriched driver mutation in Melanoma oncogenes will be covered in Chapter2 &3 of this thesis.

The mutator hypothesis serves to explain the temporal evolution of cancer and the processes that are undertaken by a cell starting with acquisition of hypermutation and/or early driver mutations that trigger clonal expansion[27, 28]. Mutations in one or more pathways of key significance such as chromosome segregation, checkpoint control and cellular responses (e.g. apoptosis) can all lead to mutator phenotype with diverse manifestations ranging from point mutations and copy number alterations to microsatellite instability. Large-scale studies across 41 cancer types with sequencing data from over 5000 patients have shown a distinct relationship between mutation frequency and cancer risk[29]. This strong correlation implies that variation in cancer risk across tissues can be mainly attributed to distinct mutation accumulation rates. It is therefore substantially more important to a) efficiently identify all somatic mutation above background mutations, and b) rank mutation hotspots and regions of hypermutation and narrow down driver mutations only.

Typically, driver mutations that undergo clonal expansion must occur in stem or proliferating cells. In the case of Melanoma, these are the stem cells in the epithelia of the skin. Driver mutations that are harbored in epithelial skin cells are then selected for clonal expansion, resulting in malignant disease[30, 31]. Selection of the aforementioned driver mutations acts in different ways in different tissues. The processes that aid in selection include increasing the relative rate of cell proliferation over differentiation and eluding quiescence, senescence, or cell death.

Most notably, eluding checkpoint and senescence has been at the forefront of the recent breakthrough therapy in cancer, immunotherapy.

As part of this thesis work, methods and a pipeline to efficiently detect mutation hotspots and accurately identify driver mutations in Melanoma are presented. A permutation-based model which takes into consideration the location of a mutation in relation to the genome to deduce the level of positive selection in cancer has been formulated and included in this thesis work. Methods and pipelines for analysis and identification of driver oncogenes in Melanoma were complemented with tools to access the pathways impacted and their manifestations as it relates to oncogenesis and disease progression.

1.2.2.2 Gene Expression Regulation& Transcriptional Activity

Patterns of gene expression underlie fundamental differences that define cell type and function and also govern the flow of countless cellular processes. Analyzing genes that are differently expressed in tumors and comparing with their normal counterparts to gain mechanistic insights into remodeling of cellular machinery in cancer cells has been focus of many studies. Gene expression analysis provides a static image of cell machinery at a given instance in time. Gene expression is mostly controlled at the level of transcription initiation and therefore transcription factors and epigenetic factors that influence transcription initiation also contribute in modulating gene expression. Gene expression profiling using high throughput techniques that allow for genome-wide analysis have opened new avenues in understanding of the tumor micro-environment and facilitate comprehensive, high-resolution studies that produce gene expression networks pertaining a disease or cell condition.

Gene regulation is a label for the cellular processes that are undertaken to control the manner and level of gene expression and the functional products of a gene. Only a small fraction of genes in a cell are expressed at a given time, and a distinct set of regulators through the process of gene expression regulation work to actively promote or suppress transcription. A series of interactions between RNA molecules, proteins, transcription factors and several other components play

an active role as part of a gene expression system to determine what gene, when, and where gets activated or expressed and the amount of RNA or protein product produced as a result. It must be noted that gene expression modulation is fairly commonplace and essential for the functioning of a normal cell. One of several mechanisms, such as regulating the rate of transcription or translation or regulating the stability of the mRNA molecule, can be employed for gene expression regulation. Gene expression modulation gives the cell the ability to control structure and function and is required in housekeeping tasks like differentiation and morphogenesis. Cancer cells posses the ability to alter the signals and underlying mechanisms in order to selectively target molecules that facilitate oncogenesis and promote disease progression and metastasis.

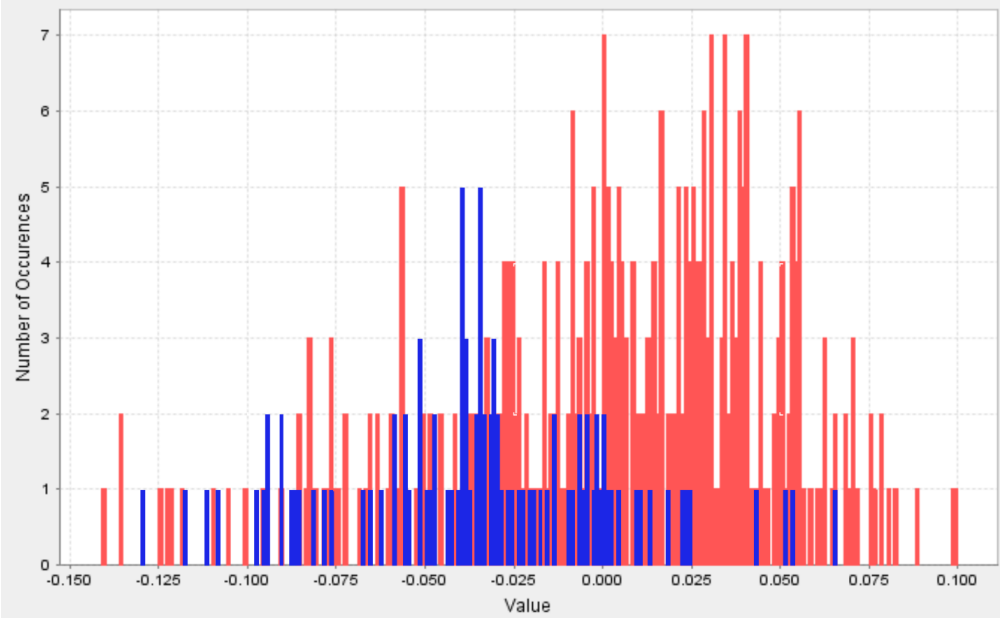


FIGURE 1.4: The gene-expression signature of DPYD in TCGA SKCM dataset shows a clear bifurcation. The blue bars depict normalized gene expression in primary Melanoma, while red bars show normalized expression in metastasized Melanoma. The normalization was performed in relation to gene-expression of metabolic genes in pyrimidine pathway [32]. DPYD signature and mutations are discussed in Chapter4.

Gene expression signatures are combined or single gene expression alterations that demonstrate validated correlation specificity in terms of diagnosis, prognosis or prediction of therapeutic response. Before the advent and wide adoption of NGS

technologies, gathering basic mechanistic, therapeutic, and functional insights of a cancer cell required long and generally expensive procedures such as clinical analysis, histological and immunohistochemical tests. The Serial Analysis of Gene Expression (SAGE) technique, based on the sequencing and quantification of mRNA in a sample as a direct measure of number of copies sequenced, was one of the first methods that enabled measurement of expression in all genes in a sample [33]. The microarray procedure, based on comparative hybridization of two cDNA strands, enabled relative quantification of the transcriptional program in a sample and gained prominence, facilitating fast and large-scale gene expression analysis[34]. The microarray technique, when used to study gene expression, was crucial in answering fundamental questions related to tumor biology, understanding disease progression, metastasis, and viable therapeutic targets.

Gene expression signatures have shown prognostic values and can serve a critical role for identification of therapeutic strategies. Enrichment Analysis (EA) reveals features of genes and pathways, whereas other techniques, such as co-expression networks and integrative clustering, reveal features of the patient subgroup. The aforementioned tools serve an exploratory analysis need when determining the cause and kind of cancer and also help narrow down the next steps in relation to identification of the most effective targeted therapy regime to prescribe based on a personalized medicine approach. Gene Set Enrichment Analysis (GSEA) serves to identify a set of genes that share common biological function, location and regulation[35]. GSEA is an analytical tool that, by interpreting gene expression data, yields indispensable insights, e.g. if a set of enriched genes are connected by a common theme, or if a pathway is over-represented. Co-expression and clustering analysis show functional relatedness and reveal functionally coherent sets of patient subclass that responds to a therapy[36]. In chapters 2,4 and 5 of this thesis, methods for performing clustering and enrichment analysis in Melanoma are outlined.

Expression profiling analysis aimed at mapping the changes a normal human melanocyte undergoes during the transformation to Melanoma have identified activation of NOTCH pathways, activation of cancer/testis antigens and down-regulation of immune modulation genes[37, 38]. Analyses aimed at studying the

molecular mechanisms that fuel the progression of Melanoma and its transformation into a malignant metastatic form with the use of expression data from DNA microarray techniques have identified enhanced expression of genes involved in extracellular assembly and genes that regulate the actin-based cytoskeleton[39]. These findings further corroborate the hypothesis that a specific subset of gene products can regulate metastasis without altering the growth properties of the tumor. Due to the heightened proliferation rate of metastasized tumor cells, re-purposing pathways to support increased demand of raw materials will be necessary without compromising growth[40].

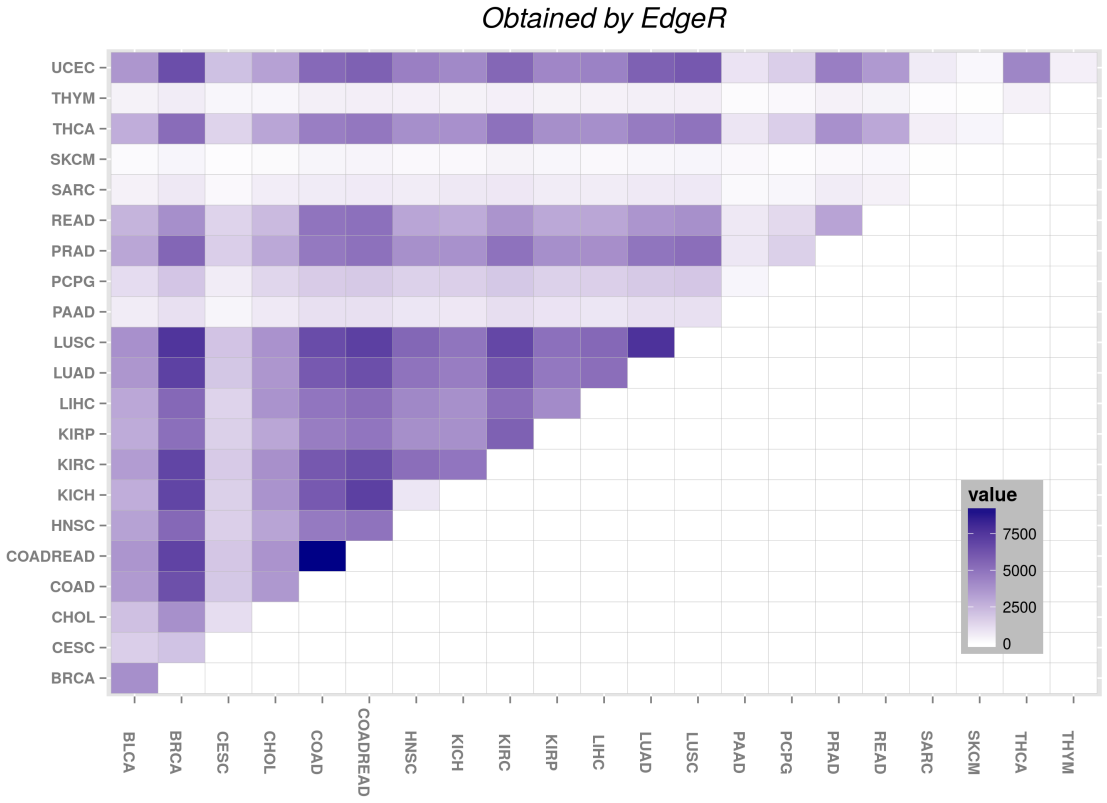


FIGURE 1.5: Gene expression profiling across 21 cancer tissues from the TCGA project. Mapping overlap of differential expression for all genes in TP, TM and NT (primary, metastatic, and normal matched tumor, respectively) cohort class (n>10000).

Subtypes based on gene expression profiling also share clinical outcomes. In a study of 57 stage IV Melanoma patients, the subtypes with shared features of low expression in immune signalling pathways had the lowest survival rate[41].

Different subtypes were also associated with distinct biological parameters such as pigmentation. In Metastatic Melanoma, expression-based subtyping through unsupervised hierarchical clustering produced two distinct groups that had distinct and statistically significant differences in survival and lesion thickness[42, 43]. Gene expression based signatures have also been validated to accurately differentiate benign nevi from Melanoma[44, 45]. One of the first prognostic expression signatures identified in cancer detailed a subset of genes in breast cancer [46, 47]. A distinct association was observed in the expression level of the gene-set with tumoral progression. More recently, predictive gene signatures are used in immunotherapy-based treatments to predict the response to therapy[48–50].

The systemic approach for identification of optimal drugs for a disease starts with computing a disease-related gene expression signature followed by comparative analysis between disease and control sample. The next step is then to identify drugs/chemical compounds that have a reverse relationship with the disease signature. This technique is widely used in the case of cancer as well, although the majority of drug-induced gene expression experiments have been conducted in cancer cell lines. This approach has led to the discovery of a number of drug candidates for various cancers; for example, in the case of lung cancer subtype Small Cell Lung Cancer (SCLC), the use of similar systemic approaches using gene expression signature profile led to the discovery of tricyclic antidepressant as a potent drug[51]. In metastatic colorectal cancer, similar analysis led to the discovery of three novel drugs, citalopram, troglitazone, and enilconazole, as viable antimetastatic compounds[52]. In the case of renal cell cancer, drug repositioning analysis has identified pentamidine as an effective anticancer agent[53]. In the case of Melanoma, drug repositioning analysis has identified HIV1-protease inhibitor nelfavir as a potent suppressor of PAX3 and MITF expression. Since acquired resistance to BRAF and MEK inhibitors is very prevalent in Melanoma [54, 55], nelfavir, which inhibits MITF, serves as an enhancer of BRAF inhibitor and can be clinically relevant in subset of Melanoma cases[56].

Major themes that have emerged from gene expression and gene signature analysis are: a) Metastatic potential is contained in the primary tumor and not acquired over time; and b) Proliferation, mitosis control, and chromosome integrity mech-

anisms play an outsized role in the development of aggressiveness as it pertains to the tumor. It must also be noted that gene expression is only a reflection of the phenotype associated with genetic alteration present in the tumor, and it is therefore vital to understand the basis of genetic alterations to design robust therapeutic strategies. Gene expression analysis based signatures, however, have a limited scope in cancer, since a surmounting body of evidence shows that the defining characteristic of most human cancers is intratumoral heterogeneity, and an expression-based signature typically accounts for macro trends and generally fails to resolve sub-clonal minutiae. A developing modality to curtail the shortcoming presented in bulk cell population analysis such as gene expression signature is single-cell sequencing analysis[57].

1.2.2.3 Somatic Copy Number Alterations

It is widely established that cancer is driven by acquisition of somatic genetic alterations that range from point substitutions to large-scale structural variations. Copy number alterations can have germline or somatic origins and are referred as copy-number alterations (CNAs) or somatic copy number alterations (SCNAs), respectively. Duplication or deletion events in the DNA that affect several base-pairs at the same time are classified as copy number alterations. SCNAs affect a larger fraction of the genome in cancers than do any other type of somatic genetic alteration and are extremely common in cancer [1, 58–61]. DNA copy number amplifications that activate oncogenes are hallmarks of nearly all advanced tumors, whereas deletion of tumor suppressors provides unencumbered proliferation potential to tumor cells[62]. Substantial evidence in support of direct and global changes in gene expression patterns as a result of DNA copy number alterations exists across several cancer tissues[63–66]. Greater understanding of the biological and phenotypic effects of SCNAs is of critical importance and has allowed for substantial advances in cancer diagnostics and therapeutics[3, 67].

Copy number amplification events cause an increase in the gene copy number of genes located in the amplified region. The phenotypic manifestation of this event is an elevated expression level. Similarly, as a result of copy number deletion

events in the DNA, a reduction or in some cases absence of expression is observed. In breast cancer, mapping amplified regions of the DNA led to the identification of crucial oncogenes such as EGFR[68, 69]. The number of overexpressed genes in the amplified regions does not directly translate to total genes in the region and varies by cancer tissue type. The most pronounced overexpression occurs in regions of high-level copy number increase[70]. In the case of chronic myeloid leukemia (CML), amplification of various drug targets is observed that confer drug resistance[71].

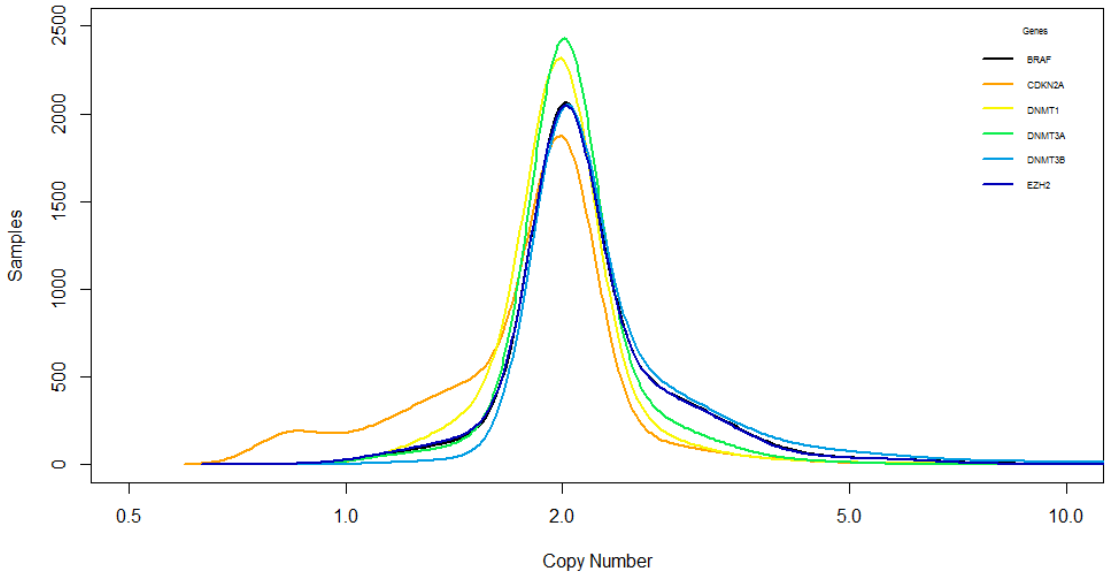


FIGURE 1.6: SCNAs may result in increase or decrease in copies of a gene which then affects gene-expression. TCGA SKCM data shows an amplification in number of copies of BRAF(oncogene) while in case CDKN2A, a tumor suppressor gene, deletion is prevalent.

Across the genome, SCNAs most commonly observed are focal (short) or span a chromosome arm (arm-level). Arm-level SCNAs are much more prevalent and observed at a rate about 30 times that of focal SCNAs[58]. The most common focal SCNAs observed in cancer are MYC amplification and CDKN2A/B deletions. c-MYC, located on chromosome 8 q24.21, is an oncogene constitutively expressed in almost all cancers and is essential for activation of cell proliferation machinery required by highly proliferative cells such as cancer. CDKN2/B act as tumor suppressors by regulating the cell cycle, disruption of which is of critical importance to a cancer cell. For arm-level SCNAs, the most commonly amplified regions

include regions of oncogenes like CDK4, EGFR, FGFR1 and KRAS, while deleted regions are most commonly shown to include genes like CDKN2A/B, PTPRD, RB1 and PTEN[58].

TABLE 1.1: The landscape of frequent somatic alterations in Melanoma. Frequencies depicted are from TCGA SKCM PanCan [12] cohort with 471 patient samples and TCGA SKCM [72] described in Chapter 3.

Gene type	Gene	Most frequent alteration	Frequency
Proto-oncogenes	*NRAS*	Mutation, amplification	20–30%
	BRAF	Mutation, amplification	60–80%
	KIT	Mutation	3–5%
	WNT11	Amplification	2–5%
	MITF	Amplification	5–10%
	c-MYC	Amplification	5–10%
	RAC1	Mutation	$\approx$ 6%
	CDK4	Mutation, amplification	3–5%
	AKT3	Amplification	$\approx$ 5%
Tumor suppressor genes	*CDKN2A*	Deletion, mutation	25–10%
	PTEN	Mutation, deletion	10–5%
	TP53	Mutation	$\approx$ 25%

Several high-throughput approaches have been applied to Melanomas to assess chromosomal aberrations and gene expression patterns in an unbiased fashion. For example, comparative genome hybridization (CGH) identified several chromosomal and genetic changes. MITF amplification events are highlights of SCNA analysis in Melanoma and are more prevalent in the metastatic form[73].Other prominent observations include losses of chromosome regions 6q, 8p, and 10 and gains in copy number of chromosome regions 1q, 6p, 7, and 8 in primary Melanomas. Frequent deletion of chromosome 13 and 17p in Melanomas arising in chronically sun-damaged skin is also commonly reported[74–79]. Structural variants and copy number alterations heavily influence the mutation frequencies of driver genes, especially in the case of NF1, TP53, PTEN and KIT. The BRAF hotspot mutation V600E is also observed to be amplified along with MET, which is located adjacent to BRAF on chromosome 7[24]. Copy number amplification of epidermal growth factor receptor (EGFR), which mediates cellular response to signal from growth factors is associated with poor prognosis in melanoma[80].

1.2.2.4 Methylation & Epigenomic Regulation

Selective modulation of genes and pathways is essential for normal functioning of a cell. Cell circuitry pertaining a cell state are routinely switched 'on' or 'off' through a myriad of processes and factors. These factors include epigenetic events occurring during cell development and proliferation that alter gene expression without changing the actual DNA sequence. Epigenomic mechanisms, with DNA methylation being the most prominently observed marker of gene modulation, are prevalent methods through which functional regulation of cell state occurs. The process of epigenomic-mediated silencing is actively repurposed in tumor cells to promote cell conditions that aid proliferation, progression and evasion of immune-mediated response. Strong evidence of extensive reprogramming of several aspects of the epigenome in cancer tissues, including Melanoma, have provided evidence that epigenetic mechanisms, such as selective methylation, modulate cell state to promote disease progression[81–84]. Silencing of tumor suppressors through targeted hypermethylation is an active machinery exploited by cancer; furthermore, aberration in DNA methylation is an epigenetic hallmark of cancer[85, 86]. Although silencing of some genes in cancer occurs by mutation, a large proportion of carcinogenic gene silencing is a result of altered DNA methylation[87].

An important type of epigenetic change in carcinogenesis is DNA methylation, a biochemical process by which a methyl (CH3) group attaches to cytosines, thereby turning off the gene so that it is no longer expressed. The most common site of DNA methylation causing silencing in cancer typically occurs at multiple CpG sites. Clusters of CpGs, referred as to as islands, are found in 5' regulatory and promoter regions of a protein-coding gene. DNA methylation in mammals is found sparsely but is globally distributed in defined CpG sequence throughout the entire genome. CpG islands are short interspersed DNA sequences that are enriched for GC. These CpG islands are normally found in sites of transcription initiation (transcription start sites; TSS), and selective methylation of these sites is known to effect gene regulation; to that end, methylation of CpG leads to gene silencing. DNA methyltransferases (DNMTs; 1, 3A, 3B) are responsible for catalyzing the transfer of a methyl group to mammalian genomic DNA and play an active role in gene silencing and repression. DNMTs induce tumor growth and aid tumor cells

by orchestrating both methylation-induced and methylation-independent changes in genes and transcription factor expression[88, 89]. Elimination of both DNMT1 and 3A nearly eliminates methyltransferase activity while disruption of DNMT3A only reduced the methyltransferase activity by 3%, indicating an enhanced role of DNMT3A in regulatory and targeted methylation[90].

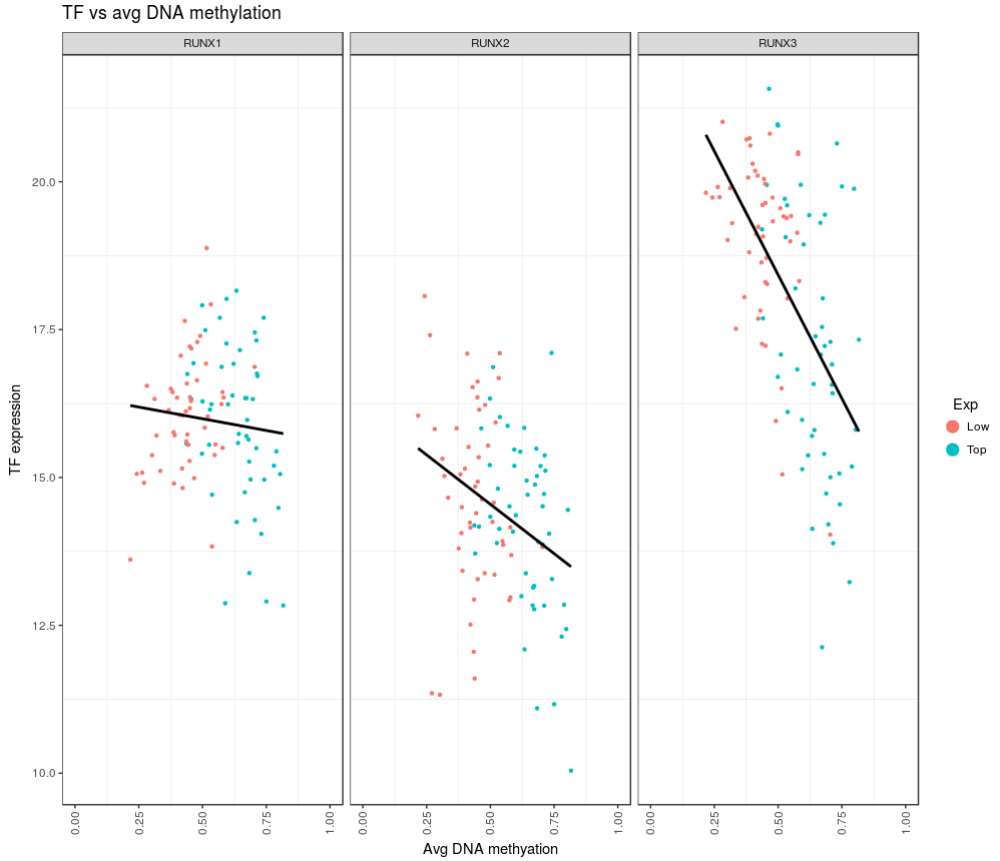


FIGURE 1.7: Methylation changes directly translate to changes in expression. A decrease in gene-expression as a result of hypermethylation in RUNX1, 2 and 3 genes is depicted in plot above. Associated methods and regulation by methylation are further explained in Chapter5 of this dissertation.

Melanoma exhibits a small degree of global hypomethylation within the bulk-genome context but focal (local) hypermethylation at the site of tumor suppressor genes and their promoter regions is responsible for most functional and phenotypic intricacies[91, 92]. Analysis of Melanoma cell lines has identified a large cohort of hypermethylated genes that are perceived to be repurposed for malignant disease progression[93]. However, the scope and causal mecha-

nisms that contribute towards pathogenesis as a result of hypermethylation remain largely unknown. The effects and functional outcome from hypomethylation have been studied much less but the phenomenon is common across several cancer tissues including Melanoma[94]. Gene-specific hypermethylation and the unique signature of methylation patterns associated with a cell state has been used to classify Melanoma. In a study with benign nevi and malignant Melanoma, 26 CpG sites and 22 genes were identified to have distinctly different methylation [95]. Methylation signatures are gaining significant prominence in prognosis and tumor grading, because methylation pattern is an accurate reflection of cellular machinery and contains much higher-resolution information of a cell state than a histopathological slide. Comprehensive pathological diagnosis in tumors of the central nervous system with use of DNA methylation data was reported to be accurate in comparison with conventional histological analysis[96].

Epigenomics encompasses the investigation of marks materialized by chemical modification of DNA and histones. DNA methylation is the most commonly observed marker of epigenetic modification in cancer, but recent studies have implicated an increasingly enhanced role of histone modification[97, 98]. Methylation is not the only form of epigenetic modification; acetylation, phosphorylation, ubiquitylation, and sumoylation are also forms of epigenetic chemical modification. Histone methylation and acetylation can occur at varying degrees; for example, mono-, di- and trimethyl histones are routinely observed and affect the chromatin accessibility and gene expression differently as well. All forms of epigenetic modification generally affect the binding of transcription factors to the transcription elements and possess the ability to modulate gene expression. Histone methyltransferases and demethylases cooperate with members of the transcriptional machinery to modulate oncogenic gene expression.

Histones are chromosomal proteins, around which genomic DNA is compactly wrapped to form the primary component of chromatin, to reduce chromosomal volume and strengthen the structure. The pattern, location and amount of histone modification play a defining role in gene expression. The histone protein family contains five major histones: H1, H2A, H2B, H3 and H4. Epigenomic modification of histones occurs through post-translational modification and alters

TABLE 1.2: Histone modification marks have a distinct fate as it pertains to transcription regulation.

Modification	Histone			
	H3K4	H3K9	H3K27	H3K36
mono-methylation	activation	activation	activation	
di-methylation		repression	repression	
tri-methylation	activation	repression	repression	activation
acetylation	activation	activation	activation	

their interaction with DNA. Alteration in histone structure due to epigenomic modification impacts chromatin structure which in turn influences gene expression. The functional interpretation of all histone modification is currently evolving, but the affect on transcription regulation by most common marks are described in the table below. Analyzing histone modification marks signals the global trend of transcription regulation in a sample[99, 100]. Histone modification marks associated with both active and repressive states of chromatin are implicated in cancer and occur simultaneously.

Chromatin represents an additional level of gene expression regulation by epigenetic mechanisms. The switching between inactive and active chromatin is closely related to the activity of histone-modifying enzymes and chromatin-remodeling complexes. Transcriptional activation of a gene is a multistep process that starts with the binding of specific transcription factors to regulatory DNA elements. Transcription factors ultimately transduce the proliferation signals elicited by growth factors. Moreover, many human oncogenes encode for transcription factors, and some of them are prevalent in particular cancer tissues (e.g., MYC, MITF). Also, some of the most prominent tumor suppressors (e.g. p53) are transcription factors. Transcription factors are therefore proteins that play a role in regulating the transcription of genes by binding to specific regulatory nucleotide sequences. Transcription factors are crucial for maintaining specific cell states and the gene regulation programs associated with them[101]. Misregulation of transcription through metabolic reprogramming is an emerging hallmark in cancer, particularly in Melanoma[62, 102].

1.3 Summary

Precise interplay between multitude of genomic events contribute to malignancies, such as cancer. In Melanoma, substantial rewiring of a cell through a diverse set of mechanisms, including genomic and epigenomic events, is pivotal for carcinogenesis, progression and therapy resistance. This thesis work outlines analysis framework and tools for investigation of Melanoma. In the chapters that follow, a pipeline to efficiently detect driver oncogenic mutations, an assessment of the impact of mutation on cancer metabolism, and overexpression-induced epigenomic rewiring that occurs in Melanoma are described.

TABLE 1.3: The Complete Picture: Landscape of major pathways alterations in cancer. These interpretation were made using TCGA [12] NGS data from over 10000 patients across multiple cancer tissues and visualized in cbioportal [103]. Tumor suppressor are depicted with green color while oncogenes are depicted in pink color.

Pathway	Gene	Fate in Cancer	Role	Other Genes
Cell cycle	CDKN2A	Deletion, Methylation	Cyclins/CDKs	CDKN2B/C
	CCND1	Amplification		
	CCNE1	Amplification		
	CDK4	Amplification		
	CDK6	Amplification		
	RB1	Deletion, mutation		
PI3K	PTEN	Mutation, Deletion	Cell growth	RICTOR
	PIK3R1	Mutation, Deletion		
	PIK3CA	Mutation, Amplification		
	STK11	Mutation		
	AKT1	Mutation, Amplification		
	AKT2	Amplification		
	AKT3	Fusion, Mutation		
p53	MDM2	Amplification	Cell survival, proliferation	ATM
	MDM4	Amplification		
	TP53	Mutation		

Notch	CREBBP	Mutation, Deletion	Cell growth, apoptosis	KDM5A
	NOTCH1	Mutation, Deletion		
	NOTCH2	Mutation, Deletion		
	NOTCH3	Mutation, Deletion		
	NOTCH4	Mutation, Deletion		
	NOTCH7	Mutation, Deletion		
	NCOR1	Mutation, Deletion		
Myc	MYC	Amplification	Cell grwoth, proliferation, apoptosis	MYCN
	MAC	Mutation		
	MGA	Methylation, Mutation		
Hippo	LATS1	Mutation, Deletion	Cell proliferation, differentiation	FAT2–4
	LATS2	Methylation		
	YAP1	Amplification		
	NF2	Mutation, Deletion		
	FAT1	Mutation, Deletion		
RTK/RAS	BRAF	Mutation, Amplification	Cyclins/CDKs	RAC1, KIT
	NRAS	Mutation, Amplification		
	EGFR	Amplification, Mutation		
	KRAS	Mutation, Amplification		
	MAPK1	Amplification		
	NF1	Deletion, Mutation		
Wnt	APC	Mutation	Cell proliferation	ZNRF3
	CTNNB1	Mutation		
	TCF7	Methylation		
TGFβ	TGFBR1	Deletion,Mutation	Proliferation, stem/progenitor phenotype	ACVR2A
	TGFBR2	Methylation, Mutation		
	SMAD2	Mutation, Deletion		
	SMAD3	Methylation, Deletion		
	SMAD4	Mutation, Deletion		

References

[1] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, April 2009.

[2] Philip J. Stephens, David J. McBride, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276):1005–1010, December 2009.

[3] Barbara A. Weir, Michele S. Woo, et al. Characterizing the cancer genome in lung adenocarcinoma. *Nature*, 450(7171):893–898, December 2007.

[4] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, October 2012.

[5] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, September 2012.

[6] T. Harada, C. Chelala, V. Bhakta, et al. Genome-wide DNA copy number analysis in pancreatic cancer using high-density single nucleotide polymorphism arrays. *Oncogene*, 27(13):1951–1960, March 2008.

[7] Wen Xue, Thomas Kitzing, et al. A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. *Proceedings of the National Academy of Sciences of the United States of America*, 109(21):8212–8217, May 2012.

[8] Serena Nik-Zainal, Peter Van Loo, et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, May 2012.

[9] Hiu Wing Cheung, Glenn S. Cowley, et al. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 108(30):12372–12377, July 2011.

[10] Mike Martin. Rewriting the mathematics of tumor growth. *Journal of the National Cancer Institute*, 103(21):1564–1565, November 2011.

[11] Peter Eirew, Adi Steif, et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426, February 2015.

[12] Cancer Genome Atlas Research Network, John N. Weinstein, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, October 2013.

[13] William D. Travis, Elisabeth Brambilla, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, 10(9):1243–1260, September 2015.

[14] Peter A. Jones and Stephen B. Baylin. The Epigenomics of Cancer. *Cell*, 128(4):683–692, February 2007.

[15] Simon A. Forbes, David Beare, Prasad Gunasekaran, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(Database issue):D805–811, January 2015.

[16] Maria Diamandis, Nicole M. A. White, and George M. Yousef. Personalized medicine: marking a new epoch in cancer patient management. *Molecular cancer research: MCR*, 8(9):1175–1187, September 2010.

[17] David E. Reich, Stacey B. Gabriel, and David Altshuler. Quality and completeness of SNP databases. *Nature Genetics*, 33(4):457–458, April 2003.

[18] Erika Maria Monteiro Santos et al. Integration of genomics in cancer care. *Journal of Nursing Scholarship: An Official Publication of Sigma Theta Tau International Honor Society of Nursing*, 45(1):43–51, March 2013.

[19] Erin D. Pleasance, R. Keira Cheetham, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196, January 2010.

[20] Benjamin Schuster-Bckler and Ben Lehner. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, 488(7412):504–507, August 2012.

[21] Michael S. Lawrence, Petar Stojanov, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, July 2013.

[22] John A. Curtin, Jane Fridlyand, et al. Distinct sets of genetic alterations in melanoma. *The New England Journal of Medicine*, 353(20):2135–2147, November 2005.

[23] Mauricio Burotto, Victoria L. Chiou, et al. The MAPK pathway across different malignancies: a new perspective. *Cancer*, 120(22):3446–3456, November 2014.

[24] Nicholas K. Hayward, James S. Wilmott, et al. Whole-genome landscapes of major melanoma subtypes. *Nature*, 545(7653):175–180, 2017.

[25] Jong Woo Lee, Young Hwa Soung, et al. Inactivating mutations of proapoptotic Bad gene in human colon cancers. *Carcinogenesis*, 25(8):1371–1376, August 2004.

[26] Ludmil B. Alexandrov, Serena Nik-Zainal, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, August 2013.

[27] I. P. Tomlinson, M. R. Novelli, and W. F. Bodmer. The mutation rate and cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 93(25):14800–14803, December 1996.

[28] Lawrence A. Loeb, Keith R. Loeb, and Jon P. Anderson. Multiple mutations and cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):776–781, February 2003.

[29] Dapeng Hao, Li Wang, and Li-jun Di. Distinct mutation accumulation rates among tissues determine the variation in cancer risk. *Scientific Reports*, 6:19458, January 2016.

[30] Allon M. Klein, Douglas E. Brash, et al. Stochastic fate of p53-mutant epidermal progenitor cells is tilted toward proliferation by UV B during preneoplasia. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1):270–275, January 2010.

[31] Elizabeth Clayton, David P. Doup, et al. A single type of progenitor cell maintains normal epidermis. *Nature*, 446(7132):185–189, March 2007.

[32] Lauren Edwards, Rohit Gupta, and Fabian Volker Filipp. Hypermutation of DPYD Deregulates Pyrimidine Metabolism and Promotes Malignant Progression. *Molecular Cancer Research*, 14(2):196–206, February 2016.

[33] V. E. Velculescu, L. Zhang, et al. Serial analysis of gene expression. *Science (New York, N.Y.)*, 270(5235):484–487, October 1995.

[34] M. Schena, D. Shalon, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235):467–470, October 1995.

[35] Aravind Subramanian, Pablo Tamayo, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[36] Homin K. Lee, Amy K. Hsu, et al. Coexpression analysis of human genes across many microarray data sets. *Genome Research*, 14(6):1085–1094, June 2004.

[37] Keith Hoek, David L. Rimm, et al. Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas. *Cancer Research*, 64(15):5270–5282, August 2004.

[38] Christopher Haqq, Mehdi Nosrati, et al. The gene expression signatures of melanoma progression. *Proceedings of the National Academy of Sciences of the United States of America*, 102(17):6092–6097, April 2005.

[39] E. A. Clark, T. R. Golub, et al. Genomic analysis of metastasis reveals an essential role for RhoC. *Nature*, 406(6795):532–535, August 2000.

[40] D. R. Welch, P. Chen, et al. Microcell-mediated transfer of chromosome 6 into metastatic human C8161 melanoma cells suppresses metastasis but does not inhibit tumorigenicity. *Oncogene*, 9(1):255–262, January 1994.

[41] Gran Jnsson, Christian Busch, et al. Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 16(13):3356–3367, July 2010.

[42] Vronique Winnepenninckx, Vladimir Lazar, et al. Gene expression profiling of primary cutaneous melanoma and clinical outcome. *Journal of the National Cancer Institute*, 98(7):472–482, April 2006.

[43] Kristen M. Carr, Michael Bittner, and Jeffrey M. Trent. Gene-expression profiling in human cutaneous melanoma. *Oncogene*, 22(20):3076–3080, May 2003.

[44] Loren E. Clarke, M. B. Warf, et al. Clinical validation of a gene expression signature that differentiates benign nevi from malignant melanoma. *Journal of Cutaneous Pathology*, 42(4):244–252, April 2015.

[45] Jennifer S. Ko, Balwir Matharoo-Ball, et al. Diagnostic Distinction of Malignant Melanoma and Benign Nevi by a Gene Expression Signature and Correlation to Clinical Outcomes. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research,*

Cosponsored by the American Society of Preventive Oncology, 26(7):1107–1113, 2017.

[46] Christos Sotiriou and Lajos Pusztai. Gene-expression signatures in breast cancer. *The New England Journal of Medicine*, 360(8):790–800, February 2009.

[47] Laura J. van 't Veer, Hongyue Dai, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, January 2002.

[48] Patrick A. Ott, Yung-Jue Bang, et al. T-Cell-Inflamed Gene-Expression Profile, Programmed Death Ligand 1 Expression, and Tumor Mutational Burden Predict Efficacy in Patients Treated With Pembrolizumab Across 20 Cancers: KEYNOTE-028. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 37(4):318–327, February 2019.

[49] Mark Ayers, Jared Lunceford, Michael Nebozhyn, et al. IFN-related mRNA profile predicts clinical response to PD-1 blockade. *The Journal of Clinical Investigation*, 127(8), June 2017.

[50] Willy Hugo, Jesse M. Zaretsky, et al. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*, 165(1):35–44, March 2016.

[51] Nadine S. Jahchan, Joel T. Dudley, et al. A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer Discovery*, 3(12):1364–1377, December 2013.

[52] Vera van Noort, Sebastian Schlch, et al. Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene-expression profiling. *Cancer Research*, 74(20):5690–5699, October 2014.

[53] Luiz Fernando Zerbini, Manoj K. Bhasin, et al. Computational repositioning and preclinical validation of pentamidine for renal cell cancer. *Molecular Cancer Therapeutics*, 13(7):1929–1941, July 2014.

[54] David S. Hong, Luis Vence, et al. BRAF(V600) inhibitor GSK2118436 targeted inhibition of mutant BRAF in cancer patients does not impair overall immune competency. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 18(8):2326–2335, April 2012.

[55] Akito Nakamura, Takeo Arita, et al. Antitumor activity of the selective pan-RAF inhibitor TAK-632 in BRAF inhibitor-resistant melanoma. *Cancer Research*, 73(23):7043–7055, December 2013.

[56] Michael P. Smith, Holly Brunton, et al. Inhibiting Drivers of Non-mutational Drug Tolerance Is a Salvage Strategy for Targeted Melanoma Therapy. *Cancer Cell*, 29(3):270–284, March 2016.

[57] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498, October 2015.

[58] Rameen Beroukhim, Craig H. Mermel, et al. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905, February 2010.

[59] Michael Baudis. Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC cancer*, 7:226, December 2007.

[60] Tae-Min Kim, Ruibin Xi, et al. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Research*, 23(2):217–227, February 2013.

[61] Graham R. Bignell, Chris D. Greenman, et al. Signatures of mutation and selection in the cancer genome. *Nature*, 463(7283):893–898, February 2010.

[62] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, March 2011.

[63] Ming-Sound Tsao, Akira Sakurada, et al. Erlotinib in lung cancer - molecular and clinical predictors of outcome. *The New England Journal of Medicine*, 353(2):133–144, July 2005.

[64] Maggie C. U. Cheang, Stephen K. Chia, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *Journal of the National Cancer Institute*, 101(10):736–750, May 2009.

[65] Edward S. Kim, Vera Hirsh, et al. Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomised phase III trial. *Lancet (London, England)*, 372(9652):1809–1818, November 2008.

[66] S. W. Lowe, S. Bodis, et al. p53 status and the efficacy of cancer therapy in vivo. *Science (New York, N.Y.)*, 266(5186):807–810, November 1994.

[67] Ruprecht Wiedemeyer, Cameron Brennan, et al. Feedback circuit among INK4 tumor suppressors constrains human glioblastoma development. *Cancer Cell*, 13(4):355–364, April 2008.

[68] J. R. Pollack, C. M. Perou, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23(1):41–46, September 1999.

[69] Jonathan R. Pollack, Therese Srlie, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12963–12968, October 2002.

[70] Elizabeth Hyman, Pivikki Kauraniemi, et al. Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Research*, 62(21):6240–6245, November 2002.

[71] Kevin M. Shannon. Resistance in the land of molecular cancer therapeutics. *Cancer Cell*, 2(2):99–102, August 2002.

[72] Jian Guan, Rohit Gupta, and Fabian V. Filipp. Cancer systems biology of TCGA SKCM: Efficient detection of genomic drivers in melanoma. *Scientific Reports*, 5:7857, January 2015.

[73] Levi A. Garraway, Hans R. Widlund, et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*, 436(7047):117–122, July 2005.

[74] Minoru Takata, Keishi Maruo, et al. Two cases of unusual acral melanocytic tumors: illustration of molecular cytogenetics as a diagnostic tool. *Human Pathology*, 34(1):89–92, January 2003.

[75] Boris C. Bastian. Understanding the progression of melanocytic neoplasia using genomic analysis: from fields to cancer. *Oncogene*, 22(20):3081–3086, May 2003.

[76] Jeff D. Harvell, Boris C. Bastian, and Philip E. LeBoit. Persistent (recurrent) Spitz nevi: a histopathologic, immunohistochemical, and molecular pathologic study of 22 cases. *The American Journal of Surgical Pathology*, 26(5):654–661, May 2002.

[77] Boris C. Bastian, Jessie Xiong, et al. Genetic changes in neoplasms arising in congenital melanocytic nevi: differences between nodular proliferations and melanomas. *The American Journal of Pathology*, 161(4):1163–1169, October 2002.

[78] M. Balzs, Z. Adm, A. Treszl, et al. Chromosomal imbalances in primary and metastatic melanomas revealed by comparative genomic hybridization. *Cytometry*, 46(4):222–232, August 2001.

[79] Boris C. Bastian, Adam B. Olshen, et al. Classifying melanocytic tumors based on DNA copy number changes. *The American Journal of Pathology*, 163(5):1765–1770, November 2003.

[80] Zsuzsa Rkosy, Laura Vzkeleti, et al. EGFR gene copy number alterations in primary cutaneous malignant melanomas are associated with poor prognosis. *International Journal of Cancer*, 121(8):1729–1737, October 2007.

[81] Peter A. Jones and Stephen B. Baylin. The fundamental role of epigenetic events in cancer. *Nature Reviews. Genetics*, 3(6):415–428, June 2002.

[82] Jordi Frigola, Jenny Song, Clare Stirzaker, et al. Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. *Nature Genetics*, 38(5):540–549, May 2006.

[83] Yasuo Koga, Mattia Pelizzola, et al. Genome-wide screen of promoter methylation identifies novel markers in melanoma. *Genome Research*, 19(8):1462–1470, August 2009.

[84] Ryan Lister, Mattia Pelizzola, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, November 2009.

[85] Goran Micevic, Nicholas Theodosakis, and Marcus Bosenberg. Aberrant DNA methylation in melanoma: biomarker and therapeutic opportunities. *Clinical Epigenetics*, 9:34, 2017.

[86] Jean-Pierre Issa. CpG island methylator phenotype in cancer. *Nature Reviews Cancer*, 4(12):988–993, December 2004.

[87] Manuel Rodrguez-Paredes and Manel Esteller. Cancer epigenetics reaches mainstream oncology. *Nature Medicine*, 17(3):330–339, March 2011.

[88] Staci L. Haney, Ryan A. Hlady, et al. Methylation-independent repression of Dnmt3b contributes to oncogenic activity of Dnmt3a in mouse MYC-induced T-cell lymphomagenesis. *Oncogene*, 34(43):5436–5446, October 2015.

[89] Qing Gao, Eveline J. Steine, M. Inmaculada Barrasa, et al. Deletion of the de novo DNA methyltransferase Dnmt3a promotes lung tumor progression. *Proceedings of the National Academy of Sciences*, 108(44):18061–18066, November 2011.

[90] Ina Rhee, Kurtis E. Bachman, Ben Ho Park, et al. DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature*, 416(6880):552–556, April 2002.

[91] Suhu Liu, Suping Ren, et al. Identification of novel epigenetically modified genes in human melanoma via promoter methylation gene profiling. *Pigment Cell & Melanoma Research*, 21(5):545–558, October 2008.

[92] Dave S. B. Hoon, Mia Spugnardi, et al. Profiling epigenetic inactivation of tumor suppressor genes in tumors and plasma from cutaneous melanoma patients. *Oncogene*, 23(22):4014–4022, May 2004.

[93] Luca Sigalotti, Alessia Covre, et al. Epigenetics of human cutaneous melanoma: setting the stage for new therapeutic strategies. *Journal of Translational Medicine*, 8:56, June 2010.

[94] Christine Guo Lian, Yufei Xu, et al. Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. *Cell*, 150(6):1135–1146, September 2012.

[95] Kathleen Conway, Sharon N. Edmiston, et al. DNA-methylation profiling distinguishes malignant melanomas from benign nevi. *Pigment Cell & Melanoma Research*, 24(2):352–360, April 2011.

[96] David Capper, David T. W. Jones, et al. DNA methylation-based classification of central nervous system tumours. *Nature*, 555(7697):469–474, 2018.

[97] Manel Esteller. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews. Genetics*, 8(4):286–298, April 2007.

[98] David B. Seligson, Steve Horvath, et al. Global histone modification patterns predict risk of prostate cancer recurrence. *Nature*, 435(7046):1262–1266, June 2005.

[99] David B. Seligson, Steve Horvath, et al. Global levels of histone modifications predict prognosis in different cancers. *The American Journal of Pathology*, 174(5):1619–1628, May 2009.

[100] M. A. Glozak and E. Seto. Histone deacetylases and cancer. *Oncogene*, 26(37):5420–5432, August 2007.

[101] Tong Ihn Lee and Richard A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, March 2013.

[102] Fabian V. Filipp, Boris Ratnikov, et al. Glutamine-fueled mitochondrial metabolism is decoupled from glycolysis in melanoma. *Pigment Cell & Melanoma Research*, 25(6):732–739, November 2012.

[103] Jianjiong Gao, Blent Arman Aksoy, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269):pl1, April 2013.

Chapter 2

Methods

Working with omics data spanning several modalities must account for measurements made across scales of time, space, and biological organization. In a similar way, computational methods associated with cancer systems biology aim to coalescence data from multiscale systems. The Cancer Genome Atlas (TCGA)[1] and International Cancer Genome Consortium (ICGC)[2] are collaborative projects with comprehensive catalogues of NGS data from over 35 cancer tissues collected from over 25000 cancer patients. The primary focus of ICGC is somatic mutation while TCGA contains multifaceted cancer-associated data. TCGA also includes clinical, biospecimen, and transcriptomic information along with genomic and epigenomic data. For Melanoma, TCGA includes data from 471 samples, of which 367 are the metastasized (TM) form of disease, 103 are solid primary tumors (TP), and 1 is a blood-derived normal sample (NB).

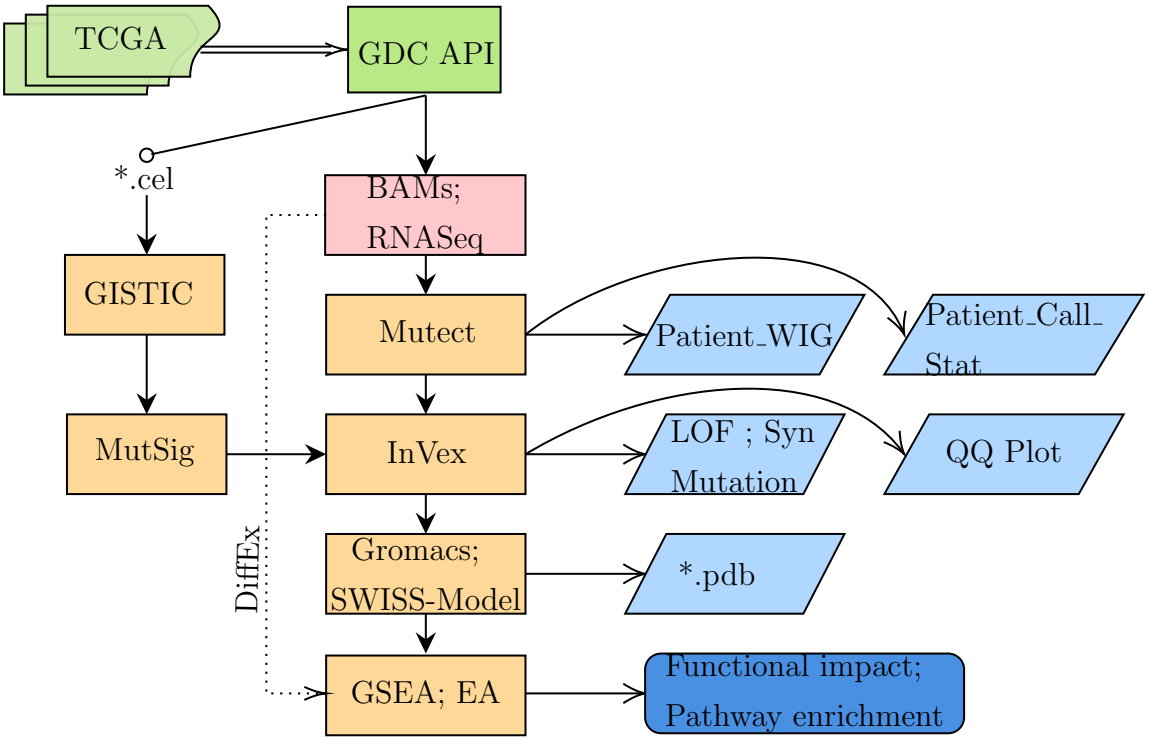
The following sections outline genetic, epigenetic and transcriptomic data analysis pipeline frameworks and methods used for performing this thesis work.

2.1 Detecting Driver Oncogenic Mutation

2.1.1 Motivation

Depending on the cancer tissue, an average whole-exome dataset may contain mutations on the order of 10^6. This is further exacerbated in Melanoma and Lung Cancer due to exogenous mutagens being the primary causes of the disease. For instance, an extremely high number of mutations is commonly observed in Melanoma; sequencing datasets therefore pose a series of challenges to analyze. An efficient mutation call tool is required to identify all instances of mutation in patient data and it is equally vital to efficiently characterize somatic mutations. In order to accurately identify downstream perturbations, formulating a faster and more efficient tool to characterize potentially cancer-driving somatic mutations above background noise is vital.

2.1.2 Pipeline



A workflow to identify and characterize driver somatic mutations and then to assess their functional impact in Melanoma as used in Chapter 3 ,4 of this dissertation. In the schematic 2.1.2, tools and methods are depicted in orange color while blue color is used for associated output.

TCGA data access

TCGA data is accessible via the NCI Genomic Data Commons (GDC) data portal, GDC Legacy Archive and the Broad Institutes GDAC Firehose. The GDC Data Portal provides access to the subset of TCGA data that has been harmonized against GRCh38 (hg38) using GDC bioinformatics pipelines. The GDC open access data does not require authentication or authorization to access it and generally includes high-level genomic data that is not individually identifiable, as well as most clinical and all biospecimen data elements. The GDC controlled access data requires dbGaP authorization and eRA Commons authentication and generally includes individually identifiable data such as low level genomic sequencing data, germline variants, SNP6 genotype data, and certain clinical data elements. This and the following study were carried out as part of IRB approved study dbGap ID 5094[3]. GDC API was accessed with TCGAbiolinks [4] package to query and download SNP6,methylation, mRNA and exome data. Whole exome and deep sequencing data was obtained from CGHub[5] through genetorrent client. Data was stored as *summarizedExperiment*[6] object class or dataframe in R [7]. Processed mutation data provided by TCGA is stored in MAF files (Mutation Annotation Format), which are derived from VCF files.

Copy number alteration

The tool GISTIC 2.0.21 [8] was used to identify genomic regions that are significantly gained or lost across a set of paired normal and tumor samples of TCGA SKCM data set. The most significant recurrent SCNAs were identified using GAIA [9], an iterative procedure where a statistical hypothesis framework is extended to take into account within-sample homogeneity.

Mutation & Filtering

Mutect, the somatic variant caller used in Chapter 3 and 4, uses tumor-matched normal samples to comprehensively identify only somatic variants. It identifies

potential variants using the tumor sample and distinguishes only somatic variants using the matched normal sample. Mutect formulates two model selection problem5[10]. The wild-type model M^0 that assumes all non-reference reads come from technical artifacts and the mutation model M^f that assumes that a variant allele is present at an unknown frequency f are two models evaluated in tumor samples. A log-likelihood ratio is computed to select the better fitted model. At potential mutation sites (high LOD score), another model selection is performed in the normal sample to compare the wide-type model M^0 and the heterozygous model $M^{0.5}$. If M^0 is strongly preferred than $M^{0.5}$, the variant is labeled as somatic.

$$LOD_{Tumor} = log_{10}\left(\frac{P(observed\,data\,in\,tumor|site\,is\,mutated)}{P(observed\,data\,in\,tumor|site\,is\,reference)}\right)$$

$$LOD_{Normal} = log_{10}\left(\frac{P(observed\,data\,in\,normal|site\,is\,reference)}{P(observed\,datain\,normal|site\,is\,mutated)}\right)$$

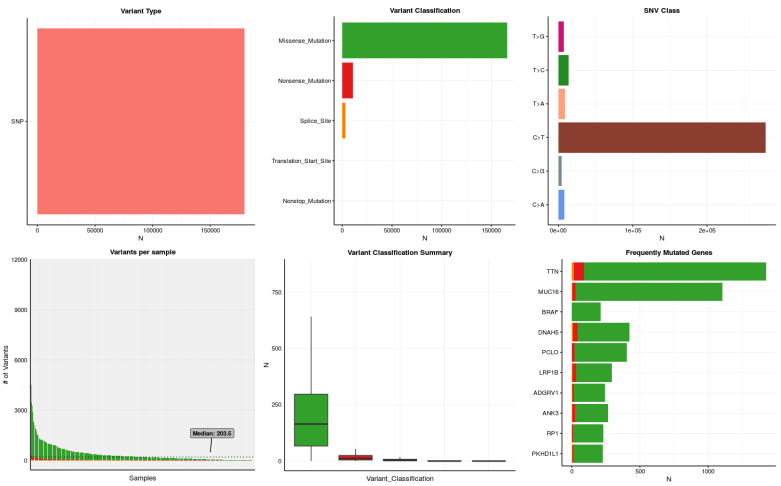


FIGURE 2.1: In case of highly mutated cancers, such as Melanoma, mutation context in addition to frequency is cruicial to efficiently identify driver mutations. This figure illustrates the rich but skewed landscape of somatic mutation in Melanoma due to UV damage signature.

MutSig then applied to somatic variants to identify: a) mutation significance, b) mutation hotspots, and c) conservation of the sites. MutSig works by creating a background model for mutations, the probability that a base is mutated by chance by permutations and comparing with observed [11]. Finally, significant

mutations and frequently-mutated genes (FMGs) are then used as input for InVEx [12]. InVEx works by creating a permutation model in order to quantify the mutation burden, which is then used to evaluate the degree of positive selection in cancer. Mutations are permuted randomly across the genes covered base pairs, respecting trinucleotide context, and the mutation burden score of the randomized instance are then calculated. The calculated mutation burden is compared with the observed burden to define the p-value for positive selection. The mutation context driven model of InVEx defines functional mutation burden by using PolyPhen2 [13] p-value in conjunction with COSMIC [14] to rank mutations with the highest functional consequence.

Structure Modeling

Homology modeling and structure simulations are useful to measure the scope of mutation as it pertains to protein structure, ligand accessibility and solvent-accessible surface area. Furthermore, comparative structure analysis between wild type and mutation hotspot can inform the stability and integrity changes that occur as a result of mutation. SWISS-MODEL [15] was used for homology modeling to determine protein structure and GROMACS [16] was used for molecular dynamics simulations. Besides quantifying protein dynamics changes such as root mean square fluctuations (RMSF), GROMACS also provides information regarding time-coarse changes in a protein structure due to mutation.

Mapping Alterations

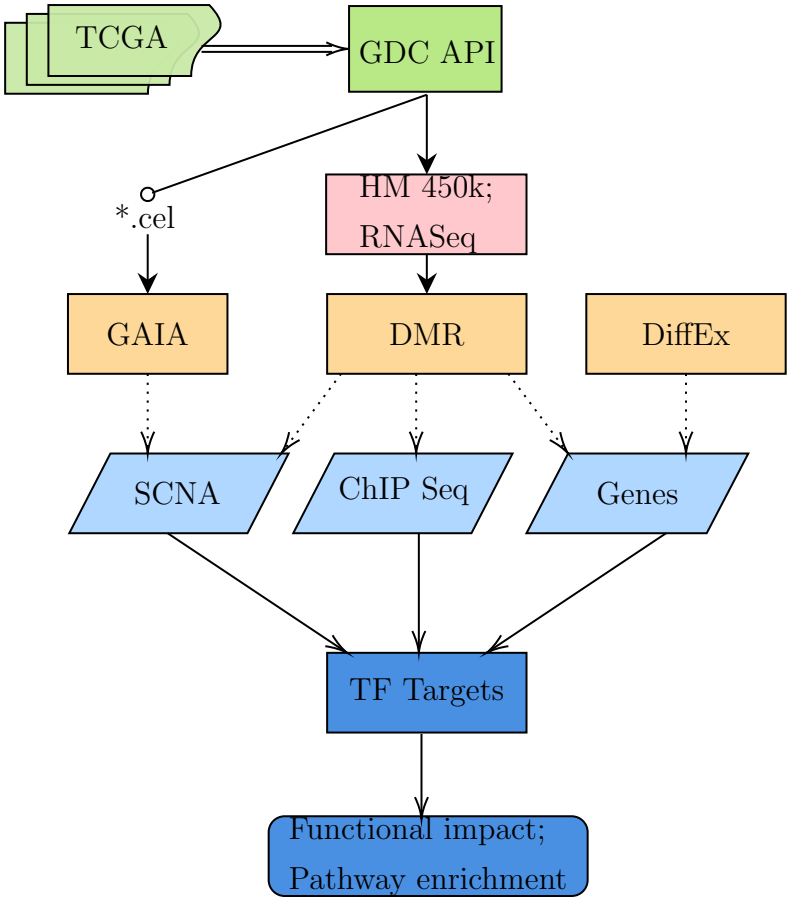
Differential expression analysis was performed using DESeq [17] and egdeR packages[18]. GSEA [19] was then used to identify and visualize KEGG [20] pathways significantly over-represented (enriched). The most statistically significant canonical pathways identified were ranked according to their p-value corrected FDR (-Log10).

2.2 Detecting Driver Epigenomic Events

2.2.1 Motivation

Epigenomic remodeling of the cellular state in a tumor directly impacts transcriptomic machinery including gene expression regulation. It is therefore crucial to define novel methods that integrate genomic data, such as gene expression, with epigenomic data such as methylation. Furthermore, upon bridging the gap in between genomic and epigenomic methods, it is also important to assess the functional scope. The following section proposes a framework, OncoBox, aimed at detecting driver epigenomic events in Melanoma.

2.2.2 OncoBox



Differential Methylation

Methylation data from TCGA was queried and downloaded using similar methods as described in TCGA data access section2.1.2. DNA methylation data is obtained from Illumina 450k arrays (containing 450,000 probes) which consists of three types of probes: cg (CpG loci), ch (non-CpG loci) and rs (SNP assay). The last type of probe can be used for sample identification and tracking and should be excluded for differential methylation analysis. All probes with at least one N/A were also removed. The DNA methylation data is presented in the form of β values that uses a scale ranging from 0.0 (probes completely unmethylated) up to 1.0 (probes completely methylated). An R summarizedExperiment object containing DNA methylation data is then used to calculate differentially methylation region (DMR). First, the difference between the mean DNA methylation (mean of the β values) between two cohorts is calculated followed by Wilcoxon test adjusting by the Benjamini-Hochberg method [21].

$$\beta_{\mathsf{i}} = \frac{\mathsf{max}(\mathsf{y}_{\mathsf{i,methyl}}, 0)}{\mathsf{max}(\mathsf{y}_{\mathsf{i,unmethyl}}, 0) + \mathsf{max}(\mathsf{y}_{\mathsf{i,methyl}}, 0)}$$

where:β_i = methylation at i^{th} cg probe.

Differential Gene Expression and DMR

After identifying differentially methylated CpG sites, in order to gauge the changes in gene-expression, information from DMRs was combined with differential gene expression data. The log_{10} (FDR-corrected P value) for DNA methylation was plotted on the x axis, and gene expression on the y axis, for each gene to observe significant association between methylation and gene-expression.

ChiP-Seq

ChIP-seq is used primarily to determine the influence of transcription factors and other chromatin-associated proteins on phenotype-affecting mechanisms. ChIP data across many tissues can be obtained from NIH Epigenome Roadmap project [22]. R bioconductor [23] package AnnotationHub [24] was used to query and download relevant ChIP data. Chipseeker package was then used for, visualization and average profile heatmap of DMRs in ChIP data [25]. Enrichment of histone modification marks was also visualized with Chipseeker package. Annotation and

R GenomicRanges data classes were created for genomic regions identified [26] in the process.

Differently Methylated Transcription Factors

To measure the functional effect of methylation on transcription factors (TF) and identify enriched TFs statistical metric to normalize over abundance of DMRs in large TF families was devised. Significant cg probes identified in step1 with p-value below 10^{-2} were grouped by TF families as classified in TFClass database [27]. The normalization parameter, Z (Equation2.3), takes into account relative size differences between TF families. The Z scores were converted to p-values using Gaussian cumulative distribution function ndtr of scipy.special python package.

$$Z_i = X - \overline{X}$$

$$Z_i = \frac{X_{iFA} - \frac{\sum_F x_{iF}}{m}}{\sigma_F}$$

where:

$$Z_i = Z\ Score$$

$$X_{iFA} = Size\ of\ a\ given\ transcription\ factor\ familyA$$

References

[1] Cancer Genome Atlas Research Network, John N. Weinstein, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, October 2013.

[2] Junjun Zhang, Joachim Baran, et al. International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database: The Journal of Biological Databases and Curation*, 2011:bar026, 2011.

[3] Matthew D. Mailman, Michael Feolo, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, 39(10):1181–1186, October 2007.

[4] Antonio Colaprico, Tiago C. Silva, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8):e71–e71, May 2016.

[5] Christopher Wilks, Melissa S. Cline, et al. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database: The Journal of Biological Databases and Curation*, 2014, 2014.

[6] Valerie Obenchain Martin Morgan. SummarizedExperiment, 2017.

[7] Ross Ihaka and Robert Gentleman. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, September 1996.

[8] Craig H. Mermel, Steven E. Schumacher, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4):R41, 2011.

[9] Sandro Morganella, Stefano Maria Pagnotta, and Michele Ceccarelli. Finding recurrent copy number alterations preserving within-sample homogeneity. *Bioinformatics*, 27(21):2949–2956, November 2011.

[10] Kristian Cibulskis, Michael S. Lawrence, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219, March 2013.

[11] Michael S. Lawrence, Petar Stojanov, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, July 2013.

[12] Eran Hodis, IanR. Watson, et al. A Landscape of Driver Mutations in Melanoma. *Cell*, 150(2):251–263, July 2012.

[13] Ivan A. Adzhubei, Steffen Schmidt, et al. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, April 2010.

[14] Simon A. Forbes, Nidhi Bindal, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research*, 39(Database issue):D945–950, January 2011.

[15] Andrew Waterhouse, Martino Bertoni, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1):W296–W303, July 2018.

[16] Sander Pronk, Szilrd Pll, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics (Oxford, England)*, 29(7):845–854, April 2013.

[17] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[18] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140, January 2010.

[19] Aravind Subramanian, Pablo Tamayo, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[20] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000.

[21] Y. Hochberg and Y. Benjamini. More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9(7):811–818, July 1990.

[22] Bradley E. Bernstein, John A. Stamatoyannopoulos, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, 28(10):1045–1048, October 2010.

[23] Robert C. Gentleman, Vincent J. Carey, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

[24] Martin Morgan , Marc Carlson, Dan Tenenbaum , Sonali Arora. AnnotationHub, 2017.

[25] Guangchuang Yu, Li-Gen Wang, and Qing-Yu He. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics (Oxford, England)*, 31(14):2382–2383, July 2015.

[26] Michael Lawrence, Wolfgang Huber, et al. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8):e1003118, 2013.

[27] Edgar Wingender, Torsten Schoeps, et al. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Research*, 46(D1):D343–D347, January 2018.

Chapter 3

Efficient detection of genomic drivers in melanoma

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
SYSTEMS BIOLOGY
CANCER
MELANOMA
CANCER GENOMICS

Received
16 July 2014

Accepted
18 December 2014

Published
20 January 2015

Correspondence and requests for materials should be addressed to F.V.F. (filipp@ucmerced.edu)

Cancer systems biology of TCGA SKCM: Efficient detection of genomic drivers in melanoma

Jian Guan, Rohit Gupta & Fabian V. Filipp

Systems Biology and Cancer Metabolism, Program for Quantitative Systems Biology, University of California Merced, Merced, CA 95343, USA.

We characterized the mutational landscape of human skin cutaneous melanoma (SKCM) using data obtained from The Cancer Genome Atlas (TCGA) project. We analyzed next-generation sequencing data of somatic copy number alterations and somatic mutations in 303 metastatic melanomas. We were able to confirm preeminent drivers of melanoma as well as identify new melanoma genes. The TCGA SKCM study confirmed a dominance of somatic BRAF mutations in 50% of patients. The mutational burden of melanoma patients is an order of magnitude higher than of other TCGA cohorts. A multi-step filter enriched somatic mutations while accounting for recurrence, conservation, and basal rate. Thus, this filter can serve as a paradigm for analysis of genome-wide next-generation sequencing data of large cohorts with a high mutational burden. Analysis of TCGA melanoma data using such a multi-step filter discovered novel and statistically significant potential melanoma driver genes. In the context of the Pan-Cancer study we report a detailed analysis of the mutational landscape of BRAF and other drivers across cancer tissues. Integrated analysis of somatic mutations, somatic copy number alterations, low pass copy numbers, and gene expression of the melanogenesis pathway shows coordination of proliferative events by Gs-protein and cyclin signaling at a systems level.

The Cancer Genome Atlas project aims at the comprehensive elucidation of genomic changes contributing to malignancies. The application of next-generation sequencing through whole-genome, whole-exome, and whole-transcriptome approaches revolutionized the resolution of cancer genome alterations, including nucleotide substitutions, small insertions and deletions, copy number alterations, chromosomal rearrangements, splice variants, regulation of gene expression, and viral or microbial interactions[1]. For melanoma patients, next-generation sequencing has already brought tangible advances. Identification of activating point-mutations in BRAF kinase (B-Raf proto-oncogene, serine/threonine kinase, Gene ID: 673) has now established a personalized medicine option with kinase inhibitors of mutated BRAF[2–5]. However, melanoma patients frequently develop resistance to BRAF inhibition[6]. In addition, melanoma subtypes with non-mutated or non-amplified BRAF, NRAS (neuroblastoma RAS viral (v-ras) oncogene homolog, Gene ID: 4893), KIT (v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog, Gene ID: 3815), or MAP2Ks (mitogen-activated protein kinase kinase, Gene IDs: 5604-5609) lack molecular targets and present a need to deepen our knowledge of the molecular signature of melanoma.

Here, we describe the genomic landscape of skin cutaneous melanoma (SKCM) based on genome-wide sequencing data from 303 TCGA malignant melanoma patients. We account for the high UV-induced basal mutation rate in skin cancers and identify genes with significantly perturbed signatures. By employing a multi-step filter we suggest a modular protocol to efficiently enrich genomic drivers in melanoma. Moreover, we characterize several novel variants of known oncogenes like BRAF and relate molecular features of new potential drivers of melanoma to recurring features observed in other cancer tissues. The comprehensive analysis provides a foundation for future functional and clinical assessment of susceptibility variants in melanoma.

Results

Patient cohort. The TCGA SKCM cohort is focused on metastatic cases (11.6% regional skin cutaneous or subcutaneous metastatic tissue, 56.4% regional metastatic lymph node, 25.1% distant or unspecified metastatic tissue) because melanoma is most often discovered after it has metastasized. We utilized files from 299 single nucleotide polymorphism (SNP) arrays, 102 whole-genome sequencing (WGS), and 276

whole-exome sequencing (WES) datasets with normal reference samples from 303 TCGA patients between 15–90 years of age (Supplementary table 1).

SCNA. Somatic copy number alterations (SCNAs) were analyzed using both SNP arrays and segmented low coverage whole-genome sequencing data; coverage of next-generation sequencing data sufficient for variant detection was set to 14 read-depth in metastases and 8 read-depth in normal blood-derived reference samples. All calls of cytobands by whole-genome sequencing were compared to level 3 segmented data at the TCGA data portal and confirmed by SNP array data (Supplementary tables 2–5). Overall, individual SNP arrays produced far fewer copy number calls compared to low coverage whole-genome sequencing experiments. Since low coverage whole-genome sequencing data produces more frequent calls, the number of significant SCNAs by whole-genome sequencing was lower than by SNP arrays.

The tool GISTIC, genomic identification of significant targets in cancer[7,8], identified 3 amplifications and 3 deletions concordantly by both SNP arrays and whole-genome sequencing; it identified 14 amplified and 13 deleted recurrent focal SCNAs detected by SNP arrays affecting 745 amplifications and 1224 deletions of genes with q-values (minimum false discovery rate at which the test may be called significant) below a threshold of 0.01 in 299 patients (Figure 1, Supplementary tables 2–5). There was significant arm-level amplification of chromosome bands 1q, 6p, 7p, 7q, 20p, and 20q detected as well as deletion of 6q, 9p, 9q, 10p, 10q, 11p, 11q, 14q, 17p with q-values below 0.01 by both SNP arrays and whole-genome sequencing. The SCNA data also revealed significant amplification of 39 miRNAs and deletion of 73 miRNAs with a gene-wise q-value below 0.01. Genes involved in pathways of mitogen-activated protein kinase (MAPK) signaling, melanogenesis, beta catenin / winglesstype (WNT), and Aurora kinase signaling are significantly deregulated, each with pathway alterations of more than 5% of the cohort size.

The most common amplification event occurs at chromosome 3 band p13, chr3:69742187-70115687, and contains MITF (microphthalmia-associated transcription factor, Gene ID: 4286) (Figure 1, Supplementary table 2–3). Focal amplification at chromosome 1 band p12, chr1:119726941-150032773, includes PHGDH (phosphoglycerate dehydrogenase, Gene ID: 26227), HMGCS2 (3-hydroxy-3-methylglutaryl-CoA synthase 2, Gene ID: 3158), NOTCH2 (notch 2, Gene ID: 4853), PDE4DIP (phosphodiesterase 4D interacting protein, Gene ID: 9659), BCL9 (B-cell CLL/lymphoma 9, Gene ID: 607). At chromosome 1 band 22q13.2, chr22:41468899-41849552, EP300 (E1A binding protein p300, Gene ID: 2033), MKL1 (megakaryoblastic leukemia 1, Gene ID: 57591), ACO2 (aconitase 2, Gene ID: 50), and RANGAP1 (Ran GTPase activating protein 1, Gene ID: 5905) are among the significantly amplified genes in melanoma with a q-value of 1.0892e-06. BRAF, EZH2 (enhancer of zeste 2 polycomb repressive complex 2 subunit, Gene ID: 2146), CREB3L2 (cAMP responsive element binding protein 3-like 2, Gene ID: 64764) at band 7q34 chr7:135929407-143664054 are amplified with

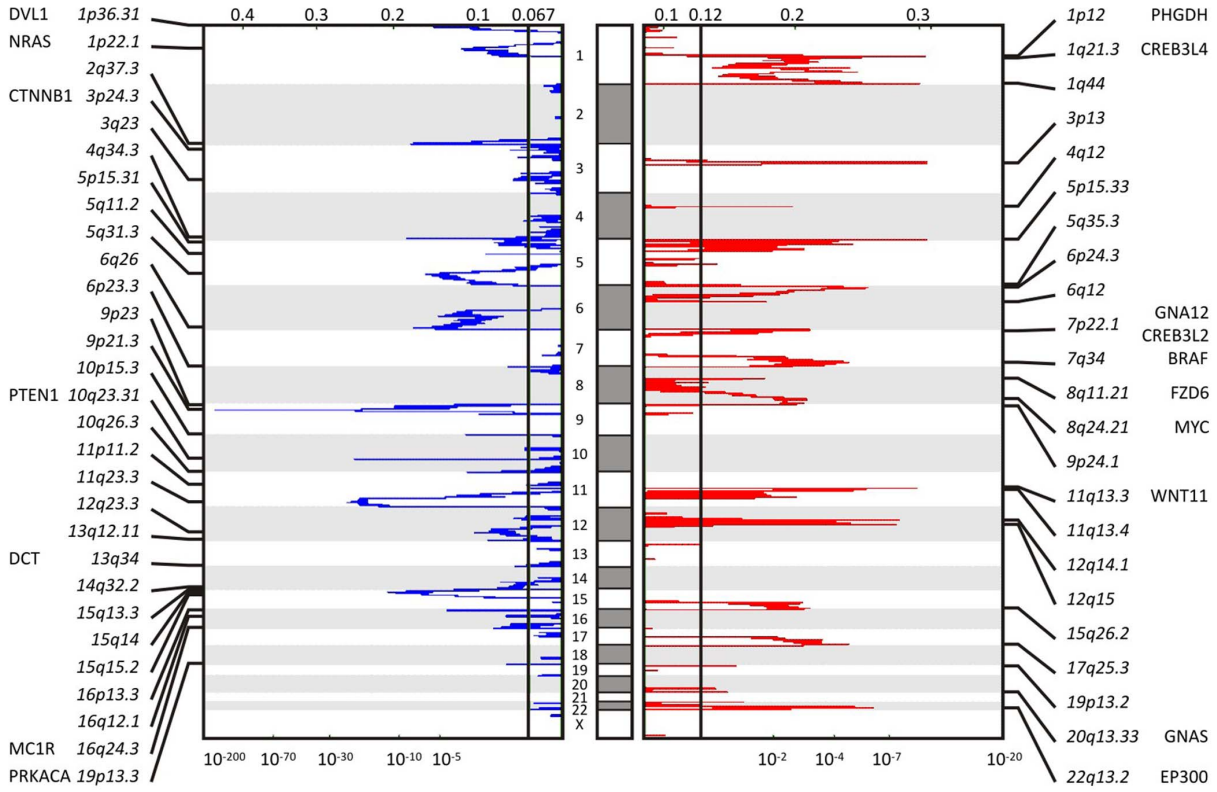


Figure 1 | Somatic copy number alteration profiling of TCGA SKCM data shows significant focal deletions and amplifications. Somatic copy number alteration analysis identifies genomic regions that are significantly gained or lost across a set of tumors. GISTIC 2.0.21 found 21 significant arm-level results, 23 significant focal amplifications, and 29 significant focal deletions in segmented SNP array data of 292 SKCM metastatic tumor samples. Among those results, amplification of BRAF, as well as reduction of NRAS and PTEN is found. The genomic position is indicated by chromosome number in the middle panel; chromosome bands and altered genes are labeled at the sides. Normalized amplifications and deletions are labeled on top and shown in red and blue, respectively.

Table 1 | Multi-step filter selects cancer drivers in genomics datasets with high mutational burden

Step	Logic	Program	Input	Output	Reference files
i	Cohort selection	TCGA Portal, CGHub, Gene Torrent	tcga_patient_list.txt	wes.bam	
ii	Mutation call	MuTect 1.1.4[28]	wes.bam	coverage.wig, call_stats.txt	HG19, COSMIC_54.vcf, dbsnp_132.vcf
iii	Recurrence, evolutionary conservation	MutSig 2.0[29]	patient.maf, coverage.wig	covariates.txt	HG19/.maf
iv	Correction for background mutation rate	InVEx 1.0.1[9]	coverage.wig, covariates.txt	significant_mutation_burden.txt, qq.png	HG19/.maf, PPH2, nucleotide_classes_HG19.txt, COSMIC, genePeptideFile_HG19
v	Mutation Signature, UV induced damage	Text editor	patient.maf	sorted_transitions.maf	nucleotide_classes_HG19.txt, uv_transitions.txt[10]
vi	Structure-activity-relationship	SWISS-MODEL 8.05, TMpred_25.0	structure.pdb	model.pdb, tm_model.txt	
vii	Pathway enrichment, Mutual exclusivity	GSEA 2.1.0[12], MEMo 1.1[11]	tcga_patient_list.txt, scna.txt, amp_del_gene.txt, covariates.txt, coverage.wig	modules.txt	
viii	Recurrence within PAN-Cancer TCGA	TCGA	covariate_target.txt, patient.maf	covariate.maf	

Somatic mutations of driver genes are called after i) cohort selection, ii) mapping of human genome and patient specific somatic references, iii) assessment of recurrence, evolutionary conversation, and iv) basal mutation rate based on frequency of mutations of introns vs exons. This first set of filters i)-iv) is necessary and sufficient to identify statistically significant enriched somatic mutations of driver genes in any dataset with high mutational burden. In a genome-wide sequencing experiment with a goal to find cancer drivers, an additional level of filters v)-viii) is advantageous. Relevance of mutations is assessed by v) nucleotide signature, vi) structure activity relationship, vii) pathway enrichment and mutual exclusivity to known cancer drivers, as well as viii) recurrence in other cancer tissues.

a q-value of 2.2567e-05. Additional focal SCNA amplifications detected by the SNP arrays are PRKAR1B (protein kinase, cAMP-dependent, regulatory, type I, beta, Gene ID: 5575), GNA12 (guanine nucleotide binding protein (G protein) alpha 12, Gene ID: 2768), RAC1 (rho family, small GTP binding protein ras-related C3 botulinum toxin substrate 1, Gene ID: 5879) at 7p22.1, FZD6 (frizzled class receptor 6, Gene ID: 8323), MYC (v-myc avian myelocytomatosis viral oncogene homolog, Gene ID: 4609) at band 8q24.21, CCND1 (cyclin D1, Gene ID: 595), WNT11 (wingless-type MMTV integration site family, member 11, Gene ID: 7481), at band 11q13.4, GNAS (guanine nucleotide binding protein, alpha stimulating, GNAS complex locus, Gene ID: 2778), AURKA (aurora kinase A, Gene ID: 6790) at band 20q13.33, and CDK4 (cyclin-dependent kinase 4, Gene ID: 1019), MDM2 (MDM2 proto-oncogene, E3 ubiquitin protein ligase, Gene ID: 4193), RAP1B (RAP1B, member of RAS oncogene family, Gene ID: 5908) at bands 12q14.1 and 12q15 and others.

The region around chromosome 9 band p21.3, chr9:21946194-21977643, includes CDKN2A (cyclin-dependent kinase inhibitor 2A, Gene ID: 1029) and shows a significant deletion with a q-value of 4.9316e-169 (Supplementary table 4–5). Other detected gene deletions include NRAS, PRKACB (protein kinase, cAMP-dependent, catalytic, beta, Gene ID: 5567), BCL10 (B-cell CLL/lymphoma 10, Gene ID: 8915), TRIM33 (tripartite motif containing 33, Gene ID: 51592) and RBM15 (RNA binding motif protein 15, Gene ID: 64783) at band 1p22, DVL1 (dishevelled segment polarity protein 1, Gene ID: 1855), RPL22 (ribosomal protein L22, Gene ID: 6146), TNFRSF14 (tumor necrosis factor receptor superfamily, member 14, Gene ID: 8764), PRDM16 (PR domain containing 16, Gene ID: 63976) at band 1p36, CTNNB1 (catenin, cadherin-associated protein, beta 1, Gene ID: 1499), RAF1 (Raf-1 proto-oncogene, serine/threonine kinase, Gene ID: 5894) at band 3p24.3, PRKAB1 (protein kinase, AMP-activated, beta 1, Gene ID: 5564) at band 12q23.3, BRCA2 (breast cancer 2, early onset, Gene ID: 675) at band 13q12.11, RB1 (retinoblastoma 1, Gene ID: 5925), DCT (dopachrome tautomerase, Gene ID: 1638) at band 13q34, AKT1 (v-akt murine thymoma viral oncogene homolog 1, Gene ID: 207) at band 14q32.3, and MC1R (melanocortin 1 receptor, alpha melanocyte stimulating hormone receptor, Gene ID: 4157) at band 16q24. The loss of chromosome 10 band q23 containing PTEN (phosphatase and tensin homolog, Gene ID: 5728) is associated with patient age (p-value of 8.81e-05).

Somatic mutations. The SKCM dataset comprises, after preprocessing of 276 patients (Supplementary table 1), a number of 55,462,639 incidences listed by MuTect, an algorithm for sensitive detection of mutations, of which 890,914 were identified as KEEP mutations. These KEEP mutations build the foundation for the gene-wise perturbation models. Exonic regions account for 220,031 mutations and classify as 60.1% missense, 3.8% nonsense, 0.5% frameshift, 0.1% in-frame insertion or deletions, 2.6% splice site, and 32.9% silent mutations. The SKCM dataset with a mutation rate of 18 mutations per mega base pairs (Mbp) is about 10-fold richer than other TCGA tissues (e.g. glioblastoma multiforme (GBM) has a rate of 1.8 mutations per Mbp in TCGA). Careful identification of non-synonymous mutations in combination with consideration of passenger events and basal mutation rate took the frequency of UV-associated gene mutations into account.

In order to identify potential melanoma driver genes, we established a multi-step filter for somatic mutations (Table 1, Methods). Steps included i) cohort selection at TCGA data portal, ii) mutation call of whole-exome sequencing data against their somatic references, iii) identification of recurrent and conserved positions, and iv) enrichment of mutations above the basal mutation rate using a permutation model[9]. This first set of filters ensures statistically significant enrichment of potential driver genes. To explore the biological impact of mutations we added a second set of filters to identify cancer drivers. Steps included v) relative frequencies of nucleotide

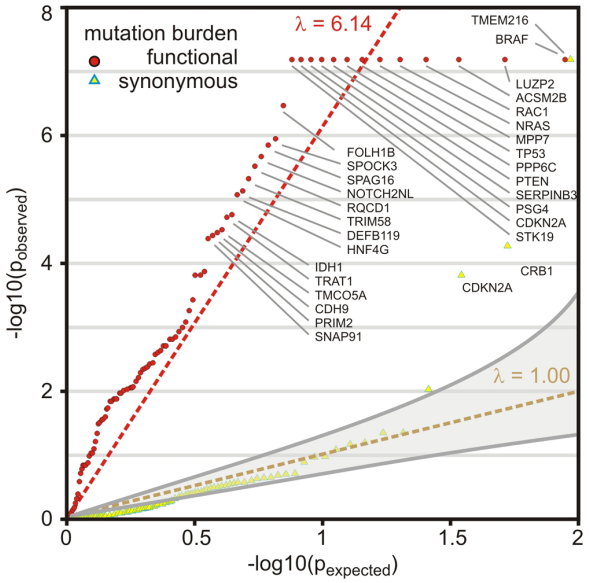


Figure 2 | Filtered genes show significant enrichment of somatic mutations above background mutation rate. QQ-plot of mutational significance analysis is based on a permutation analysis of the background mutation rate. q-values ($q=-\log_{10}(p)$) above the diagonal indicate enrichment of somatic mutation. The diagonal y=x serves as reference where observed and expected mutational burden coincide. The significantly enriched functional mutation burden of genes passed an q-value cut-off <= 0.2 is shown as red circles. The synonymous mutation burden is shown as yellow triangles. Genes with significantly enriched synonymous mutation burden passed an q-value cut-off <= 0.2 are highlighted with blue frame. Best-fit is shown as dashed-red line ($\lambda = 6.14$) and y = x as dashed-yellow line. Gray-shaded area represents 95% confidence interval for expected p-values.

mutations[10], vi) functional mutation burden using structural modeling, vii) pathway enrichment and mutual exclusivity to known cancer drivers[11], and viii) significance in multiple TCGA tissues.

In the SKCM dataset of metastatic samples with normal references, the multi-step filter analysis produced a list of 23 significantly mutated genes (22 nonsense and 1 synonymous mutation) with a q-value below 1.0e-04 (Figure 2). The multi-step filter analysis confirmed known cancer driver genes BRAF (Figure 3–4, Supplementary table 6), RAC1, NRAS, TP53 (tumor protein p53, Gene ID: 7157), CDKN2A (results in p16INK transcript), STK19 (serine/threonine kinase 19, Gene ID: 8859), PPP6C (protein phosphatase 6, catalytic subunit, Gene ID: 5537), PTEN, IDH1 (isocitrate dehydrogenase 1, Gene ID: 3417), NMS (neuromedin S, Gene ID: 129521), CDK4, and VEGFC (vascular endothelial growth factor C, Gene ID: 7424) with significantly enriched functional mutations that passed a q-value cut-off below 0.01. In addition, TMEM216 (transmembrane protein 216, Gene ID: 51259) (Figure 5, Supplementary table 7), CRB1 (crumbs family member 1, photoreceptor morphogenesis associated, Gene ID: 23418), and CDKN2A were significantly enriched genes with synonymous mutations below a q-value cut-off below 0.01. The study also highlighted genes that had never been associated with melanoma like LUZP2 (leucine zipper protein 2, Gene ID: 338645) (Supplementary information 1, Supplementary table 8), PSG4 (pregnancy specific beta-1-glycoprotein 4, Gene ID: 5672), SERPINB3 (serpin peptidase inhibitor, clade B (ovalbumin), member 3, Gene ID: 6317), SPOCK3 (sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 3, Gene ID: 50859), FOLH1B (folate hydrolase 1B, Gene ID: 219595), SPAG16 (sperm associated antigen 16, Gene ID: 79582), NOTCH2NL (notch 2 N-terminal like, Gene ID: 388677), TRIM58 (tripartite motif containing 58, Gene ID: 25893), RQCD1 (RCD1 required for cell differentiation1 homolog, Gene ID: 9125), and ACSM2B (acyl-CoA synthetase medium-chain family member 2B, Gene ID: 348158) below a q-value cut-off of 1.0e-04 (Supplementary tables 9–10).

BRAF dominates the mutational landscape of melanoma. Given the predominance of BRAF mutations in metastatic melanoma[2–5], we characterized the somatic mutation landscape of the BRAF gene in the SKCM dataset as well as in other TCGA cancer tissues.

The metastatic SKCM cohort of 276 patients with somatic controls contained 140 patients with non-silent mutations of BRAF and included 151 amino acid replacements affecting 18 unique residues (50% patient mutation frequency, p-value <1.00e-15, q-value <2.26e-12). The single most abundant protein-coding amino acid replacement observed in 119 of 276 samples is p.V600E, switching BRAF into a constitutively active protein kinase[2]. Besides V600E there are additional non-silent polar replacements in the activator loop (p.D594N, p.L597Q, p.V600K, p.V600R, and p.K601E). Next, we investigated whether such unprecedented diversity of BRAF mutations is specific to melanoma or common to other cancers. By calculating the relative frequency of mutations corrected for the cohort size, other BRAF-driven cancers were identified (Figure 3, Supplementary table 6). Thyroid cancer (THCA) stood out for containing frequent and recurrent somatic mutations of BRAF p.V600E with 249 of 350 cases (Figure 3, Supplementary table 6). In contrast, other significantly enriched datasets with BRAF mutations like colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), or SKCM showed mutations up to 37% in other conserved sections of the protein like the RAS-binding domain (RBD) (p.K183E, p.K205Q, p.E228V), the glycine-rich ATP binding site (p.G466E, p.S467L, p.G469A, p.G469E, p.G469R), or the protein surface connecting RBD and protein kinase (p.E695K) (Figure 4).

Functional analysis of somatic mutations of melanoma genes. In order to identify potential melanoma drivers, we assessed functional relevance of significant somatic mutations from nucleotide signature, structure activity relationship, mutual exclusivity to known cancer drivers, and recurrence in other cancer tissues. In the context of this study aimed at characterizing the genomic landscape of melanoma, there is space to discuss somatic mutations of two new, highly significant genes, TMEM216 (Figure 5) and LUZP2 (Supplementary information 1). Both genes display q-values below 1.0e-6 after the first four steps of the mutational analysis. Their signature of nucleotide replacement related to UV radiation deviated from the exome-wide median by more than 5%, indicative for positive selection of cancer genes (Supplementary information 2). In addition, we examined their mutational patterns in gene networks, providing important insights on gene interactions and disease drivers. We determined network associations of somatic mutations at the systems level by gene set enrichment analysis[12]. Detected somatic mutations were significantly enriched in the $G_{\alpha s}$ stimulatory heterotrimeric guanine nucleotide-binding protein (G_s-protein) pathway, M14775, with a q-value below 1.0e-6. This pathway includes BRAF, RAF1, cAMP-responsive element binding proteins CREB3 (Gene ID: 10488) and CREB5 (Gene ID: 9586), and mitogen-activated protein kinase 1, MAPK1 (Gene ID: 5594). In addition, the cyclin pathway, M1529, including mutually exclusive mutations of cyclin-dependent kinase CDK4, its inhibitor CDKN2A, proline-rich protein BstNI subfamily 1 PRB1, and tumor suppressor TP53 showed statistically significant perturbation with a q-value below 1.0e-6. The assessment of the mutational pattern in SKCM patients showed strong mutual exclusivity of TMEM216 with members of the MAPK pathway (Supplementary information 3).

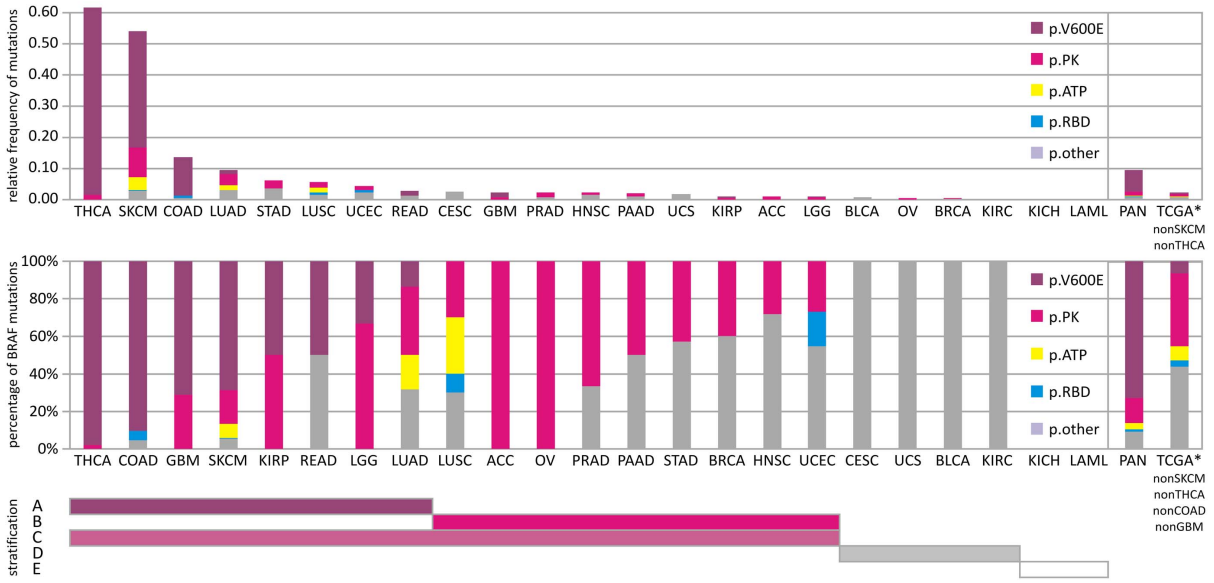


Figure 3 | Distribution of somatic mutations and mutation types in BRAF gene across The Cancer Genome Atlas Pan-Cancer analysis project. Top panel shows the relative frequency of non-silent somatic mutations (number of observed type of somatic mutation/cohort size) of detected mutations across different human cancer tissues within TCGA. Bottom panel shows fraction of mutations sorted by affected protein domains of BRAF. The analysis includes all non-silent (protein coding missense, indels, frame-shift, stop, splice site) mutations and distinguishes mutations of V600E in purple, ATP binding site in yellow, all mutations in the protein kinase (PK) domain of BRAF but not V600E or ATP binding site in pink, RAS-binding domain (RBD) in blue, and remaining protein sequence (other) in grey. The PAN-cancer analysis covers cancer tissues: adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), acute myeloid leukemia (LAML), brain lower grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), thyroid adenocarcinoma (THCA), uterine corpus endometrioid carcinoma (UCEC), and uterine carcinosarcoma (UCS). Right box within panels includes analysis for PAN-cancer cohort and sub-cohort that excludes BRAF-rich cancers of more than 0.5 relative frequency, like SKCM and THCA, or of more than 50% V600E BRAF, like THCA, COAD, SKCM, GBM (TCGA*). Patient stratification: Bars below the panels mark stratification strategy of human cancers based on their BRAF genotype. A BRAF-V600E mutation; B BRAF mutation in protein kinase domain other than V600E; C BRAF-mutation in protein kinase domain; D BRAF-mutation other than protein kinase or RBD; E no BRAF mutation.

Recurrent somatic splice site mutation of TMEM216. The transmembrane protein 216 (TMEM216) is required for tissue-specific ciliogenesis and may regulate ciliary membrane composition[13,14]. TMEM216 is the melanoma gene with the most significant synonymous somatic mutations of the TCGA SKCM dataset. TMEM216 is mutated at the highly conserved region between transmembrane helix 1 and 2 in 8 out of 276 SKCM patients (3% patient mutation frequency, p-value 6.5e-11, q-value 9.80e-8) at a single site at coding base position 138 from T to G, located at the 3′ splice site (Figure 5). The 3′ exon recognition at the acceptor splice site is critical for U2AF1 (U2 small nuclear RNA auxiliary factor 1, Gene ID: 7307) interaction[15]. The c.T138G replacement creates a mutation in the 3′ exon splice site of TMEM216. In a different TCGA dataset, the same significant, high-frequency nucleotide replacement is observed in 3 of 289 patients with lower grade glioma (LGG) (Figure 5, Supplementary table 7).

Concerted deregulation of G-protein signaling and MAPK cascade stimulates melanogenesis. Integrated systems biology analysis including somatic copy number alterations as well as pathway enrichment of somatic mutations point towards main signaling axes in melanoma; G_s-protein and MAPK signaling are frequently dysregulated in SKCM. The genomic observation in SKCM of strong mutual exclusivity of NRAS to BRAF is consistent with NRAS mutations activating both effector cascades BRAF/MEK/ERK and PI3K/Akt[11,16]. The analysis identified somatic mutations and copy number alterations affecting the G_s-protein pathway in more than 80% of the tumors. Included were activating somatic point mutations and copy number amplification of BRAF (50%, responsive to MAPK signaling pathway activator), mutation and deletion of NRAS (31%, responsive to MAPK signaling pathway activator), as well as mutation and deletion of GNAI2 (guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 2, Gene ID: 2771) (2%, responsive to GTP activator). GNAI2 proteins contribute to malignant cell growth, and its inactivation can inhibit proliferation of melanoma cells and possibly that of other malignant cells both *in vitro* and *in vivo*[17]. So far statistically significant mutation of GNAI2 in melanoma has not been reported. It is a potential therapeutic target and needs further studies to assess its clinical significance. In addition, CAMP, CREB3, CREB5, MAPK1, and RAF1 are mutated in the G_s-protein pathway in more than 10% of the tumors. Other mutations, genomic amplifications, or deletions that affect melanogenesis pathways included somatic copy number amplification in genes FZD6, GNAS, EP300, CREB3L2 and MITF. MITF is the top hit of the SCNA analysis and a critical signaling hub involved in melanocyte development, survival, and melanogenesis (Figure 6). Increased expression of MITF and its activation by phosphorylation activates

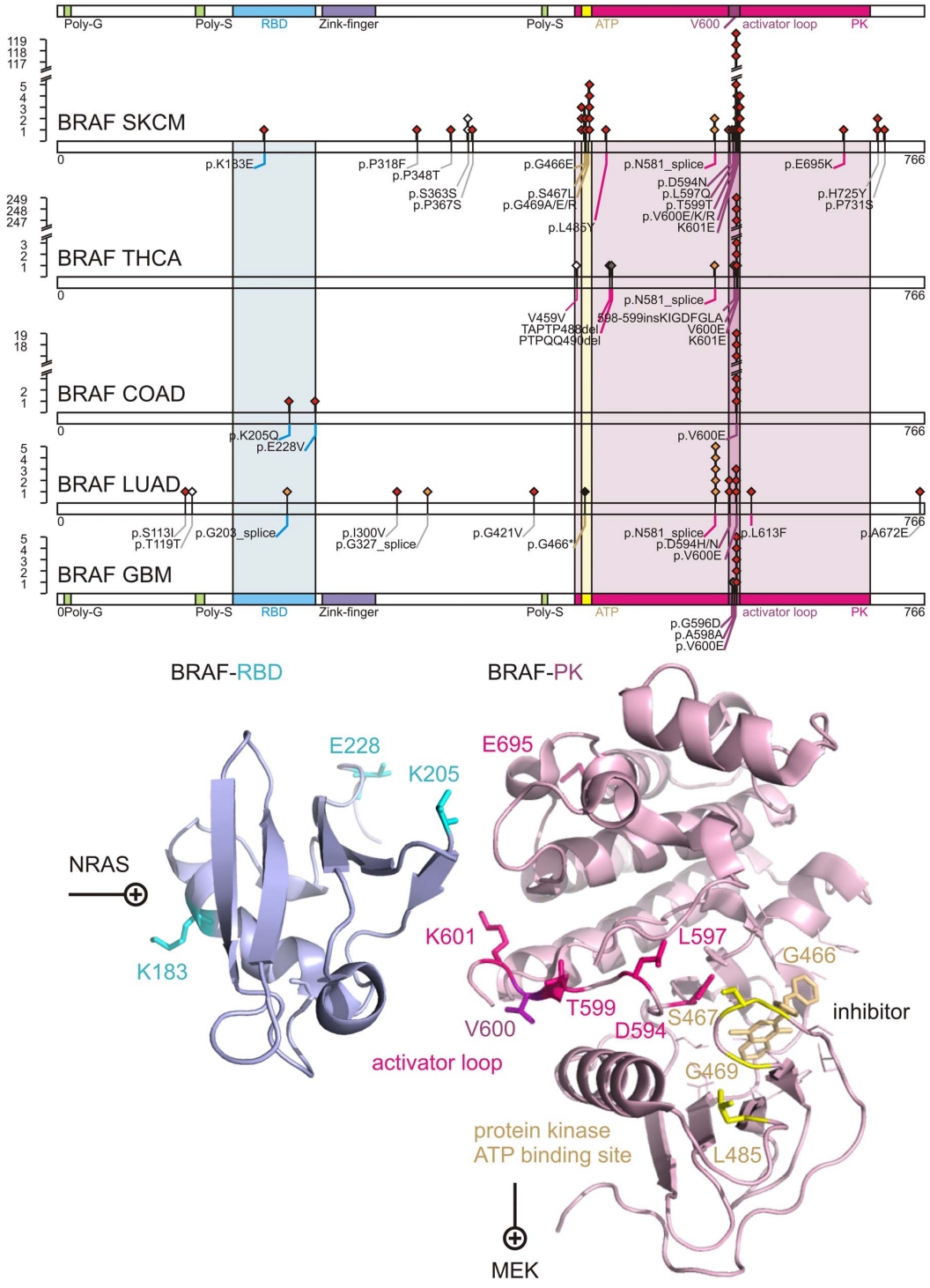


Figure 4 | Comprehensive somatic mutational landscape of BRAF as PAN-cancer driver. Distribution of somatic mutations are shown, for five TCGA tissues with significant BRAF mutations: skin cutaneous melanoma (SKCM), thyroid adenocarcinoma (THCA), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), glioblastoma multiforme (GBM). Diamonds indicate mutation type of non-synonymous mutations in red, splice-site mutations in orange, indels in brown, stop in black, and silent protein-coding mutations in white. Numbers refer to codons. Each filled circle represents an individual mutated tumour sample. The RAS binding domain (RBD) of BRAF is colored in magenta. The protein kinase (PK) domain of BRAF is colored in blue with the ATP binding site highlighted in yellow, and the activator loop around residue 600 in purple. Domains are colored accordingly. Residues affected by coding somatic mutations of BRAF, NCBI Gene ID 673, are depicted in sticks onto ribbon structure of 3ny5.pdb and 4e26.pdb.

the transcription of melanocyte-specific proteins TYR, TYRP1 and DCT. Other somatic copy number deletions occurred in genes MC1R, DVL1, CALML6, and DCT.

Discussion

Sample size of highly mutated cancer. The TCGA landmark study across many cancer types revealed that the universe of cancer mutations is much bigger than previously thought. In the case of melanoma, however, comprehensive cataloguing of low-frequency mutated genes (in ~2% patients) will require more than 5000 samples[18]. The SKCM study with about 300 samples nearly doubles the existing pool and adds value to the growing list of whole-exome sequenced melanoma[9,10,19–22].

Most genes are mutated at intermediate frequencies, creating a challenge for melanoma samples with high mutation rate of 12–18 mutations per Mbp (TCGA SKCM this study: 18/Mbp, Broad Institute's melanoma study: 12.9/Mbp[18]). A filter-based strategy helped to control for passenger mutational load (Table 1, Figure 2). Using this strategy, we were able to characterize in-depth predominant melanoma drivers—BRAF, NRAS, PTEN, TP53, CDKN2A—and also validate recently identified melanoma genes, RAC1(p.P29S)[9,10] and STK19(p.D89N)[9]. Among 10 newly identified melanoma genes, SERPINB3 is a suicide-substrate protease inhibitor, which balances cell survival and apoptosis, and is shown to be up-regulated in breast, liver, cervical, lung, and other cancers[23]. The high level of genomic instability is evidenced by significantly increased frequencies of SCNA (Figure 1). We detected concurrent arm-level alteration of both arms of chromosomes 6, 20, 9, 10, 11 indicative of aneuploidy. In addition to aneuploidy or focal SCNA events, significant enrichment of inactivating TP53 may contribute to other classes of structural variations, like breakage or fusions. While SCNAs alone cannot infer on history or heterogeneity of the tumor, future analysis on clonality status may provide additional weight to identified drivers.

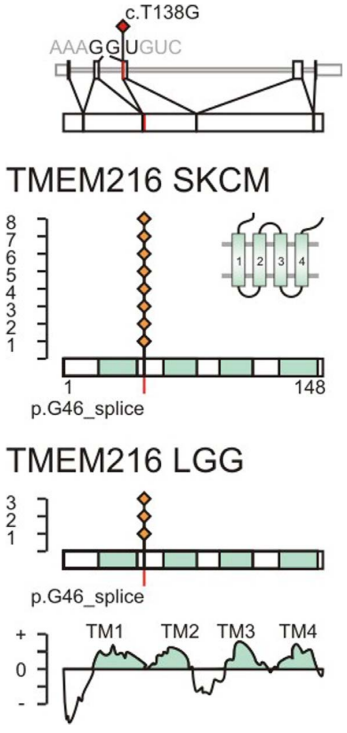


Figure 5 | The c.T138G somatic mutation is recurring and affects the splice site of TMEM216. Splice site mutations are indicated in red on the transcript and protein sequence of TMEM216, NCBI Gene ID 51259. The somatic mutation c.T138G is located on exon 3 at the second splice site. The splice site codon 46 translates to glycine 46. Observed somatic mutations in TCGA SKCM and LGG datasets are marked by red diamonds. The position of transmembrane helices is indicated in cyan and determined based on protein family PF09799, positive segments in TMPred, and uniprot entry Q9P0N5.

New cancer drivers. TMEM216 stands out among the newly discovered genes in melanoma as a low-frequency, highly statistically significant, recurrent splice site mutation. The mutation c.T138G in the splice site joining exons 2 and 3 disrupts the recognition motif of splicing auxiliary factor U2AF1[15,24]. The importance of the splice site mutation is further enhanced by the nature of the non-UV induced nucleotide change, strong mutual exclusivity with known oncogenes, and recurrence in non-melanoma cancer tissues. Germline mutations of TMEM216 (or MKS2) cause human ciliopathies like Meckel–Gruber syndrome (MKS) or Joubert syndrome (JBTS)[13] and revealed localization of TMEM216 to Golgi vesicles necessary for ciliary assembly[25]. Cilia are important organelles of cells and are involved in numerous activities such as cell signaling and processing developmental signals. Inactive TMEM216 hyper-activates RHOA signaling and increases phosphorylation of the planar cell polarity pathway of non-canonical Wnt signaling protein Dishevelled, DVL1[14,25,26]. In the context of melanoma, somatic splice-site mutation of TMEM216 suggests potential tumor suppressor function and sets the RHOA/GNA12 pathway apart from the G_s-protein/MAPK pathway by mutual exclusivity of TMEM216 towards known oncogenes NRAS and RAC1 (Supplementary information 3).

Pan-cancer. Somatic missense mutation of BRAF has been identified in roughly half of all malignant melanoma cases and at much lower frequency in all other cancers[2,9,10]. Today, whole-genome and -exome sequencing allows deep insight into molecular carcinogenesis of melanoma. The TCGA SKCM study adds to existing genome studies where BRAF(p.V600E) has been identified in 52.9% of the samples of the melanoma study of the Broad Institute[9], 64.0% in the Harvard study[22], 45.9% in the Yale study[10], as well as in 20.8% of all NCI60 cell lines[27].

Previously reported mutations rarely fell outside the kinase domain, with the coding substitution of p.V600E accounting for more than 80% of reported cases[2]. Our detailed and targeted sampling of the mutational landscape of BRAF in TCGA melanoma as well as in all other tissues of TCGA by next-generation sequencing brought three main insights forward: a) BRAF is significantly increased at a copy-number level and constitutively activated by somatic mutations; b) The whole-exome data showed unprecedented diverse mutational events within BRAF (Figure 3) aside from p.V600E preserving mutual exclusivity to activating NRAS mutations; c) Other cancers have similar or even stronger BRAF signatures than melanoma.

These observations manifest BRAF as a *bona fide* pan-cancer driver and have consequences for BRAF as an anti-cancer target and diagnostic marker. Lessons already learned from molecular studies of the BRAF pathway are relevant to almost all cancer tissues, which showed somatic missense mutations. The diverse mutational landscape challenges medicinal chemistry efforts to develop new compounds recognizing the different molecular conformations of mutated activator and ATP binding loop aside from established p.V600E binders. A single SNP test for c.1799T>A is not sufficient to capture the majority of mutational events of BRAF in its RBD or in the ATP binding site of its PK domain. Datasets like TCGA lung squamous cell carcinoma (LUSC) show mutational incidents of BRAF in more than 5% of the patients but not a single substitution of c.1799T>A. On a positive note, the diagnostic value of

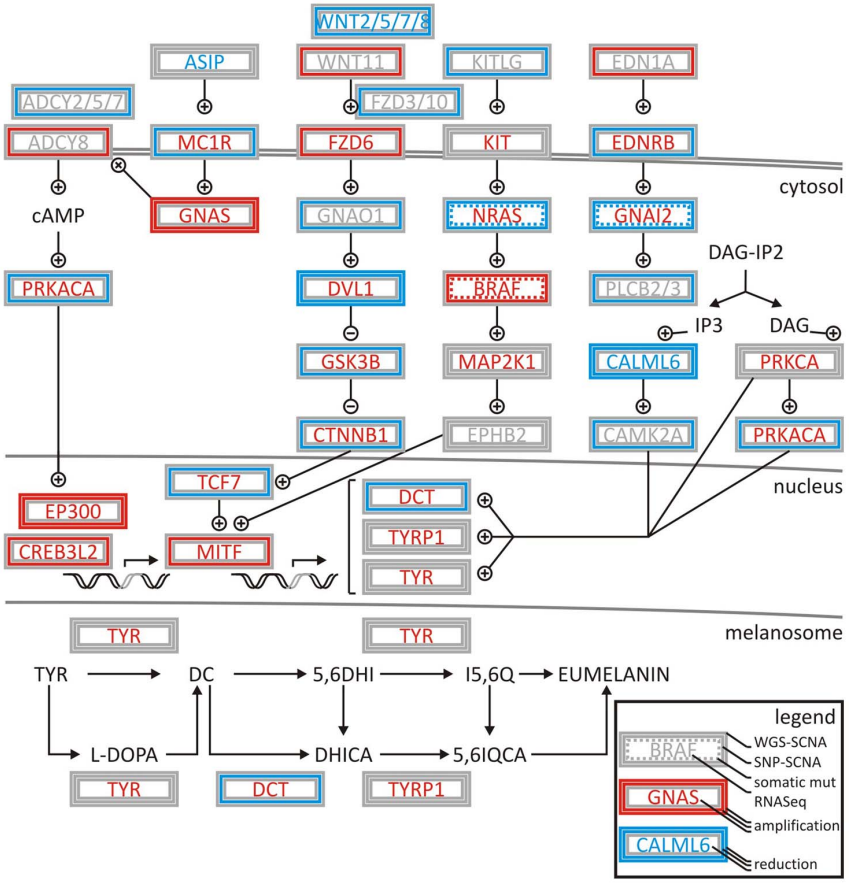


Figure 6 | Deregulation of melanogenesis by G-protein and cyclin pathway signalling in TCGA SKCM patient samples. Coordination of signaling events is demonstrated by integrating low coverage whole genome sequencing somatic copy number alteration (WGS SCNA) data, SNP somatic copy number alteration (SNP SCNA) data, somatic mutations, and RNASeq gene expression analysis. Amplifications or activations are indicated in red, deletions or reductions in blue, non-silent mutations by dashed line. Genes are boxed by analysis type. High or low activity (gene expression) is indicted by red or blue font. Grey indicated undefined state in patient cohort. + and – symbols indicate activation and inhibition of factors in the normal pathway. The network is assembled based on gene-associations of map hsa04916 and entries of M14775 and M1529 identified by systems biology gene set enrichment analysis.

c.1799T>A for BRAF diagnostics is widened to many other cancers besides SKCM that homogeneously show single p.V600E substitutions like THCA, COAD, GBM, KIRP, READ, and LGG.

Conclusion

The systems biology analysis of melanoma showed an unprecedented richness and depth of statistically significant and novel melanoma genes (Figure 2). By combining established tools in genomics, we were able to create a multi-step filter that accounts for enrichment in functional hot spots as well as the elevated and nucleotide-specific basal rate due to UV damage (Supplementary information 2). The integrated analysis of somatic mutations and structural genomic alterations in melanogenesis showed coordination of proliferative signaling events in mutually exclusive settings in melanoma (Figure 6, Supplementary information 3). If driver mutations are observed in other cancer tissues as well, lessons learned on regulation of signaling cascades and drug resistance of cancer targets might be directly translatable. Thyroid cancer showed an enrichment of the somatic melanoma driver mutation BRAF(p.V600E) that surpassed profiles of BRAF of any other tissue (Figure 3–4), while preserving mutually exclusive setting to RAS mutations (Supplementary table 6). For patients with activated BRAF pathways in their thyroid tissues, knowledge on molecular medicine of the BRAF cascade in melanoma becomes highly valuable. The systems biology integration of genomic alterations in melanoma provides a glimpse into how a spectrum of genomic aberrations contributes to melanoma genesis and progression (Figure 6). Identification of such genomic aberrations in melanoma patients contributes to new treatment regimens based on molecular understanding of driver events that govern this malignancy.

Methods

The Cancer Genome Altas project. The study was carried out as part of IRB approved study dbGap ID 5094 "Somatic mutations in melanoma". The results shown are in whole based upon data generated by the TCGA Research Network http://cancergenome.nih.gov. Restricted access whole-genome sequences and whole-exome sequences were obtained from the TCGA data portal.

Somatic copy number alterations. The tool GISTIC 2.0.21[7,8] was used to identify genomic regions that are significantly gained or lost across a set of paired normal and tumors samples of TCGA SKCM data set (for abbreviations see glossary). We executed GISTIC 2.0.21 on Illumina HiSeq data recorded with a low coverage whole-genome sequencing protocol, MD Anderson Cancer Center, TX, as well as on Agilent SNP 6.0 gene expression microarrays G4502A_07_01, UNC Chapel Hill, NC. GISTIC 2.0.21 distinguishes arm-level events from focal events at a broad length cutoff of 0.7. Events whose length was greater or less than 50% of the chromosome arm on which they resided were called arm-level or focal events, respectively, and these groups of events

were analyzed separately. The data was concordant to segmented level 3 data publically available at the TCGA data portal. Since GISTIC 2.0.21 uses ratios of segmented tumor copy number data relative to normal samples as input, segmented level 3 data aligned to HG19 served as input for analysis runs. For significant loci and genes a cutoff q-value of 0.01 was applied, and concordance determined by overlaying whole-genome sequencing and SNP data. All experiments on SCNAs were carried out at a confidence level of 0.99 according to established standards by the TCGA Research Network and compared to benchmarks established by the Broad Institute TCGA Genome Characterization Center http://www.broadinstitute.org/collaboration/gcc and Broad Institute GDAC Firehose pipeline https://confluence.broadinstitute.org/display/GDAC.

Multi-step filter for somatic mutations. We applied a multi-step filter to identify somatic mutations (Table 1). Critical components of the computational filters were i) TCGA data portal for cohort selection and CGHub for access of raw data; ii) MuTect 1.1.4 at default settings for preprocessing, alignment of reads in the tumor and normal sequencing data, and mutation calling[28]; iii) MutSig 2.0, an algorithm for identification of mutation significance, for assessing the clustering of mutations in hotspots as well as conservation of the sites[29]; and iv) InVEx 1.0.1, a permutation based Intron vs Exon algorithm for ascertaining positive selection of somatic mutation above the background level considering heterogeneity on a per-patient and per-gene level[9]. Functional impact was assessed by v) UV biased nucleotide signature[10]; vi) structural modeling using SWISS-MODEL http://swissmodel.expasy.org/ and TMpred http://www.ch.embnet.org; vii) GSEA 2.1.0 for gene set enrichment analysis[12] and MEMo 1.1 for mutual exclusivity modules analysis[11]; as well as viii) recurrence in other TCGA tissues.

i) Whole-exome sequencing files wes.bam in compressed binary version of sequence alignment map SAM (BAM) format for 276 TCGA patients (barcodes provided in Supplementary table 1) collected on Illumina HiSeq platforms recorded at the MIT Broad Institute, MA or Biospecimen Core Resource collected by TCGA consortium were downloaded from CGHub, Cancer Genomics Hub Browser, hosted at the University of California, Santa Cruz. Each sample from a tumor metastatic cancer (TM) was matched with a normal blood derived sample (NB). ii) For the MuTect 1.1.4 analysis[28] GrCh37 (Genome Reference Consortium Human Reference 37, Broad Institute variant of human genome assembly 19 (HG19)), SNP database (dbSNP) build 132.vcf, and catalogue of somatic mutations in cancer (COSMIC_54.vcf) library were referenced. dbSNP build 132.vcf is a database referenced to GrCh37 of known human germline variations derived from the 1000 genomes project. COSMIC_54.vcf is a database referenced to GrCh37 of somatically-acquired mutations found in human cancer. The call_stats.txt files containing the list of all the mutations per patient and coverage.wig in wiggle file format were generated for every matched sample as an output of MuTect 1.1.4. The mutation call_stats.txt file was queried in bash prompt to retain all the statically significant KEEP mutations under standard settings i.e. ensuring coverage of 80% power for a 0.3 allelic fraction mutation. iii) MutSig 2.0 executed on whole-exome Illumina HiSeq DNA sequencing data accesses three main sources of evidence in the data to estimate the amount of positive selection a gene underwent during tumor-igenesis: 1. Abundance of mutations relative to the background mutation rate, 2. Clustering of mutations in hotspots within the gene, and 3. Conservation of the mutated positions. MutSig 2.0 was the method of choice for the SKCM study, because it has augmented sophisticated procedures for treating the heterogeneity in per-gene, per-patient, and per-context background mutation rate. Evidence of conservation and clustering are examined by a separate part of MutSig 2.0 that performs many permutations. Mutations were inferred from raw binary alignment wes.bam files and compared to benchmarks at the cancer genome analysis multi-pipeline Firehose, which performs analyses including quality control, local realignment, single nucleotide variations identification, insertion and deletion identification, as well as inter-chromosomal and large intra-chromosomal structural rearrangement detection and mutation rate calculation. From 220,031 exonic mutations, Mutsig 2.0 produces a list of significantly mutated genes, covariates.txt. iv) The permutation algorithm InVEx 1.0.1, Intron vs Exon, was employed to efficiently model the somatic mutation rate among genes to identify the genes that most frequently harbor non-silent mutations[9]. InVEx 1.0.1 permutes coding, untranslated, and intronic mutations per nucleotide for each gene in all the patients to generate a list of genes that have the most functional impact. The polymorphism phenotyping version 2 (PPH2) library, human genome HG19, nucleotide_classes_HG19.txt, genePeptideFile_HG19, and COSMIC_54.vcf library are used as references. The coverage.wig and covariates.txt files were used as an input for InVEx 1.0.1. The QQ-plot qq.png of functional mutation burden and synonymous mutation burden significant_mutation_burden.txt were produced as an outcome of InVEx 1.0.1.

The list of enriched genes significant_mutation_burden.txt produced by filtering steps i)-iv) is analyzed for functional impact. v) Deviation from the exome-wide median of the signature of nucleotide replacement can be indicative for positive selection of cancer genes. In particular, increased transitions from cytidine to thymidine (C->T) characterize an ultraviolet-induced mutational signature (Supplementary information 2). The mutation annotation file patient.maf entries for target genes were sorted by transition type and filtered for UVA (C>A) or UVB (C>T) mediated transitions. vi) Mutations were plotted on existing experimental or modeled structures using SWISS-MODEL. In the case of transmembrane proteins, the transmembrane topology was assessed using TMpred. vii) Impact of pathways was assessed by GSEA 2.1.0 and MEMo 1.1 using scna.txt matrix file, amplified and deleted genes amp_del_gene.txt of GISTIC 2.0.21, coverage.wig file from MuTect 1.1.4, covariates.txt from Mutsig 2.0, while referencing the tcga_patient_list.txt, with a q-value threshold of 0.10 and 10 alterations. Absolute expression levels of RNASeq data from 302 patients were assessed using 5th or 95th percentile thresholds for lowly or highly expressed genes, respectively. The SKCM dataset at the TCGA datahub only contained a somatic reference file for 1 of 302 patients with metastatic melanoma at the point of the analysis. viii) For driver genes of covariates.txt in SKCM with high mutational burden in i)-iv) as well as functional impact in v)-vii), all patient.maf files in TCGA were searched for recurrence in other cancer tissues. The results are sorted covariate.maf tables for each cancer driver (Supplementary tables 9, 10).

1. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**, 685–696 (2010).
2. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).
3. Pollock, P. M. *et al.* High frequency of BRAF mutations in nevi. *Nat Genet* **33**, 19–20 (2003).
4. Chapman, P. B. *et al* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* **364**, 2507–2516 (2011).
5. Flaherty, K. T. *et al.* Improved survival with MEK inhibition in BRAF-mutated melanoma. *N Engl J Med* **367**, 107–114 (2012).
6. Sun, C. *et al.* Reversible and adaptive resistance to BRAF(V600E) inhibition in melanoma. *Nature* **508**, 118–122 (2014).
7. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
8. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
9. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
10. Krauthammer, M. *et al.* Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat Genet* **44**, 1006–1014 (2012).
11. Raphael, B. J., Dobson, J. R., Oesper, L. & Vandin, F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med* **6**, 5 (2014).
12. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–15550 (2005).
13. Hildebrandt, F., Benzing, T. & Katsanis, N. Ciliopathies. *N Engl J Med* **364**, 1533–1543 (2011).
14. Garcia-Gonzalo, F. R. *et al.* A transition zone complex regulates mammalian ciliogenesis and ciliary membrane composition. *Nat Genet* **43**, 776–784 (2011).
15. Wu, S., Romfo, C. M., Nilsen, T. W. & Green, M. R. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature* **402**, 832–835 (1999).
16. Marzese, D. M. *et al.* DNA methylation and gene deletion analysis of brain metastases in melanoma patients identifies mutually exclusive molecular alterations. *Neuro Oncol.* **16**, 1499–509 (2014).
17. Hermouet, S., Aznavoorian, S. & Spiegel, A. M. In vitro and in vivo growth inhibition of murine melanoma K-1735 cell by a dominant negative mutant alpha subunit of the Gi2 protein. *Cell Signal* **8**, 159–166 (1996).
18. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
19. Nikolaev, S. I. *et al.* Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. *Nat Genet* **44**, 133–139 (2012).
20. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
21. Wei, X. *et al.* Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat Genet* **43**, 442–446 (2011).
22. Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485**, 502–506 (2012).
23. Catanzaro, J. M. *et al.* Elevated expression of squamous cell carcinoma antigen (SCCA) is associated with human breast carcinoma. *PLoS One* **6**, e19096 (2011).
24. Brooks, A. N. *et al.* A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. *PLoS One* **9**, e87361 (2014).
25. Lee, J. H. *et al.* Evolutionarily assembled cis-regulatory module at a human ciliopathy locus. *Science* **335**, 966–969 (2012).
26. Valente, E. M. *et al.* Mutations in TMEM216 perturb ciliogenesis and cause Joubert, Meckel and related syndromes. *Nat Genet* **42**, 619–625 (2010).
27. Abaan, O. D. *et al.* The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res* **73**, 4372–4382 (2013).
28. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213–219 (2013).
29. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

Acknowledgments

We are thankful to all members of the TCGA Research Network for biospecimen collection, data acquisition, and benchmark analyses. F.V.F. is grateful for the support of grant CA154887 and CA176114 from the National Institutes of Health, National Cancer Institute.

Author contributions

F.V.F. designed the study and wrote the main article text. J.G., R.G., F.V.F. performed the data analysis of the SKCM TCGA dataset, prepared methods and supplementary information, and reviewed the final manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Guan, J., Gupta, R. & Filipp, F.V. Cancer systems biology of TCGA SKCM: Efficient detection of genomic drivers in melanoma. *Sci. Rep.* **5**, 7857; DOI:10.1038/srep07857 (2015).

Chapter 4

Hypermutation of DPYD Deregulates Pyrimidine Metabolism and Promotes Malignant Progression

Published OnlineFirst November 25, 2015; DOI: 10.1158/1541-7786.MCR-15-0403

Genomics

Molecular Cancer Research

Hypermutation of *DPYD* Deregulates Pyrimidine Metabolism and Promotes Malignant Progression

Lauren Edwards, Rohit Gupta, and Fabian Volker Filipp

Abstract

New strategies are needed to diagnose and target human melanoma. To this end, genomic analyses was performed to assess somatic mutations and gene expression signatures using a large cohort of human skin cutaneous melanoma (SKCM) patients from The Cancer Genome Atlas (TCGA) project to identify critical differences between primary and metastatic tumors. Interestingly, pyrimidine metabolism is one of the major pathways to be significantly enriched and deregulated at the transcriptional level in melanoma progression. In addition, dihydropyrimidine dehydrogenase (*DPYD*) and other important pyrimidine-related genes: *DPYS*, *AK9*, *CAD*, *CANT1*, *ENTPD1*, *NME6*, *NT5C1A*, *POLE*, *POLQ*, *POLR3B*, *PRIM2*, *REV3L*, and *UPP2* are significantly enriched in somatic mutations relative to the background mutation rate. Structural analysis of the DPYD protein dimer reveals a potential hotspot of recurring somatic mutations in the ligand-binding sites as well as the interfaces of protein domains that mediated electron transfer. Somatic mutations of *DPYD* are associated with upregulation of pyrimidine degradation, nucleotide synthesis, and nucleic acid processing while salvage and nucleotide conversion is downregulated in TCGA SKCM.

Implications: At a systems biology level, somatic mutations of *DPYD* cause a switch in pyrimidine metabolism and promote gene expression of pyrimidine enzymes toward malignant progression. *Mol Cancer Res; 14(2); 196–206. ©2015 AACR.*

Introduction

Cancer cells take advantage of distinct metabolic pathways promoting cellular proliferation or oncogenic progression. Emerging evidence highlights central metabolic pathways including glucose- and glutamine-dependent biomass production to support tumor growth (1). However, complex metabolic requirements of dividing, migrating, or nutrient and oxygen limited cancer cells suggest that tumor cells have much more complex metabolic requirements than previously appreciated (2). The Cancer Genome Atlas (TCGA) puts an even-handed view on tissue-specific genomic determinants, revolutionizing our perspective on malignancies by next-generation sequencing (3). Here we describe cross-talk between signatures of somatic mutations and gene expression of skin cutaneous melanoma (SKCM) based on RNASeq data from 471 TCGA melanoma samples. By connecting pattern of somatic mutations with responses of gene expression at a pathway level, new features of melanoma metabolism and progression are elucidated.

Systems Biology and Cancer Metabolism, Program for Quantitative Systems Biology, University of California Merced, Merced, California.

Note: Supplementary data for this article are available at Molecular Cancer Research Online (http://mcr.aacrjournals.org/).

Corresponding Author: Fabian Volker Filipp, Systems Biology and Cancer Metabolism, Program for Quantitative Systems Biology, University of California Merced, 5200 North Lake Road, Merced, CA 95343. Phone: +1-858-349-0349; E-mail: filipp@ucmerced.edu

doi: 10.1158/1541-7786.MCR-15-0403

Pyrimidine synthesis is a key metabolic bottleneck important for DNA replication in tumor cells and, therefore, represents a valuable diagnostic and therapeutic target. Early success in cancer metabolism took advantage of this characteristic by making cancer cells vulnerable to inhibition of this pathway. Heidelberger and colleagues designed fluorinated uracil-based pyrimidine analogues, which disrupted tumor DNA biosynthesis and which are to this day used to treat colorectal and breast cancer (4, 5).

To analyze TCGA SKCM dataset, we have employed a bottom-up strategy involving pathway enrichment analysis of RNASeq data and structural analysis of somatic mutations. The approach identifies *DPYD* (dihydropyrimidine dehydrogenase, Gene ID: 1806) as a pivotal factor of pyrimidine metabolism and offers a comprehensive view on how a hypermutated metabolic gene deregulates pyrimidine and nucleic acid synthesis and promotes malignant progression of melanoma.

Methods

Patient cohort

The TCGA SKCM cohort includes RNASeq data for 471 samples allowing us to extract statistical significant pattern of differential expression between solid primary tumors (TP; 103 patients) and metastatic tumors (TM; 367 patients), while there is only one dataset for blood-derived normal tissue (NB; 1 patient; Supplementary Table S1). In addition, we utilized files from whole-exome datasets of 339 patients (61 TP; 278 TM) (Supplementary Table S2; ref. 6). Clinical data including a history of drug treatment was available for 447 patients (Supplementary Table S3). The study was carried out as part of Institutional review board approved study dbGap ID 5094 "Somatic mutations in melanoma" and conducted in accordance with the Helsinki Declaration

Published OnlineFirst November 25, 2015; DOI: 10.1158/1541-7786.MCR-15-0403

of 1975. The results shown are based upon next-generation sequencing data generated by the TCGA Research Network http://cancergenome.nih.gov. Restricted access clinical, RNASeq, and whole-exome sequences were obtained from the TCGA genome data access center and the data portal.

Identification of somatic mutations

Identification of somatic mutations took advantage of components of the modular multistep filter as described (6). TCGA data portal was used for cohort selection and CGHub for access of raw data. Whole-exome sequencing data for 339 patients with primary tumor or metastatic tumor were matched with blood-derived normal reference. For the MuTect 1.1.4 analysis (7) GrCh37 (Broad Institute variant of HG19), dbSNP build 132.vcf, and COSMIC_54.vcf library were referenced. Somatic incidences file was queried in bash prompt to retain all the statically significant KEEP mutations. The coverage.wig files served as input to model and account for Intron versus Exon functional mutation burden in InVEx 1.0.1 (8). In addition, MutSig 2.0 assessed the clustering of mutations in hotspots as well as conservation of the sites (9). It is noted that the SKCM cohort contains an interesting case, patient TCGA-FW-A3R5, who has more than 20,000 mutations and an APOBEC signature (10). This patient shows multiple missense mutations in *DPYD* with nucleotide transitions according to canonical UVB signature, C>T and G>A. Including or excluding this patient had no implications on the outcome of this study.

Structural model and molecular dynamics simulation

The structural model of human DPYD was based on *Sus scrofa* X-ray structure (PDB entry 1gth) using swiss-model. Mutations were plotted on the modeled human structure and ligand proximity was evaluated by a 5Å cut-off. The solvent accessible surface of each residue of DPYD was determined on the basis of a molecular dynamics simulation over a 5 ns trajectory using GROMACS 5.0.2 (11).

Gene expression analysis and statistical analysis

Level 3 RNASeq Log2-transformed expression levels for 18,086 genes were collected for each sample. Differential expression was determined by DESeq in the R package and Student *t* test was used to determine significant differences in expression between TP and TM samples and onto metabolic pathways (12). The probability of the test statistics (*P* values) were adjusted for multiple hypotheses testing (13). When referred to genomic information, gene symbols are italicized and upper case, while protein names are upper case but not italicized. All used gene symbols are listed with gene description in the glossary in the Supplementary tables.

Results

Pathway enrichment of differential RNASeq gene expression data identifies shift in metabolism

Differential expression analysis by DESeq showed 4383 and 4811 to be significantly down- and upregulated, respectively. KEGG Pathway enrichment analysis highlights three distinct sets of pathways, metabolism, cancer signaling, and epidermal developmental markers, to be central to the changes occurring in the metastatic transition. Metabolic pathways include global metabolism (KEGG ID:01100), oxidative phosphorylation (ID:00190), pyrimidine metabolism (ID:00240), purine metabolism (ID:00230), glycosphingolipid biosynthesis (ID:00601), metabolism of cytochrome P450 (ID:00980), tyrosine metabolism (ID:00350), as well as glutathione metabolism (ID:00480) to be significantly enriched pathways with deregulated gene expression with *P* values lower than 0.001. Interestingly, metabolic pathways show comparably high enrichment as pathways known to be closely associated with an invasive, metastatic phenotype. Next to pyrimidine metabolism, focal adhesion, actin cytoskeleton regulation, and tight junctions are highly enriched in the metastatic melanoma cohort with *P* values below 1.0E−04. Pyrimidine metabolism stands out as highly enriched pathway (enrichment ratio down 3.60, ratio up 2.19, adjusted *P* value down 3.49E−10 and adjusted *P* value up 4.00E−04). There are 34 and 23, in total 57 genes in pyrimidine metabolism, which are significantly down- and upregulated, respectively (Supplementary Table S4). Pyrimidine enzymes undergoing differential expression between skin cutaneous primary and metastatic tumors include all steps in nucleoside triphosphate synthesis, DNA and RNA polymerases, as well as pyrimidine degradation. The top downregulated enzyme of pyrimidine metabolism is CDA (cytidine deaminase, Gene ID: 978, log2 change between primary and metastatic tumor: 1.66, *P* 5.06E−15), the highest upregulated enzyme is DPYD (log2 difference between primary and metastatic tumor: +1.43, *P* 2.85E−13). Genes differentially expressed in pathways in cancer include important signaling molecules in MAPK signaling like *BRAF*, *NRAS*, *KRAS*, *MAPK8*, *MAP2K1*, in WNT signaling, *CTNNB1*, *FZD1/3/4/8*, in STAT signaling, *PIAS1*, *STAT1*, *STAT5B*, in AKT signaling *AKT3*, *MTOR*, and others like *RB1*, *NFKB1*. Another remarkable theme of enriched genes during metastatic progression is dedifferentiation of melanogenesis, keratinocytes, and Wingless (WNT) signaling (Table 1). Such gene sets are important for cell differentiation of normal melanocytes, epidermal development, and pigmentation.

Hypermutation of dihydropyrimidine dehydrogenase in pyrimidine metabolism

To identify potential melanoma driver genes in pyrimidine metabolism, we assessed recurrence and statistical enrichment of somatic mutations for all pyrimidine genes in melanoma. *DPYD* stands out for its highest mutation rate above 20% for all melanoma patients and significant enrichment of somatic mutations above background mutation rate with a *q* value of 4.40E−06 (Table 2). Twelve other pyrimidine genes show high somatic mutation rates, *P* values of recurrence and conservation, or *q* values of enrichment above background mutation rate including *DPYS* (dihydropyrimidinase, Gene ID: 1807), *AK9* (adenylate kinase domain containing 1, Gene ID: 221264), *CAD* (carbamoylphosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase, Gene ID: 790), *CANT1* (calcium activated nucleotidase 1, Gene ID: 124583), *ENTPD1* (ectonucleoside triphosphate diphosphohydrolase 1, Gene ID: 953), *NME6* (NME/NM23 nucleoside diphosphate kinase 6, Gene ID: 10201), *NT5C1A* (5′-nucleotidase, cytosolic IA, Gene ID: 84618), *POLE* (polymerase (DNA directed), epsilon, Gene ID: 5426), *POLQ* (polymerase (DNA directed), theta, Gene ID: 10721), *POLR3B* [polymerase (RNA) III (DNA directed) polypeptide B, Gene ID: 55703], *PRIM2* [primase, DNA, polypeptide 2 (58 kDa), Gene ID: 5558], *REV3L* [REV3-like, polymerase (DNA directed), zeta, Gene ID: 5980], and *UPP2* (uridine phosphorylase 2, Gene ID: 151531). The majority of these genes show statistically significant somatic mutations in other cancers of the TCGA Pan-cancer cohort (Table 2). For example, *PRIM2* has recurrent somatic mutations

Published OnlineFirst November 25, 2015; DOI: 10.1158/1541-7786.MCR-15-0403

Table 1. KEGG pathway enrichment analysis of RNASeq data of 470 TCGA SKCM patients with adjusted *P* values below 1.00E−03

Pathway name	KEGG ID	Number of deregulated genes in TCGA SKCM RNASeq		Expected number		Ratio of enrichment		*P* value from hypergeometric test		Adjusted *P* value by multiple test adjustment		Number of reference genes in pathway
		Down	Up	Down	Up	Down	Up	Down	Up	Down	Up	
Metabolic pathways	ID:01100	359	216	107.77	119.87	3.33	1.80	3.16E−99	5.34E−18	6.79E−97	9.19E−17	1130
Pathways in cancer	ID:05200	76	108	31.09	34.58	2.44	3.12	1.64E−13	3.70E−28	5.88E−12	1.94E−26	326
Pyrimidine metabolism	ID:00240	*34*	*23*	*9.44*	*10.50*	*3.60*	*2.19*	*1.30E−11*	*2.00E−04*	*3.49E−10*	*4.00E−04*	*99*
Purine metabolism	ID:00230	42	44	15.45	17.19	2.72	2.56	1.37E−09	3.14E−09	3.27E−08	1.35E−08	162
Melanogenesis	ID:04916	31	27	9.63	10.71	3.22	2.52	2.41E−09	4.49E−06	5.18E−08	1.18E−05	101
Endocytosis	ID:04144	47	60	19.17	21.32	2.45	2.81	5.73E−09	5.40E−14	1.03E−07	4.73E−13	201
Phagosome	ID:04145	38	50	14.59	16.23	2.60	3.08	2.86E−08	1.40E−13	4.10E−07	1.09E−12	153
Peroxisome	ID:04146	24	21	7.53	8.38	3.19	2.51	1.84E−07	5.44E−05	2.20E−06	1.00E−04	79
Spliceosome	ID:03040	32	34	12.11	13.47	2.64	2.52	2.41E−07	2.68E−07	2.73E−06	8.28E−07	127
Focal adhesion	ID:04510	43	68	19.07	21.22	2.25	3.21	3.12E−07	5.81E−19	3.35E−06	1.11E−17	200
Wnt signaling pathway	ID:04310	35	46	14.31	15.91	2.45	2.89	5.01E−07	1.59E−11	4.49E−06	1.04E−10	150
MAPK signaling pathway	ID:04010	52	86	25.56	28.43	2.03	3.02	6.02E−07	1.11E−21	5.18E−06	2.91E−20	268
Neurotrophin signaling	ID:04722	31	36	12.11	13.47	2.56	2.67	7.79E−07	2.45E−08	6.44E−06	8.72E−08	127
Actin cytoskeleton regulation	ID:04810	43	68	20.31	22.60	2.12	3.01	1.86E−06	2.51E−17	1.45E−05	3.51E−16	213
Protein processing in ER	ID:04141	36	58	15.74	17.50	2.29	3.31	1.91E−06	3.70E−17	1.45E−05	4.86E−16	165
Tight junction	ID:04530	30	32	12.59	14.00	2.38	2.29	5.59E−06	6.03E−06	3.43E−05	1.54E−05	132
Ubiquitin-mediated proteolysis	ID:04120	30	43	12.87	14.32	2.33	3.00	9.00E−06	1.80E−11	5.09E−05	1.15E−10	135
Calcium signaling pathway	ID:04020	36	52	16.88	18.78	2.13	2.77	1.03E−05	4.91E−12	5.68E−05	3.33E−11	177
GnRH signaling pathway	ID:04912	24	27	9.63	10.71	2.49	2.52	2.13E−05	4.49E−06	1.00E−04	1.18E−05	101
Small-cell lung cancer	ID:05222	21	26	8.11	9.02	2.59	2.88	3.67E−05	4.06E−07	2.00E−04	1.24E−06	85
Oocyte meiosis	ID:04114	25	29	10.68	11.88	2.34	2.44	4.48E−05	4.05E−06	2.00E−04	1.10E−05	112
ECM-receptor interaction	ID:04512	21	24	8.11	9.02	2.59	2.66	3.67E−05	5.36E−06	2.00E−04	1.39E−05	85
N-Glycan biosynthesis	ID:00510	15	16	4.67	5.20	3.21	3.08	3.24E−05	2.71E−05	2.00E−04	6.39E−05	49
Renal cell carcinoma	ID:05211	18	33	6.68	7.43	2.70	4.44	7.37E−05	1.18E−14	3.00E−04	1.13E−13	70
Acute myeloid leukemia	ID:05221	16	23	5.44	6.05	2.94	3.80	5.84E−05	5.29E−09	3.00E−04	2.10E−08	57
VEGF signaling pathway	ID:04370	19	25	7.25	8.06	2.62	3.10	7.13E−05	1.41E−07	3.00E−04	4.42E−07	76
Fc gamma phagocytosis	ID:04666	21	38	8.96	9.97	2.34	3.81	2.00E−04	5.66E−14	7.00E−04	4.75E−13	94
RNA transport	ID:03013	29	49	14.40	16.02	2.01	3.06	2.00E−04	3.31E−13	7.00E−04	2.48E−12	151
Adipocytokine signaling	ID:04920	17	22	6.49	7.21	2.62	3.05	2.00E−04	1.07E−06	7.00E−04	3.16E−06	68
Phosphatidylinositol signaling	ID:04070	18	31	7.44	8.27	2.42	3.75	3.00E−04	1.94E−11	9.00E−04	1.20E−10	78
Gap junction	ID:04540	20	30	8.58	9.55	2.33	3.14	3.00E−04	5.82E−09	9.00E−04	2.26E−08	90
Melanoma	ID:05218	17	20	6.77	7.53	2.51	2.66	3.00E−04	3.29E−05	9.00E−04	7.51E−05	71
Natural killer cell cytotoxicity	ID:04650	—	61	—	14.43	—	4.23	—	2.13E−24	—	7.46E−23	136
Cell adhesion molecules	ID:04514	—	54	—	14.11	—	3.83	—	2.46E−19	—	5.17E−18	133
Jak-STAT signaling pathway	ID:04630	—	59	—	16.44	—	3.59	—	2.37E−19	—	5.17E−18	155
Toll-like receptor signaling	ID:04620	—	41	—	10.82	—	3.79	—	7.37E−15	—	7.37E−14	102
TGFβ signaling pathway	ID:04350	—	31	—	8.91	—	3.48	—	1.81E−10	—	9.27E−10	84
NOD-like receptor signaling	ID:04621	—	24	—	6.15	—	3.90	—	1.34E−09	—	5.99E−09	58
Allograft rejection	ID:05330	—	18	—	3.93	—	4.59	—	6.72E−09	—	2.52E−08	37
Antigen processing presentation	ID:04612	—	26	—	8.06	—	3.22	—	3.22E−08	—	1.13E−07	76
Basal transcription factors	ID:03022	—	15	—	3.93	—	3.82	—	2.26E−06	—	6.33E−06	37
Oxidative phosphorylation	ID:00190	76	—	12.59	—	6.04	—	5.25E−43	—	5.64E−41	—	132
Proteasome	ID:03050	19	—	4.20	—	4.53	—	5.22E−09	—	1.02E−07	—	44
RNA polymerase	ID:03020	14	—	2.77	—	5.06	—	9.76E−08	—	1.31E−06	—	29
Glycosphingolipid biosynthesis	ID:00601	13	—	2.48	—	5.24	—	1.67E−07	—	2.11E−06	—	26
Metabolism of xenobiotics P450	ID:00980	22	—	6.77	—	3.25	—	4.02E−07	—	4.12E−06	—	71
Tyrosine metabolism	ID:00350	16	—	3.91	—	4.09	—	4.56E−07	—	4.40E−06	—	41
Glutathione metabolism	ID:00480	17	—	4.77	—	3.57	—	1.95E−06	—	1.45E−05	—	50
Arginine and proline metabolism	ID:00330	17	—	5.15	—	3.30	—	6.44E−06	—	3.85E−05	—	54
Glycosylphosphatidylinositol	ID:00563	11	—	2.38	—	4.61	—	7.33E−06	—	4.26E−05	—	25
Drug metabolism P450	ID:00982	20	—	6.96	—	2.87	—	1.09E−05	—	5.86E−05	—	73
Fructose mannose metabolism	ID:00051	12	—	3.43	—	3.50	—	7.82E−05	—	3.00E−04	—	36
Sulfur relay system	ID:04122	6	—	0.95	—	6.29	—	1.00E−04	—	4.00E−04	—	10
Arachidonic acid metabolism	ID:00590	16	—	5.63	—	2.84	—	9.20E−05	—	4.00E−04	—	59
Sphingolipid metabolism	ID:00600	12	—	3.81	—	3.15	—	2.00E−04	—	7.00E−04	—	40
Glycosaminoglycan degradation	ID:00531	8	—	1.81	—	4.41	—	2.00E−04	—	7.00E−04	—	19
Hedgehog signaling pathway	ID:04340	15	—	5.34	—	2.81	—	2.00E−04	—	7.00E−04	—	56
Drug metabolism	ID:00983	14	—	4.96	—	2.82	—	3.00E−04	—	9.00E−04	—	52

NOTE: Pathway enrichment on pyrimidine metabolism is highlighted in italic.

Published OnlineFirst November 25, 2015; DOI: 10.1158/1541-7786.MCR-15-0403

Table 2. Analysis of somatic mutations in pyrimidine metabolism of 339 whole-exome sequenced TCGA SKCM patients (KEGG pathway ID:00240)

Symbol	Gene ID	Mutation rate	*P* value MutSig	*q* value MutSig	*q* value InVEx	TCGA Pan-cancer Tissue (*P* and *q* value MutSig)
AK9	221264	5.00%	*4.30E−01*	*2.66E−01*	8.80E−04	LAML (P = NA; q = 1.97E−02)
CAD	790	5.00%	6.77E−03	4.06E−02	2.71E−02	LGG (P = 2.87E-02; q = 8.81E−02)
CANT1	124583	3.00%	4.45E−02	2.86E−02	2.71E−02	PRAD (P = 4.13E−02; q = 4.62E−03)
DPYD	1806	20.50%	*1.00E+00*	*7.13E−01*	4.40E−06	HNSC (P = 1.40E−02; q = 7.36E−02)
DPYS	1807	7.00%	*4.87E−01*	*5.32E−01*	8.25E−03	PAAD (P = 3.02E−02; q = 3.14E−07)
ENTPD1	953	4.00%	*1.53E−01*	*1.61E−01*	1.58E−02	—
NME6	10201	1.00%	*1.22E−01*	*1.76E−01*	1.58E−02	COAD (P = 1.77E−01; q = 3.68E−02)
NT5C1A	84618	1.00%	2.60E−02	9.72E−02	*6.44E−01*	STAD (P = 2.68E−02; q = 9.90E−02)
POLE	5426	6.00%	*2.69E−01*	*6.22E−01*	1.51E−03	ACC (P = 5.86E−03; q = 2.02E−02) KIRC (P = 2.85E−02; q = 9.37E−02) UCEC (P = 1.29E−02; q = 6.90E−02)
POLQ	10721	8.00%	*3.55E−01*	*7.01E−01*	8.14E−04	LUSC (P = 3.40E−02; q = 1.43E−01)
POLR3B	55703	5.00%	*7.59E−01*	*9.47E−01*	1.32E−03	UCEC (P = 2.24E−03; q = 3.96E−03)
PRIM2	5558	6.00%	3.52E−02	1.00E−05	7.63E−04	HNSC (P = 5.58E−03; q = 3.34E−06) LUSC (P = 3.40E−04; q = 2.27E−04)
REV3L	5980	6.00%	*1.47E−01*	*4.29E−01*	*2.20E−01*	BRCA (P = 6.48E−03; q = 2.62E−02)
UPP2	151531	2.00%	1.22E−02	2.99E−02	7.63E−04	LUAD (P = 3.90E−01; q = 3.40E−02)

NOTE: *P* values above 0.05 or *q* values above 0.10 are in italic. The column TCGA Pan-cancer lists significant somatic events in pyrimidine metabolism in other cancer tissues.

at residue E221, G334, P391, and is also significantly mutated in HNSC (P = 0.00558; q = 3.34E−06) or LUSC (P = 0.000339; q = 2.27E-04). Somatic *DPYD* mutations coincide with deleterious mutations of gatekeeper and caretaker genes *TP53*, *BRCA1*, *FAT3*, *FAT4*, *PTPRD*, and *SPEN* with *P* values below 1.0E−06 and *q* values below 1.0E−04 connecting to DNA maintenance and stability. In comparison with other TCGA tissues, *DPYD* is the top somatically mutated gene in pyrimidine metabolism, affecting 67 patients of 278 whole-exome sequenced metastatic melanoma (Fig. 1A, Supplementary Table S5). There are in total 74 non-synonymous mutations in *DPYD* detected, including incidents where two or three residues of the same polypeptide chain are affected. Examples of multiple mutations coinciding in *DPYD* are S204F and D949N in patient TCGA-EE-A2MI, or V396I, G851R, E937K in patient TCGA-FW-A3R5. The nucleotide signature of somatic transitions of *DPYD* tracks with the validated mutational signature of melanoma identified across human cancers (6, 10), and is governed by UVB-associated C>T/G>A transitions (Supplementary Table S6; Fig. 1B).

Structural hotspots of somatic hypermutation of DPYD in ligand-binding sites as well as interfaces of protein domains

To decipher functional implications of somatic mutations of DPYD, we analyzed the domain distribution, polymorphism phenotyping v2 (PPH2) scores, solvent accessible surface, proximity to ligands, and mutational recurrence of all identified somatic mutations (Supplementary Table S7). The cytosolic dihydropyrimidine dehydrogenase (EC 1.3.1.2; OMIM 612779 and 274270) is the initial and rate-limiting enzyme in the catabolism of pyrimidines. It reduces the pyrimidine bases thymine and uracil in a NADPH-dependent manner. The highly conserved homodimeric 1025-residue protein contains four 4Fe-4S-clusters, one FAD, and one FMN in the active site cavity of each subunit (Figs. 2 and 3A and B). A special electron transfer pathway involves the 4Fe-4S-clusters of both subunits, so that DPYD comprises two independent electron transfer chains and is active just as a dimer (14, 15). The somatic mutations affect 153 residues of 1025 in the TCGA Pan-cancer dataset (TCGA: 181 mutation affecting 153 unique residues; SKCM: 74 mutations affecting 60

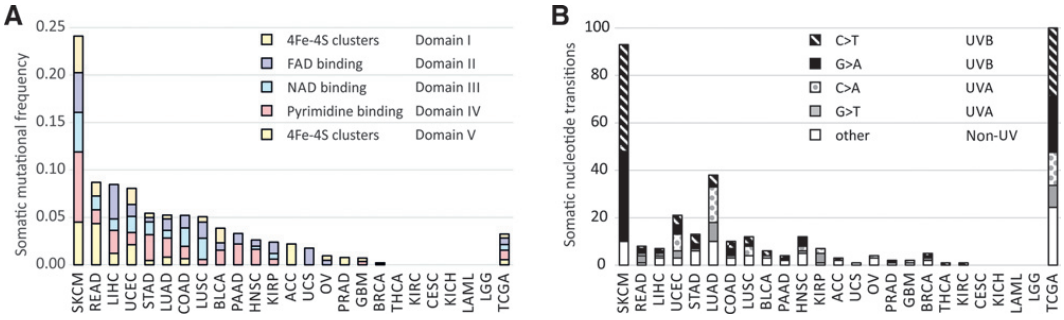


Figure 1.
SKCM stands out in frequency and nucleotide signature of somatic mutations of dihydropyrimidine dehydrogenase (*DPYD*) across Pan-cancer patients of The Cancer Genome Atlas (TCGA). A, somatic mutational frequency of *DPYD* mutations in melanoma and across TCGA PAN-cancer patients. B, nucleotide signature of somatic transitions of *DPYD* mutations analyzed by UV-type for melanoma (UVB associated with C>T and G>A, and UVA-type with G>T and C>A) shown as absolute count of mutational incidences. Fraction of nucleotide signature of somatic transitions of *DPYD* mutations is given at the right across all TCGA Pan-cancer tissues in TCGA.

Published OnlineFirst November 25, 2015; DOI: 10.1158/1541-7786.MCR-15-0403

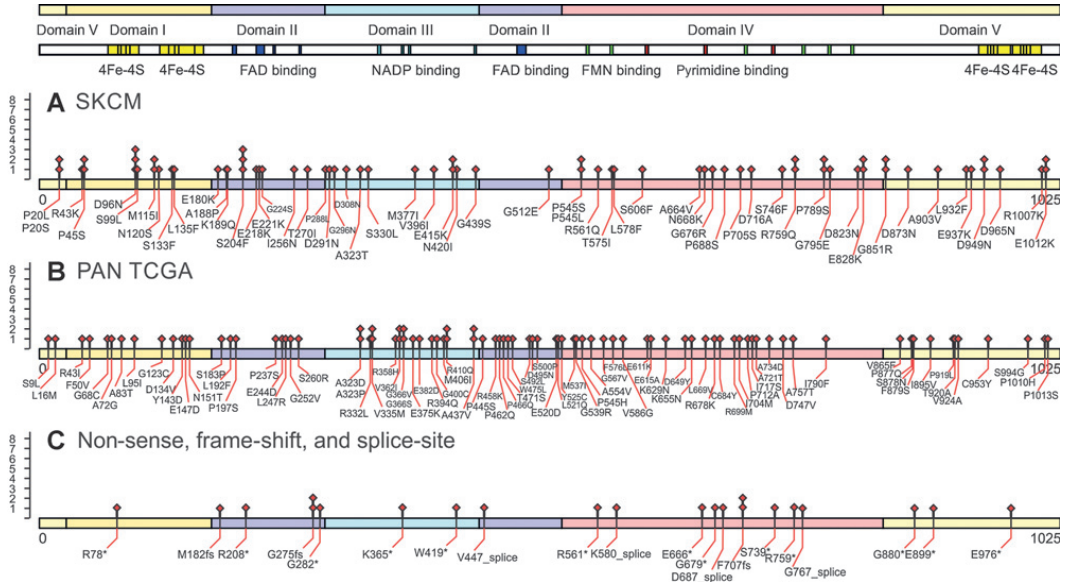


Figure 2.
Somatic mutational landscape of *DPYD* mutations in melanoma and across TCGA PAN-cancer patients. Somatic mutations are indicated on the protein sequence of *DPYD*, NCBI Gene ID 1806, A, for skin cutaneous melanoma (SKCM) and B, for 24 TCGA tissues with missense *DPYD* mutations. C, non-sense, frame-shift, and splice-site mutations are indicated separately. Functional domains are annotated according uniprot entry Q12882 and 11179210: Domain I N-terminal 4Fe-4S clusters (27-172, yellow); domain II FAD-binding domain (173-286, 442-524, blue); domain III NADPH-binding domain (287-441, cyan); domain IV FMN and pyrimidine-binding domain (525-848, red; FMN binding in green); domain V C-terminal 4Fe-4S clusters (1-26, 848-1025, yellow).

unique residues). Recurrent somatic mutations are detected in all five functional domains of DPYD (Fig. 2). While the pyrimidine-binding domain has the highest mutational count with 55 mutations in total, correction for domain length shows that domains II–IV, which bind metabolite substrates and cofactors, show enriched mutation frequency (37 mutations over 197 residues in FAD-binding domain II domain; 33 mutations over 155 residues in NADPH-binding domain III domain; 55 mutations over 323 residues in the FMN and pyrimidine-binding domain IV). The functional impact of somatic missense mutations was quantified using PPH2 scores and plotted onto the protein structure (Fig. 3C). Molecular dynamics simulations in combination with computation of solvent accessible surface revealed an accumulation of damaging, missense somatic mutations in the core of the protein (PPH2 scores > 0.95; making up 52% of somatic mutations; solvent accessible surface values of affected residues below 0.50). In contrast, possibly benign somatic changes are surface-bound (PPH2 scores < 0.50; making up 37% of somatic mutations; elevated average value above 0.65 for the solvent accessible surface, Fig. 3C). The location of somatic mutations predicted to be damaging based on high functional PPH2 scores and lack of surface accessibility coincides with detected somatic recurrence in SKCM as well as in the TCGA Pan-cancer cohort (Fig. 3A, C and D).

Recurring somatic missense mutations D291N, V335M, and A437 frame the nucleotide-binding site (Fig. 4A). In addition, the NADPH-binding domain houses mutations A323D/P/T, V362I, G366I/S/V, as well as V365 nonsense mutation. Somatic mutations involved in the hydrogen bond network and within less than 3.5Å to the FAD ligand are P197S, E218K, G224S/V, S260R, and S492L. The mutation L135F is located at the catalytic route between FAD and N-terminal 4Fe-4S cluster (Fig. 4A). There are 20 non-recurring mutations that populate residues involved in the electron transfer between the four 4Fe-4S cluster. Between C-terminal 4Fe-4S cluster and electron entry site of FMN, there is the somatic mutation E611K (Fig. 4B). Mutations D949N, D965N, A554V, and E615A line up between C-terminal 4Fe-4S domain V and pyrimidine binding domain IV. The pyrimidine substrate binding pocket is framed by recurrent somatic mutations T575I, E611K, N668K, and G795E (Fig. 4B). A hotspot of recurring somatic mutations D96N (3x), S204F (3x), M115I (2x), G851R (2x), E828K (2x), and P545H/L/S is at the interface between domain I, II, and IV (Fig. 3C and D). S204F is a structural residue of FAD binding domain II, linking the three domains forming the large cleft of DPYD. S204F is part of an α-helix which directly bridges to residues L95I and N1200S, affected by somatic mutations. Similarly, Q828 is an anchor at domain IV and spans a hydrogen bond network to domain I via D96N, S99L, and M115I, affected by somatic mutations.

Cross-talk between *DPYD* mutations and gene expression of the pyrimidine pathway

Next, we addressed whether the mutational and transcriptional signature of pyrimidine metabolism in melanoma follows a distinct pattern. More than half of the pyrimidine enzymes are differentially expressed between primary and metastatic tumor, showing distinct clusters in key steps of pyrimidine nucleoside triphosphate synthesis, DNA and RNA

Published OnlineFirst November 25, 2015; DOI: 10.1158/1541-7786.MCR-15-0403

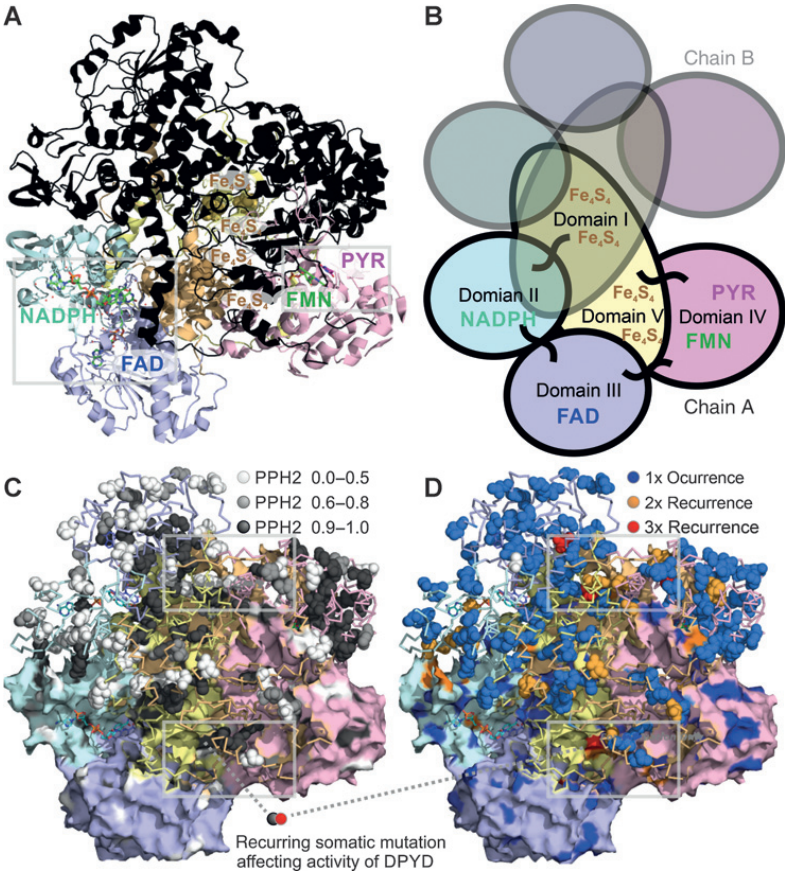


Figure 3.
Structural analysis of DPYD dimer reveals hotspots of somatic mutations in ligand binding sites as well as interfaces of protein domains. A, The 4Fe-4S cluster domain I of DPYD (uniprot entry Q12882 and PDB entry 1gth) is shown in orange, the FAD-binding domain II in blue, the NADPH-binding domain III in cyan, the pyrimidine- and FMN-binding domain IV in red, and the 4Fe-4S cluster domain V is shown in yellow. The lower monomer is color-coded according to domains and contains small-molecule ligands as sticks, the upper monomer is overlaid in black. B, the five domains of each DPYD monomer are color coded according to protein domains showing that the electron transfer chain crosses the dimer interface twice. The 4Fe-4S cluster domains I and V are intertwined and form extended inter-domain contacts. C, functional impact of somatic missense mutations plotted as PPH2 scores onto surface (lower monomer; showing surface-bound mutations) and ribbon of DPYD structure (emphasizing accumulation of deleterious mutations in protein core in black). D, recurrent missense somatic mutations in ligand-binding sites as well as interfaces of protein domains of DPYD. Gray frames indicate regions of coincidence of high functional PPH2 scores, lack of surface accessibility coincides, and somatic recurrence shown enlarged in Fig. 4.

synthesis, as well as pyrimidine degradation (Fig. 5A). At the mutational level there is also a progressive enrichment of somatic mutations in the SKCM cohort comparing primary and metastatic tumors. *DPYD* is mutated in 11.5% of primary tumors, while in metastatic tumors somatic mutations are detected in 22.5% of all whole exome–sequenced samples (Fig. 5B). Comparison of gene expression and mutational data of *DPYD* and other key pyrimidine enzymes in SKCM shows that enrichment of somatic mutations in metastatic tumors coincides with elevated expression levels (Fig. 5). The gene expression signature is significantly enhanced by somatic *DPYD* mutation. The expression level of pyrimidine enzymes changes in SKCM metastatic tumor samples with *DPYD* wild-type status in comparison with metastatic tumor samples with *DPYD* mutation with *P* value below 0.05. In addition, the direction of expression change in melanoma progression (up or down from tumor to metastasis) is the same as the difference between *DPYD* mutation and wild-type (up or down from wild-type to mutation, respectively). The observed deregulated gene expression of pyrimidine enzymes, including *DPYD* itself, correlates with metastatic progression and is enhanced by somatic *DPYD* mutations (Fig. 5B and C). Almost all differentially expressed pyrimidine enzymes (with the exception of *POLR3D*), which show up- or downregulation between primary and metastatic tumors, show progressive increase or decrease with *DPYD* mutation, respectively (Supplementary Table S8). Somatic mutations of *DPYD* enhance the metastatic progression signature of melanoma (Fig. 5D).

Bifurcation of pyrimidine metabolism in metastatic melanoma

Somatic mutations and differential gene expression have severe implications for the metabolic network of pyrimidine metabolism. Mapping of gene expression data onto a pathway map of pyrimidine metabolism (modeled after KEGG pathway ID:00240) revealed a 2-fold separation. Pyrimidine degradation initiated by enzymes DPYD and DPYS is significantly upregulated (Fig. 6A). Enzymes TYMP (thymidine phosphorylase, Gene ID: 1890), UPP1 (uridine phosphorylase 1, Gene ID: 7378), CDA, TK1 (thymidine kinase 1, soluble, Gene ID: 7083), TK2 (thymidine kinase 2, mitochondrial, Gene ID: 7084), UCK1 (uridine-cytidine kinase 1, Gene ID: 83549), and DTYMK [deoxythymidylate kinase (thymidylate kinase), Gene ID: 1841] salvaging pyrimidines are significantly downregulated. Enzymes DCK (deoxycytidine kinase, Gene ID: 1633),

Published OnlineFirst November 25, 2015; DOI: 10.1158/1541-7786.MCR-15-0403

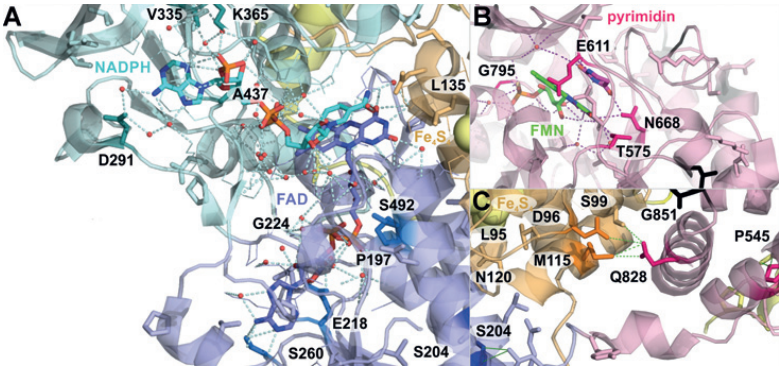


Figure 4.
Recurring somatic missense mutations frame ligand–binding sites of DPYD modulating its enzymatic activity. A, recurring somatic mutations in ligand-binding sites of NADPH, FAD, and 4Fe-4S clusters at the interface of domain I–III. B, recurring somatic mutations in pyrimidine and FMN-binding site affect hydrogen bond network of enzymatic effector domain IV. C, accumulation of recurring somatic mutations at domain interface shows hinge-residue Q828 in domain IV and its connectivity to domain I via hydrogen bonds. Location of expanded regions on global map of protein structure of the DPYD dimer is indicated as gray frames in Fig. 3.

CANT1, AK9, and NME7 (NME/NM23 family member 7, Gene ID: 29922) providing pyrimidine nucleoside triphosphate and nucleic acid building enzymes POLA1 [polymerase (DNA directed), alpha 1, Gene ID: 5422], POLK [polymerase (DNA directed) kappa, Gene ID: 51426], POLQ, POLR1A [polymerase (RNA) I polypeptide A, Gene ID: 25885], POLR1B [polymerase (RNA) I polypeptide B, Gene ID: 84172], POLR2B [polymerase (RNA) II (DNA directed) polypeptide B, Gene ID: 5431], POLR2D [polymerase (RNA) II (DNA directed) polypeptide D, Gene ID: 5433], POLR3A [polymerase (RNA) III (DNA directed) polypeptide A, Gene ID: 11128], POLR3D, REV3L, PRIM1 (primase, DNA, polypeptide 1, Gene ID: 5557), PNPT1, and TWISTNB (TWIST neighbor, Gene ID: 221830) are upregulated. This is enforced by a significant downregulation of pyrimidine nucleoside triphosphate–degrading enzymes ENTPD3 (ectonucleoside triphosphate diphosphohydrolase 3, Gene ID: 956), ENTPD8 (ectonucleoside triphosphate diphosphohydrolase 8, Gene ID: 377841), ITPA [inosine triphosphatase (nucleoside triphosphate pyrophosphatase), Gene ID: 3704], NT5C (5′, 3′-nucleotidase, cytosolic, Gene ID: 30833), and NT5M (5′,3′-nucleotidase, mitochondrial, Gene ID: 56953). In addition, enzymes RRM1 (ribonucleotide reductase M1, Gene ID: 6240), RRM2B [ribonucleotide reductase M2 B (TP53 inducible), Gene ID: 50484] required for anabolic conversion of uracil nucleosides to thymidine diphosphate nucleosides are upregulated, while enzymes TYMP- or CDA-mediating production of uracil, uridine, and deoxyuridine are downregulated. At a pathway level, somatic mutations in SKCM patients are most frequently observed in enzymes DPYD, DPYS, ENTPD1, CANT1, and UPP2 of pyrimidine degradation, as well as degradation of pyrimidine nucleoside triphosphates (q value below 0.03; Fig. 6B; Supplementary Table S5). Somatic mutations of *DPYD* significantly enhance the signature of pyrimidine nucleoside triphosphates and nucleic acid–generating enzymes CMPK1, AK9, NME7, POLA1, POLD3, POLK, POLR3F, POLR3G, PRIM2, and TWISTNB ($P < 0.05$; Fig. 6C, Supplementary Table S8). In addition, *DPYD*-mutated samples show significantly increased *DPYD* transcript levels ($P < 0.05$). Taken together, the combined mutational and gene expression analysis shows a shift towards pyrimidine nucleoside triphosphates and nucleic acid synthesis, and disconnection from pyrimidine salvage and degradation (Fig. 6D; Supplementary Table S4).

Discussion

Enzymes in pyrimidine metabolism undergo a significant deregulation at the gene expression level in the transition from skin cutaneous primary tumors toward metastatic tumors (Table 1; Fig. 5). This transition is accompanied by an enrichment of somatic mutations of *DPYD* (Table 2; Figs. 1–4).

The mutational analysis identified more than 130 unique and novel recurrent somatic mutations in *DPYD*, including recurrent missense, nonsense and splice site mutations (Supplementary Table S7). In addition, we were able to confirm frequently recurring deleterious mutations S204F and G275 frame shift (Fig. 4A and C; refs. 8, 16). The mutational burden of *DPYD* after correction for background rate is equally high as established melanoma drivers and shows significant enrichment with a q value of 4.40E−06 (Table 2; ref. 8). An emerge theme in cancer genomics, facilitated by the advent of deep sequencing data of large patient cohorts, is that the mutational landscape of proto-oncogenes and tumor suppressors is more diverse than anticipated (6). Structural analysis of cancer driver *BRAF* showed unprecedented events in the RAS-binding domain interface and the ATP-binding pocket aside from established p.V600E/K/R/D substitutions. Detailed topological analysis of the DPYD dimer reveals structural hotspots in ligand-binding sites and interfaces of protein domains of DPYD. Events with three-time recurrence are detected in each of the functional domains with p.D96N (Fe-S cluster I–II), p.S204F (FAD-binding domain), p.A323T/P/D (NADPH-binding domain), and p.P545S/L/H (pyrimidine-binding domain). There are distinct areas of interest with high density of somatic recurrence of mutations in DPYD (Fig. 4). Two NADPH-binding loops between V335 and G366 positions the nucleotide and initiate the electron transfer. Somatic mutations V335M, A437M, D291N, V362I, and G366I/S/V closely frame the nucleotide-binding site and are expected to have reduced NADPH binding, similarly to the reduced affinity of reported variant G366A (Fig. 4A; ref. 17). Recurring somatic mutations T575I, E611K, N668K, and G795E in pyrimidine and FMN-binding site affect hydrogen bond network of enzymatic effector domain IV (Fig. 4B).

The interface between FAD-binding domain II, N-terminal 4Fe-4S cluster domain I, and pyrimidine-binding domain IV stands out for high-frequency recurrences of somatic mutation (Fig. 3C and D). E828 is engaged in a tight hydrogen bonding network to

Published OnlineFirst November 25, 2015; DOI: 10.1158/1541-7786.MCR-15-0403

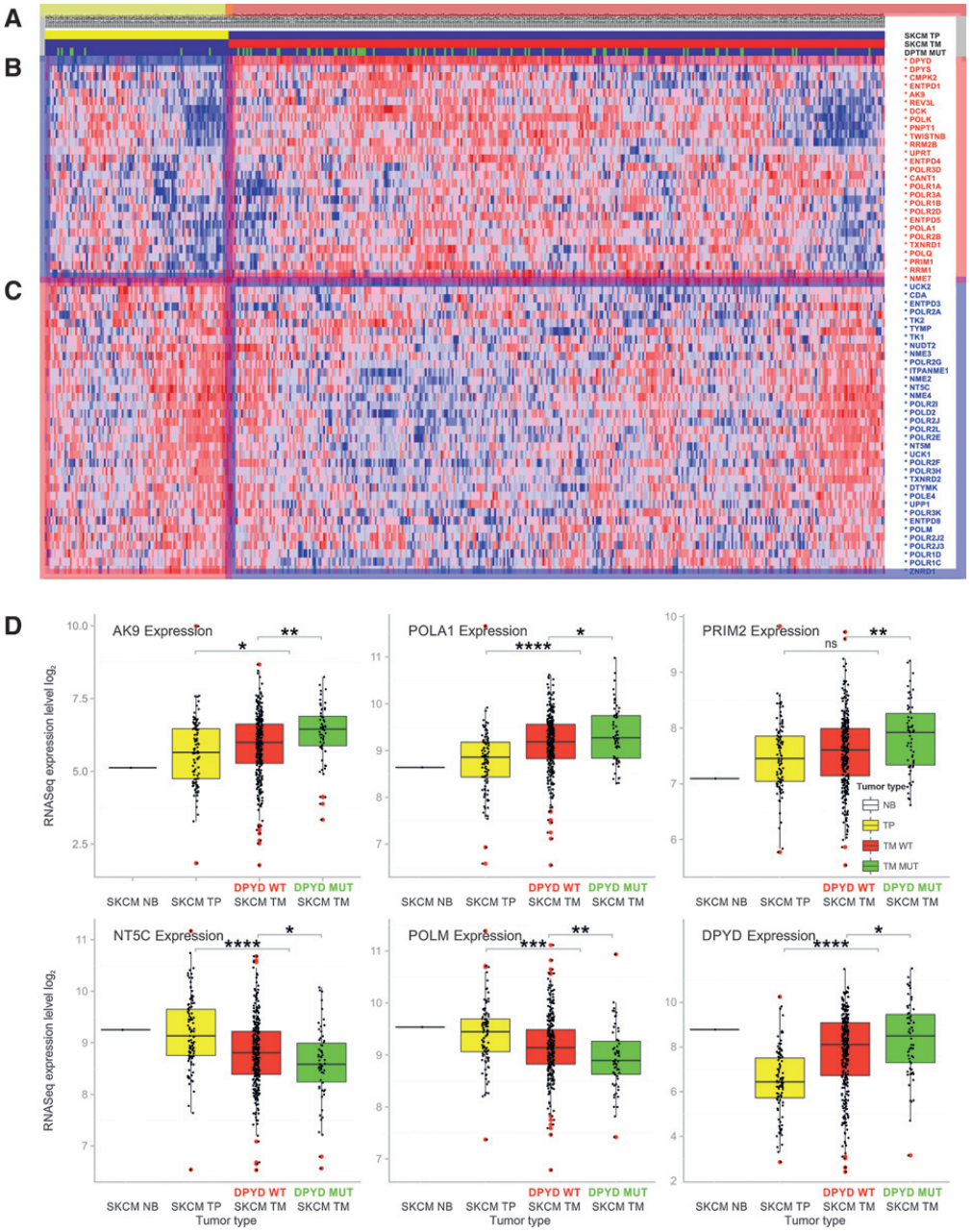


Figure 5.

Gene expression signature of pyrimidine metabolism in the progression of metastatic melanoma. A, tumor status of TCGA SKCM patients. Solid primary tumors (TP, yellow) are marked in the first row; metastatic tumors (TM, red, second row) are marked in the second row. Mutational status of DPYD is indicated in the third row (*DPYD* MUT, green). B, significant upregulation and downregulation (C) of genes in pyrimidine metabolism between skin cutaneous primary and metastatic tumors. Genes with *P* values below 0.05 are marked with an asterisk next to the gene symbol. D, impact of *DPYD* mutations on gene expression of pyrimidine enzymes is shown in green. Tumor progression of skin cutaneous melanoma (SKCM) cohort is shown for normal tissue (NB, black), solid primary tumor (TP, yellow), metastatic tumor with *DPYD* WT status (TM, *DPYD* WT, red), and metastatic tumor with *DPYD* mutations (TM, *DPYD* MUT, green). Box plots depict data distributions through quartiles. Asterisks above plots indicate results of statistical significance tests (Student *t* test: *, $P \leq 0.05$; *, $P \leq 0.01$; *, $P \leq 0.001$; *, $P \leq 0.0001$; ns, $P \geq 0.05$).

Published OnlineFirst November 25, 2015; DOI: 10.1158/1541-7786.MCR-15-0403

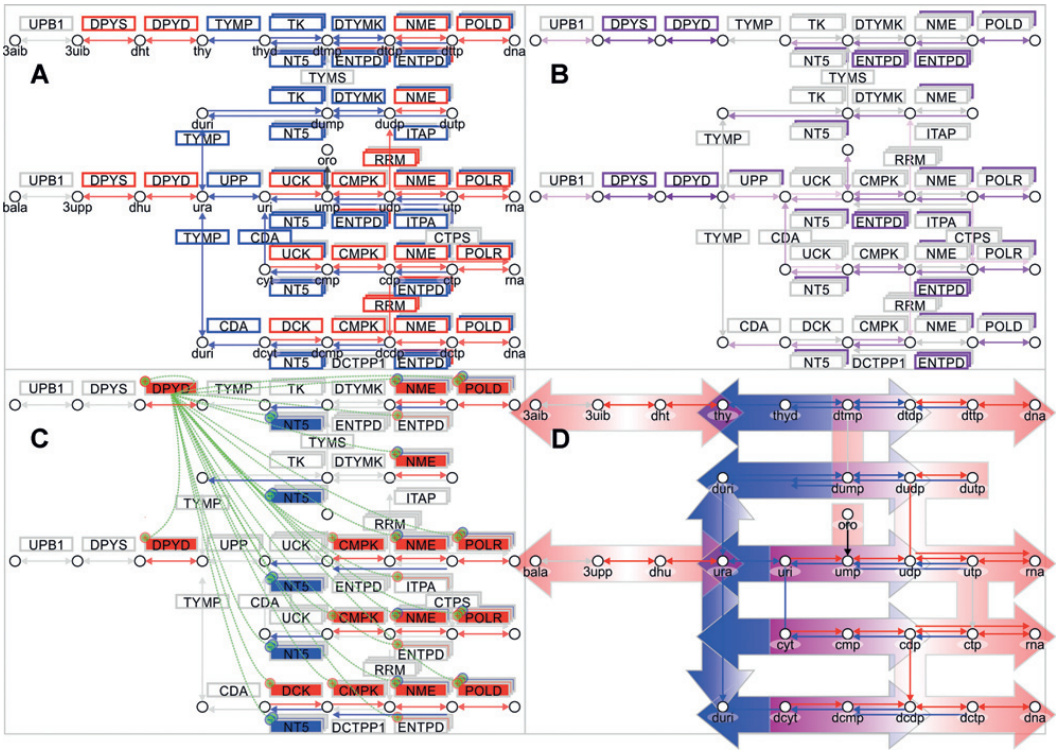


Figure 6.
Somatic mutations of *DPYD* in TCGA SKCM enhance metastatic signature of melanoma and promote deregulation of the pyrimidine pathway toward malignant cancer progression. A, gene expression signature is plotted onto pathway map of pyrimidine metabolism in metastatic melanoma (KEGG pathway ID:00240). On the left side, enzymes DPYD, DPYS, and UPB1 are responsible for pyrimidine degradation, in the center TYMP connects uracil derivatives, and at the right pyrimidine kinases DTYMK and CMPK, nucleoside diphosphate kinases NME, and polymerases POLD and POLR provide synthesis of DNA and RNA nucleic acids. B, frequency of somatic mutations in pyrimidine enzymes is color coded from 0% in gray to 25% in purple in metastatic melanoma (Supplementary Table S5). C, impact of *DPYD* mutations on enzymes of pyrimidine metabolism is indicated by regulatory symbols and shading of enzyme boxes (red, plus) for enhancement and suppression (blue, minus; Supplementary Table S8). Somatic frequency of mutations, enhanced gene expression signature of patients with *DPYD* mutations are provided in Supplementary tables. D, pathway map shows separation and directionality of upregulated pyrimidine degradation (red, left) and nucleic acid synthesis (red, right) by downregulated uracil and thymidine salvage (blue, center). Concerted dysregulation of metabolic enzymes at a pathway level contributes to bifurcation of pyrimidine metabolism. The systems biology maps depict metabolites as circles, reactions with their respective directionality as arrows, and enzymes as boxes. Enzymes of anabolic direction are shown above reactions and enzymes of catabolic direction below reactions. Staggered boxes indicate metabolic redundancy that multiple genes encode enzymes for the reaction.

D96, S99, and M115, which are also affected by somatic mutations (Fig. 4C). E828K has demonstrated higher DPYD activity (18), stressing the importance of this hydrogen bonding network. Somatic mutations in highly recurring sites D96N, S204F, P545L/S/H is associated with high PPH2 values of 1.0 indicating possibly damaging outcome of DPYD function. In contrast, the somatic mutation A323T has been shown to be benign with enzymatic activity close to wild-type (18). Overall, the functional analysis of somatic mutations shows strong agreement with a comparative *in vitro* analyses of *DPYD* variants that somatic mutations have reduced DPYD activity (Supplementary Table S7; refs. 17, 18).

Pyrimidine enzymes undergo differential upregulation between skin cutaneous primary and metastatic tumors in key steps of nucleoside triphosphate, DNA and RNA synthesis, as well as pyrimidine degradation. Somatic mutations can generate metabolic bottlenecks and reroute metabolic paths. Visualization of somatic incident at a pathway level helps identifying such bottlenecks (Fig. 6). The intricate network pyrimidine metabolism has built-in redundancy, where enzymatic steps can be encoded by different genes or enzymes can recognize and process multiple substrates. Furthermore, the pathway contains steps for conversion between uridine and thymidine nucleotides as well as for salvage of pyrimidine bases. However, if the transition between skin cutaneous primary and metastatic tumors relies on distinct isoenzymes, new therapeutic targets might open. Distinct overexpression of nucleoside diphosphate kinases AK9 or NME7 in metastatic cancer puts emphasis on pyrimidine nucleotide synthesis while pyrimidine deamination is downregulated (Figs. 5B and 6C). On the basis of the metabolic maps, another potential

Published OnlineFirst November 25, 2015; DOI: 10.1158/1541-7786.MCR-15-0403

melanoma drug target is RRM1. Established efficacy of nucleoside analogues in acute leukemias might facilitate new treatment regimens in skin cancer (19). Given a strong reliance on biosynthetic building blocks, the upregulation of pyrimidine degradation lowers the pool of nucleotide bases available for salvage. The ribonucleotide reductase RRM1 responds to DPYD alteration, is significantly upregulated in metastatic melanoma, and bridges bottlenecks between deoxyribo- and ribonucleotides (Supplementary Table S8; Fig. 6C and D).

The systems biology analysis of melanoma data in TCGA revealed a strong separation of pyrimidine degradation and nucleotide synthesis, which is important for effective nucleic acid synthesis (Fig. 6D). The mechanism of *DPYD* controlling pyrimidine metabolism is unknown (17). A likely possibility is the existence of metabolic feedback loops of other enzymes shifting metabolism into a different gear within the progression of cancer (20). Elevated *DPYD* expression results in low metabolite pools of the pyrimidine nucleobases thymine and uracil, which could allosterically bind metabolic enzymes or signaling molecules. Moreover dihydropyrimidines and deoxypyrimidines are allosteric inhibitors of thymidine kinase (21, 22), which enhance the importance of TYMS for *de novo* pyrimidine synthesis.

Nucleotide synthesis is closely linked to production as well as stability of nucleic acids. Not surprisingly, purine metabolism scored equally high in the enrichment study, as both pathways share important enzymes in nucleoside salvage and nucleic acid processing (Table 1). However, mutational signature, correlation between somatic alterations and gene expression, and metabolic bottlenecks were unique to pyrimidine metabolism motivating further studies of *DPYD*. Remarkably, mutated *DPYD* was found to be overexpressed in metastatic cells promoting synthesis of DNA and RNA (Fig. 5). In addition, somatic *DPYD* alterations cooccurred with mutations of tumor suppressors and DNA caretakers in melanoma patients. Deregulated pyrimidine catabolism may not only be connected to nucleotide anabolism but also negatively affect DNA maintenance and stability. Further experiments will be needed to decipher the cellular mechanisms responsible for the development and the progression of melanoma.

In addition to executing the epithelial–mesenchymal transition program, metastatic cells acquire traits associated with high-grade malignancy, including resistance to apoptosis and chemotherapy. Patients with a complete or partial *DPYD* deficiency have been reported as suffering from lethal toxicity after the administration of 5-FU (5-fluorouracil, PubChem CID:3385) (23). On the basis of the pathway analysis of SKCM samples, we established a gene expression signature of pyrimidine enzymes, which grants drug sensitivity while limiting toxicity. 5-FU has to be processed by TYMP, TK, CMPK, and NME enzymes to produce the active drug-metabolite FdUMP (5-fluorodeoxyuridine monophosphate, PubChem CID:8642), which is a tight-binding inhibitor of TYMS. As TYMS represents the sole intracellular source of *de novo* TMP, the inhibition of TS exploits a metabolic bottlenecks in the biosynthesis of DNA (Fig. 6). In addition, UPP, UCK, CMPK, NME enzymes facilitate production of 5-FUTP (5-fluorouridine triphosphate, PubChem CID:10255482) causing nucleic acid damage and apoptosis. Low levels of NT5 support accumulation of 5-FUMP (5-fluorouridine monophosphate, PubChem CID:150856) and FdUMP and cell toxicity. Despite DPYD degrades 5-FU to DHFU (5-dihydrofluorouracil, PubChem CID:121997), pyrimidine degradation is necessary and causes systemic failure if absent or partially dysfunctional (24). None of the metastatic melanoma patients show compatible signatures of gene expression (Figs. 5A and B and 6A). On the basis of the mutation rate of *DPYD* of more than 20% in melanoma in combination with downregulation of TK and UPP, we predict high risks of 5-FU toxicity in melanoma. For these reasons, fluorinated uracil-based pyrimidine analogues cannot be considered to be a safe treatment regime for melanoma patients. While knockdown experiments will be necessary to identify more efficient therapeutic regimen in the pyrimidine pathway, the systems biology analysis provides a diagnostic insight at the pathway level. Importantly, the increased genotyping coverage achieved by a comprehensive description of the mutational landscape of *DPYD* improves predictive value for 5-FU toxicity.

Conclusion

The structure-based analysis of detected somatic events highlights vulnerabilities in *DPYD*. Recurring missense mutations accumulate in ligand-binding sites as well as at domain interface between Fe4S4 clusters, FAD, and pyrimidine binding. The transcriptional data shows that mutated *DPYD* selectively activates components of pyrimidine metabolism. The cross-talk between somatic mutations and gene expression promotes proliferative aggressiveness. Taken together, the transition from primary to metastatic tumors reconfigures the pyrimidine metabolism and emphasizes nucleic acid synthesis required for rapid cellular proliferation.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: F.V. Filipp
Development of methodology: F.V. Filipp
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): L. Edwards, R. Gupta, F.V. Filipp
Writing, review, and/or revision of the manuscript: F.V. Filipp
Study supervision: F.V. Filipp

Acknowledgments

The authors thank all the members of the TCGA Research Network for biospecimen collection, data acquisition, and benchmark analyses.

Grant Support

F.V. Filipp is grateful for the support of grants CA154887 and CA176114 from the NIH, NCI.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received September 24, 2015; revised October 27, 2015; accepted November 11, 2015; published OnlineFirst November 25, 2015.

References

1. Boroughs LK, DeBerardinis RJ. Metabolic pathways promoting cancer cell survival and growth. Nat Cell Biol 2015;17:351–9.
2. Filipp FV, Ratnikov B, De Ingeniis J, Smith JW, Osterman AL, Scott DA. Glutamine-fueled mitochondrial metabolism is decoupled

Published OnlineFirst November 25, 2015; DOI: 10.1158/1541-7786.MCR-15-0403

from glycolysis in melanoma. Pigment Cell Melanoma Res 2012; 25:732–9.
3. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 2010;11:685–96.
4. Heidelberger C, Chaudhuri NK, Danneberg P, Mooren D, Griesbach L, Duschinsky R, et al. Fluorinated pyrimidines, a new class of tumour-inhibitory compounds. Nature 1957;179:663–6.
5. Wilson PM, Danenberg PV, Johnston PG, Lenz HJ, Ladner RD. Standing the test of time: targeting thymidylate biosynthesis in cancer therapy. Nat Rev Clin Oncol 2014;11:282–98.
6. Guan J, Gupta R, Filipp FV. Cancer systems biology of TCGA SKCM: efficient detection of genomic drivers in melanoma. Sci Rep 2015; 5:7857.
7. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213–9.
8. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, et al. A landscape of driver mutations in melanoma. Cell 2012;150:251–63.
9. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8.
10. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature 2013; 500:415–21.
11. Pronk S, Pall S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 2013;29:845–54.
12. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
13. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9:811–8.
14. Podschun B, Cook PF, Schnackerz KD. Kinetic mechanism of dihydropyrimidine dehydrogenase from pig liver. J Biol Chem 1990;265:12966–72.
15. Dobritzsch D, Ricagno S, Schneider G, Schnackerz KD, Lindqvist Y. Crystal structure of the productive ternary complex of dihydropyrimidine dehydrogenase with NADPH and 5-iodouracil. Implications for mechanism of inhibition and electron transfer. J Biol Chem 2002; 277:13155–66.
16. Krauthammer M, Kong Y, Ha BH, Evans P, Bacchiocchi A, McCusker JP, et al. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. Nat Genet 2012;44:1006–14.
17. Amstutz U, Froehlich TK, Largiader CR. Dihydropyrimidine dehydrogenase gene as a major predictor of severe 5-fluorouracil toxicity. Pharmacogenomics 2011;12:1321–36.
18. Offer SM, Fossum CC, Wegner NJ, Stuflesser AJ, Butterfield GL, Diasio RB. Comparative functional analysis of DPYD variants of potential clinical relevance to dihydropyrimidine dehydrogenase activity. Cancer Res 2014;74:2545–54.
19. Bonate PL, Arthaud L, Cantrell WR Jr, Stephenson K, Secrist JA III, Weitman S. Discovery and development of clofarabine: a nucleoside analogue for treating cancer. Nat Rev Drug Discov 2006;5:855–63.
20. Filipp FV. Cancer metabolism meets systems biology: pyruvate kinase isoform PKM2 is a metabolic master regulator. J Carcinog 2013;12:14.
21. Lin J, Roy V, Wang L, You L, Agrofoglio LA, Deville-Bonne D, et al. 3′-(1,2,3-Triazol-1-yl)-3′-deoxythymidine analogs as substrates for human and Ureaplasma parvum thymidine kinase for structure-activity investigations. Bioorg Med Chem 2010;18:3261–9.
22. Shaul YD, Freinkman E, Comb WC, Cantor JR, Tam WL, Thiru P, et al. Dihydropyrimidine accumulation is required for the epithelial-mesenchymal transition. Cell 2014;158:1094–109.
23. van Kuilenburg AB, Haasjes J, Richel DJ, Zoetekouw L, Van Lenthe H, De Abreu RA, et al. Clinical implications of dihydropyrimidine dehydrogenase (DPD) deficiency in patients with severe 5-fluorouracil-associated toxicity: identification of new mutations in the DPD gene. Clin Cancer Res 2000;6: 4705–12.
24. van Kuilenburg AB. Dihydropyrimidine dehydrogenase and the efficacy and toxicity of 5-fluorouracil. Eur J Cancer 2004;40:939–50.

Chapter 5

Overexpression of DNMT3a 3b in TCGA SKCM induces targeted regulation led by selective methylation

Overexpression of *DNMT3a & 3b* in TCGA SKCM induces targeted regulation led by selective methylation.

Systems Biology and Cancer Metabolism, University of California Merced, Merced, CA 95343, USA

Rohit Gupta and Fabian V. Filipp

e-mail: filipp@ucmerced.edu

Abstract

Aberrant DNA methylation is an epigenetic hallmark of Melanoma. Focal DNA hypermethylation of tumor suppressor gene promoters is observed across many cancer tissues. Activation and repression of transcription factors that play a detrimental role in oncogenesis and tumor suppression through targeted methylation led by DNMTs is vital for understanding the mechanisms employed by tumor cells to reprogram cell machinery. In this study, genomic analysis was performed to assess the scope of overexpression in key methylation-inducing genes, DNA methyltransferases 3a & 3b (DNMT3a, DNMT3b), through subtyping based on gene expression data. A large cohort of human skin cutaneous melanoma (SKCM) patient data from the The Cancer Genome Atlas (TCGA) was used for the analysis. Differential methylation in DNMT expression-based patient cohort revealed selective hyper- and hypomethylation in gene and promoter regions of key tumor suppressors and oncogenes, respectively. In addition, transcription factors of RUNX family and HOX family were also targeted for hypermethylation. Overall, at a systems biology level, a pattern of targeted methylation induced as a factor of DNMT3a & 3b gene expression directs repression of key tumor suppressors to promote malignant progression.

Introduction

DNA methylation in mammals is found sparsely but is globally distributed in definite CpG sequences throughout the entire genome. CpG islands are short interspersed DNA sequences that are enriched for GC. These CpG islands are normally found in sites of transcription initiation (transcription start sites; TSS), and selective methylation of these sites has been studied to affect gene regulation; to that end, methylation of CpG leads to gene silencing. It is crucial to study the methylation pattern of CpG islands in cancer, as silencing of tumor suppressors through targeted hypermethylation to silence tumor suppressors has been found to be a prominent machinery exploited by cancer; furthermore, aberration in DNA methylation is an epigenetic hallmark of cancer [1, 2].

DNA methyltransferases (DNMTs; 1, 3A, 3B) are responsible for catalyzing the transfer of methyl groups to mammalian genomic DNA and play an active role in gene silencing and repression. DNMTs have been found to induce tumor growth and aid tumor cells by orchestrating both methylation-induced and methylation-independent changes in genes and transcription factor expression [3-5]. Elimination of both DNMT1 and 3A nearly eliminates methyltransferase activity, while disruption of DNMT3A only reduced the methyltransferase activity by 3%, indicating an enhanced role of DNMT3A in regulatory and targeted methylation [5]. DNMT3A deletion promotes tumor growth, but not initiation, in mouse models of lung cancer. DNMT3A-deficient tumors had high proliferation and significantly fewer differentially-methylated regions across genomic DNA, indicating DNMT3A-mediated targeted promoter methylation for gene repression [4]. DNMT3B silencing also inhibits proliferation and stimulates apoptosis in hepatocellular carcinoma (HCC) and its siRNA-mediated knockdown led to locus-specific hypomethylation and increase in gene expression [6]. Analysis of melanoma cell lines has identified a large cohort of hypermethylated genes that are perceived to be repurposed for malignant disease progression [7]. However, the scope and causal mechanisms that contribute towards pathogenesis as a result of hypermethylation remain largely unknown. The effects and functional outcome from hypomethylation have been studied much less but the phenomenon is common across several cancer tissues, including Melanoma [8].

Subtyping is an effective way to determine common set of changes that occur in a subset of patients with a specific cancer. In Melanoma subtyping based on mutation, gene expression, methylation, UV signature etc. has been performed to identify underlying factors that play an enhanced role in promoting cancer phenotype [9,10]. Methylation-based subgrouping study in 50 metastatic Melanoma patients revealed hypermethylation in PRC2 target gene-sets and 3 different methylation groups [11]. The BRAF V600E mutation in melanoma has also been studied to direct widespread promoter methylation and epigenetic silencing and BRAF-directed pathways also mediate epigenetic silencing in colorectal cancer [12, 13]. While epigenetic silencing and selective induction by methylation has been studied in many cancer tissues, the effect and scope of DNMT enzymes' transcriptional activity as it relates to driving these changes in cancer is not explicitly established. In this study we performed subtyping in TCGA-SKCM patients on the basis of DNMT 1, 3A & 3B transcriptional activity to map out global and targeted methylation that are directed by DNMTs in Melanoma.

Overexpression of DNMT3B and its splice variant DNMT3B4 has been studied in breast cancer, clear-cell-renal-cell carcinoma (ccRCC) and non-small-cell ling cancer (NSCLC). Aberrant methylation and hypermethylation of PRC2 targets as a result of overexpression was observed [14-16]. In Melanoma, DNMT3B is also overexpressed, and its knockdown markedly suppresses Melanoma formation and proliferation [17] but the underpinning mechanisms and host of genes and transcription factors that govern this are poorly understood. Furthermore, combining transcription activity of DNMT1, 3A and 3B with methylation changes will provide a window to understand the alterations in Melanoma machinery that stem from epigenetic changes led by the methyltransferase group of enzymes.

Transcription factors are crucial for maintaining specific cell states and gene-regulation programs associated with them[18]. Misregulation of transcription through metabolic reprogramming is an emerging hallmark in cancer, particularly in Melanoma[19]. Previous studies across several cancers including Melanoma have provided substantial evidence depicting epigenetic mechanisms, such as selective methylation, as the primary methods by which modulation of cell state occurs to promote disease progression [20-23].

Materials and Methods

Genomic Analysis of TCGA-SKCM

We used National Cancer Institute (NCI) Genomic Data Commons (GDC) portal to access TCGA-SKCM Level 2 and 3 methylation, mutation, transcriptomic, and clinical datasets used in this study. GDAC Firehose along with GDC data portal were used to query GISTIC2 processed Copy Number Variation (CNV) data. DNA methylation data obtained from GDC portal was acquired through Illumina Infinium HumanMethylation450k beadchip (HM450k) which contains 485579 probes per patient sample. Methylation levels are reported as a beta score (ß) between 0 and 1, with a score of 1 denoting highest methylation.

RNAseqV2 with gene expression normalization acquired by Illumina HiSeq2000 was used for gene expression quantification. We used TCGABiolinks package to process the data and convert each dataset into SummarizedExperiment class object to add subtyping data and clinical information to the data.

ChiP data used in this study was obtained from foreskin primary melanocyte cells (skin03; EpigenomeID – E061), through Roadmap epigenomics project database and retrieved using AnnotationHub package in ENCODE narrowPeak format as BED files.

Statistical methods for epigenomic analysis

Patient cohort based on DNMTs (1, 3a & 3b) transcriptional activity were created by selecting 50 patients, each with highest and lowest expression of DNMT1, DNMT3a and DNMT3b genes (Supplement Table 1). Three cohorts comprising 100 patients each (100/471 from TCGA-SKCM) with 50 subjects in “high” and 50 subjects in “low” transcription activity group were created and then used to identify regions of DNA with differential methylation as a direct cause of varied transcription levels.

Differentially-methylated CpG sites on the DNA were identified by a two-step process that first calculated the difference between mean DNA methylation for each patient cohort for each probe in all patients in the cohort. We removed probes with NA values and a minimum absolute beta-value difference less than 0.15. Finally, Wilcoxon test adjusted by the Benjamini-Hochberg method was performed and probes with adjusted p-values of <0.01 were selected as significantly hypermethylated probes.

To assess the effect of selective methylation on genes we integrated differential gene expression data with differential methylation. Differential expression analysis (DEA) between patient cohorts previously created was carried using edgeR package which performs pair-wise Fisher's exact test by comparing and performing statistical tests in each gene pair between the cohorts. Gene-pairs with log fold change (logFC) of above 10^-5 and FDR corrected p-value of above 10^-5 were selected. Next, DMR with FDR corrected p-value of 10^-2 and above and mean methylation of at least 0.15 were merged with the gene-pairs from previous step.

Regulatory motif search was performed using rGADEM on selectively DMR identified in previous steps. A window of 100 bases (upstream and downstream of the probe) was created at each differentially-methylated CpG site. Starting position weighted matrix (PWM) was created with equal probability distribution for each nucleotide. rGADEM uses a genetic algorithm (GA) with an embedded expectation-maximization algorithm to improve starting PWM based on the sequence encountered in the window created in earlier step. We used motifStack to visualize regulatory motifs identified. To identify transcription factors that most closely match with the regulatory motifs we used MotIV, an R Bioiconductor package that matches motif PWM against JASPAR database using alignment.

To annotate and visualize histone methylation marks and the average profile peaks in and around hypomethylated and hypermethyalted regions in ChiP-Seq data we used an R Bioconductor package called Chipseeker. A window of 3000bp (+/- 3 kbp) was created both upstream and downstream of each differently-methylated probe to create average profile enrichment plots of histone methylation marks. A heatmap of peak binding regions of histone marks was made using ggplot. R package TxDb (object TxDb.Hsapiens.UCSC.hg19.knownGene), which contains transcript-related features of a genome, was used in conjunction with Chipseeker for annotation of peaks. Enrichment analysis (EA) to assess KEGG pathways and genes affected by CG probes with significant hypermethylation and hypomethylation was performed using annotated profile peaks obtained in the previous step of differential methylation analysis. The Chipseeker package was then used to create an EA plot for KEGG Pathway enrichment.

To aid identification of distal regulatory enhancers we used R package ELMER (Enhancer Linked by Methylation/Expression Relationship). To filter HM450k probes that function as distal regulatory probes and are located at least +/- 2kb away from transcription start sites (TSS) we used comprehensive list of TSS annotated by ENSEMBL and accessed by biomaRt package. A multiassayexperiment (MEA) object containing DNA methylation data of distal enhancer and gene expression values was created to be used in ELMER. For each distal probes a rank based on Beta methylation is created which is then used to identify hyper/hypo methylation based on unpaired one-tailed t-test between between patient cohorts. Next, each enhancer probe with methylation changes was tested for correlation with 10 upstream and 10 downstream genes.

To analyze the role and scope of each transcription factor family in promoting differential methylation and thereby transcriptional misregulation to promote tumor phenotype as indicated by EA plot, we used Python 2.7.16 in Ipython shell. BeautifulSoup python package was then used to parse TFClass database[24] which contains exhaustive classification of transcription factors

and associated families. Significant CG probes (10^-2) with high degree of hyper-/hypomethylation were grouped by transcription factor families and total number of hyper and hypo probes were computed. Relative significance score was then calculated as factor of transcription factor family size and total number of significant probes across all transcription factors for a given family. Relative significance score was however an inaccurate indicator of significance as relatively larger TF families harbored many more CG probes but had little to no functional influence. To correct for size biases and take only significant probes in functionally relevant regions into consideration, CG probes in distal intergenic regions were ignored while probes in regions of exons, promoter and enhancer were retained. The normalization parameter, Z-score, was devised to take into account relative size differences between TF families. Z-score of significance was calculated using formula (1). The Z-scores were converted to p-values using Gaussian cumulative distribution function ndtr of scipy.special python package. Transcription factors meeting p-value threshold of <= 0.1 were then selected.

$$Z_i = X - \overline{X}$$

$$Z_i = \frac{X_{iFA} - \frac{\sum x_{iF}}{m}}{\sigma_F}$$

where: Z_i = Z-score

X_{iFA} = Number of CG probes in a given transcription factor

X_{iF} = Number of CG probes in TF Family

m = Number of members for a gen TF family

To perform downstream analysis to assess the scope of selective methylation in a transcription factor family tftargets R package was used to obtain activated and repressed target genes for transcription factor family identified in preceding step. To visualize changes in target genes in low and high DNMT expression patient cohorts R library ggplot2 was used.

Results

DNMT3a and 3b transcriptional activity drive methylation in key tumor suppressor genes and pathways like Wnt and TGFbeta.

473/473 TCGA-SKCM patients expressed DNMT1, DNMT3a and DBNT3b at varying levels. DNMT1 showed overall high expression relative to all other genes and highest within methyltransferase genes. Genomic coordinates of differentially-methylated CpG regions were then mapped to gene locations. DNMT1 had 106 differentially-methylated probes when comparing methylation (ß value) between DNMT1 high expression and low expression cohort. These 106 differentially-methylated probes spanned 68 genes, with some genes harboring multiple differentially-methylated probes, and 15 probes were located in intergenic regions.

DNMT3A contained 3714 differentially-methylated probes spanning 1518 genes with some genes and transcription factors. 1338/3714 differentially-methylated probes were located in intergenic regions. Differential methylation analysis in DNMT3B produced a total of 1222 DMR spanning 567 genes, 404/1222 probes were located in intergenic region while genes and transcription factors like HCK (chemokine signaling pathway) , EXOC3L2, PLEC1, TET1 and RUNX1 containing at least 5 differentially-methylated probes each.

Homeobox family of transcription factors (HOXB13, HOXD12, HOXA13, HOXA4, HOXB9, HOXD13) were significantly hypermethylated and contained about 30 hypermethylated probes in total between DNMT3a and 3b patient cohort. HOX genes have been studied to be tumor suppressors and targets of selective methylation in colorectal cancer [25]. Differential expression and correlation between expression and metastasis in HOX genes has been observed in Melanoma. In this study we identify hypermethylation in several HOX genes; most noticeably, HOXB13 contained 9 and 3 hypermethylated probes in DNMT3a and DNMT3b respectively. HOXB13 and other members of HOX family of transcription factors were identified by ChiP antibody to be hypermethylated in colorectal cancer and are believed to function as tumor suppressors [26], while HIXD9 has promoter methylation and is associated with clinical prognosis in Melanoma Brain Metastasis (MBM)[26]. HOX transcript antisense RNA (HOTAIR) originates from the HOXC cluster, and is shown to have pro-metastasis activity in breast and pancreatic cancer [27,28]. HOTAIR recruits PRC2 to specific target genes[29,30].

The Runt-domain containing family of transcription factors (RUNX1, RUNX2, RUNX3) conserve the ability to counter oncogenic signals through oncogene-induced senescence and function as tumor suppressors [31]. All three RUNX members are integral components in the activation of the TGF-ß and Wnt signaling pathways [32]. In our data-set we observe significant hypermethylation in DNMT3A and 3B high transcription activity cohort. RUNX1, studied to regulate E-cadherin and mediating TGF-ß led tumor suppression [33], contained 7 hypermethylated probes in DNMT3A and 4 hypermethylated probes in DNMT3B high activity cohort respectively. RUNX3, also a known tumor suppressor [34], contained 4 hypermethylated probes in DNMT3A and 2 hypermethylated probes in DNMT3B high activity cohort respectively. RUNX3 and RUNX1 interact with FOXOA3 to induce BCL2 [35,36]. We also observed hypermethylation in the distal enhancer site of FOXD2, driving its suppression. RUNX2 is overexpressed in Melanoma and is reported to mediate migration and invasion in Melanoma cell lines [37]; we observed 3 hypermethylated probes in DNMT3A and 1 probe in DNMT1 high activity cohort respectively.

Hypermethylation of SFRP2 has been found to silence Wnt/ß-catenin pathway in gastric cancer[38]. 12 significantly hypermethylated probes across DNMT3A and DNMT3B high activity cohorts were associated with SFRP2 in our dataset. We believe that a combination of epigenetic silencing through hypermethylation and copy number deletion (CNV) contribute to inactivation/silencing of the TGF-ß and Wnt signaling pathways in Melanoma. Sclerostin domain containing 1 (SOSTDC1) protein, an antagonist and modulator of the BMP [39] and Wnt

signaling pathways in breast cancer [40], contained 4 and 6 significantly hypomethylated CG probes in DNMT3A and DNMT3b patient cohorts, respectively.

Highly selective significant hypermethylation of ZFHX3 was observed in DNMT3a and DNMT3b which contained 8 and 4 hypermethylated CG probes, respectively. ZFHX3 has been found to be a tumor suppressor in several cancers and functions by negatively regulating c-Myb and trans-activation of cyclin-dependent kinase inhibitor 1A (p21CIP1). Tumor necrosis factor receptor 1 (TNFRSF1A), a substrate of TNFalpha, leading to its activation, was significantly hypermethylated and contained 4 hypermethylated CpG probes in patients with high DNMT3a transcriptional activity. It should be noted that a battery of immune response, apoptosis and inflammation response genes and transcriptional factors were also selectively hypermethylated in patient cohorts with high DNMT3a and 3b activity.

Genes and targets of genes in Polycomb repressive complex 2 (prc2) were also observed to be selectively methylated. Most noticeably, JARID2 contained 3 hypomethylated probes in its promoter region in patient cohort with high DNMT3a expression and 2 hypomethylated probes in DNMT3b high expression patient cohort. Considerable overlap between targets of PRC2 and selectively methylated genes in DNMT3a and DNMT3b high activity cohort was also observed. Key genes (PIK3CA, PIK3CD, PIK3CG) in PI3K-AKT signaling pathway were selectively hypomethylated and thereby transcriptionally activated to promote the cancer phenotype. Selective hypomethylation of proto-oncogenes like PTPRN2 (increased expression confers resistance to apoptosis in breast cancer), CACNA1H (MAPK signaling pathway), ARC (apoptosis repressor), GSE1 (oncogene in breast cancer), ELANE (inhibited apoptosis by activating PIK/Akt activation) and NCOR1/SMRT corepressors (Notch signaling pathway) was also observed.

Copy number loss and gain correlate with hypermethylation and hypomethylation respectively.

Recurrent copy number variations (CNV) with significant amplification and deletion and significance threshold of q-value <= 10^-2 were identified using GAIA in TCG-SKCM data. Gain or loss of chromosome regions as identified by CNV analysis correlated with regions of hypermethylation and hypomethylation, indicating mutually-exclusive mechanisms undertaken by the cancer cell for deletion and amplification of gene/chromosome regions with oncogene or tumor suppressors, respectively. Copy number gain is associated with increases in gene expression due to extra copies of the gene and, similarly, copy number losses tend to translate into loss of gene expression. Unsurprisingly, regions of hypermethylation that overlapped with copy number deletion events had a substantial decrease in expression. The chronological order or mechanisms that aid concurrent copy number deletion/amplification and hypermethylation/hypomethylation are not included in this study and further investigation would be required. This overlap between targets of selective methylation and somatic copy number alteration further highlights the robust nature of a tumor cell and its capacity to recruit diverse sets of mechanisms to promote malignant disease progression.

Enrichment of histone methylation mark H3K27me3, associated with polycomb repression

To understand the scope of varied methylation on targeted activation/repression of transcription factors and other chromatin-associated proteins we used ChIP-seq analysis in tandem with the DNA methylation and gene expression data. ELMER analysis was used to infer the effect of DNA methylation on distal regulatory enhancers and to map the effect of targeted methylation at distal enhancers on regulation of upstream regulators like transcription factors.

The p53/RUNT DNA binding transcription factor superfamily, which is found in the p53 and the RUNT families of transcription factors, was one of the most significant targets of selective methylation, regulation by enriched motif and distal enhancers. RUNX1, RUNX2, RUNX3 had 19 significantly hypermethylated probes within the gene-body and 35 significantly hypermethylated probes in their distal enhancer sites. To access the role of hypermethylation in distal enhancer sites we evaluated correlation in expression of RUNX1, 2 and 3 with DNA methylation at those sites. This analysis clearly demonstrated a significant decrease in expression with increase in methylation at distal enhancer sites (>2kb from TSS). Consequently, higher transcriptional activity in DNMT1, 3a & 3b drove hypermethylation in RUNX transcription factors which resulted in an active repression of their expression. Analysis of regulatory motifs in differently methylated regions to identify transcription factors that can bind to them revealed TP53 as one of the most significant (8e-02) target of motif CTGCGCCAGGC found in hypermethylated CG probes in DNMT3B.

Selective methylation at distal enhancer sites resulted in a targeted regulation of WNT9B gene expression, this along with promoter methylation observed at other members of Wnt and MAPK signalling pathway point towards an active involvement of DNMTs in regulation of MAPK and Wnt pathways. We also observed significant association of expression in key regulatory TFs like FOS,FOSB, FOSL1, FOSL2, E2F5, E2F7, JUN, JUNB with DNA methylation at enriched motif sites of distal enhancers.

ChiP-Seq analysis to assess how differently-methylated regions identified in previous steps influenced chromatin accessibility and histone methylation showed a significant enrichment of H3K4me1 (Histone H3 lysine monomethylation), H3K4me3 (Histone H3 lysine 4 trimethylation), H3K27me3 (Histone H3 lysine 4 trimethylation), H3K27ac (Histone H3 lysine 27 trimethylation), H3K9me3 (Histone H3 lysine 9 trimethylation) in the DMR CpG regions of DNMT3A while H3K4me1 (Histone H3 lysine monomethylation), H3K4me3 (Histone H3 lysine 4 trimethylation), H3K27me3 Histone H3 lysine 27 trimethylation), H3K27ac (Histone H3 acetylated at lysine 27) were significantly enriched in DMR CpG regions of DNMT3B. Enrichment of histone methylation mark H3K27me3 which is mediated by polycomb repressive complex (PRC) for gene silencing [41-43] was in line with selective hypomethylation of PRC genes like JARID2.

Effect of selective methylation in transcription factors translated to their targets as well.

KEGG pathways classified as associated with transcription misregulation in cancer were observed to have a significant overlap with hypermethylated gene regions as identified in our previous analysis. These findings indicate re-purposing at the transcription factor and

transcription factor family level to promote malignant progression, and not limited to targeted genes and promoters. In order to map out transcription factor families with the highest degree of differential methylation, we implemented a statistical parameter to rank transcription factor families that harbored the most functionally relevant and significantly differentially-methylated regions. To select only the functionally relevant probes, we annotated DMR probes identified in previous steps and first observed the relative distribution of probes in introns, exons, promoter, transcription start sites (TSS) and distal intergenic regions (Supplementary Figure1 & 2). Probes in all regions except distal intergenic were then selected for further analysis. Total number of functionally significant CG probes for every TF family as classified by tfclass mammalian database was then calculated. Finally, to correct for large transcription factor families size normalization using Z-score normalization was performed.

Transcription factor family with the highest number of significant, functionally relevant hypermethylated probes were the HOX-related factor family, with 53 probes across 52 transcription factors in its family, followed by the paired-related HD family (PPRX2) containing 50 significant probes. The SOX-related factor family contained 34 significant probes across 23 members. The Runt-related factor family and TBX2-related factor family harbored 31 probes each across 3 and 4 transcription factors in their family respectively. Upon performing size normalization, Runt-factor family followed by TBX2 and Friend of GATA protein factor (FOG 1, 2) families were found to be harbor most significant probes. The Runt-related transcription factor family was the most significantly enriched family with p-value of 0.000391, followed by TBX2 related factors with a p-value of 0.011. Interestingly, individual members of Runt-related transcription factor family were also target of selective hypermethylation and substantial enrichment at the transcription factor family level further validates our finding.

Second degree effects of hypermethylation in transcription factor families propagated to their target genes. Across all members of transcription factor families with enriched hypermethylated probes, effector actions were either limited or inversed. In the case of target genes that are known to be activated, either no increase and in some cases repression of mRNASeq expression was observed, and in the case of genes repressed by hypermethylated targets, an increase or no change in expression was generally observed. Runt-related transcription factor family with Z-enrichment p-value of 0.000391 had up to 1.5 fold change in IL2, a target generally activated by RUNX1 transcription factor, but as a result of hypermethylation in RUNX1 and Runt-factor family an inverse action was observed. A general trend of decrease in expression across targets activated by RUNX1, RUNX2 and RUNX3 was observed. Most notably, genes associated with immune and inflammation response, CCL3, IL2, JUN, CDKN1A exhibit constitutive inhibition in the DNMT high expression cohort, while constitutive activation of targets otherwise repressed by Runt-family transcription factor was observed as well. Genes serving functions related to oncogenesis and actively promoting the tumor phenotype, including VEGFA, MYC, JAG1 for example, were seen to be up-regulated.

Discussion

The high degree of methylome-led regulatory mechanisms are implicated in cancer progression, especially in the case of Melanoma. The underpinning mechanisms that enable mis-direction of routine cellular processes, like methylation, by a cancer cell is of principal research interest among many groups seeking to better therapeutic strategies to treat Melanoma. A series of reprogramming in cellular machinery through selective methylation led by DNMTs is observed throughout this study. Highly targeted changes in DNA methylation are repurposed by melanoma cells to promote constitutive activation and repression of genes and transcription factor activity. The effects of selective methylation observed in transcription factors propagated to their target genes as well. We observed consistent increase in mRNAseq expression of target genes generally known to be repressed by a hypermethylated transcription factor and similarly, decrease in expression of target genes otherwise known to activated.

Analysis of gene expression data and differential methylation patterns in select genes and regions of promoters point towards highly localized regulation tied to DNMT transcriptional activity. Mapping targets subject to differential methylation onto pathways shows a pattern of enrichment or repression. Highly significant focal hypomethylation of oncogenes and transcription factors and hypermethylation of tumor suppressors shows that re-molding of cell metabolism is routine in cancer cells. Similarly, several members of pathways required for cancer proliferation, like immune modulation, were hypermethylated and as a result the overall pathway was observed to be repressed in order to further promote the tumor phenotype. Focal regulation as a result by DNMT3a and 3b is the first study in the TCGA SKCM dataset to our knowledge that establishes a link between direct regulation potential of DNMTs as a factor of their expression. Interestingly, comparative analysis between methylation patterns of primary and metastasized Melanoma did not show significant enrichment in gene of significance in metastasis. This result further shows that signals of metastasis are harbored early in the primary Melanoma and largely have mutation and copy number alteration origins.

Both the Runt-related transcription factor family and its individual member genes, RUNX1, RUNX2, RUNX3, were targeted for significant hypermethylation. RUNX genes and transcription factor families function as important tumor suppressors by activating a host of immune-related pathways and help counter oncogenic signals. While the tumor suppressor role and hypermethylation of RUNX genes has been studied in other cancer tissues, this study demonstrates the mechanism and scope of DNMT3a & 3b over-expression that drives this hypermethylation event in Melanoma.

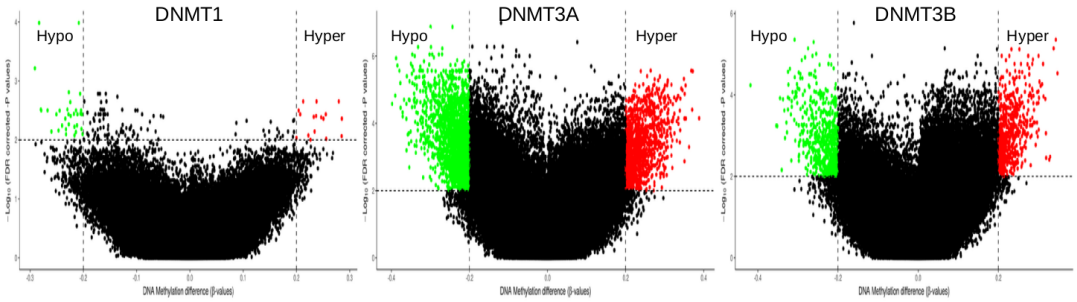
An enrichment of histone methylation marks, H3K27me3, associated with polycomb repression (PRC2) and other mono-, di- and tri-histone methylation marks known to be associated with transcription repression observed in this study further show large-scale repression. Taken together with focal methylation, this finding is in line with overall transcription mis-regulation of pathways through repression commonly observed in cancer. Distal enhancers of transcription factors like JUN and FOS that are crucial for oncogenesis were targeted for hypomethylation, while tumor suppressors, including TP53, significantly aligned with the consensus regulatory motif detected in the hypermethylated CpG region of DNMT3B.

Consistent with the tumor suppressor function of KLF4 we saw a hypermethylation-driven decrease in its expression by its promoter and enhancer methylation driven my DNMT 3a & 3b.

While the role of methylation in cancer is well established, very little is known of the impact on methylation as a result of transcription activity of the key methylation genes, DNMT1, 3a & 3b. We demonstrate a clear model of regulation tied to transcription activity in TCGA-SKCM patient cohort. Taken together, it can be concluded that DNMT (1, 3A &3B) transcriptional activity that drives selective methylation plays a pivotal role in the regulation of genes, distal enhancers, transcription start sites and transcription factors. Overexpression of DNMT3a plays an enhanced role in marking sites for selective methylation. These insights, combined with our data, support the rationale of a diagnostic signature for Melanoma based on methylation. ß

Figures

Figure1: Differential methylation and somatic copy number alteration analysis in overexpression patient cohort of TCGA SKCM.



1A. differentially-methylated regions (DMR) identified in high and low transcriptional activity patient class in DNMT1, DNMT3a and DNMT3b. Selective methylation in regions include notable tumor suppressors TP53, Tumor necrosis factor receptor 1 (TNFRSF1A) a substrate of TNFalpha and proto-oncogenes like ARC (apoptosis receptor) and several polycomb group of proteins (PRC1 & PRC2).

1B. Somatic copy number analysis (outermost ring) with amplified and deleted regions in TCGA-SKCM, density of hypermethylated (middle ring) and hypomethylated regions (innermost ring) as identified by differently methylated CpG probes. Co-occurence of hypermethylation with deletion events and hypomethylation with amplification events is observed.

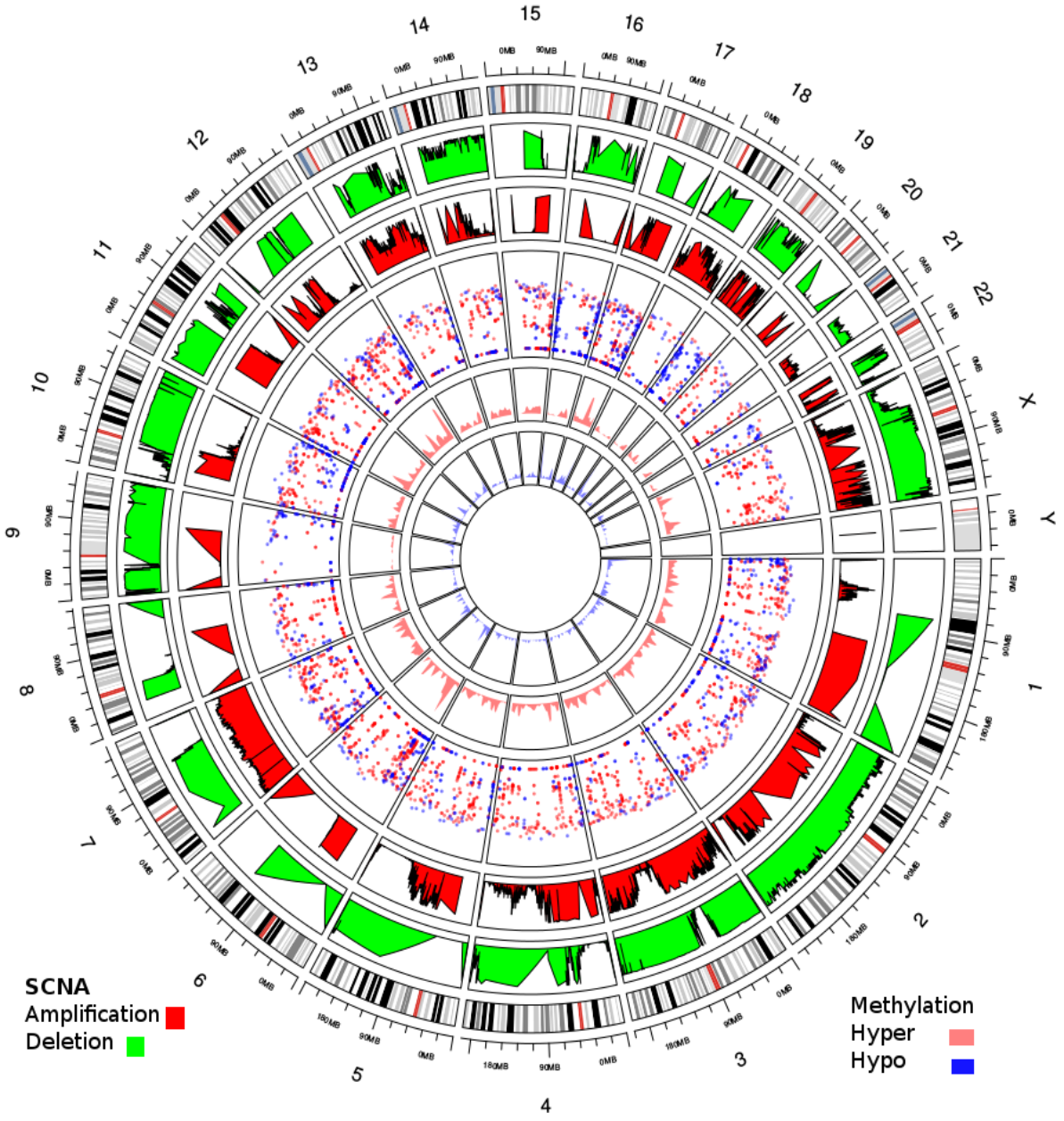
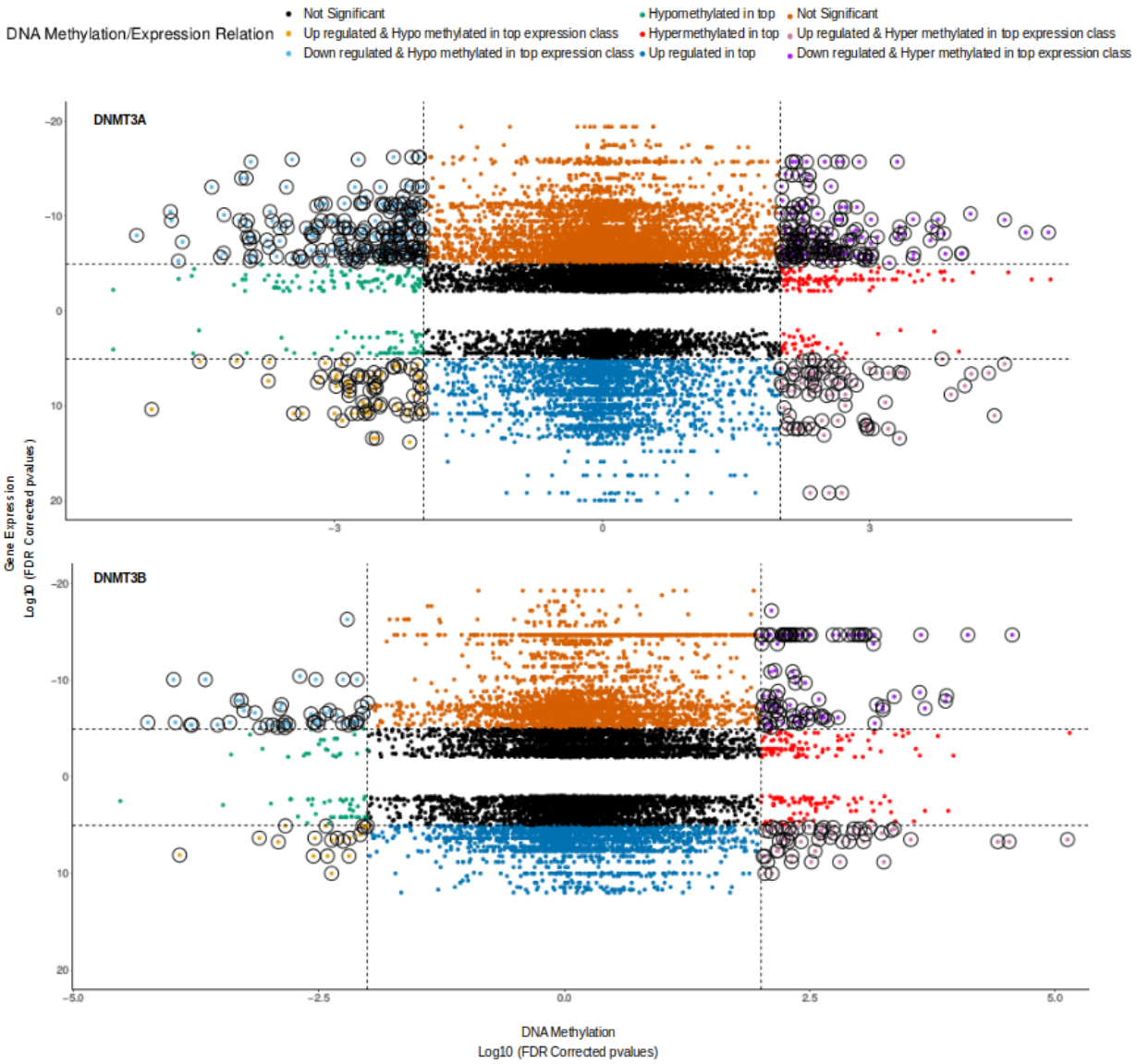
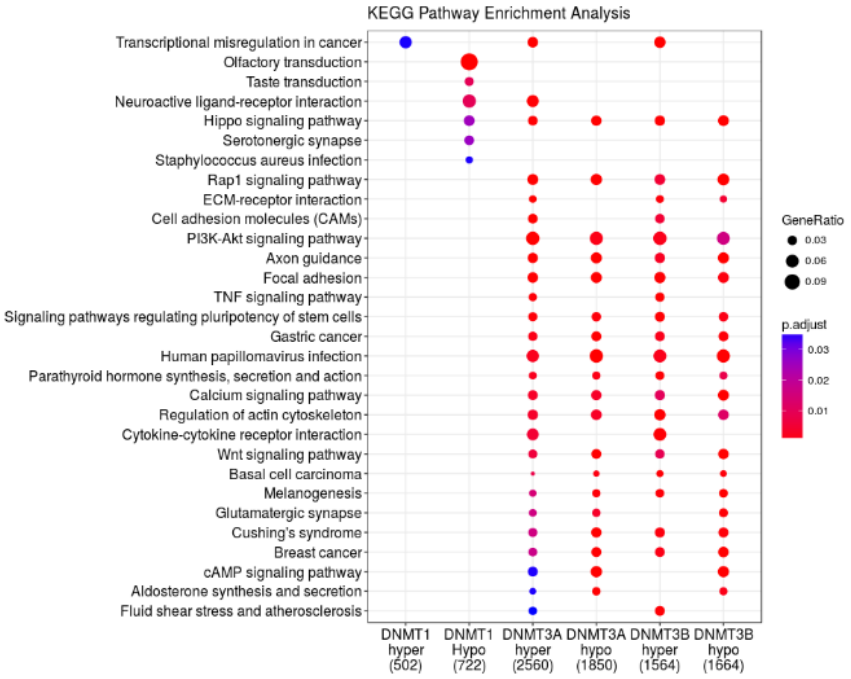

SCNA
Amplification
Deletion
Methylation
Hyper
Hypo

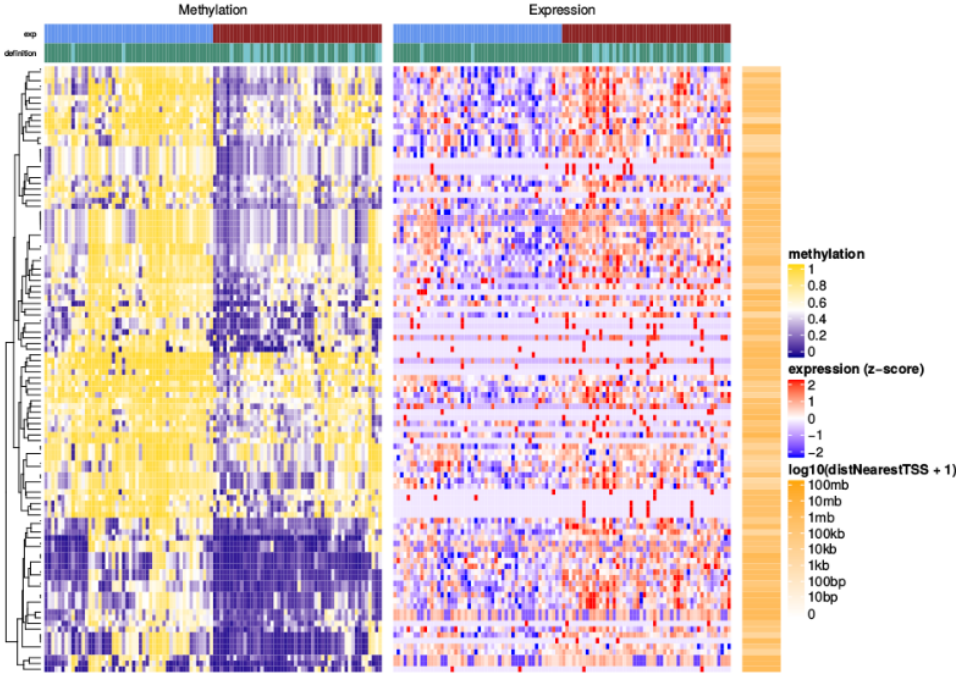
Figure 2: Comparing differential targets with gene expression shows prominently affected pathways and gene. Transcription mis-regulation related pathways were observed to be most significantly enriched in hypertmethylation.



2A. Combining differential expression with differential methylation in DNMT3A and DNMT3B to observe effect of selective methylation on gene expression



2B. KEGGpathways enriched by selective methylation.



2C: Heatmap depicting changes in gene expression in members of enriched pathways as identified in figure 2b.

Figure 3: ChiP-Seq analysis to assess the impact on chromatin accessibility due to histone modification mark shows an enrichment of marks associated with transcription repression. Enrichment of histone methylation mark H3K27me3, mediated by polycomb repressive complex (PRC) for gene silencing, was in-line with selective hypomethylation of PRC genes like JARID2 observed in our data. Distribution of histone modification marks in relation to differently methylated CpG sites in DNMT3A.

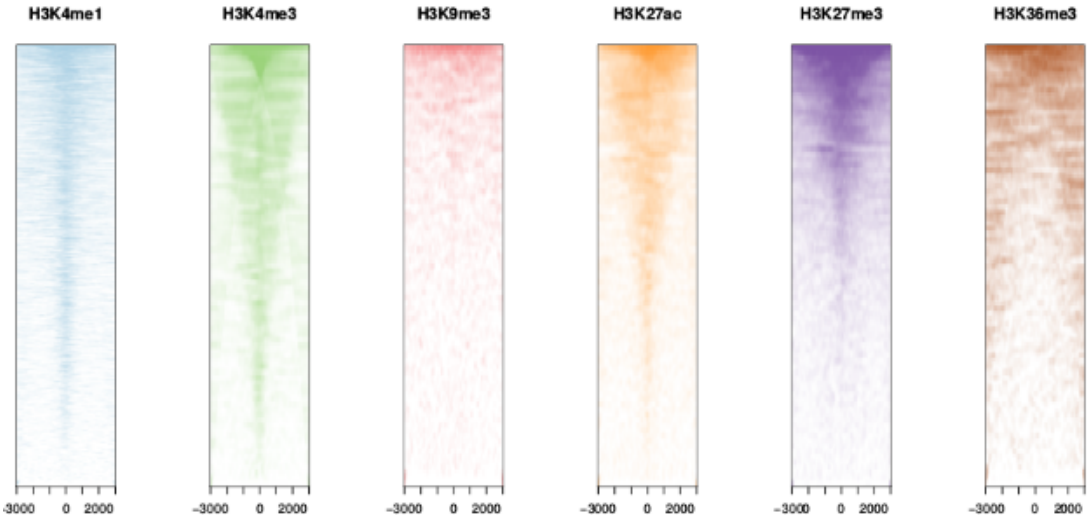
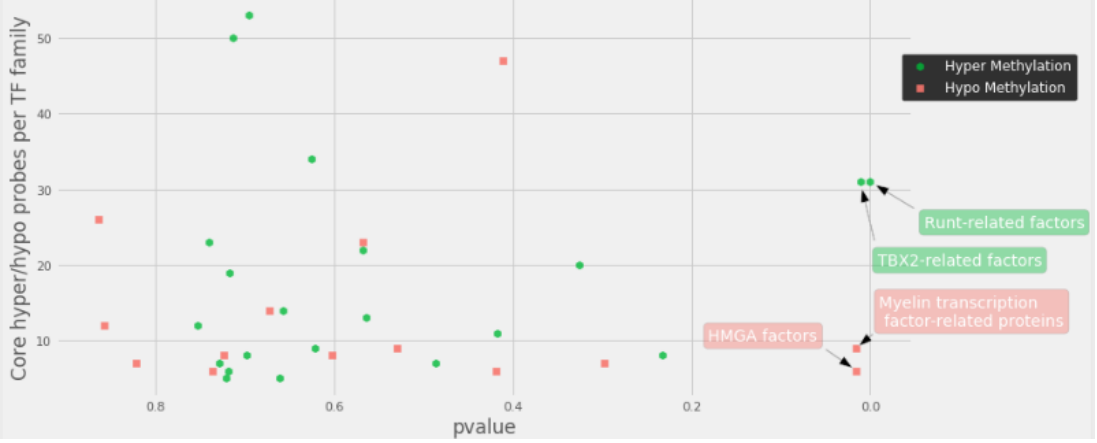
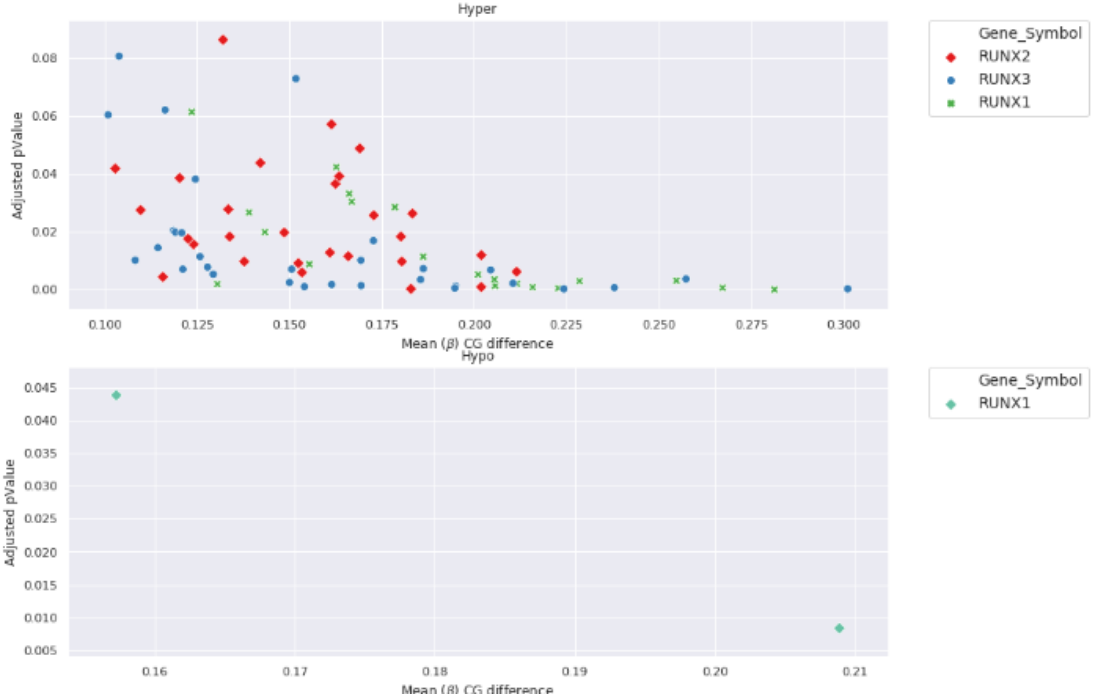


Figure 4: Transcription factor families ranked by p-value to identify the families harboring the most number of functionally valuable CG probes. Size normalization in transcription factor families was performed to correct for large transcription factor families harboring relatively large number of significant CG probes. Only CG probes found in functional location were included for normalization step and significant CG probes in distal location in relation to closes genes were ignored.

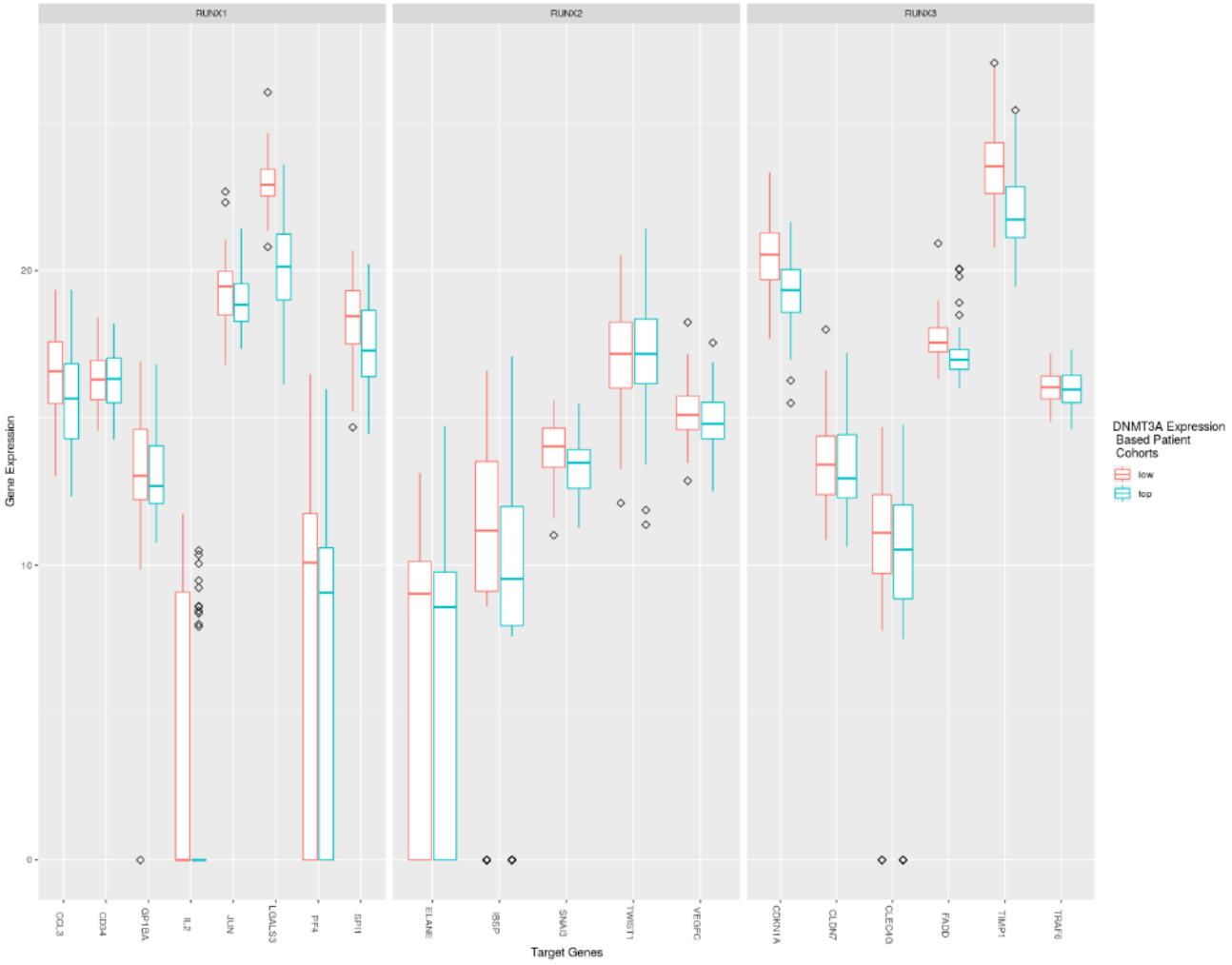


4A: Runt-related transcription factor family (p-value 0.000391) and TBX2-related factor transcription family (p-value 0.011132)harbored most significant and functionally relevant hypermethylated CG probes.



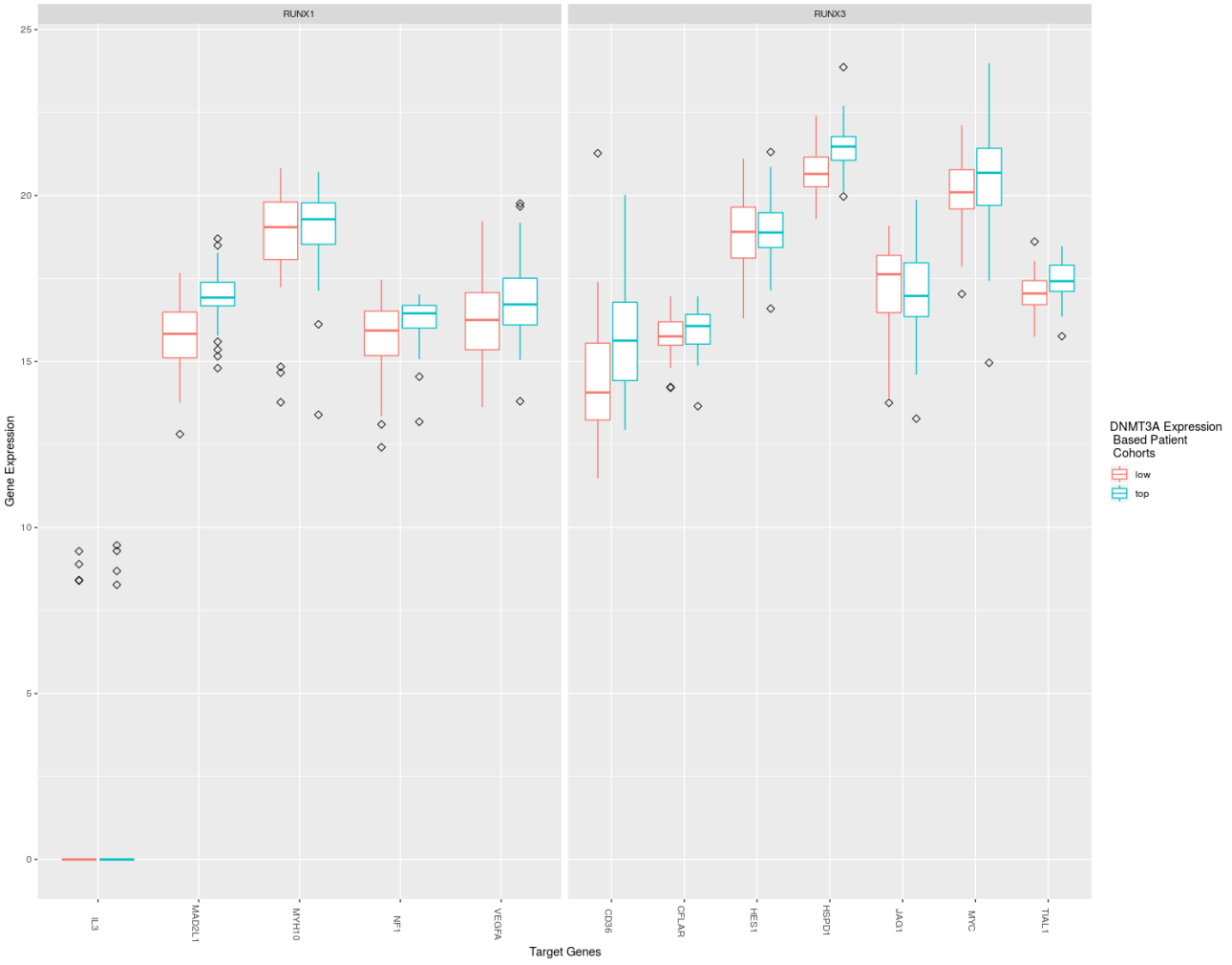
4B: Distribution of functionally signficant CG probes across members of Runt-related transcription factor family. As expected, there were almost no hypomethylated signficant CG probes due to substantial enrichment of hypermethylated CG probes.

Figure 5: Mapping the effect of selective methylation onto their targets. Gene expression changes in targets of RUNX1, RUNX2 and RUNX3 that are conventionally activated(5A) and repressed(5B) were plotted to observe changes as a result of hypermethylation in individual genes.



5A: Gene targets activated by RUNX1, 2 & 3 are prominently immune modulation related factors. As a result of selective hypermethylation a reversal of conventional function was observed.

5B: Oncogenic factors, like MYC, that are repressed by transcription factor RUNX3 in its function as a tumor suppressor were activated in DNMT3A hypermethylated-overexpression cohort.


RUNX1
RUNX3
Gene Expression
25
20
15
10
5
0
IL3
MAD2L1
MYH10
NF1
VEGFA
CD36
CFLAR
HES1
HSPD1
JAG1
MYC
TIAL1
Target Genes
DNMT3A Expression Based Patient Cohorts
low
top

References

1. Micevic, Goran, et al. “Aberrant DNA Methylation in Melanoma: Biomarker and Therapeutic Opportunities.” Clinical Epigenetics, vol. 9, 2017, p. 34. doi:10.1186/s13148-017-0332-8.

2. Issa, Jean-Pierre. “CpG Island Methylator Phenotype in Cancer.” Nature Reviews. Cancer, vol. 4, no. 12, 2004, pp. 988–93. doi:10.1038/nrc1507.

3. Haney, Staci L., et al. “Methylation-Independent Repression of Dnmt3b Contributes to Oncogenic Activity of Dnmt3a in Mouse MYC-Induced T-Cell Lymphomagenesis.” Oncogene, vol. 34, no. 43, Oct. 2015, pp. 5436–46. doi:10.1038/onc.2014.472.

4. Gao, Qing, et al. “Deletion of the de Novo DNA Methyltransferase Dnmt3a Promotes Lung Tumor Progression.” Proceedings of the National Academy of Sciences of the United States of America, vol. 108, no. 44, Nov. 2011, pp. 18061–66. doi:10.1073/pnas.1114946108.

5. Rhee, Ina, et al. “DNMT1 and DNMT3b Cooperate to Silence Genes in Human Cancer Cells.” Nature, vol. 416, no. 6880, Apr. 2002, pp. 552–56. doi:10.1038/416552a.

6. Wang, Jia-Chen, et al. “DNA Methyltransferase 3b Silencing Affects Locus-Specific DNA Methylation and Inhibits Proliferation, Migration and Invasion in Human Hepatocellular Carcinoma SMMC-7721 and BEL-7402 Cells.” Oncology Letters, vol. 9, no. 6, June 2015, pp. 2499–506. doi:10.3892/ol.2015.3077.

7. Sigalotti, Luca, et al. “Epigenetics of Human Cutaneous Melanoma: Setting the Stage for New Therapeutic Strategies.” Journal of Translational Medicine, vol. 8, June 2010, p. 56. doi:10.1186/1479-5876-8-56.

8. Lian, Christine Guo, et al. “Loss of 5-Hydroxymethylcytosine Is an Epigenetic Hallmark of Melanoma.” Cell, vol. 150, no. 6, Sept. 2012, pp. 1135–46. doi:10.1016/j.cell.2012.07.033.

9. Cancer Genome Atlas Network. “Genomic Classification of Cutaneous Melanoma.” Cell, vol. 161, no. 7, June 2015, pp. 1681–96. doi:10.1016/j.cell.2015.05.044.

10. Guan, Jian, Rohit Gupta, and Fabian V. Filipp. "Cancer systems biology of TCGA SKCM: efficient detection of genomic drivers in melanoma." Scientific reports 5 (2015): 7857. doi:10.1038/srep07857

11. Lauss, Martin, et al. “DNA Methylation Subgroups in Melanoma Are Associated with Proliferative and Immunological Processes.” BMC Medical Genomics, vol. 8, Nov. 2015, p. 73. doi:10.1186/s12920-015-0147-4.

12. Fang, Minggang, et al. “Common BRAF(V600E)-Directed Pathway Mediates Widespread Epigenetic Silencing in Colorectal Cancer and Melanoma.” Proceedings of the National Academy of Sciences of the United States of America, vol. 113, no. 5, Feb. 2016, pp. 1250–55. doi:10.1073/pnas.1525619113.

13. Hou, Peng, et al. “The BRAF(V600E) Causes Widespread Alterations in Gene Methylation in the Genome of Melanoma Cells.” Cell Cycle (Georgetown, Tex.), vol. 11, no. 2, Jan. 2012, pp. 286–95. doi:10.4161/cc.11.2.18707.

14. Teneng, I., et al. “Global Identification of Genes Targeted by DNMT3b for Epigenetic Silencing in Lung Cancer.” Oncogene, vol. 34, no. 5, Jan. 2015, pp. 621–30. doi:10.1038/onc.2013.580.

15. Liu, You, et al. “An Association between Overexpression of DNA Methyltransferase 3B4 and Clear Cell Renal Cell Carcinoma.” Oncotarget, vol. 8, no. 12, Mar. 2017, pp. 19712–22. doi:10.18632/oncotarget.14966.

16. Plourde, Karine V., et al. “Genome-Wide Methylation Analysis of DNMT3B Gene Isoforms Revealed Specific Methylation Profiles in Breast Cell Lines.” Epigenomics, vol. 8, no. 9, 2016, pp. 1209–26. doi:10.2217/epi-2016-0013.

17. Micevic, Goran, et al. “DNMT3b Modulates Melanoma Growth by Controlling Levels of MTORC2 Component RICTOR.” Cell Reports, vol. 14, no. 9, Mar. 2016, pp. 2180–92. doi:10.1016/j.celrep.2016.02.010.

18. Lee, Tong Ihn, and Richard A. Young. “Transcriptional Regulation and Its Misregulation in Disease.” Cell, vol. 152, no. 6, Mar. 2013, pp. 1237–51. doi:10.1016/j.cell.2013.02.014.

19. Filipp, Fabian V., et al. “Glutamine-Fueled Mitochondrial Metabolism Is Decoupled from Glycolysis in Melanoma.” Pigment Cell & Melanoma Research, vol. 25, no. 6, Nov. 2012, pp. 732–39. doi:10.1111/pcmr.12000.

20. Jones, Peter A., and Stephen B. Baylin. “The Fundamental Role of Epigenetic Events in Cancer.” Nature Reviews. Genetics, vol. 3, no. 6, June 2002, pp. 415–28. doi:10.1038/nrg816.

21. Frigola, Jordi, et al. “Epigenetic Remodeling in Colorectal Cancer Results in Coordinate Gene Suppression across an Entire Chromosome Band.” Nature Genetics, vol. 38, no. 5, May 2006, pp. 540–49. doi:10.1038/ng1781.

22. Koga, Yasuo, et al. “Genome-Wide Screen of Promoter Methylation Identifies Novel Markers in Melanoma.” Genome Research, vol. 19, no. 8, Aug. 2009, pp. 1462–70. doi:10.1101/gr.091447.109.

23. Lister, Ryan, et al. “Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences.” Nature, vol. 462, no. 7271, Nov. 2009, pp. 315–22. doi:10.1038/nature08514.

24. Wingender, Edgar, et al. “TFClass: Expanding the Classification of Human Transcription Factors to Their Mammalian Orthologs.” Nucleic Acids Research, vol. 46, no. D1, Jan. 2018, pp. D343–47. doi:10.1093/nar/gkx987.

25. Ghoshal, Kalpana, et al. “HOXB13, a Target of DNMT3B, Is Methylated at an Upstream CpG Island, and Functions as a Tumor Suppressor in Primary Colorectal Tumors.” PloS One, vol. 5, no. 4, Apr. 2010, p. e10338. doi:10.1371/journal.pone.0010338.

26. Marzese, Diego M., et al. “Epigenome-Wide DNA Methylation Landscape of Melanoma Progression to Brain Metastasis Reveals Aberrations on Homeobox D Cluster Associated with Prognosis.” Human Molecular Genetics, vol. 23, no. 1, Jan. 2014, pp. 226–38. doi:10.1093/hmg/ddt420.

27. Rinn, John L., et al. “Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs.” Cell, vol. 129, no. 7, June 2007, pp. 1311–23. doi:10.1016/j.cell.2007.05.022.

28. Gupta, Rajnish A., et al. “Long Non-Coding RNA HOTAIR Reprograms Chromatin State to Promote Cancer Metastasis.” Nature, vol. 464, no. 7291, Apr. 2010, pp. 1071–76. doi:10.1038/nature08975.

29. Kim, K., et al. “HOTAIR Is a Negative Prognostic Factor and Exhibits Pro-Oncogenic Activity in Pancreatic Cancer.” Oncogene, vol. 32, no. 13, Mar. 2013, pp. 1616–25. doi:10.1038/onc.2012.193.

30. Wu, Liang, et al. “Binding Interactions between Long Noncoding RNA HOTAIR and PRC2 Proteins.” Biochemistry, vol. 52, no. 52, Dec. 2013, pp. 9519–27. doi:10.1021/bi401085h.

31. Kilbey, Anna, et al. “Oncogene-Induced Senescence: An Essential Role for Runx.” Cell Cycle , vol. 7, no. 15, Aug. 2008, pp. 2333–40. doi:10.4161/cc.6368.

32. Ito, Yoshiaki, and Kohei Miyazono. “RUNX Transcription Factors as Key Targets of TGF-Beta Superfamily Signaling.” Current Opinion in Genetics & Development, vol. 13, no. 1, Feb. 2003, pp. 43–47.

33. Hong, Deli, et al. “Runx1 Stabilizes the Mammary Epithelial Cell Phenotype and Prevents Epithelial to Mesenchymal Transition.” Oncotarget, vol. 8, no. 11, Mar. 2017, pp. 17610–27. doi:10.18632/oncotarget.15381.

34. Li, Qing Lin, et al. “Causal Relationship between the Loss of RUNX3 Expression and Gastric Cancer.” Cell, vol. 109, no. 1, Apr. 2002, pp. 113–24.

35. Yamamura, Yasuko, et al. “RUNX3 Cooperates with FoxO3a to Induce Apoptosis in Gastric Cancer Cells.” The Journal of Biological Chemistry, vol. 281, no. 8, Feb. 2006, pp. 5267–76. doi:10.1074/jbc.M512151200.

36. Zhang, Zhizhong, et al. “Prognostic Significance of RUNX3 Expression in Human Melanoma.” Cancer, vol. 117, no. 12, June 2011, pp. 2719–27. doi:10.1002/cncr.25838.

37. Boregowda, Rajeev K., et al. “RUNX2 Is Overexpressed in Melanoma Cells and Mediates Their Migration and Invasion.” Cancer Letters, vol. 348, no. 1–2, June 2014, pp. 61–70. doi:10.1016/j.canlet.2014.03.011.

38. Wang, Hao, et al. “Concurrent Hypermethylation of SFRP2 and DKK2 Activates the Wnt/β-Catenin Pathway and Is Associated with Poor Prognosis in Patients with Gastric Cancer.” Molecules and Cells, vol. 40, no. 1, Jan. 2017, pp. 45–53. doi:10.14348/molcells.2017.2245.

39. Laurikkala, Johanna, et al. “Identification of a Secreted BMP Antagonist, Ectodin, Integrating BMP, FGF, and SHH Signals from the Tooth Enamel Knot.” Developmental Biology, vol. 264, no. 1, Dec. 2003, pp. 91–105.

40. Clausen, Kathryn A., et al. “SOSTDC1 Differentially Modulates Smad and Beta-Catenin Activation and Is down-Regulated in Breast Cancer.” Breast Cancer Research and Treatment, vol. 129, no. 3, Oct. 2011, pp. 737–46. doi:10.1007/s10549-010-1261-9.

41. Ferrari, Karin J., et al. “Polycomb-Dependent H3K27me1 and H3K27me2 Regulate Active Transcription and Enhancer Fidelity.” Molecular Cell, vol. 53, no. 1, Jan. 2014, pp. 49–62. doi:10.1016/j.molcel.2013.10.030.

42. Ku, Manching, et al. “Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains.” PLoS Genetics, vol. 4, no. 10, Oct. 2008, p. e1000242. doi:10.1371/journal.pgen.1000242.

43. Peters, Antoine H. F. M., et al. “Partitioning and Plasticity of Repressive Histone Methylation States in Mammalian Chromatin.” Molecular Cell, vol. 12, no. 6, Dec. 2003, pp. 1577–89

Chapter 6

Conclusion

Understanding how diverse genomic events converge to support cell fate decisions is an important question in oncology. It is essential to understand the processes that govern the reshaping of the cellular state in order to promote a disease state. Greater emphasis on studying the genomic basis of the tumor cell is required, to gain insights into mechanistic principles which would pave the way for the design of efficient, personalized, and targeted therapies. Significant work needs to be done to formalize the comprehensive map of changes that aid in the development and progression of tumors. Genomic analysis allows leveraging of modern computing frameworks for faster results, producing valuable, high-resolution, and large-scale data of the cancer cell state. In this dissertation, I have focused on the development of detection methods and workflows to study the oncogenic factors that drive the malignant progression and metastasis in Melanoma. A significant component of this thesis relates to the application of systems biology approaches through the invention and use of bioinformatics-based methods on Melanoma datasets from The Cancer Genome Atlas (TCGA).

In the first project, listed in Chapter 3 of this dissertation, the genomic landscape of primary and metastatic Melanoma was characterized. This project further emphasizes that the universe of somatic mutation in Melanoma is more expansive than previously identified. This work, with nearly 300 whole-exome samples, nearly doubled the sample pool at the time. Furthermore, several novel

driver oncogenic mutations and 10 novel oncogenes that play a significant role in Melanoma oncogenesis and progression were identified. This body of work once again establishes the prevalence of UV-induced damage and BRAF hotspot mutations in the mutation landscape of Melanoma. The novelty introduced by this study is the multi-step pipeline (Section 2.1.2) to identify driver oncogenic mutations with the highest mutation burden, indicative of positive selection in Melanoma. The use of a probabilistic model for permutation in tandem with multi-step filter enables: a) the detection of predominant hotspot mutation regions in the genome, b) the ranking of mutations by degree of positive selection, and c) validation of pathway enrichment. The novel oncogenic factors identified in this project work depict the extensive gene/pathway network remodeling in Melanoma due to accumulated driver oncogenic mutations. The discovery of novel oncogenic driver mutations at a splice site of the TMEM216 gene presents a new class of Melanoma driver mutations. The concerted deregulation caused by somatic mutations and copy number alterations provide further evidence of the robust nature of Melanoma and tumors in general.

Future studies could expand this knowledge base by use of recent genomic technologies, such as single-cell genomic analysis from circulating tumor DNA (ctDNA), for discovery of somatic mutations at low and ultra-low frequencies. Additionally, single-cell genomics would also be instrumental in resolving tumor heterogeneity because each cell is considered a separate entity and whole-genome and whole-exome data fail to capture intricacies within each cell of a population.

The metabolism of cancer cells is known to support continued growth under a diverse set of cellular conditions. The complex metabolic requirements of dividing, proliferating, or nutrient-limited cancer cells illustrates that tumor cells utilize a diverse and extensive metabolic rewiring. However, being able to effectively translate this knowledge of strict nutrient requirements in a rapidly growing tumor cell mass to a therapeutic regimen is not widespread. Furthermore, in-depth understanding of the synergy between genomic events such as somatic mutations and factors that govern tumor cell metabolism is lacking. The second project, described in this thesis work as part of Chapter 4 , addresses the question of influence of somatic mutations and gene signature associated with the DPYD gene

to promote pyrimidine metabolism in Melanoma. Chapter 4 utilizes the detection pipeline (Section2.1.2) from the previous project listed in Chapter 3 of this thesis, and integrates other genomic components such as gene expression, gene-expression signature, and structural impact of somatic mutations. This project revealed the process of hypermutation in a metabolic gene (DPYD) that ultimately leads to deregulation of pyrimidine and nucleic acid pathways to promote malignant progression of Melanoma. The TCGA SKCM cohort of 471 Melanoma patients used in this project had a statistically significant enrichment of DPYD mutations for positive selection. The novelty of this work is integration of several facets of genomic data to propose a possible causal link between driver oncogenic mutations that perturb pyrimidine metabolic pathways for malignant disease progression. The signature of bifurcation in pyrimidine metabolism might be useful as a prognostic signature in Melanoma to predict the risks of 5-FU toxicity.

Chapter 5 of this dissertation proposes a framework and associated tools to comprehensively map the oncogenic scope of epigenetic events linked with Melanoma. This project outlined systemic regulatory methylation in Melanoma tied with over-expression in key methyl enzymes, specifically, a disruption of tumor suppressor genes and transcription factors through focal hypermethylation events to limit gene expression and actively promote malignant progression of Melanoma. In addition, this project provides an example of a congruent relationship between two distinct genomic events: somatic copy number alterations and methylation. Upon examining the genomic regions of focal hypermethylation, silencing of tumor suppressor genes and transcription factors was found to be the de facto cause of the malignant state of the cell. The novelty of this work is the bundled package that enables similar epigenomic analysis across any cancer tissue with use of data from TCGA. The findings of this project further illustrate active modulation by the recruitment of epigenomic mechanisms known to be associated with cancer. Together, these results indicate that over-expressed DNMT3A & DNMT3B enzymes function as potent regulators and play a substantial role in promoting malignant progression of Melanoma. This strategy of combined analysis of genomic and epigenomic datasets is not just limited to Melanoma; the underlying methods used in this project are broadly applicable for resolving questions pertinent to

other malignancies.

Together these results represent a cancer systems biology investigation of Melanoma datasets and explain the genomic basis of the defining characteristics of a tumor cell. An extensive genomic and epigenomic network was found to regulate several aspects of cell state. Moreover, genomic mechanisms were found to be one of the primary modes of propagation for the transformational regulatory signals essential for the development and continued growth of tumor cells. Further development of genomic methods in conjunction with other omics strategies will permit cancer systems biologists to better characterize key cellular states and processes. The application of systems biology principles to clinical problems will continue to usher in an era of precision-guided, personalized medicine. Furthermore, bringing quantitative measures and systems-wide genomics approaches to a clinically-relevant field like cancer will provide researchers with a rational, targeted approach to cancer. Overall, I envision my future efforts to focus on the development of novel cancer genomic methods that further delineate the tumor cell state, which inevitably shall provide better therapeutic strategies.