Identify Genetic and Epigenetic Abnormalities Associated with Therapeutic Vulnerability in Human Cancers

Wen-Wei Liang
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

Part of the Bioinformatics Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
WASHINGTON UNIVERSITY IN ST. LOUIS
Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:
Li Ding, Chair
Christopher Maher
Nancy Saccone
Tim Schedl
Jieya Shao

Identify Genetic and Epigenetic Abnormalities Associated with Therapeutic Vulnerability in Human Cancers
by
Wen-Wei Liang

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the requirements for the degree of Doctor of Philosophy

January 2021
St. Louis, Missouri
# Table of Contents

List of Figures..................................................................................................................v
Acknowledgments ...........................................................................................................vi
Abstract of the Dissertation.............................................................................................ix
Chapter 1: Introduction ....................................................................................................1
  1.1 Cancer is a heterogeneous and complex disease driven by genetic mutation........2
  1.2 Genomics empowers the elucidation of the molecular architecture and driver
candidates in human cancers.........................................................................................4
  1.3 Figures ...................................................................................................................8
  1.4 Reference ..............................................................................................................9
Chapter 2: Diverse impacts of microsatellite instability across 33 cancer types.............11
  2.1 Abstract ..............................................................................................................13
  2.2 Introduction .........................................................................................................13
  2.3 Results ................................................................................................................15
    2.3.1 Identify canonical and noncanonical MSI tumors across 33 cancer types....15
    2.3.2 Canonical and noncanonical MSI tumors are associated with elevated mutation rate and
    MMR deficiency in a lineage-dependent manner.....................................................15
    2.3.3 Microsatellite-derived neoantigen from canonical and noncanonical MSI tumors are
    immunogenic ........................................................................................................17
  2.4 Discussion .............................................................................................................17
  2.5 Methods ..............................................................................................................18
    2.5.1 Data Preparation for TCGA samples..............................................................18
    2.5.2 Data Preparation for CPTAC samples.............................................................19
    2.5.3 Creation of MMR Gene Set ............................................................................20
    2.5.4 Estimate MSI Status By Using MSIsensor .....................................................21
    2.5.5 HLA Typing and Neoantigen Prediction From Microsatellite Sites ............21
  2.6 Figures ................................................................................................................23
  Reference .....................................................................................................................28
Chapter 3: Driver Fusions and Their Implications in the Development and Treatment of Human
Cancers ..............................................................................................................................30
  3.1 Abstract ..............................................................................................................33
  3.2 Introduction .........................................................................................................33
  3.3 Results ................................................................................................................35
Chapter 4: Integrative multi-omic profiling reveals tumorigenic DNA methylation associated with therapeutic vulnerability

4.1 Abstract .................................................................................................................. 68
4.2 Highlights ................................................................................................................ 68
4.3 Introduction .............................................................................................................. 69
4.4 Results ...................................................................................................................... 71
  4.4.1 Integrative multi-omic profiling empowers aberrant DNA methylation discovery .................................................................................................................. 71
  4.4.2 cis-acting aberrant DNA methylation on driver genes can be tumorigenic .......................................................................................................................... 73
  4.4.3 Hypomethylated RTKs are newly identified driver events .................................. 74
  4.4.4 Transcription factor STAT5A hypermethylation is associated with pervasive changes in STAT5A regulome activity ........................................................................ 76
  4.4.5 Hypermethylated STAT5A is associated with immune cells depletion in HNSCC ........................................................................................................ 77
  4.4.6 Aberrant methylation associated with therapeutic vulnerability in cancer ........................................................................................................ 79
4.5 Discussion ................................................................................................................. 81
4.6 STAR Methods ........................................................................................................ 83
  4.6.1 CPTAC datasets description .............................................................................. 83
  4.6.2 DNA methylation data preprocessing ................................................................ 84
  4.6.3 Defining aberrant DNA methylation using RESET ........................................... 85
  4.6.4 Defining cancer-associated genes ...................................................................... 85
List of Figures

Figure 1.1 Dissertation Overview ........................................................................................................8
Figure 2.1 Relationship among MMR deficiency, mutation frequency, and MSI status across 21 cancer types. .......................................................................................................................... 23
Figure 2.2 Quantitative MSIsensor scores correlated with qualitative results from gel-assay .... 24
Figure 2.3 Exome-based and WGS-based MSIsensor scores are correlated. ................................. 25
Figure 2.4 Canonical and noncanonical MSI tumors are significantly associated with high mutational burden. .......................................................................................................................... 26
Figure 2.5 Canonical and noncanonical MSI tumors are validated in independent dataset. ...... 27
Figure 2.6 Microsatellite-derived neoantigens are potentially immunogenic............................ 28
Figure 3.1 Fusion detection and landscape in cancer. ..................................................................... 52
Figure 3.2 Fusion expression outliers ............................................................................................... 54
Figure 3.3 Protein kinase fusions .................................................................................................... 56
Figure 3.4 Kinase gene expression regulated by fusion ................................................................. 58
Figure 3.5 Mutual exclusivity between driver mutations and driver fusions ............................... 60
Figure 3.6 Druggable fusion targets ............................................................................................... 61
Figure 4.1 The cancer methylome landscape associated with transcriptomic and proteomic change .................................................................................................................................. 88
Figure 4.2 Characterization of aberrant methylation in driver genes .......................................... 90
Figure 4.3 Collaborative effects of FGFR2 mutations and hypomethylation on FGFR2 upregulation .................................................................................................................................. 91
Figure 4.4 STAT5A hypermethylation associated with pervasive STAT5A regulome changes. .... 93
Figure 4.5 Functional impact of STAT5A hypermethylation on HNSCC tumor immunosuppression .................................................................................................................................. 94
Figure 4.6 Summary of the cancer methylome for tumor signatures and therapy ..................... 95
Figure 5.1 The cancer methylome landscape across different cancer types ............................... 108
Acknowledgments

Years ago, I spent numerous afternoons in a public library after school, reading page by page for the knowledge underlying common nature exists in all of the things spread before me yet I didn’t understand. Years after, what I understand is still limited, but I am grateful I am becoming a scientist spending my career on pursuing knowledge. I won’t be able to become the person and scientist I am today without the support and friendship from everyone standing by me for the past six years.

First I would like to thank my committee members, Dr. Nancy Saccone, Dr. Christopher Maher, Dr. Tim Schedl, and Dr. Jieya Shao. Every time when I presented my works to them or ran into each other on campus, their support and constructive feedback really helped to shape the works.

Next I would like to thank my thesis mentor, Dr. Li Ding. Li is definitely a role model as a female scientist, especially for international students like me. Most of the days in PhD life are a haze of frustration, while Li’s passion toward science and big visions in cancer research drive away the pain. Thanks for Li’s invitation for joining the Ding lab. It turns out joining Ding lab is the best decision I have ever made.

Everyone in the Ding lab is very talented and works very hard to pursue science as a team. We help each other to overcome every deadline and struggle. In the meantime, we can quickly adapt and evolve under stress, just like tumors - bear with me for my bad analogy! I would like to thank my wonderful members to help me get through all the struggles, especially for Yige, Reyka, and Steven. I cannot imagine that I will be here today without you guys around. You guys are always so encouraging and have so much confidence in me. Outside of the lab, we celebrated each other’s important moments, enjoyed the food, and were being fancy or weird together. I’ve enjoyed every moment in Ding lab. I hope I have passed along the vibe to the younger generation in the Ding lab.

I still remember the moment when I received the admission letter from DBBS/MGG, which is the dream school in my application list. In St. Louis, I have met lots of good friends from different places and accomplished lots of first-timer things due to the supportive environment provided by Wash U. Although we haven’t met each other for a long time due to pandemic, I missed the times when we ran into each other at Kaldi’s Café and talked about all the happy things
and struggles. I am so lucky to have all the wonderful peers to grow together at the early stage as a scientist.

I also want to thank the Taiwan - Washington University in St. Louis Fellowships to support my study and all the Taiwanese organizations in St. Louis. As a Taiwanese, we naturally connected and recognized each other when we ran into each other in St. Louis. It is wonderful to have all the Taiwanese friends to hang out, enjoy Taiwanese food, and promote Taiwanese culture together, especially for traditional holidays like Lunar New Year and Mid-Autumn Festival. With all my Taiwanese folks around, it makes St. Louis my second home.

I would like to specially thank my GUZI friends. When we met each other in 2014 in St. Louis, most of us were wondering what kind of the career trajectory we were going to have in the next few years. Now we are in 2021 and four of us become a PhD, and three of us are or are going to be a medical doctor. I am grateful to have my GUZI friends all the time to walk through all the ups and downs. Hopefully we can soon have a gathering after the pandemic.

Finally, I would like to thank my loving family, my sisters Wen-Ching and Wen-Rou. Without their support, I won’t be able to pursue what I love in the States. I thank my parents for their unconditional love and support. I know they are looking after me all the time as always. And last but not least, thank my cats Pearl and Milktea. Most of the time I had been seeing everything around me through a haze, but my furry family just give off light every time when I get home.

Wen-Wei Liang

Washington University in St. Louis
January 2021
Dedicated to my family, for their support and love
Abstract of the Dissertation

Identify Genetic and Epigenetic Abnormalities Associated with Therapeutic Vulnerability in Human Cancers

by

Wen-Wei Liang

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2021

Professor Li Ding, Chair

Human cancer is a complex and dynamic disease with mutational, spatial, and temporal heterogeneity. There has been a concerted effort to address the heterogeneity by profiling large-scale multiomic datasets with an emphasis on abnormalities such as DNA mutations at coding regions and structural variants. However, a systematic analysis of alterations at “dark matter” is lacking – including DNA slippage events at microsatellite sequences, gene fusions arising from genomic rearrangements, and aberrant DNA methylations at promoter regions. This dissertation focuses on developing data-driven computational analysis pipelines to study the patterns of microsatellite instability, gene fusions, and aberrant DNA methylation, and their functional impacts within and across cancer types, utilizing integrative system-biology approach that combines multi-dimensional genomic, epigenomic, proteomic and clinical data. First we estimated the mutational load and microsatellite instability (MSI) status of 10,980 tumors across 33 cancer types from The Cancer Genome Atlas (TCGA). Beyond the well-characterized canonical MSI-prone tumor types, we identified additional MSI-high tumors in other non-canonical MSI tumor
types and further validated in another independent data set from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). A survey of the 993 CPTAC tumors across 7 cancer types revealed tumors with high MSI were associated with high number of predicted microsatellite-derived neoantigen, suggesting that the aberrant expansion and deletion of microsatellite sequence is immunogenic even the non-canonical MSI-high tumor. Next, we focuses on investigating gene fusions in 9,624 TCGA tumor samples across 33 cancer types predicted by multiple RNA-sequencing-based fusion calling tools and validated by orthogonal whole-genome sequencing-based approach. We demonstrated that gene fusions are mutually exclusive with the other driver mutations in most of the cancer types, and function as the sole driver in more than 1% of cancer cases. Lastly, we leveraged the complementary nature of RNA-seq and proteomic data to identify aberrant DNA methylation leading to both transcriptional and translational changes. The integrated multi-omic profiling cataloged epigenomic aberrations of 506 CPTAC tumors across five cancer types, highlighting key changes to driver genes that affect cancer hallmark pathways in a coordinated manner. Overall, our systematic pan-cancer studies uncover determinants and consequences of genetic and epigenetic variation beyond the conventional mutational profiling, revealing potential new disease mechanisms and therapeutic opportunities.
Chapter 1: Introduction
Cancer is the second leading cause of deaths globally, nearly 1 in 6 deaths is due to cancer. With decades of effort in learning cancer genome, precious lessons from genome sequencing studies show that each cancer type has distinct mutation pattern and treatment responsiveness\textsuperscript{1,2}. While breathtaking “mountains”, such as recurrently mutated genes, has greatly advanced our knowledge of human cancer, little is known about the “hidden hills” that are located outside of protein coding regions\textsuperscript{3}. New strategies for the prevention and control of cancer will rely on a thorough understanding of the genetic factors both at the coding and noncoding regions in human genome.

1.1 Cancer is a heterogeneous and complex disease driven by genetic mutation

Cancer cells are able to undergo unrestrained proliferation, invade surrounding tissue, and metastasize to distant organs. The molecular mechanisms that underlie these characteristics of cancer cells are the sustained proliferative signaling, evasion of growth suppressors, resistance to cell death, replicative immortality, induction of angiogenesis, activation of invasion and metastasis, reprogramming of energy metabolism, and evasion of immune destruction\textsuperscript{4}. Generally, these hallmarks of cancer cells are driven by genetic variation, which convert a normal cell to a benign tumor, or a benign tumor to a malignant one.

In the process of tumor development, tumor cells undergo continual evolution to establish malignancy. Various mutations are acquired and selected against during this process. The mutations that confer a selective growth advantage to the tumor cell over other cells are driver mutations, which are thought to have a causal role in cancer initiation and progression. On the other hand, co-occurring mutations that do not affect the fitness of cancer cells are passenger mutations, which are generated by chance, as a result of the constant background mutation rate\textsuperscript{3,5}.
Cancer cells usually carry high numbers of passenger mutations during tumorigenesis because they are neutral with respect to selection. Therefore, distinguishing driver mutations from commonly abundant passenger mutations is critical for the understanding of tumor development.

A typical solid tumor has been estimated to contain about 30 to 70 nonsynonymous mutations that have accumulated during development, while it requires only 2 to 8 driver mutations for tumorigenesis. Most of these driver mutations are found to either activate oncogenes or inactivate tumor suppressor genes. Oncogene activation or overexpression usually drives the growth of the cell in which it resides. One prominent example of an oncogene in breast cancer is human epidermal growth factor receptor 2 (HER2). It is a tyrosine kinase receptor that can dimerize with itself or other membrane receptors. The dimerization/oligomerization of HER2 initiate various proliferation-promoting signaling pathways, including mitogen-activated protein kinase (MAPK) and the phosphatidylinositol 3-kinase (PI3K) pathways. Gain-of-function mutations in HER2 account for approximately 20% of breast cancer tumors. In addition, inhibition of the aberrant or overabundant HER2 proteins is able to reverse the oncogenic effect of HER2, which makes HER2 an attractive therapeutic target in HER2-positive patients. Indeed, HER2 targeted therapy has improved the prognosis and outcome for breast cancer patients in the past decade.

Tumor suppressor genes typically function as negative regulators of cancer cell proliferation, which is critical for preventing normal cells from malignant transformation. Inactivation or reduction of tumor suppressor genes that result from loss-of-function mutations contributes to cancer progression. The most frequently mutated tumor suppressor gene across all cancer types, including breast cancer, is tumor protein p53 (encoded by TP53). Upon DNA damage or other stress, p53, a transcription factor, activates a wide range of cellular pathways to
maintain genome integrity and prevent tumor formation. The trans-activated pathways include apoptosis, cell cycle arrest, autophagy, anti-oxidant defense, etc. The loss of p53 function through global transcriptional and translational repression leads to tumorigenesis\textsuperscript{12}. Although inactivation of tumor suppressor genes is more predominately found than over activation of oncogenes in tumors, it is more difficult to attack cancer cells with loss-of-function mutations than those with gain-of-function mutations. Restoring the function of tumor suppressor genes remains a difficult task in cancer therapy.

Cancer is a complex and dynamic ecosystem with mutational, spatial, and temporal heterogeneity. The heterogeneity can be exemplified by how passenger mutations can be a double-edged sword in the therapeutic resilience of cancer. Passenger mutations are usually hidden inside a tumor before treatment, but some of these mutations may convert to driver mutations and promote cancer regrowth once the selective environment is changed\textsuperscript{5}. On the other hand, passenger mutations can serve as patient- and tumor-specific neoantigens that boost the ability of endogenous T cells to destroy cancer cells\textsuperscript{13}. The mixed roles of passenger mutations are just the tip of the iceberg in regards to the complexity of cancer. To comprehensively address the heterogeneity and complexity of human cancer, conducting genome-wide analysis in human cancer is imperative.

1.2 Genomics empowers the elucidation of the molecular architecture and driver candidates in human cancers

Advances in high-throughput sequencing and bioinformatic technologies have dramatically improved our understanding of molecular aberrations in human malignancies in recent years. When the first complete draft of the human reference genome was released in 2004\textsuperscript{14}, it became possible to assess cancer genomes, at single-base resolution. In general, genomic DNA or RNA isolated from tumor samples are sequenced and compared with those from matched
normal tissues. Through comparison, different types of genetic alteration are identified, including single nucleotide variants (SNVs), insertions or deletions (INDELs), structural rearrangements, and copy-number alterations. Moreover, regions of genomes that differ frequently enough are considered as potential causal variants that lead to malignancy. Since the first whole-genome comparison of an acute myeloid leukemia patient was published in 2008\textsuperscript{15}, more and more cancer genomes have been sequenced and are available for further investigation.

Extensive national or international collaboration projects, such as The Cancer Genome Atlas (TCGA) or the Clinical Proteomic Tumor Analysis Consortium (CPTAC), have been established to promote the generation, analysis, and sharing of cancer genome sequencing data. The US-based TCGA aims to assess the genomic, transcriptomic, and epigenomic aberrations in 100 to 1,000 tumors from each of 33 cancer types. The CPTAC project also seeks to study those aberrations in cancer cells across 10 different cancer types with an emphasis of large-scale mass spectrometry-based proteome data. These large consortia have made a giant leap in cancer genome analysis in several ways. First, the well-coordinated workflows for sequencing and processing avoid redundant and laborious effort in generating omics data. Second, processed data are widely accessible to all investigators, facilitating data sharing and interpretation. Third, the multiple data types that were extracted from the same set of samples allow researchers to perform integrative analysis easily in a quality-controlled fashion. Finally, by providing a large number of samples for each cancer type, the consortia improved the statistical power in detecting driver mutations. As a result, the large-scale projects from TCGA and CPTAC are invaluable resources for systematic characterization of cancer genomes\textsuperscript{16}.

Analyses of the sequenced cancer genome have generated several statistics on mutation patterns of human cancer. Each solid tumor has an average of 33 to 66 genes having non-
synonymous somatic mutations, and about 95% of these mutations are SNVs. There are at least 299 genes identified as significantly mutated genes across 33 cancer types, suggesting that those genes are associated with tumorigenesis in a particular tumor or multiple tumors. The number of cancer-associated genes is still growing as samples and cancer types are added. The compiled statistics also revealed that different cancers have distinct genomic properties in terms of mutation frequency of single-point mutations, collection of frequently mutated genes, and mutational spectrum, to name a few. For example, in regards to mutation frequency, melanomas and lung tumors contain approximately 200 nonsynonymous mutations per tumor due to the involvement of potent mutagens, while pediatric tumors and leukemias have around 10 mutations per tumor. Those studies highlight the complex nature of human cancers and serve as foundation for subsequent in-depth investigations of mutational patterns both across and within individual tumor types.

The use of algorithms to pinpoint cancer-driving variants before experimental validation is important in the mining of cancer genomes. With the knowledge inferred from evolutionary conservation, protein domain analysis, and cancer-associated cellular pathways, prioritization of non-synonymous variants based on their likely damaging effect is well established in variant interpretation. The most common strategy for identifying coding driver candidates is first to map non-synonymous variants to protein-coding regions, and then search recurrently mutated genes in a cohort of cancer patients. A recurrence pattern suggests that those mutations underwent positive selection during tumorigenesis. Other strategies to detect signals of cancer gene positive selection includes conservation of the nucleotide or amino acid sequence relative to other sequenced species, functional impact prediction on the encoded protein, and combination of mutations at the pathway/network level, etc. Integrating two or more signals of positive selection
based on their complementary biological features has been demonstrated to improve the prediction accuracy of each individual approach, and holds great potential for finding reliable driver genes.$^2$1

With the progress in sequencing technique, collaborative analyses, and bioinformatic development, a genomic era of cancer studies is developing rapidly, providing an unprecedented opportunity to comprehensively characterize the human malignancies. However, most of the studies concentrate on recurrent and somatic mutations in the coding fraction of the cancer genome, which represents only 2% of the human genome. Although the work remains incomplete, understanding the cancer features outside of coding regions is an important next step for gaining deeper insight into the cancer genome landscape. Therefore, we set out to systematically elucidate several less explored features in cancer genome, including DNA slippage events at microsatellite sequences, gene fusions arising from genomic rearrangements, and aberrant DNA methylations at promoter regions (Figure 1).
1.3 Figures

Characterize the hidden landscape of cancer genome

The Cancer Genome Atlas (TCGA)
• 10,980 TCGA tumors
• 33 cancer types
• WXS, RNA-seq, DNA methylation

Clinical Proteomic Tumor Analysis Consortium (CPTAC)
• 506 CPTAC tumors
• 5 cancer types
• WXS, RNA-seq, DNA methylation, Proteomic data

Figure 1.1 Dissertation Overview
1.4 Reference


Chapter 2: Diverse impacts of microsatellite instability across 33 cancer types
Preface

Portions of this chapter have been reproduced and adapted from the following published manuscripts:


&


*indicates co-first authors
2.1 Abstract
The DNA mismatch repair (MMR) system corrects erroneous insertions, deletions, and substitutions during DNA replication. MMR deficiency is the canonical cause of microsatellite instability (MSI) in hereditary and sporadic colorectal and endometrial cancers, but the molecular and clinical consequences of MMR deficiency in other cancer types have not been extensively characterized. Here, we present an analysis of genomic and epigenomic alterations in 39 MMR-related genes and their downstream effects in over 10,000 human cancers. We find that mutation and methylation status of MMR genes has cancer type-specific effects on point mutation frequency and gene expression. Further, using an NGS-based tool to determine microsatellite instability (MSI) status of diverse cancers, we relate MMR status to MSI. Finally, we relate immune response and clinical outcome to MMR status, and suggest a possible mechanism by which MMR deficient cancers demonstrate a brisk response to immunotherapy.

2.2 Introduction
The DNA mismatch repair (MMR) machinery maintains genomic stability by correcting erroneous insertions, deletions, and base substitutions during DNA replication. A set of evolutionarily conserved core proteins, including MLH1, MLH2, PMS2, and MSH6, recognize the distortions of the helical structure caused by mismatches, direct MMR machinery onto the freshly replicated DNA strand that contains the errors, remove several nucleotides of DNA containing the mismatch, and correctly resynthesize and ligate the DNA\(^1\). When genes in the MMR machinery malfunction, mutations accumulate, and a phenotypic hallmark of MMR deficiency occurs, that is microsatellite instability (MSI).

Microsatellite are repetitive sequences of DNA with a unit length ranging from one to six bases in discrete sites throughout the genome. Since those sites are particularly vulnerable to errors
during DNA replication, MSI is introduced when MMR machinery is dysfunctional. MSI is an abnormal extension and shortage of tandem repeats when compared to normal, which has been identified and characterized as an important biomarker in several cancer pathological associations. MSI is observed in about 20% of gastrointestinal, 30% of endometrial, and 15% of colorectal cancers\(^1\).

MSI has traditionally been assessed using the reference Bethesda panel of 5 or 7 repeat markers. While PCR-based methods using Bethesda criteria are the gold standard for MSI detection, they are expensive, optimized for colorectal and endometrial cancers only, and limited to a small subset of microsatellites. Next-generation sequencing has enabled genome-wide MSI profiling. Because MSI and other hypermutator phenotypes are predictors of response to immunotherapy, there has been renewed interest in characterizing MSI across cancer types using NGS tools. For example, a recent effort profiled instability signatures in microsatellites across coding regions from 6,747 human tumors. The number and patterns of microsatellites can distinguish MSI-high tumors from microsatellite stable tumors\(^2\).

Using MSIsensor\(^3\), a software tool that quantifies MSI in paired tumor-normal genome sequencing data, we characterized the mutational landscape and MSI status across cancer types, and relate these patterns to somatic and germline mutations in MMR genes. We then identified the contributions of mutations and epigenetic changes in individual MMR genes to the MSI and hypermutator phenotypes. Finally, we examined the clinical consequences of MMR-deficient tumors as it relates to immunogenicity.
2.3  Results

2.3.1  Identify canonical and noncanonical MSI tumors across 33 cancer types

We determined the MSI landscape in cancer patients by analyzing the TCGA exome sequencing data from 9,423 tumors and matched normal pairs across 33 cancer types. MSI status was first quantified by MSIsensor score and characterized as MSI-High (score > 10), MSI-Low (3.5 < score < 10), and microsatellite stable (score < 3.5). As expected, most of the MSI-High/MSI-Low tumors are from the five well-characterized canonical MSI-prone tumor types (Figure 1), namely UCEC (31.1%), STAD (18.8%), COAD (17.8%), READ (6.8%), and ESCA (1.6%). MSIsensor results in these cancers agreed well with orthogonal MSI-PCR clinical status reported by TCGA (p=1.9e-198, Fisher's exact test, Figure 2). Notably, we also identified a few MSI-High/MSI-Low tumors in other non-canonical MSI tumor types, such as KICH (12.1%), PAAD (6.0%), OV (3.4%), KIRC (1.8%) and BRCA (1.3%). These observations suggest the more general importance of taking MSI status into consideration, even for noncanonical cancer types, which currently appear to be overlooked in analysis and clinical prognosis.

In order to assess the difference between exome-based and WGS-based MSIsensor scores, we also inferred the latter from available WGS data from 932 tumors and matched normal pairs in parallel (Figure 3). Score correlations varies by cancer type, with UCEC, STAD, DLBC, SARC, SKCM, LUSC, KIRP, and COAD showing good agreement (R² ranges from 0.94 to 0.47) and others showing mediocre to poor agreement (R² < 0.42).

2.3.2  Canonical and noncanonical MSI tumors are associated with elevated mutation rate and MMR deficiency in a lineage-dependent manner

To further distinguish the determinants and consequences between canonical and noncanonical MSI tumors, we investigated the mutation rates and its relationship with MMR
deficiency. We systematically profiled point mutations in coding regions to estimate the incidence of somatic mutation (Figure 4A). As expected, most of the canonical MSI tumors have significantly higher number of point mutations (median = 965 point mutations) than noncanonical MSI tumors and non-MSI tumors, confirming canonical MSI tumors are hypermutators. Although the number of point mutations of noncanonical MSI tumors (median = 677 point mutations) is significantly lower than canonical MSI tumors, it remains significantly higher than that of non-MSI tumors. Our result suggests that noncanonical MSI tumors are associated with hypermutators, indicating a lineage-dependent response of MMR deficiency.

Next, we south out to relate the deleterious mutations on MMR genes to MSI tumors without MLH1 hypermethylation. The majority of canonical MSI tumors have mutated POLE (n=25) or POLD1 (n=17) (Figure 4B), in line with the finding that POLE and POLD1 are consistently more prevalent in MSI tumors. Notably, the majority of noncanonical tumors have mutated MLH1 (n=11), POLE (n=6), LIG1 (n=5), and MLH3(n=5) (Figure 4C). Our analysis shows that canonical and noncanonical MSI tumors have distinct mutational patterns of MMR genes, suggesting that the determinants of the two MSI tumor groups are likely lineage-dependent.

As an independent validation of the canonical and noncanonical MSI tumors, we used the Clinical Proteomic Tumor Analysis Consortium (CPTAC) exome data of 993 samples across 7 cancer types. MSIsensor estimation of MSI status in tumors versus normal tissue confirmed the identification of canonical (COAD and UCEC) and noncanonical (BRCA) MSI tumors, where those MSI tumors are associated with high mutation rate (Figure 5). The result independently validating the presence of not only canonical but also noncanonical MSI tumors.
2.3.3 Microsatellite-derived neoantigen from canonical and noncanonical MSI tumors are immunogenic

MSI tumors show improved response to immune checkpoint therapy, independent of histology\(^5\). In order to estimate the immunotherapy efficacy of canonical and noncanonical MSI tumors, we identified the microsatellite-derived neoantigens predicted to bind with MHC proteins by using pVAC-seq\(^6\) (Figure 6A). The number of neoantigens varied between canonical and noncanonical MSI tumors (Figure 6B), where correlated positively with MSIsensor score in canonical MSI tumors and trended positive in BRCA but not GBM or KIRC. This result suggests that microsatellite instability leads to excess neoantigen load both in canonical and noncanonical MSI tumors, but varies based on tissue of origin.

2.4 Discussion

Despite the use of immune checkpoint inhibitors to treat cancers in recent years, it is still difficult to precisely select patients who will benefit from immunotherapy, even among those patients with overexpressed PD-L1. MSI might be an important molecular feature for responsiveness to immune checkpoint inhibition. Our comprehensive analysis of MSI across multiple cancer types corroborates the known relationship among microsatellite instability, tumor mutational burden, and expression of immune modulators, but also suggests additional groups of MSI tumors from noncanonical cancer types. The modest MSI phenotype of noncanonical cancer types indicates potential therapeutic implications, which has been partially tested in the neoantigen load analysis section. The excess neoantigen load of MSI tumors predicted by genomics-based methods could be further validated by orthogonal proteomic evidence\(^7\). Our results also show that there are several subtypes of MSI tumors, i.e. canonical and noncanonical MSI tumors, that may be distinguished by the particular MMR component that is defective, which in turn could influence MSI and responsiveness to therapy. Together, the increased number of microsatellite-associated
neoantigens resulting from MMR deficiency may be more permissive to anti-PD-1 therapy in this genetically defined subset of cancers.

2.5 Methods

2.5.1 Data Preparation for TCGA samples

A publicly available MAF file was compiled by the MC3 Working Group and is annotated with filter flags to highlight potential artifacts or discrepancies. This dataset represents the most uniform attempt to systematically provide mutation calls for TCGA tumors. The MC3 effort provided consensus calls from 7 software packages. Flagged artifacts include: non-exonic regions, whole-genome amplified (WGA) samples, exclusion lists, blood/tumor derived pairs, strand-bias, contamination estimations, oxo-guanine artifacts, low normal read depth, polymorphisms common in EXAC, mutations present in a panel of normal samples, non-preferred tumor normal pairs, and mutations outside the regions of interest for any caller. If a mutation was not assigned any flag and was called by 2 or more variant calling software packages, it received a ‘PASS’ identifier. We restricted our analysis to PASS calls with the exception of samples from OV and LAML, which were some of the earliest sequenced by TCGA. Preparations for these samples utilized whole genome amplified (WGA) DNA, an important factor in that the WGA process can induce artifactual mutations. Of the 412 OV and 141 LAML samples present in our data 347 (84%) and 141 (100%), respectively, had variants derived from WGA DNA. In order to maintain sample sizes and uniformity in mutation calling, we did not filter mutations containing only ‘wga’ filter tags from these two cancer types. The final driver-discovery dataset consisted of 9,079 samples having a total of 791,637 missense mutations, 323,884 silent mutations, 96,196 3’UTR mutations, 57,900 nonsense mutations, 42,251 intronic mutations, 42,251 Frameshift deletions, 34,266 5’UTR,
21,804 splice site mutations, 19,856 RNA mutations, 11,305 frameshift insertions, 7,622 3′flanking mutations, 6,419 5′flanking mutations, 6,144 in-frame deletions, 1,362 translation start site mutations, 964 nonstop mutations, and 632 in-frame insertions.

These 33 cancer types included in this study are adrenocortical carcinoma [ACC], bladder urothelial carcinoma [BLCA], brain lower grade glioma[LGG], breast invasive carcinoma [BRCA], cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC], cholangiocarcinoma [CHOL], colon adenocarcinoma [COAD], esophageal carcinoma [ESCA], 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], liver hepatocellular carcinoma [LIHC], lung adenocarcinoma [LUAD], lung squamous cell carcinoma [LUSC], lymphoid neoplasm diffuse large B cell lymphoma [DLBC], mesothelioma [MESO], ovarian serous cystadenocarcinoma [OV], pancreatic adenocarcinoma [PAAD], pheochromocytoma and paraganglioma [PCPG], prostate adeno-carcinoma [PRAD], rectum adenocarcinoma [READ], sarcoma [SARC], skin cutaneous melanoma [SKCM], stomach adenocarcinoma [STAD], testicular germ cell tumors [TGCT], thymoma [THYM], thyroid carcinoma [THCA], uterine carcinosarcoma [UCS], uterine corpus endometrial carcinoma [UCEC], and uveal melanoma [UVM].

2.5.2 Data Preparation for CPTAC samples

We collected the processed and normalized genomic and proteomic data generated for the analysis from the publications of each cohort. Our data collection includes:

CPTAC2 BR prospective (submitted)

CPTAC2 CO prospective⁹
CPTAC2 OV prospective
CPTAC3 ccRCC discovery
CPTAC3 LUAD discovery
CPTAC3 GBM discovery (submitted)
CPTAC3 UCEC discovery

Some preprocessing steps were applied to harmonize the data across cohorts. Samples excluded in the original publications were removed. The protein abundance was kept unique per gene symbol. Peptides that don’t match the site location or the exact protein sequence were removed. Modification sites per peptide were named using the pattern “<gene symbol>:<protein id>:<sites>:<duplication>”, where duplicated sites were ordered by the detection percentage across the cohort. Peptides detected in < 20% of the samples were removed.

The peptide search databases of these cohorts include RefSeq 20160914 (CPTAC2 BR and OV prospective), RefSeq 20171003 (CPTAC2 CO perspective), and RefSeq 20180629 (all CPTAC3 cohorts). To facilitate the cross-data interpretation, we improved the annotation of the databases by recovering HGNC Gene IDs from NCBI/Entrez Gene IDs using HGNC BioMart (https://biomart.genenames.org/) and retrieving the UniParc IDs and checksums for all the protein sequence in all databases (DOI: 10.1093/bioinformatics/bth191). Note that we keep the gene symbols identical to the RefSeq records at the time of download.

2.5.3 Creation of MMR Gene Set
We listed all the 39 genes involved in the MMR pathway by manually literature review. The genes were grouped based on functional complexes as follows.
MutS complex: MSH2, MSH6, MSH3, MSH4, MSH5.
MutL complex: MLH1, PMS2, MLH3.
Polymerase: POLD1, POLD3, POLE, POLE2, POLE3, POLE4, POLK
Others: MGMT, ERCC1, ERCC2, ERCC3, ERCC4, ERCC5, ERCC6, PMS1, PMS2L3, PMS2P, RFC1, RFC2, RFC3, RFC4, RFC5, RPA1, RPA2, RPA3, RPA4, RPA, LIG1, PCNA, RFC, EXO1.

2.5.4 Estimate MSI Status By Using MSIsensor
MSIsensor is a C++ algorithm that distinguishes microsatellite unstable tumors from microsatellite stable samples based on tumor/normal sequence data. Homopolymer regions of 5 or more nucleotides in length are aggregated separately in tumor/normal pairs and compared using a chi-square statistic. MSI-high was calculated as an MSI score greater than 3.5. Parameters for running MSIsensor “msi” command are as follows: –l (minimal homopolymer size) = 1 and –q (minimal microsatellitesize) = 1. These settings are not a minimal number of repeats, but rather the minimal number of nucleotides to consider within the repeat. 357 scores were generated from BAM files other than those used for variant calling by the MC3 Working group.

2.5.5 HLA Typing and Neoantigen Prediction From Microsatellite Sites
HLA class I typing of CPTAC samples was determined by using OptiType tool. OptiType was run under its default parameters for RNA sequencing FASTA files. For each predicted microsatellite passed chi-square test, we obtained translated protein sequences for novel transcripts from VEP v97 (Ensembl Variant Effect Predictor) using default setting. We constructed different epitope lengths (8-11-mer) from the translated protein sequence. We predicted the binding affinity between epitopes and the major histocompatibility complex (MHC) using pVAC-seq docker
version⁶. Epitopes with binding affinity smaller than 500nM which are also not present in the wild-type transcript are reported as neoantigens.
2.6 Figures

Figure 2.1 Relationship among MMR deficiency, mutation frequency, and MSI status across 21 cancer types. Tumor mutational burden (upper) and MSI scores (middle) segregated by cancer types. The fraction of MSI prone tumors explained by MMR deficiency is displayed on the lower panel. Each dot represents one cancer sample. Dots are color-coded based on the presence of one or more somatic or germline mutations in an MMR gene, or on MLH1 promoter hypermethylation. MSI score threshold is displayed with a dashed horizontal line.
Figure 2.2 Quantitative MSIsensor scores correlated with qualitative results from gel-assay
Histogram shows distribution of log2-transformed MSI score. Samples are colored by the clinical classification experimentally validated by TCGA.
Figure 2.3 Exome-based and WGS-based MSIsensor scores are correlated. Correlation between exome-based and WGS-based MSIsensor scores segregated by cancer types.
Figure 2.4 Canonical and noncanonical MSI tumors are significantly associated with high mutational burden.

(A) Number of point mutations stratified by MSI status. Boxplots indicate median number of point mutations with 25th and 75th percentile hinges and whiskers that extend to 1.5*IQR.

(B) Number of canonical MSI tumors with indicated mutations (missense: left panel, truncation: right panel) segregated by MMR genes.

(C) Number of noncanonical MSI tumors with indicated mutations (missense: left panel, truncation: right panel) segregated by MMR genes.
Figure 2.5 Canonical and noncanonical MSI tumors are validated in independent dataset.
Number of point mutations stratified by MSI status in canonical and noncanonical cancer types. Boxplots indicate median number of point mutations with 25th and 75th percentile hinges and whiskers that extend to 1.5*IQR.
**Reference**


Chapter 3: Driver Fusions and Their Implications in the Development and Treatment of Human Cancers
Portions of this chapter have been reproduced and adapted from the following published manuscripts:


&


*indicates co-first authors
3.1 Abstract
Gene fusions represent an important class of somatic alterations in cancer. We systematically investigated fusions in 9,624 tumors across 33 cancer types using multiple fusion calling tools. We identified a total of 25,664 fusions, with a 63% validation rate. Integration of gene expression, copy number, and fusion annotation data revealed that fusions involving oncogenes tend to exhibit increased expression, whereas fusions involving tumor suppressors have the opposite effect. For fusions involving kinases, we found 1,275 with an intact kinase domain, the proportion of which varied significantly across cancer types. Our study suggests that fusions drive the development of 16.5% of cancer cases and function as the sole driver in more than 1% of them. Finally, we identified druggable fusions involving genes such as TPRSS2, RET, FGFR3, ALK, and ESR1 in 6.0% of cases, and we predicted immunogenic peptides, suggesting that fusions may provide leads for targeted drug and immunotherapy.

3.2 Introduction
The ability to determine the full genomic portrait of a patient is a vital prerequisite for making personalized medicine a reality. To date, many studies have focused on determining the landscape of single nucleotide polymorphisms, insertions, deletions, and copy number alterations in cancer genomes\textsuperscript{1–6}. While such genomic alterations make up a large fraction of the typical tumor mutation burden, gene fusions also play a critical role in oncogenesis. Gene fusions or translocations have the potential to create chimeric proteins with altered function. These events may also rearrange gene promoters to amplify oncogenic function through protein overexpression or to decrease the expression of tumor suppressor genes.

Gene fusions function as diagnostic markers for specific cancer types. For example, a frequent translocation between chromosomes 11 and 22 creates a fusion between EWSRI and FLII
in Ewing’s sarcoma. Also, the Philadelphia chromosome 9-22 translocation is characteristic of chronic myeloid leukemia, resulting in the fusion protein \( BCR--ABL1 \). This fusion leads to constitutive protein tyrosine kinase activity and downstream signaling of the PI3K and MAPK pathways, which enables cells to evade apoptosis and achieve increased cell proliferation\(^7\)–\(^{10}\). Fibrolamellar carcinoma (FLC) in the liver is characterized by a \( DNAJB1--PRKACA \) fusion. A recent study of TCGA tumors revealed this fusion transcript is specific to FLC, differentiating it from other liver cancer samples\(^6\). In contrast, \( FGFR3--TACC3 \) is an inframe activating kinase fusion found in multiple cancer types, including glioblastoma multiforme (GBM)\(^{11,12}\) and urothelial bladder carcinomas (BLCA)\(^{13}\). Other recurrent fusions have also been reported in multiple cancer types\(^{14–16}\), and functional characterization of a few selected fusion genes in cellular model systems has confirmed their oncogenic nature\(^{17}\).

Recently, large-scale genomic studies have utilized the TCGA RNA-Seq data corpus to systematically identify and compile fusion candidates across many cancer types. For example, as part of its goal to develop a comprehensive, genome-wide database of fusion genes, ChimerDB\(^{18}\) has analyzed RNA-Seq data of several thousand TCGA cases. Giacomini et al. performed breakpoint analysis on exon microarrays across 974 cancer samples and identified 198 candidate fusions in annotated cancer genes\(^{19}\). A searchable portal of TCGA data includes 20,731 fusions called from 9,966 cancer and 648 normal samples\(^{20}\). Some studies focus on important classes of genes, such as kinase fusions\(^{21}\), which may have particular structural properties that are selected for during oncogenesis and cancer progression. However, most efforts have utilized only a single fusion calling algorithm. Since disagreements among different callers are common, there is a need to develop a comprehensive approach that combines the strengths of various callers to achieve
higher fusion calling accuracy. Further, large-scale analyses are likely to expand the targetable landscape of fusions in cancer, revealing potential treatment options for patients.

Here, we leverage multiple newly-developed bioinformatic tools to methodically identify fusion transcripts across the TCGA RNA-Seq data corpus using the ISB Cancer Genomics Cloud. These tools include STAR-Fusion, Breakfast, and EricScript (STAR Method). Fusion calling across 9,624 TCGA tumor samples from 33 cancer types identified a total of 25,664 fusion transcripts, with 63.3% validation rate for the samples having available whole genome sequencing data. Further, we investigated the relationship between fusion status and gene expression, the spectrum of kinase fusions, mutations and fusions found in driver genes, and fusions as potential drug and immunotherapy targets.

### 3.3 Results

#### 3.3.1 Fusion detection pipeline and WGS-based validation of a subset of fusion predictions

We analyzed RNA-Seq data from 9,624 tumor samples and 713 normal samples from The Cancer Genome Atlas (TCGA) using STAR-Fusion (STAR Method), EricScript\textsuperscript{22}, and Breakfast (STAR Methods, Table S1). A total of 25,664 fusions were identified after extensive filtering using several panel-of-normals databases, including fusions reported in TCGA normal samples, GTEx tissues\textsuperscript{23} and non-cancer cells\textsuperscript{24} (STAR Methods, Fig. 1A, and Table S1). Our pipeline detected 405 out of 424 events curated from individual TCGA marker papers (Table S1) (95.5% sensitivity).

We further cross-confirmed our transcriptome sequencing-based fusion detection pipeline by incorporating whole genome sequencing (WGS) data, where available. WGS paired-end reads aligned to the partner genes of each fusion were used to validate fusions detected using RNA-Seq. Using all available whole-genome sequencing, including both low-pass and high-pass data, from
1,725 of the 9,624 cancer samples across 25 cancer types, we were able to evaluate 18.2% (4,675 fusions) of our entire fusion call set. Of that subset, WGS validated 63.3% of RNA-Seq based fusions by requiring at least three supporting discordant read pairs from the WGS data (Figure S1).

### 3.3.2 Fusion landscape across 33 cancer types
Categorizing the 25,664 fusions based on their breakpoints, we found that the majority of breakpoints are in coding regions (CDS) of both partner genes (Fig. 1B). Surprisingly, there are many more fusions in 5’ UTRs compared to 3’ UTRs for both partner genes, given that 3’ UTRs are generally longer (Mann-Whitney U Test, p<2.2e-16). This could be explained by having more open chromatin in the 5’ UTR region\(^2\), the larger number of exons in 5’ UTRs than 3’UTRs (Mann-Whitney U Test, p<2.2e-16)\(^3\), but could also indicate some regulatory mechanisms, e.g. alternative usage of the promoter region of a partner gene.

For different cancer types, the total number of fusions per sample varies from 0 to 60, with a median value of one (Figure S1). Cancer types having the fewest number of fusions per sample are kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), low grade glioma (LGG), pheochromocytoma and paraganglioma (PCPG), testicular germ cell tumors (TGCT), thyroid carcinoma (THCA), thymoma (THYM), and uveal melanoma (UVM), each with a median of zero. Other cancer types show a range of medians between 0.5 and 5 fusions per sample, although most samples demonstrate zero or only one inframe, disruptive fusion relevant to oncogenesis.

Frequencies of recurrent fusions found in each cancer are illustrated in Figure 1C (Table S1). The most recurrent example within any cancer type was TMPRSS2--ERG in prostate adenocarcinoma (PRAD, 38.2%). We found FGFR3--TACC3 to be the most recurrent fusion in BLCA (2.0%), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC, 1.7%),
and lung squamous cell carcinoma (LUSC, 1.2%). Other top recurrent fusions include \textit{EML4--ALK} in lung adenocarcinoma (LUAD, 1.0%), \textit{CCDC6--RET} in THCA (4.2%), and \textit{FGFR2--BICC1} in cholangiocarcinoma (CHOL, 5.6%).

### 3.3.3 Fusion gene expression in oncogenes and tumor suppressors

Fusion events may be associated with altered expression of one or both of the fusion gene partners, a well-known example being multiple myeloma tumors in which translocation t(4;14) fuses the highly-expressed IGH locus with the tyrosine protein kinase \textit{FGFR3} \textsuperscript{27}. We integrated gene expression, copy number, and fusion annotations to systematically test for associations between gene expression and fusion status.

For each fusion having an oncogene, kinase, or tumor suppressor (Table S2), we determined whether that sample was an expression outlier for that gene and subsequently examined resulting percentages of both under- and overexpressed genes in each cancer type (Table S3). Figure 2A shows that between 6% (mesothelioma, MESO) and 28% (KIRP) of kinase fusions displayed outlier overexpression of the kinase partner. Oncogenes tended to show higher likelihoods of overexpression, while tumor suppressors displayed lower likelihoods. Between 3% (breast invasive carcinoma, BRCA) and 38% (PCPG) of tumor suppressor gene fusions showed outlier under expression, generally higher than both oncogenes and kinases.

Figure 2B illustrates the median percentile expression level of the most highly recurrent oncogenes and tumor suppressors involved in fusions (Table S3). Samples with fusions involving oncogenes, such as \textit{EGFR}, \textit{ERBB2}, and \textit{RET}, showed increased expression of those genes relative to samples without fusions across cancer types. Most tumor suppressor genes (TSGs) showed
inconsistent patterns of expression across cancer types. However, the global trend for TSGs is decreased expression compared to non-fusion samples.

We also examined the relationship between TSG mutations and fusions to determine whether frequently-fused TSGs were also disrupted by other mutation types. A variety of patterns were noted. For example, TP53 is affected by mutations rather than fusions in most cancer types. However, in sarcoma (SARC), both fusions and mutations affecting TP53 were detected. In acute myeloid leukemia (LAML), several CBFB fusions but no mutations were observed, yet other cancer types also exhibited CBFB mutations (Table S3, Figure S2). Our results suggest that alternative mechanisms are utilized by tumor cells in a cancer type-specific manner.

We also observed associations between fusion status and expression level in well-known fusions (Table S3), such as RET--NTRK1 in thyroid cancer, EML4--ALK in lung cancer, and DNAJB1--PRKACA in the fibrolamellar carcinoma subtype of liver cancer. RET fusions in thyroid carcinoma (THCA) and lung adenocarcinoma (LUAD) are inframe protein kinase fusions with overexpression of the 3' RET oncogene (Fig. 2C). Recurrent CBFB--MYH11 fusions in LAML are significantly associated with decreased expression of the tumor suppressor CBFB, which functions as a transcriptional regulator (Fig. 2D).

In breast cancer, copy number amplification is a well-known mechanism of ERBB2 overexpression and treatment of these HER2+ patients with trastuzumab is an established and effective targeted therapy. Interestingly, three out of four samples with ERBB2 fusions and two samples without a called fusion showed HPV integration within 1Mb of ERBB2. ERBB2 fusion gene partners PPP1R1B and IKZF3 are genomic neighbors of ERBB2, suggesting that these fusions could be a by-product of local instability, produced by the viral integration and subsequent breakage
fusion events. By careful analysis of the association between fusions and expression, we have identified strategies for improving both sensitivity and specificity of fusion calls.

3.3.4 Structure and spectrum of kinase fusions
Some oncogenic kinase fusions are susceptible to kinase inhibitors\textsuperscript{21}, suggesting that additional therapeutic candidates might be discovered by examining fusion transcripts involving protein kinase genes. In total, we detected 2,892 such events, comprising 1,172 with kinase at the 3’ end (3’-kinase), 1,603 with kinase at the 5’ end (5’-kinase), and 117 with both partners being kinases (both-kinase) (Fig. 3A and Table S4). Analysis of the catalytic kinase domains using the UniProt/PFAM domain database (STAR Methods) showed that 1,275 (44.1\%) kinase fusions retained an intact kinase domain (Fig. 3A). We further predicted open reading frames for these fusions and separated them into three categories with respect to the frame of the 3’ gene: inframe, frameshift, and no frame information (e.g. breakpoint at UTR, intron, or non-coding RNA). In general, there were more inframe fusions than frameshift fusions, especially for 3’-kinase fusions, because preserving the reading frame is required to keep the kinase domain intact. For subsequent kinase analyses, we focused only on those 1,275 fusions having intact domains, further classifying the both-kinase group into 3’-kinase or 5’-kinase based on the position of the intact domain.

Comparison of kinase fusions across different cancer types indicated that kinase fusions are significantly enriched in thyroid carcinoma (THCA, 35.6\%, Fisher’s Exact Test, \(p < 2.2\times 10^{-16}\)) (Fig. 3B). Moreover, the majority were 3’-kinase fusions (94.0\%), a significantly higher percentage than what we observed in other cancer types (Fisher’s Exact Test, \(p < 2.2\times 10^{-16}\)). We further divided these fusions into eight categories based on different kinase groups, including AGC, CAMK, CK1, CMGC, STE, TK, TKL. In general, we found that the percentages of different
categories vary across cancer types (Fig. 3B). For example, there are more TK fusions in THCA and GBM, more CK1 fusions in uterine corpus endometrial carcinoma (UCEC), colon adenocarcinoma (COAD), and esophageal carcinoma (ESCA), and more AGC fusions in liver hepatocellular carcinoma (LIHC). Across different cancer types, we found an enrichment of TK and TKL kinase fusions for 3’-kinases, but no strong preference for 5’-kinases (Figure S3).

Recurrent kinase fusions are of great interest as potential drug targets. Overall, we detected 744 5’-kinase and 531 3’-kinase fusions. Of these, 147 and 99 were recurrent, respectively, mostly across cancer types rather than within cancer types (Figure S3). As expected, fusions in the FGFR kinase family (FGFR2 and FGFR3) are the most frequent 5’-kinase fusions, given their high recurrence in individual cancer types (Fig. 3C). WNK kinase family fusions (WNK1 and WNK2) were also detected in multiple cancer types. The WNK family is phylogenetically distinct from the major kinase families, and there is emerging evidence of its role in cancer development. Here, we found a total of 23 WNK family fusions, most of which resulted in higher expression of WNK mRNA (Figure S4). The increased expression was not generally accompanied by copy number amplification; for example, neither WNK1 nor WNK2 were amplified in ESCA or LIHC. Incidentally, ERC1--WNK1 was also detected recently in an independent Chinese esophageal cancer cohort. For 3’-kinase fusions, all the top 10 kinase genes are tyrosine kinases, most of which are enriched in THCA, including RET, BRAF, NTRK1, NTRK3, ALK, and REF1 (Fig 3C).

FGR fusions were found in 7 samples the same partner gene WASF2, 5 of which showed higher expression of FGR gene. In these five samples, the breakpoints for the two genes are the same (5’UTR of both genes) resulting in usage of the stronger WASF2 promoter for the FGR gene. Interestingly, recurrent MERTK fusions are singletons in each individual cancer type with TMEM87B and PRKACA fusions are only observed in liver cancer with DNAJB1 (Figure S3).
To further understand the regulation of kinase fusions, we compared the gene expression patterns between the kinase gene and partner gene. There are in total 1,035 kinase fusions with both gene expression and copy number data available. To control for the effect of copy number amplification on gene expression, we focused on the fusions with copy numbers between 1 and 3, including 439 5’-kinase and 339 3’-kinase fusions (Fig. 4A-B). For 5’-kinase fusions, the kinase gene expression quantiles are uniformly distributed, indicating that the kinase gene expressions in the samples with fusion are not significantly different from the samples without fusion (Fig. 4A). However, 3’-kinase genes tend to show higher expression in samples with a fusion compared to the ones without. To explain this, we classified the fusion events into three categories based on the relative expression pattern between the kinase gene and its partner in samples from the same cancer type. Most (66.7%, 293/439) 5’-kinase fusions showed lower expression in the partner gene compared to the kinase. In contrast, 70.5% (239/339) of 3’-kinase fusions showed higher partner expression (Fig. 4A-B). Moreover, those 3’-kinase fusions involving a more highly expressed 5’ partner also show higher kinase expression (Fig. 4C). For example, we found a TRABD--DDR2 fusion in one head and neck squamous cell carcinoma (HNSC) sample, which fused the stronger TRABD promoter with DDR2, resulting in its overexpression (Fig. 4D). This patient could potentially be treated using dasatinib, which targets overexpressed DDR2 in HNSC33. DDR2 fusions were also detected in another 9 samples from 5 different cancer types, which could be treated similarly given sufficient DDR2 overexpression (Table S1).

**3.3.5 Mutual exclusivity between fusions and mutations**

While mutations in oncogenes or tumor suppressors may lead to tumorigenesis, fusions involving those genes are also an important class of cancer driver events. We systematically profiled mutations and fusions in 299 cancer driver genes34 (Table S2) to assess the contributions...
of fusion genes in carcinogenesis in the 8,963 TCGA patients that overlap between the mutation call set (Public MC3 MAF\textsuperscript{35}, Key Resources Table) and our fusion call set. We characterized patients as having a driver mutation, a mutation in a driver gene, and/or a driver fusion (fusion involving a driver gene).

Although the majority of cancer cases have a known driver mutation (48.6%, mean 6.8 mutations) or mutations in a driver gene (28.1%, mean 4.2 mutations), we found 8.3% have both a driver mutation and driver fusion event (mean 5.5 mutations and 1.2 fusions), 6.4% have both a mutation and fusion in a driver gene (mean 4.2 mutations and 1.3 fusions), and 1.8% have a driver fusion only (mean 1.1 fusions) (Fig. 5A). This distribution is consistent with the notion that only a few driver events are required for tumor development\textsuperscript{2}.

We further examined the total number of mutations for samples and observed a low mutational burden in the group with driver fusion only, which is comparable with the group with no driver alterations (Fig. 5B). The significant decrease in the numbers of mutations (Mann-Whitney U Test, \(p<2.2\times10^{-16}\)) reflects the functionality of fusions across multiple cancer types. Moreover, within cancer types, we observed a range of 0.2% (HNSC) to 14.0% (LAML) of tumors with fusions but no driver gene mutations. Among those LAML tumors that have fusions and no driver gene mutations, we identified several well-recognized fusions relevant to leukemia, such as \textit{CBFB--MYH11} (number of samples=3), \textit{BCR--ABL1} (n=2), and \textit{PML--RAR} (n=2). We also identified the leukemia-initiating fusion \textit{NUP98--NSD1} in two LAML tumors\textsuperscript{36}.

We then examined the relationship of fusions and mutations in the same driver gene (Fig. 5C). The result shows that when fusion events are present in a gene, mutations in the same gene are rarely found, supporting a pattern of mutual exclusivity of the two types of genomic alteration.
This trend was observed across many patients and many cancer types. Our results suggest that a considerable number of tumors are driven primarily or solely by fusion events.

3.3.6 Contributions of fusions to cancer treatment

We investigated potentially druggable fusion events in our call set using our curated Database of Evidence for Precision Oncology (DEPO; Sun, et al. submitted) (Table S5). We defined a fusion as druggable if there is literature supporting the use of a drug against that fusion, regardless of cancer type (allowing for “off-label” drug treatment). We found potentially druggable fusions across 29 cancer types, with major recurrent druggable targets in PRAD (TMPRSS2, 205 samples), THCA (RET, 33 samples), and LAML (PML--RARA, 16 samples) (Fig. 6A). FGFR3 was a potential target (both on-label and off-label) in 15 cancer types. Overall, we found 6.0% of samples (574/9,624 samples) to be potentially druggable by one or more fusion targeted treatments. Further study of fusions in human cancer will facilitate the development of precision cancer treatments.

We analyzed patterns of fusion druggability in LUAD, stratifying by smoking status. In this data set, 15% of LUAD samples (75 out of 500 samples with known smoking status) were never smokers, while a significantly higher percentage of never smokers (15 out of 75 samples) vs. smokers (9 out of 425 samples) were found to have a druggable fusion (Chi-square test, p<1e-6) (Fig. 6B). Several FDA approved drugs exist to target ALK fusions in lung and other cancer types. We observed ALK fusions in 20 samples from 8 cancer types (5 samples in LUAD). In most cases, fusion status corresponded to copy number neutral overexpression of ALK (Fig. 6D). In 17 out of 20 cases, ALK was the 3’ partner of the fusion pair, with EML4 being the most frequent 5’ partner (7 out of 17).
ESR1 encodes an estrogen receptor with important and druggable relevance to breast cancer\textsuperscript{37}. We detected ESR1 fusions in 16 samples from 5 different cancer types (9 samples from BRCA). Of the 9 BRCA samples, 8 are known be from the Luminal A or B subtypes. We observed strict mutual exclusivity between ESR1 mutations and fusions (Fig. 5C). Of the 16 fusions, 11 have ESR1 at the 5’ end, and 5 at the 3’ end. When ESR1 is the 5’ gene in the fusion, the transactivation (AF1) domain is always included (Fig. 6D). When ESR1 is the 3’ gene, the transactivation (AF2) domain is always included. Those samples with ESR1 fusion tend of have higher ESR1 expression, especially in the 9 BRCA samples (Figure S5). Similarly, ESR1 expression is higher when ESR1 is mutated in BRCA, CESC, and UCEC, which are all hormone receptor related cancer types\textsuperscript{36,38,39}. Further functional study to determine the mechanism of ESR1 fusions could suggest drug development directions.

Immunotherapy based on tumor-specific neoantigens shows promise in treating cancer patients\textsuperscript{40}. Gene fusions found in tumor cells can generate peptides, which may serve as neoantigen candidates. However, patients with known driver fusions may be poor candidates for immunotherapy due to their reduced mutational burden, especially without clear evidence of immune cell infiltration and overall immunogenicity. As an exploratory and speculative analysis, we investigated neoantigens produced by gene fusions\textsuperscript{41}. On average, there were 1.5 predicted neoantigens per fusion across different cancer types (Figure S6 and Table S5). The mean number of predicted neoantigens per fusion ranged from 0.33 in KICH to 2.88 in THYM. We also compared the number of neoantigens for inframe and frameshift fusions (Figure S6). Results show that frameshift fusions can generate more immunogenic epitopes than inframe fusions (mean value: 2.2 vs 1.0), though nonsense mediated decay might reduce some of this potential difference.
We further investigated seven fusions for which there were at least four samples having one or more neoantigen candidates (Figure S6). In particular, $\text{TMPRSS2}--\text{ERG}$, $\text{CCDC6}--\text{RET}$, and $\text{FGFR3}--\text{TACC3}$ have the highest number of samples with predicted neoantigen candidates. Our results show that the fusion product is only immunogenic in a small subset of patients, especially for $\text{TMPRSS2}--\text{ERG}$ fusions. Again, without clear evidence of immune cell infiltration and overall immunogenicity, any fusion neoantigen analysis remains exploratory and speculative.

### 3.4 Discussion

In this study, we applied multiple RNA-Seq fusion callers, namely STAR-Fusion, EricScript, and Breakfast, followed by a stringent filtering strategy, to identify potential driver fusion events across 33 cancer types. We were able to successfully identify 95.5% of fusions reported in TCGA marker papers. While existing studies have published fusion calls across the TCGA cancer cohort\textsuperscript{20,21}, we have improved on prior analyses by integrating results across multiple fusion callers and by applying stringent filtering to derive a confident dataset of fusion events from 9,624 tumor samples. Importantly, we investigated the biology and evaluated the significance of fusions in the cancer context. Of the 25,664 fusions we detected, 18.2% could be tested for validation using available whole-genome sequencing data, leading to a 63.3% validation rate.

By integrating gene expression, copy number, and fusion annotation data, we evaluated the biological and therapeutic implications of fusion events. Kinase and oncogene related fusions tended to be overexpression outliers, while fusions involving tumor suppressor genes showed the opposite effect overall. When comparing fusion events to the remainder of the cancer cohort, fusions involving oncogenes such as $\text{EGFR}$, $\text{ERBB2}$, and $\text{RET}$ had increased expression.
Overexpressed fusions, especially inframe kinase fusions, are commonly targeted for therapy due to their susceptibility to kinase inhibitors.

For all 2,892 kinase fusions, we translated the resulting peptide sequence, finding that 1,275 had functional catalytic kinase domains. Comparison of kinase fusions across different cancer types showed that THCA has significantly more kinase fusions, most of which were 3’ kinase fusions. In addition to well-known recurrent fusions like FGFR3--TACC3, we also detected 245 kinases with recurrent fusions to different partner genes, which may ultimately prove to be successful drug targets.

We showed that a meaningful percentage of patients (16.8%) harbor fusions involving cancer driver genes but have no driver gene mutations. Notably, 6.0% of cancer patients could potentially benefit from existing drugs targeting fusion products. Moreover, our analysis also highlights an important consideration for immunotherapy treatment in patients with fusions. The significant decrease in mutational burden observed in patients with fusions in driver genes points toward a reduced efficacy of immunotherapy in these patients, despite fusion peptides themselves potentially being good immunogenic targets. Many fusions are already known to be drug targets.

Our study demonstrates the necessity of performing fusion analysis across multiple cancer types. Our approach integrated the results of multiple fusion calling algorithms, lending confidence to fusions with lower levels of RNA-seq read support that might otherwise have been discarded. We sought to prioritize fusions relevant to cancer by highlighting their association with gene expression, potential for targeted therapy, and role in cancer hallmark pathways. Fusion allele frequency is an elusive measure from RNA-Seq data and tracking the clonal evolution of fusions within a tumor remains an exciting opportunity for study. Fusions play an increasingly appreciated role in tumorigenesis and progression and represent an important source of improved treatment.
options. Ultimately, our multi-tool, integrative bioinformatic detection approach helps to define the universe of fusions in cancer. Further, it reminds us that developing robust and widely applicable clinical diagnostic approaches that can document fusions across cancer types is vital. Such approaches are critical to identifying those patients who can benefit from both established treatments and clinical trials.

3.5 STAR Methods

3.5.1 Dataset description

Aligned RNA-Seq bam files were analyzed using the ISB Cancer Genomics Cloud (https://isb-cgc.appspot.com/). These 33 cancer types included in this study are adrenocortical carcinoma [ACC], bladder urothelial carcinoma [BLCA], brain lower grade glioma [LGG], breast invasive carcinoma [BRCA], cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC], cholangiocarcinoma [CHOL], colon adenocarcinoma [COAD], esophageal carcinoma [ESCA], 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], liver hepatocellular carcinoma [LIHC], lung adenocarcinoma [LUAD], lung squamous cell carcinoma [LUSC], lymphoid neoplasm diffuse large B-cell lymphoma [DLBC], mesothelioma [MESO], ovarian serous cystadenocarcinoma [OV], pancreatic adenocarcinoma [PAAD], pheochromocytoma and paraganglioma [PCPG], prostate adenocarcinoma [PRAD], rectum adenocarcinoma [READ], sarcoma [SARC], skin cutaneous melanoma [SKCM], stomach adenocarcinoma [STAD], testicular germ cell tumors [TGCT], thymoma [THYM], thyroid carcinoma [THCA], uterine carcinosarcoma [UCS], uterine corpus endometrial carcinoma [UCEC], and uveal melanoma
[UVM]. The sample set consists of 10,331 total TCGA samples, 9,624 tumor samples, and 713 normal samples.

Level-3 gene expression (RSEM) and segment-based copy number data were downloaded from Broad GDAC firehose ([https://gdac.broadinstitute.org](https://gdac.broadinstitute.org)) (version: 2016_01_28). Gene-based copy number data were obtained by intersecting with RefSeq gene annotation bed file (version: 2013-07-27). Mutation calls were provided by the Multi-Center Mutation Calling in Multiple Cancers (MC3) working group within TCGA

3.5.2 **Fusion detection and filtering**  
TCGA RNA-Seq data were downloaded from Cancer Genomics Hub (CGHub, [https://cghub.ucsc.edu](https://cghub.ucsc.edu)) and analyzed using the ISB Cancer Genomics Cloud ([https://isb-cgc.appspot.com/](https://isb-cgc.appspot.com/)). For each sample, the fastq file was mapped to the human genome (build 38) followed by fusion calling using STAR-Fusion (parameters: --annotation --coding-effect), EricScript (default parameters) ([https://sites.google.com/site/bioericscript/](https://sites.google.com/site/bioericscript/)) and BREAKFAST (two different minimum distance cut-offs were used: 5 kb and 100 kb) ([https://github.com/annalam/breakfast](https://github.com/annalam/breakfast)). STAR-Fusion showed higher sensitivity in detecting the fusions reported in previous TCGA studies. Therefore, we focused on the STAR-Fusion output and integrated EricScript and BREAKFAST output in one of the following filtering steps: 1) an exclusion list of genes was curated, including uncharacterized genes, immunoglobulin genes, mitochondrial genes, etc. Fusions involving these genes were filtered; 2) Fusions from the same gene or parologue genes (downloaded from [https://github.com/STAR-Fusion/STAR-Fusion_benchmarking_data/tree/master/resources](https://github.com/STAR-Fusion/STAR-Fusion_benchmarking_data/tree/master/resources)) were filtered; 3) Fusions reported in normal samples were filtered, including the ones from TCGA normal samples, GTEx tissues, and non-
cancer cell study\textsuperscript{24}) For the fusions reported by only STAR-Fusion, a minimum value of FFPM > 0.1 (fusion fragments per million total reads) was required, as suggested by the authors; for the fusions reported by two or more callers, no minimum FFPM was required. 5) Finally, fusions with the same breakpoints in \( \geq 10 \) samples across different cancer types were removed unless they were reported in previous TCGA studies.

\textbf{3.5.3 Validation of fusion transcripts}

For fusion events where low-pass whole genome sequencing data or whole genome sequencing (WGS) data were available from the ISB Cancer Genomics Cloud (https://isb-cgc.appspot.com/), we obtained high quality (-q 20) reads mapping to each partner gene and the 100kb region up and downstream using SAMtools. At least 3 discordant reads from WGS were required to determine if the fusion prediction was validated.

\textbf{3.5.4 Gene expression analysis}

We collected gene expression, copy number, and fusion annotations to test for associations between gene expression and fusion status. We used Tukey’s definition of outliers to determine if the expression level at a given gene was an outlier or not. An overexpression outlier means the sample’s expression level at a given gene was greater than (75th percentile) + 1.5*IQR, where IQR is the interquartile range. An underexpression outlier means the sample’s expression level at that gene was less than (25th percentile) - 1.5*IQR. To test for a significant association between expression and fusion status, we calculated p-values using both a t-test and Fisher’s Exact Test. If either of those results passed stringent FDR multiple test correction, three or more fusions were reported, and if the median expression of the fusions was in the top or bottom decile of the data, we reported those genes for manual review.
3.5.5 Protein kinase fusion analysis

We curated a list of kinase genes from previous publications and public databases (Table S5). Then we compared this list with UniProt/PFAM domain database (http://www.uniprot.org/database/DB-0073) to retain the ones with an annotated kinase domain. For the fusions involving kinase genes, we used AGFusion (https://github.com/murphycj/AGFusion) to check whether the annotated kinase domain was still present in the fusion transcript to separate them into fusions with an intact kinase domain versus those with a disrupted kinase domain. We compared the breakpoint positions in each fusion with the annotation file to check whether the breakpoint was in the 5’UTR, CDS, or 3’UTR region. Kinase genes are classified into eight groups: AGC, CAMK, CK1, CMGC, STE, TK, TKL, and others based on the PhosphoSite Database. The percentage of kinase genes in each group across different cancer types was defined as the number of kinase genes with fusions in each group divided by their sum, denoted as \( p_{g} \). For each cancer type, the number of kinase genes in each group was first normalized by \( p_{g} \), denoted as \( n_{g} \). Then each number was divided by their sum \( n_{g}/\sum n_{g} \) to calculate a normalized percentage of kinase genes in each group.

3.5.6 Neoantigen prediction

For each predicted fusion, we obtained translated protein sequences for novel transcripts from STAR-Fusion. The wild-type protein sequences are obtained from Ensembl Database. We constructed different epitope lengths (8-11mer) from the translated protein sequence. Each sample’s HLA type comes from the TCGA Pan-Cancer Immune Group (Synapse ID: syn5974636). We predicted the binding affinity between epitopes and the major histocompatibility complex
(MHC) using NetMHC4\textsuperscript{41}. Epitopes with binding affinity \( \leq 500\text{nM} \) which are also not present in the wild-type transcript are reported as neoantigens. We required at least 5 splitting reads for supporting junctions to filter fusions with low expression.

3.5.7 Mutual exclusivity analysis

For TCGA tumor samples where both MC3\textsuperscript{35} (Key Resources Table) mutation calls and gene fusion calls were available, we obtained the genetic alteration events, including fusion, inframe deletion, inframe insertion, missense mutation, nonsense mutation, nonstop mutation, splice site mutation, and translation start site mutation in 299 driver genes. We separated all the genomic alterations and events into “driver mutation”, “mutation”, and “fusion” categories, and compiled a genomic alteration profile for each sample. To test if the total number of mutations are significantly different among groups, we took samples without mutations in the following genes: \textit{POLE, MLH1, MLH3, MGMT, MSH6, MSH3, MSH2, PMS1,} and \textit{PMS2}, to exclude the confounding factor stemming from microsatellite instability. We then calculated p-values by using Mann-Whitney U Test.

3.5.8 DEPO

DEPO is a curated list of druggable variants filtered such that each variant corresponds to one of several categories: single nucleotide polymorphisms or SNPs (missense, frameshift, and nonsense mutations), inframe insertions and deletions (indels), copy number variations (CNVs) or expression changes. Each variant/drug entry in DEPO was paired with several annotations of potential interest to oncologists. DEPO is available as a web portal (http://dinglab.wustl.edu/depo).
3.6 Figures

(A) Fusion calling and filtering pipeline.

(B) Cartoon overview of fusion gene partner breakpoints. Purple indicates the 5′ gene partner and green indicates the 3′ gene partner. For both the 5′ and 3′ gene partner, fusion gene breakpoints can occur in the following genomic regions: 5′ untranslated region, coding region, and 3′ untranslated region.

Figure 3.1 Fusion detection and landscape in cancer.
region (5’UTR, triangle), coding sequence (CDS, rectangle), 3’UTR (circle), and noncoding region (rounded rectangle). For each fusion event, a dotted line connects the breakpoints in the 5’ and 3’ gene partners to create the predicted fusion and the circle size, while number represents the total fusion events classified into the associated fusion category.

(C) The dot plot shows the frequency of recurrent fusions found in each cancer type. The most recurrent fusion in each cancer type is labeled. Cancer types without recurrent fusions are not shown.
Figure 3.2 Fusion expression outliers

(A) The dot plot indicates the percentage of fusions called in which one of the partner genes is an expression outlier (overexpression or underexpression). The size of the dot corresponds to the number of fusions called in each cancer type. Color corresponds to genes of interest coming from lists of oncogenes, protein kinases, and tumor suppressor genes.

(B) The dot plot shows the relative expression level of samples with fusions compared to those without fusions. Each sample has a particular expression percentile at a given gene, and color indicates the median percentile of samples with a fusion in that gene. Genes are the fifteen most recurrent oncogenes and tumor suppressor genes. Size corresponds to the number of samples in each cancer type with a fusion at that gene.
(C) Expression of samples at \textit{RET} in thyroid carcinoma (THCA). Color indicates a categorical copy number ranging from deep deletion to high amplification.

(D) Expression of samples at \textit{CBFB} in acute myeloid leukemia (LAML). Color indicates a categorical copy number ranging from deep deletion to high amplification.
Figure 3.3 Protein kinase fusions.

(A) The bar chart indicates the number of protein kinase fusions with the kinase at the 5’ or 3’ end, inframe or frameshift, and kinase domain intact or disrupted.

(B) The left bar plot shows the percentage of samples with kinase fusions across different cancer types. The number of samples with a kinase fusion is also indicated at the end of each bar. 5’ kinase and 3’ kinase fusions are marked in light green and blue, respectively.
The right bar plot shows the normalized percentage of kinase fusions broken down by kinase groups.

(C) The dot plot shows the numbers of samples for recurrent fusions across different cancer types. 5’ kinase and 3’ kinase fusions are marked in light green and blue, respectively.
Figure 3.4 Kinase gene expression regulated by fusion

(A) The scatterplot shows the gene expression quantile (y-axis) for the 5’-kinase without copy number variation (between 1 and 3 copies, x-axis). All genes are classified among three categories: kinase expression higher, equal, and lower, as compared to partner expression, marked in blue, grey, and red, respectively. The density plot for expression quantile is also shown on the right panel.

(B) The scatterplot shows the gene expression quantile (y-axis) for the 3’-kinase without copy number variation (between 1 and 3 copies, x-axis). The colors represent the same three categories as (A). The density plot for expression quantile is also shown.

(C) Boxplot comparing the distribution of kinase gene expression quantile between the three groups defined in (A) for 5’-kinase and 3’-kinase, respectively.
(D) Schematic of *TBABD--DDR2* fusion gene structure in a HNSC sample, and scatter plot of *DDR2* copy number versus mRNA expression in HNSC. The samples with and without this fusion are marked in red and blue, respectively.
Figure 3.5 Mutual exclusivity between driver mutations and driver fusions.

(A) The bar plot shows the percentages of samples with driver mutations only (green), mutations only (orange), driver mutation and fusion (blue), mutation and fusion (pink), or fusion only (light green) events in 299 cancer driver genes.

(B) Distribution of mutation burden across each alteration group designated in all figures.

(C) All samples with fusions or mutations in any of the genes indicated on the left are displayed on the x-axis. For each gene, samples are clustered by the alteration group. Bottom bar indicates cancer type.
Figure 3.6 Druggable fusion targets.

(A) The bar chart indicates the number of samples potentially treatable based on their fusion status.

(B) Percentages of LUAD samples with known smoking status.

(C) *ESR1* domains kept in *ESR1* fusions across cancer types.

(D) *ALK* expression across cancer types indicating *ALK* fusion status.
Reference


Chapter 4: Integrative multi-omic profiling reveals tumorigenic DNA methylation associated with therapeutic vulnerability
4.1 Abstract

Gene DNA methylation is frequently dysregulated in human cancers, however such consequential deregulations are usually undelineated because they are inextricably linked with other neutral alterations during tumor development. Here, we quantify multi-omic interactions of 506 CPTAC tumors from kidney, brain, lung, head and neck, and endometrium, to identify aberrant methylation associated with RNA and protein abundance changes. Beyond the conventional genetic studies, we uncover cancer-specific epigenetic drivers such as hypomethylated $FGFR2$ in endometrial cancer. Tumors exhibiting hypermethylated $STAT5A$ are associated with pervasive regulome downregulation and immune cells depletions, suggesting epigenetic silencing of $STAT5A$ is a molecular switch that can be targeted to prevent immunosuppression. Finally, we demonstrate that methylation subtype-enrichment information can explain intra-tumor heterogeneity and phenotypes. Overall, our study provides a landscape of cis-acting DNA methylation associated with transcriptional and translational changes in tumors, illuminating how epigenetic regulation can act upon molecular mechanisms to promote tumorigenesis.

4.2 Highlights

1. Integrated omic profiling catalogs epigenomic aberrations associated with cis-transcriptional and translational changes across five cancer types.

2. $FGFR2$ and $EGFR$ hypomethylation are bona fide driver DNA methylation events.

3. $STAT5A$ methylation may serve as an epigenetic switch for tumor immune response in HNSCC.
4. Methylation subtype-enriched RNA and protein signatures provide insights into tumor phenotypes and tumor heterogeneity.

4.3 Introduction

Methylation of cytosines is an epigenetic modification adding information onto DNA without changing the genetic sequence, conferring stability and flexibility in spatiotemporal gene regulation of many biological processes. Aberrant DNA methylation is a hallmark of the development and progression of human cancers\textsuperscript{1-3}, which has primarily been observed in the global hypomethylation of mostly repetitive sequences and gene-specific hypermethylation of numerous CpG islands (CGI)\textsuperscript{4,5}. Epigenetic modifications within promoter regions can alter the expression of proximal genes, leading to transcriptional silencing of tumor suppressor genes or a loss of regulation of oncogenes. The widespread and pervasive changes in the patterns of DNA methylation arise in the early stages of cancer formation\textsuperscript{6}. Given the reversible and dynamic nature of DNA methylation, treating cells with DNA demethylating agents can reprogram neoplastic cells toward a normal state\textsuperscript{7}. Delineating the functional consequences of aberrant DNA methylation is critical for improving cancer diagnosis, prognosis, and treatment.

Despite the growing number of genome-wide methylation studies across various cancer types, it remains challenging to distinguish the relatively small number of functional DNA methylations that are responsible for the development and progression of cancer from the large number of co-occurring neutral methylations\textsuperscript{8}. Various systematic analytical approaches have been developed to explore the connection between DNA methylation and tumorigenic outcomes by integrating multi-omic data\textsuperscript{9,10}. However, studies exploring the downstream effects of aberrant DNA methylation have largely relied on tumor RNA expression without corresponding adjacent
normal samples. We aimed to define the extent to which information contributed by DNA methylation is transcribed to RNA expression, translated to protein abundance, and how this may lead to various tumor outcomes. Since proteomic data provides a direct measure of biological activity, profiling the functional changes of DNA methylation directly from transcriptomic and then proteomic data could complement each other, enhance interpretation of driver aberrations, and provide clinically relevant information.

Here, we investigate aberrant DNA methylation associated with transcriptional and translational changes using multi-omics resources from the Clinical Proteomic Tumor Analysis Consortium (CPTAC)\textsuperscript{11}. By combining DNA methylation, RNA expression, and proteomic data from 506 patients across five cancer types, we can systematically elucidate the information flow, characteristics, and functional impacts of \textit{cis}-acting aberrant DNA methylation in tumorigenesis for the first time. Our study provides a comprehensive catalogue of putative DNA methylation events with specific functional consequences, highlighting key changes to driver genes that affect cancer hallmark pathways in a coordinated manner. We also reveal distinct methylation subtypes among cancer types and compile a set of RNA and protein signatures that closely reflect the molecular features associated with each subtype. Finally, we pinpoint the clinically-actionable genes regulated by DNA methylation, which is critical for finding subtype-specific diagnosis and treatment strategies. The extensive functional catalogue established in this study of human DNA methylomes is an important step towards a deeper understanding of DNA methylation-mediated tumorigenesis and tumor progression, offering new opportunities for epigenetic therapy.
4.4 Results

4.4.1 Integrative multi-omic profiling empowers aberrant DNA methylation discovery

To create a comprehensive functional landscape of cancer methylomes across cancer types, we assembled 506 human tumors with available DNA methylation (Infinium EPIC array), gene expression levels (RNA-seq), and protein abundance level (mass spectrometry), across five cancer types from CPTAC including 107 clear cell renal cell carcinoma (ccRCC), 97 glioblastoma (GBM), 106 head and neck squamous cell carcinoma (HNSCC), 108 lung adenocarcinoma (LUAD), and 88 uterine corpus endometrial carcinoma (UCEC). For cohorts without DNA methylation data from normal adjacent tissues, GBM and UCEC, we collected DNA methylome from The Cancer Genome Atlas (TCGA) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) datasets (Table S1, Methods). These TCGA and TARGET datasets both include DNA methylation data from corresponding normal adjacent tissues in the form of Infinium HumanMethylation BeadChip (HM450), giving us an opportunity to estimate the methylation variance among reference DNA methylation data and thus to test the association in tissue-specific context with negligible batch effects (Figure S1).

We detected recurrent and deregulated DNA methylation in tumors compared to adjacent normal tissues, and tested the association between DNA methylation and mRNA expression change or protein abundance change separately using RESET\textsuperscript{10}. Our analysis aimed to profile the aberrant DNA methylation for regions actively contributing to gene regulation\textsuperscript{12} thereby limiting the probe sets to those within promoter regions (Figure S1). Overall, we detected 2,236 hypermethylated CpG sites associated with mRNA downregulation of 978 genes, and 895 CpG sites with protein downregulation of 434 genes; while 333 hypomethylated CpG sites are associated with mRNA upregulation of 256 genes, and 176 with protein upregulation of 137 genes.
(FDR < 10%, Figure 1A). Consistent with previous studies, DNA hypermethylation is more frequent than hypomethylation events in all cancer types\textsuperscript{10,13}. Our results confirmed several classic epigenetically-regulated genes in tumors, including \textit{MLH1}\textsuperscript{14}, \textit{MGMT}\textsuperscript{15}, \textit{WIF1}\textsuperscript{16}, \textit{CASP8}\textsuperscript{17}, \textit{HIC1}\textsuperscript{18}, and \textit{GSTP1}\textsuperscript{19}. However, the vast majority of identified genes had not been previously implicated in DNA methylation-mediated regulation (Table S1).

Next, we examined the concordance between mRNA and protein changes of individual aberrant DNA methylation events. Out of 1,051 hypermethylated genes and 225 hypomethylated genes with both mRNA and protein abundance data, we identified 387 hypermethylated genes and 83 hypomethylated genes associated with both RNA expression and protein abundance changes (Figure 1B). Overall changes were moderately or strongly correlated across cancer types (Spearman’s rho ranging from 0.30 to 0.82, Spearman’s P value < 0.001; Figure 1C, Figure S1), suggesting transcriptional and translational changes from the same aberrant DNA methylation are comparable between the two separate association tests. Genes with a negative correlation between transcript and protein abundance may be attributed to biological factors such as translational regulation, tissue-specific expression, and mRNA degradation or technical factors due to low-abundance proteins or transcripts\textsuperscript{20,21}.

Having identified a comprehensive map of \textit{cis}-acting cancer methylome, we were able to delineate the functional impacts of those deregulated DNA methylations on cancer cell development in a set of genes directly related to tumorigenesis (Table S2). Figure 1D illustrates the median expression and methylation differences between samples with or without aberrant DNA methylation in cancer-associated genes. In line with previous studies, we observed 43 of the aberrations are cancer type-specific events\textsuperscript{13,19,22}, such as tumor necrosis receptor \textit{FAS} in GBM\textsuperscript{23} and homeobox genes associated transcription factor \textit{HNF1B} in UCEC\textsuperscript{24}, highlighting the context-
dependent complexity of DNA methylation regulation and tissue-specific carcinogenesis\textsuperscript{13}. Conversely, only 3 of them are common events across cancer types, including hypomethylated \textit{TRIP10}, \textit{MGMT}, and \textit{STAT5A}. The recurrence of those aberrant DNA methylation patterns across cancer types suggests their importance in cancer cell development.

4.4.2 \textit{cis}-acting aberrant DNA methylation on driver genes can be tumorigenic

We reasoned that DNA methylation modulating driver gene expression could be important in cancer progression, so we characterized \textit{cis}-acting aberrant DNA methylation on 299 driver genes\textsuperscript{25} to understand the nature of tumorigenic DNA methylation. Since DNA methylation can affect the binding of transcription factors and \textit{vice versa}\textsuperscript{26}, we examined the number and enrichment of transcription factor binding sites (TFBS) in loci associated with expression changes. All the loci are associated with at least one TFBS, and hypomethylated CGI sites were characterized by a higher number of TFBS (mean 25.1 TFBS) than hypermethylated CGI sites (mean 15.8 TFBS) (Figure 2A). \textit{STAT5A} and \textit{HIF1B}, which are both master regulators involved in tumorigenesis in various cancer types, were the most enriched TFBS in hypermethylated and hypomethylated loci, respectively. The results support the notion that transcription factors might be the readers and effectors of aberrant DNA methylation\textsuperscript{26}, leading to altered expression as revealed by transcriptomic and proteomic data.

Next, we wanted to explore the relationship between genetic alterations and DNA methylation for each gene and cancer type. Of the 7 driver genes with \textit{cis}-acting DNA methylations, most of the methylations were mutually exclusive with genomic alterations as shown in Figure 2B. For example, the correlation between \textit{B2M} expressions and \textit{B2M} promoter DNA methylation
suggests methylation is the main factor regulating its expression (Figure 2C). Our results suggest aberrant DNA methylation resulting in functional changes undergoes positive selection and thus likely sufficient to promote tumorigenesis.

Interestingly, DNA hypomethylation is the main perturbation occurring in the \textit{IDH2} gene and was found in 7.9\% (17 of 215 tumors) of the total tumors in ccRCC (n=13) and LUAD (n=4). Promoter DNA methylation is strongly correlated with \textit{IDH2} expression (Figure 2D). Since overexpression of \textit{IDH2} contributes to altering energy metabolism\textsuperscript{27,28}, we examined the \textit{IDH2}-related cancer metabolism activities between samples with and without \textit{IDH2} hypomethylation. The results suggest a general upregulation of metabolic gene expression including \textit{KDMs}, \textit{ALKBHs}, \textit{TETs}, and \textit{MTOR}, while the genes and extent vary slightly by cancer types (Figure 2D). Also, we detected a compelling enrichment of samples with hypomethylated \textit{IDH2} in metabolic processes directly associated with tumor development (Hypergeometric P < 0.05, Figure S2). \textit{IDH2} hypomethylation is consistently correlated with upregulation of \textit{MTOR} across cancer types, suggesting its prevalent oncogenic function through \textit{MTOR} pathway activation.

4.4.3 Hypomethylated RTKs are newly identified driver events

We found that several receptor tyrosine kinases (RTKs) (e.g. \textit{FGFR2} and \textit{EGFR}) are frequently hypomethylated across cancer types (Figure 1D). To better dissect the contribution of hypomethylated RTKs to oncogene activation, we systematically examined promoter methylation and genetic alterations, including mutations, fusions, and CNVs, and their effects on RTKs RNA and protein levels.

While 66.7\% (12 of 18 mutations) identified \textit{FGFR2} missense and indel mutations were activating mutations that enabled high-grade inflammation and cell proliferation\textsuperscript{29–34}(Figure 3A),
we discovered 7 UCEC tumors carrying cis-acting hypomethylated FGFR2, 5 of which had co-occuring genomic alterations. Unsupervised clustering of DNA methylation data across 101 CPTAC UCEC tumors and 43 TCGA normal samples revealed that FGFR2 hypomethylated cases form a distinct cluster with lower DNA methylation than normal cases (Figure 3B). Specifically, one CGI (cg10314760) within the FGFR2 promoter displayed a strong correlation between promoter hypomethylation and active gene expression both at RNA and protein levels (Figure 3C). Our results suggest that promoter methylation is a major factor modulating FGFR2 expression and FGFR2 hypomethylation represents another main mechanism of RTK activation comparable to activating mutations.

To separate the oncogenic effects of FGFR2 hypomethylation from co-occurring aberrations, we stratified UCEC tumors by the FGFR2 genomic alteration status and examined FGFR2 expression within each stratum. We found that tumors with FGFR2 hypomethylation are correlated with FGFR2 upregulation in each stratum (Figure 3D). The result not only reveals the dominant gain-of-function role of promoter hypomethylation, but also suggests that whereas activating mutations, amplifications, and promoter hypomethylation enable FGFR2 upregulation to different extents, co-occurring FGFR2 hypomethylations result in similar, even more profound, expression changes than either genomic alteration alone.

Similarly, Hypomethylated EGFR is associated with EGFR upregulation in GBM (Figure S3). Overall, although recurrent gain-of-function genomic alterations in RTKs have long been known to promote a variety of cancers\textsuperscript{35}, our results reveal that RTK hypomethylation is a \textit{bona fide} epigenetic driver across multiple cancer types.
4.4.4 Transcription factor STAT5A hypermethylation is associated with pervasive changes in STAT5A regulome activity

Altered expression of a single transcription factor (TF) usually disrupts downstream regulome activity, the interplay between TF and its products, and leads to neoplastic transformation\textsuperscript{36}. Therefore, to identify deregulated TF genes modulated by promoter methylation in tumors, we tested the association between TFs with \textit{cis}-acting aberrant DNA methylation and their regulome gene expression changes at both RNA and protein levels.

\textit{STAT5A} controls a wide range of cellular processes such as cell identity, cytotoxicity, and cell survival; dysregulation of those pathways contributes to tumorigenesis\textsuperscript{37}. Unsupervised clustering of \textit{STAT5A}-interacting gene expression data from 101 HNSCC tumors divided samples into two groups: those with high regulome activity, and those with low regulome activity (Figure 4A). Notably, samples with hypermethylated \textit{STAT5A} were significantly enriched in the regulome-low group (Fisher’s exact test, \(P=0.001\)). The same pattern was observed in protein abundance (Figure 4B, Fisher’s exact test, \(P=0.028\)). The enrichments were not correlated with tumor purity (Figure S4). Exome sequencing of these tumors did not identify any distinct, recurrent coding sequence mutations in \textit{STAT5A}-interacting genes (Table S3), suggesting additional genetic drivers were not involved. \textit{STAT5A} phosphorylation is not significantly associated with either \textit{STAT5A} methylation status or \textit{STAT5A} regulome activity (Figure S4). The enrichments we observed may indicate \textit{STAT5A} is aberrantly methylated in HNSCC tumors and may subsequently lead to pervasive regulome changes.

To delineate the contribution of aberrantly methylated \textit{STAT5A} to downstream regulome changes, we compared HNSCC tumors by \textit{STAT5A} methylation status. Since samples with hypermethylated \textit{STAT5A} are associated with lower regulome activity both at RNA and protein
levels (Figure 4C), we hypothesized that STAT5A-interacting components would be downregulated in samples with hypermethylated STAT5A, including receptors, kinases, repressors, co-activators, and target genes (Figure 4D). Among the target genes, we observed significant downregulation of IRF1, PRF1, IFNG, IL2RA, and IL6ST (Wilcoxon P<0.05) both at RNA and protein levels, suggesting a direct regulatory role of hypermethylated STAT5A in cytokine production, cytotoxicity, cell proliferation, and interferon signaling. Similar to HNSCC tumors, samples with hypermethylated STAT5A are associated with low regulome activity in ccRCC tumor (Figure S4). In line with our previous finding that STAT5A binding motif is enriched within driver genes with aberrant DNA methylation (Figure 2A), our results suggest STAT5A plays an important role in DNA methylation-mediated tumorigenesis.

4.4.5 **Hypermethylated STAT5A is associated with immune cells depletion in HNSCC**

STAT5A signaling represents a critical signaling pathway in the pathogenesis of human cancers. Recent data have shown that STAT5A-mediated interferon signaling regulates the expression of PD-L1 and PD-L2, reflecting the clinical significance of STAT5A signaling in immunotherapy^38^. The lowly expressed STAT5A target genes from our analysis are directly implicated in immune response, such as IL2RA, IRF1, and IFNG, which cooperate to maintain normal immune function and homeostasis^39^. Thus, we focused on characterizing the immune component of HNSCC tumors to understand how hypermethylated STAT5A affects the tumor microenvironment.

To construct the microenvironment of HNSCC tumors, we deconvoluted cell mixtures into multiple immune cell types from transcriptome data by using xCell^40^. Consensus clustering of
thirty different immune-related cell types identified three major immune clusters, and \( \text{STAT5A} \) hypermethylated samples were significantly enriched in one of the immune subtypes (Fisher’s exact test, \( P=0.002 \)) (Figure 5A). Interestingly, the Immune Subtype 2 of HNSCC had significantly lowest xCell immune score (Figure 5B, Wilcoxon \( P < 0.01 \)) and was enriched with late-stage tumors (Figure 5A, top panel, Fisher’s exact test, \( P=0.007 \)). Deconvolution of tissue-infiltrating immune and stromal populations revealed that samples with hypermethylated \( \text{STAT5A} \) had limited numbers of CD4+ and CD8+ T cells, macrophages, dendritic cells, and stroma cells relative to samples with normal methylated \( \text{STAT5A} \). In line with the enrichment of samples with hypermethylated \( \text{STAT5A} \) in the immune-low group, \( \text{STAT5A} \) hypermethylated samples, which adjusted for tumor purity, displayed lower immune scores (Figure 5B), decreased expression of genes associated with immune effectors and dendritic cells at both gene and protein levels (Figure 5C and 5D).

Our finding of lower tumor microenvironment factors in \( \text{STAT5A} \)-hypermethylated samples is consistent with previous studies in which hematopoietic stem cell proliferation was severely impaired in the \( \text{Stat5A} \)-deficient mice\textsuperscript{41–43}. Epigenetically silenced \( \text{STAT5A} \) has been identified as a key tumor suppressor in lymphoma cancer cell lines, contrasting to the role of its closely related paralog, \( \text{STAT5B} \), in malignant transformation\textsuperscript{44}. In addition, studies have indicated that the development of HNSCC is closely related to immunosuppression and immune escape\textsuperscript{45}, suggesting that \( \text{STAT5A} \) hypermethylation might be a mechanism to mediate the disease dependent specificity of \( \text{STAT5A} \) among \( \text{STATs} \) family\textsuperscript{46}, where the alteration disallow access to distinct panels of genes involved in tumor immunogenicity.
4.4.6 Aberrant methylation associated with therapeutic vulnerability in cancer

Understanding the global methylation patterns for various tumors helps tailor treatment to specific tumor subtypes. To identify subtypes based on methylation patterns, we used uniform manifold approximation and projection (UMAP) to reduce methylation signals from 340,000 CGIs into two dimensions. Cancers form well-separated clusters by the organ system, such as brain (GBM), kidney (ccRCC), lung (LUAD), head and neck (HNSCC), and uterus (UCEC) (Figure 6A). The results suggest that the UMAP projection of DNA methylation data reflects cell-of-origin faithfully.

Next, we took CGIs showing significant difference between tumor samples and normal tissues, then performed unsupervised classification of tumors within each cancer type using consensus clustering on the most variable 8,000 CGIs. We identified between 2 and 6 clusters from each cancer type, and re-colored the UMAP projection according to our subtype classifications (Figure 6A, third panel). The methylation subtype result captured several important clinical characteristics. For example, UCEC subtype C3 is enriched with tumors displayed high frequency microsatellite instability and has higher mean MSIsensor score compared to other UCEC groups (Figure S6). The GBM subtype C6 features CpG island methylator phenotype associated with IDH1 mutation. Our result suggests that clustering of cancer samples based on variable DNA methylation can lead to identification of molecularly and clinically relevant subtypes.

To further explore the biological differences between methylation subtypes, we performed over-representation pathway analysis using differentially expressed genes and proteins in each of the per-cancer methylation subtypes (Figure 6B and Table S4, Hypergeometric P < 0.05). Significant subtype-specific tumorigenic signatures were consistently observed at the
transcriptomic and proteomic level, such as UCEC-C2 being enriched for NOTCH signaling, and LUAD-C2 with RNA metabolism\textsuperscript{48,49}. Despite having very different cancer type compositions, HNSCC-C1 and GBM-C3 were characterized by immune-related signatures both at RNA and protein levels, and were strongly associated with the immune-high groups identified by cell-type enrichment score (Fisher’s exact test, $P < 0.001$). The correlation suggests signaling convergence among various cancers, in line with the previous studies that some methylation subtypes are significantly associated with immune signature\textsuperscript{49–51}. We also identified four druggable genes that are significantly overexpressed within specific methylation subtypes compared to the other subtypes from each cancer type (Figure S6 and Table S5, Wilcoxon $P < 0.005$). The results not only demonstrate the heterogeneity within cancer types, but also provide the new clues for understanding how distinct methylation patterns may lead to various cancer phenotypes across cancer types.

Finally, we investigated potentially druggable targets with \textit{cis}-acting deregulated DNA methylation, which may confer a therapeutic vulnerability to drugs targeting aberrant gene expression (Table S6). We integrated \textit{cis}-acting DNA methylation with the Clinical Interpretation of Variants in Cancer (CIViC)\textsuperscript{52}, prioritizing for target genes with outlier expression for which pharmacological intervention might be available. Allowing for “off-label” drug treatment, we found 31.6% of samples (160 of 506 tumors) likely to benefit from one or more treatments targeting genes altered by DNA methylation (Figure 6C). The most frequent druggable DNA methylation events across the five cancer types are those on \textit{EGFR} (number of tumors = 55), \textit{NAPRT} (n=52), and \textit{MGMT} (n=33). Tumor-specific loss of \textit{NAPRT} mediated by promoter hypermethylation is synthetically lethal with NAMPT inhibitor treatment in multiple cancer types, resulting in inactivations of nicotinic acid salvage pathways\textsuperscript{53}. Collectively, the characterization
of cis-acting aberrant DNA methylation and cancer methylome reveals potential new directions for treatment optimization.

### 4.5 Discussion

Our multi-omics analysis of 506 tumors greatly boosts our ability to identify functional impacts of epigenetic changes in tumors, providing unprecedented insights into both cancer type-specific and common cis-acting DNA methylation and methylation-based stratification of cancer patients. Based on genome-wide DNA methylation patterns, we identified and characterized twenty methylation subtypes enriched with various RNA and protein signatures showing potential therapeutic and prognostic implications for cancer management. Of these, one multi-cancer group consists of different cancer types enriched with a convergent immune-related signature. Therefore, the cancer methylome might be an important factor in determining the efficacy of potential immunotherapies. We observed significantly elevated expression levels of putative drug targets in four methylation subtypes including two LUAD subtypes and two GBM subtypes (Figure 6B), implying that the differential expression of target genes is related to distinct methylation patterns. Moreover, we observed clinically-relevant alterations with important therapeutic potential in 160 out of 506 tumors. Targeting those common aberrant methylation events could broaden the therapeutic reach of existing drugs by including more patients and a broader range of tumor types. To maximize this benefit, future studies optimizing epigenetic therapies should be performed.

We uncovered several bona fide DNA methylation drivers with functional consequences, such as hypomethylated FGFR2 and EGFR (Figure 3 and Figure S3). Excluding mutations and copy number variations in FGFR2, 7 out of 95 (7.4%) of UCEC tumors harbored hypomethylated
FGFR2, which could expand conventional genotype-directed clinical trials. In addition, the apparent co-occurrence of FGFR2 hypomethylation and genomic alterations within the same tumors suggests an selective advantage to development of a second FGFR2 alteration, perhaps from enhanced FGFR2 signaling in these cells. It is also possible that the genomic alterations disrupt the reading, writing, or maintaining of DNA methylation machinery in tumor cells, subsequently leading to an aberrant decrease of methylation within the FGFR2 promoter. However, since our result shows UCEC tumors harboring epigenetic and genetic alterations are significantly associated with FGFR2 upregulation, it is likely that FGFR2 hypomethylation works in concert with other genomic alterations at FGFR2 to promote tumorigenesis.

Studies have shown HNSCC tumors evade the host immune system by manipulating their own immunogenicity\textsuperscript{45}, suggesting that STAT5A hypermethylation uncovered in our study is a novel mechanism to promote immunosuppression. Furthermore, our findings reinforce the critical nature of STAT5A as a signaling hub in modulating tumor immunogenicity, shaping immune response, and orchestrating immune cell differentiation across multiple cancers, and may suggest opportunities for therapeutic intervention through reversing epigenetic alterations within the STAT5A promoter. DNMT inhibitors such as 5’-aza have been shown to reduce methylation of STAT5A promoter in cell lines\textsuperscript{44}, and was an FDA-approved therapy in treating myelodysplastic syndrome\textsuperscript{54}. Furthermore, activation of STAT5A signaling may transform an immunologically cold, or inactive, tumor into a hot, inflamed tumor and thus increase anti-tumor immune response. Understanding the full extent of mechanisms responsible for the effect of STAT5A hypermethylation on the immune-related signaling pathways, the interaction between them, and the impact on therapeutic sensitivity requires additional investigation.
Overall, our results help deconvolute the contribution of DNA methylation in tumorigenesis and delineate what roles it has in initiating and maintaining malignancies. This deep characterization of cis-acting events and cancer methylome will inform functional explorations of aberrant DNA methylation in a systematic way, revealing potential new disease mechanisms and therapeutic opportunities.

### 4.6 STAR Methods

#### 4.6.1 CPTAC datasets description

We aggregated somatic variants, copy number variations, transcriptomic and proteomic data generated by the National Cancer Institute CPTAC from CPTAC data portal, Genomic Data Commons (GDC), and published studies (See Data and Code Availability). The datasets includes CPTAC Clear Cell Renal Cell Carcinoma (ccRCC) Discovery Study\(^5^5\), CPTAC Glioblastoma (GBM) Discovery Study, CPTAC Lung Adenocarcinoma (LUAD) Discovery Study\(^5^6\), CPTAC Head and Neck Cancer (HNSCC) Discovery Study, and CPTAC Uterine Corpus Endometrial Carcinoma (UCEC) Discovery Study\(^5^7\).

Generally, all the whole exome sequencing and RNA sequencing data was harmonized by National Cancer Institute Genomic Data Commons (GDC) [https://gdc.cancer.gov/about-data/gdc-data-harmonization](https://gdc.cancer.gov/about-data/gdc-data-harmonization), which included alignment to GDC’s hg38 human reference genome (GRCh38.d1.vd1) and additional quality checks. Somatic mutations were called by the SomaticWrapper pipeline, which includes four different callers: Strelka\(^2^5^8\), MUTECT v1.1.7\(^5^9\), VarScan v.2.3.8\(^6^0\), and Pindel v.0.2.5\(^6^1\). Gene fusions in RNA-Seq samples were called using three callers: STAR-Fusion, EricScript, and Integrate, with fusions reported by at least 2 callers or reported by STAR-Fusion being retained\(^6^2\). For transcriptomic data, We obtained the gene-level
read count, Fragments Per Kilobase of transcript per Million mapped reads (FPKM), and FPKM Upper Quartile (FPKM-UQ) values by following the GDC’s RNA-Seq pipeline (Expression mRNA Pipeline) [https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/](https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/), except running the quantification tools in the stranded mode. We used HTSeq v0.11.2 to calculate the gene-level stranded read count (parameters: -r pos -f bam -a 10 -s reverse -t exon -i gene_id -m intersection-nonempty --nonunique=none) using GENCODE v22 (Ensembl v79) annotation downloaded from GDC (gencode.gene.info.v22.tsv). The read count was then converted to FPKM and FPKM-UQ using the same formula described in GDC’s Expression mRNA Pipeline documentation. The peptide search database of those cohorts is RefSeq 20180629. To facilitate the cross-data interpretation, we improved the annotation of the proteomic databases by recovering HGNC Gene IDs from NCBI/Entrez Gene IDs using HGNC BioMart ([https://biomart.genenames.org/](https://biomart.genenames.org/)) and retrieving the UniParc IDs and checksums for all the protein sequence in all databases (DOI: [10.1093/bioinformatics/bth191](https://doi.org/10.1093/bioinformatics/bth191)). Note that we keep the gene symbols identical to the RefSeq records at the time of download.

### 4.6.2 DNA methylation data preprocessing

Raw methylation image files generated by Illumina Infinium EPIC BeadChip were downloaded from the CPTAC GDC (See Data Availability). We calculated and analyzed methylated(M) and unmethylated (U) intensities for tumor and normal adjacent tissue samples as described previously. We flagged locus as NA where probes did not meet a detection p-value of 0.01. Probes with MAF more than 0.1 were removed, and samples with more than 85% NA values were removed. Resulting beta values of methylation were utilized for subsequent analysis.
The raw methylation image files of TARGET-NBL, TCGA-GBM, and TCGA-UCEC datasets generated by Illumina Infinium HumanMethylation450 BeadChip assays, used here for normal adjacent tissue approximation, were downloaded from TCGA FireHose (https://gdac.broadinstitute.org/) or TARGET data matrix (https://ocg.cancer.gov/programs/target/data-matrix) and processed as aforementioned steps.

4.6.3 Defining aberrant DNA methylation using RESET
We mapped CpG probes to canonical transcriptional start sites and unconventional exonic TSS as defined by FANTOM5 consortium, and then applied RESET algorithm separately on each tumor type using the corresponding adjacent normal tissue samples. For identifying aberrant DNA methylation events associated with transcriptional or translational changes, aggregated mRNA expression and global proteome data were used as an input, respectively. We considered the association as significant only events with a false discovery rate FDR < 10% and a RESET score > 1.

4.6.4 Defining cancer-associated genes
Cancer-associated genes were compiled from genes defined by Bailey et al and cancer-associated genes listed in Mertins et. al and adapted from Vogelstein et al.

4.6.5 Cell type enrichment deconvolution using gene expression
We used the FPKM-UQ expression matrix as the input, and the abundance of each cell type was inferred by the xCell web tool. xCell is a gene signatures-based method learned from thousands of pure cell types from various sources, which performed the cell type enrichment
analysis from gene expression data for 64 immune and stromal cell types (default xCell signature). xCell generated an immune score per sample that integrates the enrichment scores B cells, CD4+ T-cells, CD8+ T-cells, DC, eosinophils, macrophages, monocytes, mast cells, neutrophils, and NK cells; a micro-environment score which was the sum of the immune score and stroma score.

4.6.6 Immune clustering using cell type enrichment scores
We used the FPKM-UQ expression matrix as the input, and the abundance of each cell type was inferred by the xCell web tool. xCell is a gene signatures-based method learned from thousands of pure cell types from various sources, which performed the cell type enrichment analysis from gene expression data for 64 immune and stromal cell types (default xCell signature). xCell generated an immune score per sample that integrates the enrichment scores B cells, CD4+ T-cells, CD8+ T-cells, DC, eosinophils, macrophages, monocytes, mast cells, neutrophils, and NK cells; a micro-environment score which was the sum of the immune score and stroma score.

Immune subtypes of each of HNSCC tumors were generated based on the consensus clustering of the cell type enrichment scores by xCell. Among the 64 cell types tested in xCell, we selected immune-related cell types. We performed consensus immune clustering based on the z-score normalized xCell enrichment scores. The consensus clustering was determined by the R package ConsensusClusterPlus(parameters: reps = 2000, pItem = 0.9, pFeature = 0.9, clusterAlg = "kmdist", distance = "spearman").

4.6.7 Pathway over-representation analysis
To designate the representative pathways of methylation subtypes from transcriptomic and proteomic data, we used the Wilcoxon rank sum test to select the top 250 differentially expressed
features (mRNA and proteins) for each subtype. We then performed hierarchical clustering on these features. Each set of clustered features underwent pathway enrichment analysis using Reactome\textsuperscript{69}. Pathways with p-value smaller than 0.05 were manually reviewed and highlighted in Figure 6B.

4.6.8 Identification of druggable gene with aberrant methylation
CIViC is a curated list of druggable variants describing their therapeutic, prognostic, diagnostic and predisposing relevance. We downloaded the list and intercept with the list of genes with aberrant methylation. The potential druggability of each gene was manually reviewed to see if the altered expression of such gene is associated with therapeutic relevance supported by literature.
4.7 Figures

Figure 4.1 The cancer methylome landscape associated with transcriptomic and proteomic change
(A) Significance of aberrant methylation associated with transcriptomic (upper) and proteomic (lower) changes. The statistical significance (false discovery rate (FDR)<0.1)
and median difference of expression between aberrant and normal samples are displayed in volcano plot. Representative genes are colored based on methylation status: yellow, hypermethylation; blue, hypomethylation. Dot size indicates the number of CGIs associated with expression changes.

(B) Venn diagrams of the number of genes with hypermethylation (upper) and hypomethylation (lower) having significant transcriptomic and/or proteomic changes.

(C) Correlation between RNA expression and protein abundance of hypermethylation (blue), normal methylation (grey), and hypomethylation (yellow).

(D) Common and cancer type-specific aberrant methylations. The mean methylation values difference (upper), RNA difference (middle), and protein difference (lower) between aberrant and normal samples at significant CpG sites are indicated by shading of the filled circle. The methylation difference is derived from the mean beta values differences between samples with or without aberrant methylation. For RNA and protein differences, the values are derived from the mean differences of scaled RNA sequencing data or proteomic data between samples with or without aberrant methylation.
Figure 4.2 Characterization of aberrant methylation in driver genes

(A) The Distribution of the number of transcription factor binding sites for functional hypermethylation (yellow) and hypomethylation (blue).

(B) Mutual exclusivity and co-occurrence of genomic and epigenomic alterations in driver genes in different cancer types. Each column represents a different tumor.

(C) Example of correlations between promoter methylation and gene expression (upper) and protein abundance (lower) in selected driver genes. Samples are colored based on genetic and/or epigenetic alterations of the gene.

(D) Correlations between promoter methylation and gene expression (upper) and protein abundance (lower) in IDH2. Pathway diagram representing the mean expression difference between IDH2 hypomethylated samples and normal methylated samples as indicated by shading of the filled squares.
Figure 4.3 Collaborative effects of *FGFR2* mutations and hypomethylation on *FGFR2* upregulation

(A) Lolliplot showing somatic mutations of *FGFR2* in UCEC samples. The amino acids and types of mutations are labelled. Positions which are recurrently mutated are highlighted with the number of occurrences. The *FGFR2* functional domains are colored. Ig: Immunoglobulin.

(B) DNA methylation of the *FGFR2* promoter in UCEC tumors (upper) and normal adjacent tissues (lower).

(C) Correlation between methylation at significant CpG sites and gene expression (upper) and protein abundance (lower). Samples are colored based on genetic and/or epigenetic alterations of *FGFR2*. Tumors harboring *FGFR2* hypomethylation are highlighted by large dot size.
(D) Effect of \textit{FGFR2} hypomethylation on RNA and protein levels stratified by \textit{FGFR2} genomic alterations. Tumors harboring \textit{FGFR2} hypomethylation are highlighted by large dot size.
Figure 4.4 STAT5A hypermethylation associated with pervasive STAT5A regulome changes

(A) Unsupervised clustering of STAT5A regulome genes using Pearson correlation of scaled RNA sequencing data and annotated with STAT5A expression and methylation levels. Mean activity indicates the overall sum of regulome activity.

(B) Unsupervised clustering of STAT5A regulome genes using Pearson correlation of scaled global proteome data.

(C) Violin plot showing samples with hypermethylated STAT5A have lower regulome activity than the samples with normal methylated STAT5A.

(D) Pathway members and interactions in the STAT5A regulome. The mean expression differences between STAT5A hypermethylated samples and normal methylated samples are indicated by shading of the filled squares.
Figure 4.5 Functional impact of STAT5A hypermethylation on HNSCC tumor immunosuppression.

(A) Distinct immune subtypes of HNSCC tumors identified by consensus clustering of 106 HNSCC tumors using xCell enrichment scores. Top panel shows the immune score, DNA methylation status of STAT5A, immune subtype, and tumor stage. The heatmap shows the xCell enrichment scores deconvoluted from RNA-seq data.

(B) Violin plot showing samples with hypermethylated STAT5A are enriched in Immune Subtype 2 and have decreased immune score.

(C) Violin plot showing samples with hypermethylated STAT5A have decreased expression of immune effectors.

(D) Violin plot showing samples with hypermethylated STAT5A have decreased expression of dendritic cells.
Figure 4.6 Summary of the cancer methylome for tumor signatures and therapy

(A) Projection of the 506 cancer methylomes using UMAP. Each point is a sample and is colored based on the cancer type (left), sample type (middle), or methylation subtype (right).

(B) Relationship of per-cancer methylation subtypes (upper), RNA expression signature (middle), and protein signature (lower). Potentially druggable genes that are significantly upregulated within subtypes are labeled in inset.

(C) Breakdown of potentially druggable genes whose expression is altered by tumorigenic DNA methylation.
Reference


Chapter 5: Conclusion and Future Directions
5.1 Experimental validation can confirm and complement large-scale multi-omic analysis.

Our work in previous chapters illustrate the power of large-scale multi-omic analysis in understanding both heterogeneity and complexity of human cancers. We believe that our analytical framework and systematic pan-cancer studies uncover determinants and consequences of genetic and epigenetic variation beyond the conventional mutational profiling, revealing potential new disease mechanisms and therapeutic opportunities. To further validate the predicted therapeutic effect identified in our data-driven analysis results, one of the future directions that would be interesting to pursue is to conduct wet lab experimental validation. Take the tumorigenic DNA methylation for example, we can introduce site-directed methylation or demethylation at a particular gene promoter in the corresponding cell line or mouse model by using CRISPR/Cas9 system\(^1\). With confirming \(FGFR2\) upregulation after introducing site-specific demethylation at \(FGFR2\) promoter in endometrial cell line\(^2\), we can perform growth assay and colony formation assay to determine the oncogenic role of hypomethylated \(FGFR2\) in UCEC. This will help us to elucidate the underlying mechanism of tumor development regulated by RTK hypomethylation. Furthermore, we can apply the established cell line model in drug screening aiming to revert this aberrant methylation in tumor cells. By harnessing the power of large-scale multi-omics analysis, scientists can better prioritize the therapeutic candidate and devise the experimental validation accordingly.

5.2 Combining bulk and single-cell technology can enhance the identification of tumor origin.

Following the large-scale multi-omics bulk sequencing methods, a number of single-cell technology that can better dissect the heterogeneous subpopulations of tumors emerged recently\(^3\). Another future direction is to expand the depth of large-scale multi-omics analysis using single
cell technologies. One question we are particularly interested is to identify the cell-of-origin of
tumors, i.e. the normal cell that acquired the first cancer-promoting mutations, by leveraging the
complementary nature of high-throughput DNA methylation array and single-cell nuclei RNA
sequencing for chromatin accessibility. As part of the effort, we have done extensive work of using
methylation data to identify global methylation patterns for 6000 human tumors across 15 cancer
types from CPTAC, TCGA, and TARGET (Figure 1). By reducing methylation signals from
340,000 CGIs into two dimensions, we observed that most tumors form well-separated clusters by
the organ system, such as brain (GBM and NBL), kidney (ccRCC), blood (AML), large
intestine(COAD and READ), and uterus (UCEC), while the squamous cell cancers, i.e LSCC,
HNSCC and CESC, form a distinct cluster that partially clustered with other cancer types. The
results suggest that the projection of DNA methylation data reflects cell-of-origin faithfully. To
further dissect the particular lineage trajectories of tumors, we will use paired snATAC-seq and
snRNA-seq data for 50+ cases across 10 cancer types. First, we will reduce the snATAC-seq data
from thousands peaks into two dimensions, and identify common and distinct sub-clusters among
different groups based on the profiles of chromatin accessibility. Next, we will quantify the global
epigenetic patterns, the transcriptional network differences, and clinical relevance for those
heterogeneous sub-clusters by connecting differences in transcriptional regulation with differential
gene expression across clusters. The overall clustering for various diseases should mostly reflect
cell-of-origin faithfully, while particular sub-clusters will reveal unique lineage differentiation
trajectories. Our in-depth analysis will provide high-resolution insights to transcriptional and
cellular dynamics among tumors with common and distinct cell-of-origins.
5.3 High-throughput functional genomics empowers genomics-guide cancer treatment

The current paradigm of genomic-guide cancer medicine research has been focused on targeting single gene critical for optimal cancer cell fitness, such as oncogene or tumor suppressor gene. However, the concerted effort of large-scale cancer multiomic profiling so far has shown that there are finite and diminishing opportunities to cure cancer merely by targeting one single gene. Synthetic lethality is the genetic interaction between two or more genes where only their co-alteration results in cell death, emerging as a promising therapeutic strategy against cancer recently. This is exemplified by PARP inhibitor, which is an effective therapy for cancer patients with BRCA1 or BRCA2 mutations in tumors across multiple cancer types. When BRCA mutant tumors are treated with a PARP inhibitor, the two complementary DNA repair pathways, DNA homologous recombination repair and single-strand break repair, are co-targeted in tumors, resulting in therapeutic vulnerability in tumor cells. Such synthetic lethal interactions remain to be discovered in cancer. Therefore, the other future direction I would like to pursue is to create new tumor vulnerability from synthetic lethal interactions among genes that lead to cancer by using high-throughput CRISPR/Cas9-based mutagenesis studies. To take advantage of the considerable progress in accumulated cancer sequencing data, cancer multiomics, and functional genomic assay using CRISPR/Cas9, I propose to test the synthetic lethality interaction in a comprehensive way. Findings will advance our understanding of the interactions among various abnormalities in tumorigenesis, provide guidance for effective treatments based on molecular features, and potentially help to expand the repertoire of genomics-guided therapy. My long-term goal is to effectively identify clinical actionable targets in cancer cells by employing data-driven analysis combined with patient-guided CRISPR mutagenesis.
5.4 Concluding remarks

We believe that our research will have significant impact on different areas of genomic-guided cancer treatment research: (1) elucidating the patterns and potential therapeutic effects of microsatellite instability, gene fusions, and aberrant DNA methylation, (2) demonstrating how large-scale multi-omics analysis can expand our understanding of human cancers, and (3) laying the groundwork for how we effectively interpret tumorigenic abnormalities and associated clinical relevance. Together, we hope that our research will bridge the bench research to the bedside clinical application in cancer treatment.
5.5 Figures

Figure 5.1 The cancer methylome landscape across different cancer types
Projection of the 6000 cancer methylomes using UMAP. Each point is a sample and is colored based on the sample subtype (left), cancer type (middle), or data source (left)
Reference