PAN-CANCER ANALYSIS OF DRIVER MUTATIONS AND MODULES IN CANCER GENOMES

By

CARLOS AREVALO

UNIVERSITY OF CALIFORNIA SANTA CRUZ

Molecular, Cell and Developmental Biology

MARCH 2021

SUMMARY

Cancer driver mutations and modules are central to understanding tumorigenesis, tumor malignancy, recurrence, and drug resistance. However, driver mutations and modules across tumor types are not completely characterized nor understood due to limitations of tumor samples and the complexity of biological pathways. Here, we present a comprehensive analysis of driver mutations and modules in cancer genomes across 49 cancer types from The Cancer Genome Atlas (TCGA) Pan-Cancer studies, and Cancer Cell Line Encyclopedia (CCLE). We used whole-exome somatic mutation variants from TCGA and CCLE to characterize a landscape of driver alterations across cancers in protein coding genomic regions. We identified well-known and studied driver mutations and revealed previously unannotated driver genes. *TP53* was the most significant and mutated driver across tumors, followed by *KRAS*, *PTEN*, *RB1*, *CDKN2A*, *NRAS*, *CASP8*, and *STK11*. *DSPP* and *PTPRQ* were the only unannotated driver gene in the Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC), which were highly significant in both TCGA and CCLE pan-cancer analysis. We also observed highly significant and potentially unannotated driver genes such as *ELAVL1*, *FHL2*, *GNB2L1*, *DNAH8* and *DNAH9*. Many of these possible unannotated drivers, such as *DSPP*, have been previously suggested as potential cancer driver genes using experimental approaches, however, they have not been previously characterized as drivers across all cancer types. This study provides a model and comprehensive landscape of driver genes and mutations across cancer type genomes which might serve as an asset for future research and clinical endeavors.

ACKNOWLEDGEMENTS

During the last year I had the opportunity to discover more about the mechanisms that lead to genetic malignancies such as cancer.

I am forever indebted to Dr. Angela Brooks who has advised me since junior year, carefully guiding me through the realism of bioinformatics, computational biology, transcriptomics, and cancer research. Dr. Brooks, through your motivation and example, patience and dedication, you inspire and have shaped me into a scientist that looks beyond science, but also to solve the social issues that affect our surrounding communities. I am also indebted to Jon Akutagawa who has provided great advice and with patience and dedication has instructed me computational and theoretical skills throughout this project. Jon, your support, encouragement, and dedication have helped me to learn so much about bioinformatics and computational biology and inspired me to work in my first research manuscript.

I am also thankful to all past mentors including Dr. Cameron Soulette and Dr. Eva Robinson. Dr. Soulette taught me the theory and process of using nanopore long-read sequencing to discover novel isoforms associated with splicing factor alterations in cancer, and experimental approaches to validate the results. Dr. Soulette, who I spent weeks trying to validate U2AF1 S34F-associated novel isoforms, inspired me to work with patience and motivation and always trust the scientific process. Dr. Robinson, who I used to ask any question related to both my experimental work and general scientific curiosity, has been a great mentor that with patience and enthusiasm guided me through the process of failed experiments.

My deepest gratitude to Alison Tang, Brandon Saint-John, Alexis Thornton, and Dennis Mulligan for instructing me more on how to use certain computational methods and approaches.

I am also thankful to the UCSC Genomics Institute Research Mentoring Institute (RMI) program for providing great academic and financial support throughout the last two years. I am indebted to Zia Isola for her great academic support and dedication to the RMI cohort.

I am forever grateful to the Broad Institute of MIT and Harvard Summer Research Program (BSRP) for providing mentorship and training in computational genomics. I am indebted to Dr. Anna Greka, who has provided great academic support and many times taken the time to discuss my future in science. Dr. Greka, who I study lipotoxic signatures that lead to diabetes, you are an inspiration and hope one day I can help so many people directly through science and medicine like you do.

CONTENTS

INTRODUCTION

The Historic Race to Cancer Omics

Cancer is an ancient set of diseases of the *-ome*: genome, transcriptome, proteome, and metabolome, and as breakthroughs advance, we discover even more about the audacity of cancer cells. A century before DNA sequencing and the genomic revolution started, we first learned to study the hallmarks of cancer cells, biological capabilities acquired during the many steps of tumor development, such as cell proliferation, and the metabolic processes that allow cancer cells to obtain resources under various conditions (Hanahan, et al. 2011). Despite not being a universal feature of proliferating cancer cells, the *Warburg Effect* taught us the ways most cancer cells adapt to different environments (Warburg O., 1925; Warburg O., 1956). Otto Warburg demonstrated that unlike normal differentiated cells, which rely on mitochondrial oxidative phosphorylation to generate energy needed for cellular processes, most cancer cells instead rely on aerobic glycolysis (Warburg O., 1956; Vander Heiden et al., 2009). Aerobic glycolysis allows cancer cells to convert nutrients such as glucose and glutamine more efficiently than normal cells, thus having the resources to produce the building blocks of new cells and promote anabolism and proliferation (Liberti et al., 2016).

A better understanding of cancer emerged after Watson and Crick published the molecular structure of nucleic acids (Watson & Crick, 1953). This work, which includes significant findings from Rosalin Franklin, revolutionized biology and medicine, by unlocking a better understanding of the molecular mechanisms of cells. In addition, incredible progress has been made in cancer research since the publication of the first draft of the human genome (International Human Genome Sequencing Consortium., et al. 2001; Venter, et al. 2001; International Human Genome Sequencing Consortium, 2004; Hood et al., 2004). The sequencing

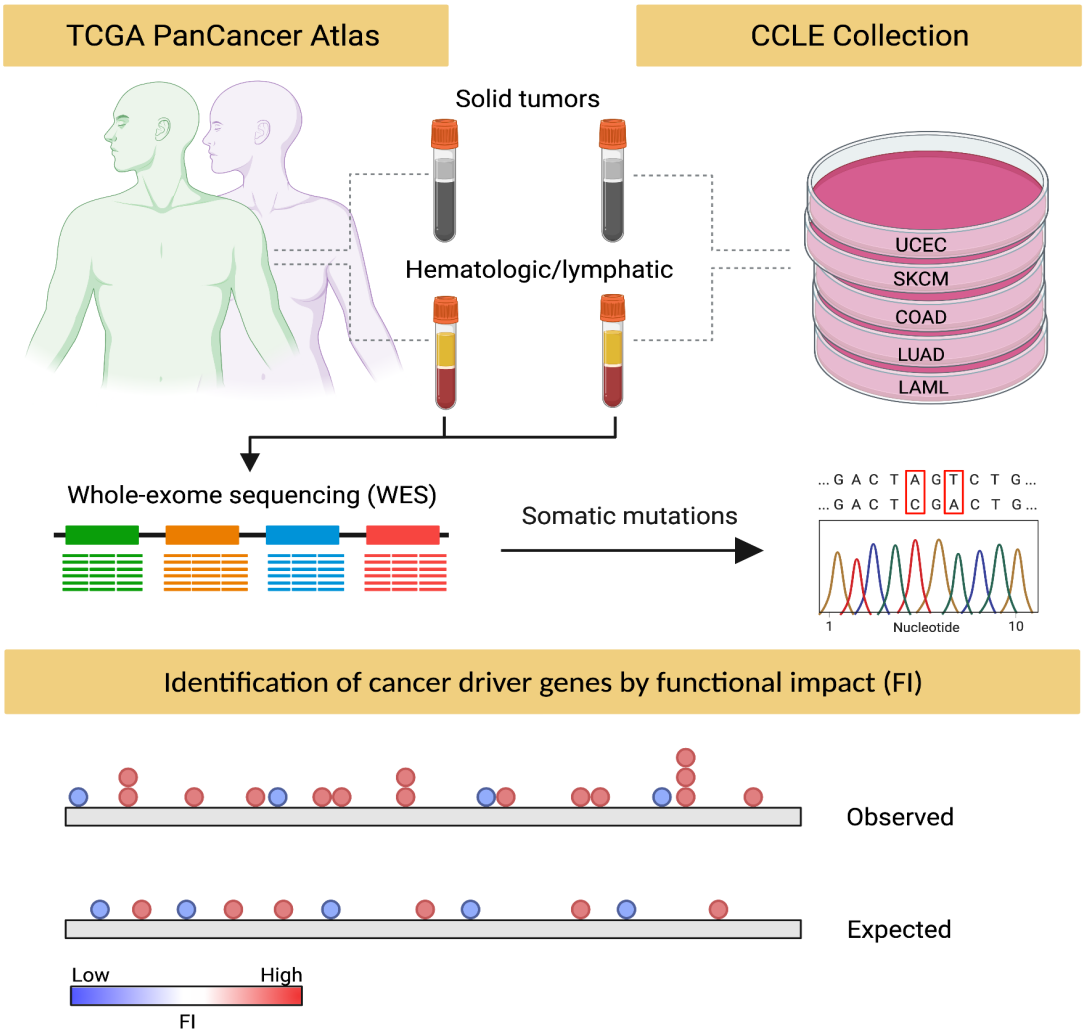
of the human genome jumpstarted the field of cancer genomics and allowed the study of cancer types at the single-nucleotide level (Gibbs et al., 2020). These new discoveries and advances led to the development of new computational tools and multi-institutional efforts to characterize cancer genomes at large scale.

Over the past two decades, The Cancer Genome Atlas (TCGA) project (Cancer Genome Atlas Research Network et al., 2013), a joint effort by the National Institute of Health (NIH) and National Human Genome Research Institute (NHGRI) to transform cancer research and discovery through technological advances, DNA sequencing, and data sharing, provided significant insights into the quest for the eradication of cancer. Numerous contributions have been made through the TCGA project, from better understanding tumorigenesis and cancer alterations across the whole genome to developing better technologies to sequence DNA and analyze data (Hoadley et al., 2018; Ding et al., 2018; Saltz et al., 2018; Malta et al., 2018; Thorsson et al., 2018). The TCGA PanCancer Atlas project has characterized thousands of novel somatic mutations, structural variations, altered biological pathways, and epigenetic changes across tumor types, therefore providing the most comprehensive catalog on cancer to date (The Cancer Genome Atlas Research Network., Genome Characterization Center, Chang, K. et al., 2013; Ding et al., 2018; Hoadley et al., 2018; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium., Campbell, P.J. et al., 2020).

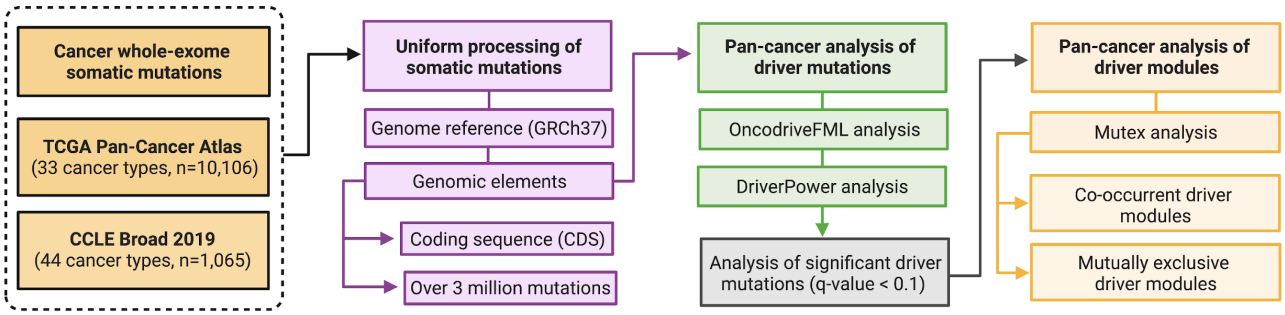
Figure 1. Functional impact of mutations reveals driver genes. A. Overview of TCGA PanCancer Atlas and CCLE sample collections and processing. Different cancer types were sequenced at TCGA consortium members and contributing centers (Hoadley et al., 2018). CCLE cell lines were sequenced at both the Broad Institute and Sanger Institute (Ghandi *et al., 2019*).

B. Pan-cancer analysis workflow to characterize a whole cancer exome landscape of driver mutations across cancer types.

A.



B.



The need to characterize genetic variations, candidate targets, and therapeutics, and to identify novel marker-driven cancer dependencies gave birth to the Cancer Cell Line Encyclopedia (CCLE) project (Ghandi et al., 2019). More than a thousand cancer cell lines have been studied and characterized in the CCLE providing a framework in which to study genetic variants and therapeutics for human cancers. To improve the understanding of molecular features that contribute to cancer phenotypes, the CCLE expanded the characterizations of cancer cell lines to include genetic array data for 1749 cell lines from individuals of various lineages and ethnicities (Ghandi *et al.,* 2019) and , provided the most comprehensive catalog of human cancer models. The TCGA and CCLE projects are two of the most important scientific advances for cancer research. However, the analysis of cancer genomes at large scale continues to provide significant information with clinical and therapeutic implications.

Recent pan-cancer studies of whole genomes mapped the pattern in somatic structural variations, RNA alterations, non-coding somatic mutations, chromosomal rearrangements, and evolution of thousands of cancers (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020; Li et al., 2020; PCAWG Transcriptome Core Group et al., 2020; Rheinbay et al., 2020; Cortés-Ciriano et al., 2020; Gerstung et al., 2020). These studies have provided significant insights into improving existing clinical treatments, novel early detection advances, and has further driven the quest for a universal cure for these malignancies (Li et al., 2014). Despite these remarkable accomplishments in cancer research and *omics*, the driver mechanisms and modules that drive tumor malignancy, progression, recurrence, and help tumors achieve drug resistance after treatments are not completely characterized nor understood due to limitations of tumor samples and data.

Nature of Cancer

Cancer occurs when an accumulation of somatic mutations and other genetic alterations cause cells to multiply out of control and form a tumor that might eventually metastasize. Metastatic tumors are termed malignant due to the fact that they have the ability to spread through the blood or lymph system, and form new tumors in other organs or tissues throughout the body. Malignant tumors are characterized by somatic mutations, non-heritable mutations that affect the overall survival of cancer patients. For example, in a recent study in patients with resected stage I lung adenocarcinoma (LUAD), *KRAS* mutation was an independent prognostic factor for overall survival and recurrence (Kadota et al., 2016). Kadota et al. demonstrated that the five years overall survival in patients with *KRAS* mutant tumors (n = 124, 63%) was significantly worse than those with *KRAS* wild-type (n = 339; 77%; p-value < 0.001), and in solid predominant tumors, *KRAS* mutations correlated with worse overall survival (p-value = 0.008) and increased recurrence (p-value = 0.005).

The Drivers of Cancer

Cancer genomes contain two kinds of mutations: passenger mutations and driver mutations. Passenger mutations arise randomly from sequences that do not seem to contribute directly to cancer progression or cancer malignancy (Wodarz et al., 2018). However, mutations that contribute to tumor development and metastasis, such as *KRAS* mutant, are known as *drivers.* Despite the fact that *drivers* are central to understanding tumor malignancy, it is not well understood how driver alterations occur and systematically contribute to the hallmarks of cancer. Cancer driver genes can be functionally classified as tumor suppressor genes (TSGs) or oncogenes (OGs) based on their role in tumorigenesis (Waks et al., 2016). In normal form TSGs

suppress and control cellular growth, however when mutated, they are inactivated and lead to cancer (Krug et al. 2002). OGs, which seem to be antithetical to TSGs, make normal cells grow out of control and become tumors via mutations that activate *proto-oncogenes,* normal genes that code for proteins that regulate cellular grow and proliferation and when mutated become oncogenes (Shen et al., 2018; Waks et al., 2016; Cline MJ. 1987). Studies show that both the loss of function in TSGs and gain of function in OGs are essential for tumorigenesis and tumor progression (Shen et al., 2018; Krug et al. 2002). However, a complete landscape of tumor suppressor and oncogenic drivers and how they systematically drive tumors is not yet characterized.

The Mysterious Roles of Driver Associations and Modules

When considering pairs of driver mutations within a given tumor type, two basic patterns emerge: co-occurrence or mutual exclusivity (Campbell PJ., 2017). Knowledge of signaling pathways and/or protein alterations networks is required to understand co-occurrent and mutually exclusive driver associations (Zhang et al., 2014). *Co-occurrent driver associations*, when two driver alterations are significantly observed in the same tumor or tumor stage, work together to benefit the hallmarks of cancer. In co-occurrent driver events one mutation ameliorates deleterious consequences of the other, therefore, increasing cell proliferation and survival (Campbell PJ., 2017). For example, *MYC* amplifications have been found to have significantly increased *MYC* expression in the presence of *TP53* mutations across tumors (Ulz et al. 2016). *MYC* encodes a nuclear phosphoprotein that plays a role in cell cycle progression, apoptosis, and cell proliferation, differentiation and survival (Chen et al. 2018). *TP53* encodes the p53 protein which plays essential functions in cellular response to diverse stresses and maintenance of genomic integrity (Aubrey et al. 2016). As co-occurrent driver genes, *MYC*-associated tumors

often require disruption of the apoptosis pathway to promote cell proliferation, thus inducing *TP53*-dependent apoptosis (Ulz et al. 2016; Hermeking et al. 1994).

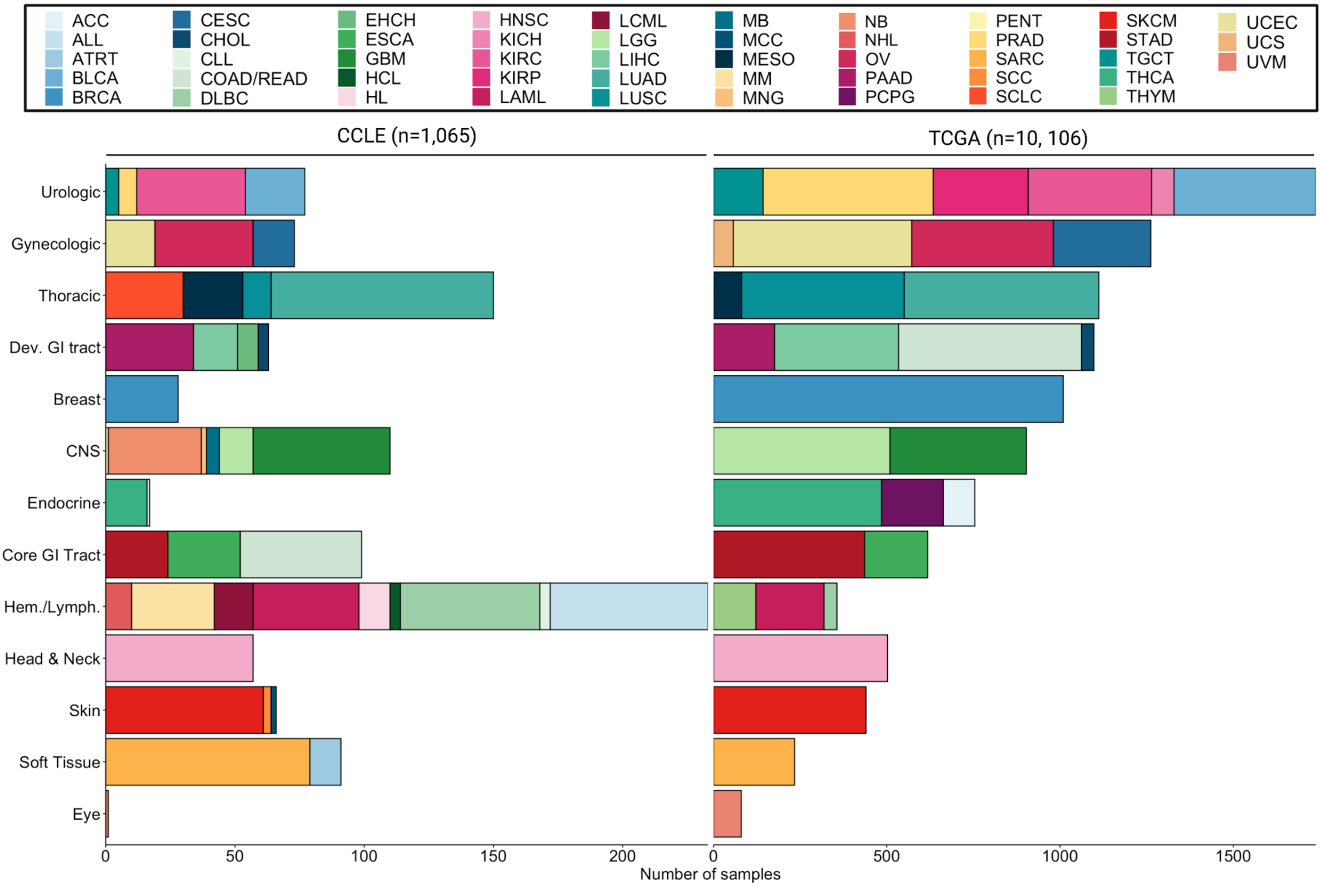
Mutually exclusive driver associations, two driver mutations that are rarely observed in the same tumor or tumor stage, usually involve driver alterations in the same signaling pathway leading to functional redundancy (Campbell PJ., 2017). For example, one potential explanation for oncogenic *EGFR* and *KRAS* mutual exclusivity is that the two activating mutations are functionally redundant, per se, their coexistence does not benefit the evolutionary process of the cancer cell or clone (Cisowski et al., 2017). Another explanation for mutually exclusive driver events is synthetic lethality, which means, having two driver alterations in the same pathway is toxic to the cancer cell and leads to eventual death (Campbell PJ., 2017).

Driver modules consist of many of these co-occurrent and mutually exclusive driver associations, and these associations are the grammar for the hallmarks of cancer. Driver modules include genetic alterations that confer fitness advantage to cancer cells such as single-nucleotide variants (SNVs), copy number alterations (CNAs), changes in the transcriptional activity of genes, and changes in protein concentration (Silverbush et al., 2019). In order to develop more efficient treatments for malignant and difficult to treat tumors, we need to better understand the networks or modules that characterize tumor malignancy and how they change after medical interventions or treatments (Vogelstein et al., 2013; Mateo et al., 2020). To do so, we need to characterize a complete landscape of driver mutations and modules across cancer genomes.

Figure 2. Landscape of 49 cancer types from TCGA Pan-Cancer Atlas and CCLE. A. 49 cancer types analyzed from TCGA Pan-Cancer Atlas and CCLE across 13 anatomical cohorts. TGCA shows the number of samples and CCLE the number of cancer cell lines from all WES

data included in analysis. Note: COAD and READ were combined in both TCGA and CCLE analyses, however, they are counted as two separate cancer types.

A.



From a Pan-cancer Analysis to a Landscape of Driver Mutations

A recent study led by TCGA Research Network characterized the most comprehensive cancer genome landscape of human tumors to date from 2,658 whole-cancer genomes and their matching normal tissues across 38 tumor types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and TCGA (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020). This study showed that on average, cancer genomes contain 4–5 driver mutations when

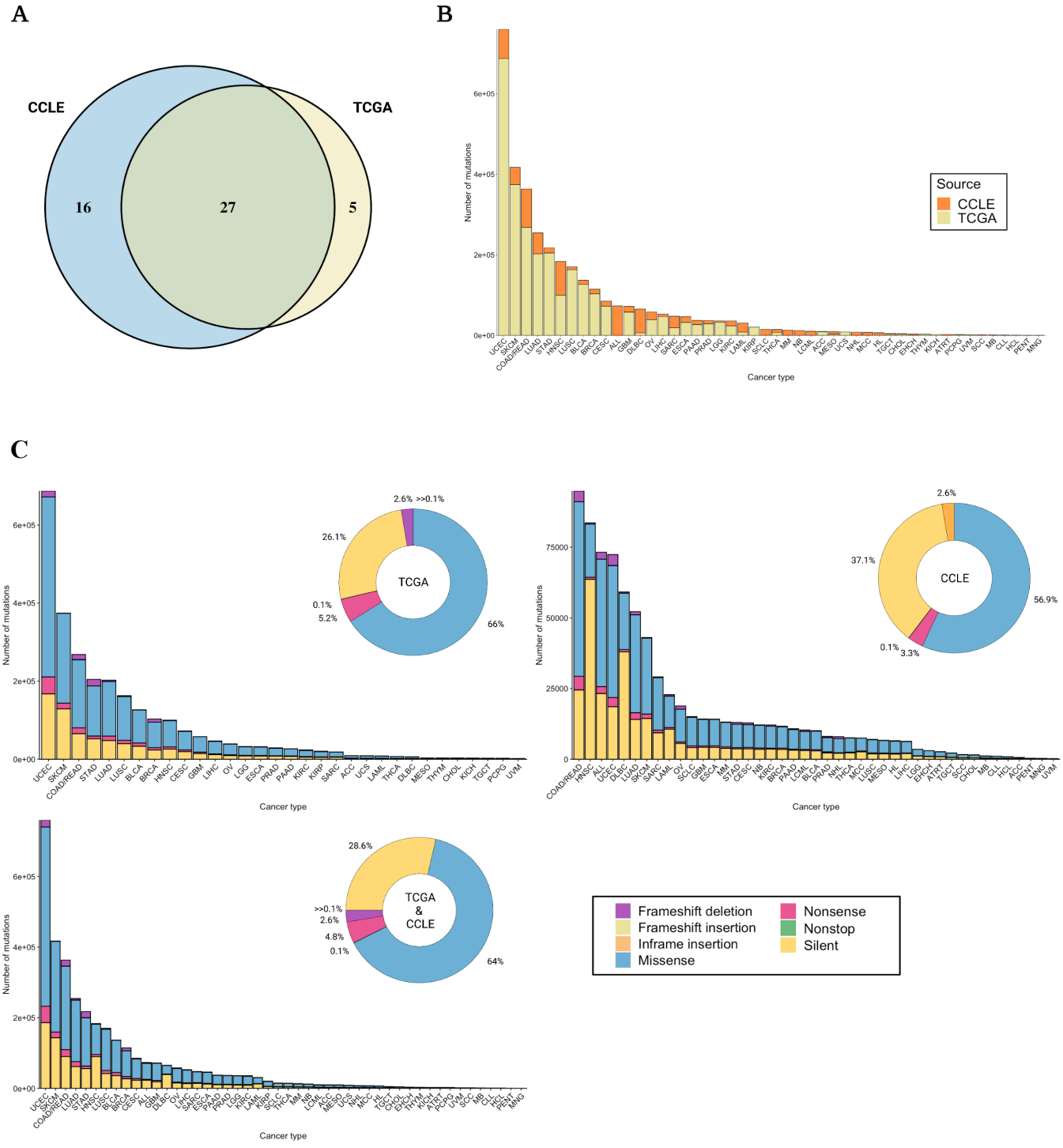
combining coding and non-coding genomic elements; however, in around 5% of the cases no driver mutations were identified, demonstrating that cancer driver discovery is not yet complete (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020).

A complete landscape of driver mutations across cancer genomes might unravel the systemic and mechanistic events that characterize tumorigenesis, tumor malignancy, and mutational commonalities across tumor types. Importantly, it will serve as an asset for future research and clinical endeavors. Pathways involved in tumorigenesis are complex and interconnected, and it is difficult to define the borders of a signaling pathway, and the notion of parallel or common pathways (Remy et al., 2015). However, a complete characterization of the mutational landscape of driver alterations across cancer types might elucidate altered processes and pathways involved in cancer formation, metastasis and drug resistance.

Here, we took on the challenge to develop a comprehensive analysis of driver mutations and modules from 10,106 tumor samples and 1,065 cancer cell lines across 49 cancer types from TCGA PanCancer Atlas studies (Hoadley et al., 2018) and CCLE (Ghandi *et al., 2019*) (Fig. 1, Fig. 2). We aim to build a driver mutations landscape of cancer exomes across tumor types using *OncodriveFML* (Mularoni et al., 2016) and *DriverPower* (Babur et al., 2015) to identify novel driver alterations that might characterize the hallmarks of malignant tumors. In addition, we aim to develop a comprehensive analysis of cancer driver modules to characterize the systemic networks that write the hallmarks of tumors.

Figure 3. Mutation count and effect of 49 cancer types from TCGA Pan-Cancer Atlas and CCLE. A. Venn Diagram showing the number of mutual cancer types in TCGA and CCLE. **B.** Landscape of processed somatic mutations in each cancer type in TCGA Pan-Cancer Atlas and CCLE. **C.** Fractions of variant effects across tumors in TCGA Pan-Cancer Atlas and CCLE

compendiums. Seven effects were mutual across somatic mutation variants in TCGA and CCLE from which missense and silent mutations predominate. Donut plots showing the percentage of mutational effects included for TCGA, CCLE and both TCGA and CCLE combined.



METHODS

Ethical Review

Human tumor tissue, adjacent normal tissue, and normal whole blood samples were de-identified from original samples in this study. All samples were obtained and sequenced by the TCGA consortium members and contributing centers with informed consent from all patients according to the local Institutional Review Boards (IRB) protocols as described by Hoadley et al. (Hoadley et al., 2018). Biospecimens were centrally processed and distributed to TCGA analysis centers. TCGA Project Management collected necessary human subjects documentation to ensure the project complies with 45-CFR-46 (the "Common Rule"). In addition, the program obtained documentation from every contributing clinical site to verify that IRB approval was obtained to participate in TCGA.

TCGA WES

WES somatic mutation calls for 10,124 donors were obtained from TCGA Pan-Cancer Atlas (https://gdc.cancer.gov/about-data/publications/pancanatlas). Somatic mutation variants for 33 cancer types were downloaded from cBioPortal (https://www.cbioportal.org/datasets, human genome assembly GRCh37 (hg19) and sequencing method Illumina HiSeq). WES TCGA Pan-Cancer Atlas variant calls can also be found at the National Cancer Institute Genomic Data Commons (GDC) (https://gdc.cancer.gov/about-data/publications/pancanatlas), University of California, Santa Cruz (UCSC) Xena Browser (https://xenabrowser.net/datapages/), and Broad Institute GDAC Firehose (https://gdac.broadinstitute.org/). Notice that data from the cBioPortal and GCD were aligned to human genome assembly GRCh37, while data from UCSC Xena

Browser in GRCh38. In this study, all data sets were aligned to the human genome assembly GRCh37.

CCLE WES

WES for 326 cell lines was performed at the Broad Institute Genomics Platform as described by Ghandi et al. (Ghandi *et al., 2019)*. Libraries were constructed and sequenced on either Illumina HiSeq 200 or Illumina GAIIX, with 76-bp paired-end reads. Outputs from Illumina software were processed by the Picard data-processing pipeline to yield BAM files containing well-calibrated and aligned reads. WES variant calls for 1,072 cell lines were based from the Sanger Institute, indicated as "Sanger WES" and were constructed and sequenced on Illumina HiSeq 2000 (COSMIC: http://cancer.sanger.ac.uk/cell_lines, EGA accession number: EGAD00001001039). Sanger WES sequence data was reprocessed using CCLE pipelines and merged into a single data set which includes WGS data for 329 cell lines (CCLE: https://portals.broadinstitute.org/ccle/data, File: CCLE_DepMap_18q3_maf_20180718.txt). CCLE mutation calls data can also be found in the cBioPortal, UCSC Xena Browser, and the Broad Institute Cancer Dependency Map (DepMap) project (https://depmap.org/portal/download/).

TCGA PanCancer Atlas Somatic Variant Calling

TCGA whole exome somatic mutation calling was performed using MutSigCV (version 1.4) to determine significantly mutated genes (Lawrence et al., 2013) and GISTIC2.0 to identify recurrent deletion and amplification peaks (Mermel et al., 2011) as described by the Zheng et al. (Zheng et al., 2016). More details for each TCGA PanCancer Atlas data can be found at NIH

website by selecting each individual study (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/studied-cancers).

CCLE Somatic Variant Calling

A variant calling pipeline was designed to process all sequencing data generated by the CCLE project as described by Ghandi et al. (Ghandi et al., 2019). Mutation analysis for SNVs was performed using MuTect v1.1.6 (Huang et al. 2015) in single sample mode with default parameters (Ghandi et al., 2019). In addition, short indels were detected using Indelocator (http://archive.broadinstitute.org/cancer/cga/indelocator) in single sample mode also with the default parameters. To ensure high-quality variant calls, it was required a minimum coverage of 4 reads with a minimum of two reads supporting the alternate allele. Variants with an allelic fraction below 0.1 and variants outside the protein CDS were excluded. In order to remove germline-like variants, any variant with a normal allelic frequency greater than 10^-5 as described by the Exome Aggregation Consortium (ExAC) project (Lek et al. 2016) was removed with exception of any cancer-recurrent variants with minimum frequency of 3 in TCGA or a frequency of 10 in COSMIC (Lek et al. 2016).

Processing of TCGA Somatic Mutation Calls

All pre-processed TCGA Pan-Cancer Atlas whole exome somatic mutation calls for 10,124 samples across 33 cancer types were merged into a single data set for uniform processing and analysis. In addition, all biased variants, multiple nucleotide variants (MNVs), introns variants, and any other variants not associated with protein coding sequences such as splice sites, 5' untranslated region (UTR), and 3' UTR variants were removed from the data set. After

applying these metrics, only 10,106 tumor samples were included in the eventual analysis making a mutational landscape of 2,685,974 SNVs across 33 cancer types. This collection includes a sample distribution as follow: ACC (n=91) (Zheng et al., 2016), BLCA (n=409) (Hoadley et al., 2018), BRCA (n=1009) (The Cancer Genome Atlas Network et al., 2012; Ciriello at el., 2015), COAD/READ (n=528) (The Cancer Genome Atlas Network et al., 2012), CHOL (n=36) (Farshidfar et al., 2017), CESC (n=281) (The Cancer Genome Atlas Research Network et al., 2017), DLBC (n=37) (Hoadley et al., 2018), ESCA (n=182) (The Cancer Genome Atlas Network, Analysis Working Group: Asan University et al., 2017), GBM (n=394) (Brennan et al., 2013), HNSC (n=502) (The Cancer Genome Atlas Network., Genome sequencing centre: Broad Institute et al., 2015), KIRC (n=356) (The Cancer Genome Atlas Research Network., Analysis working group: Baylor College of Medicine et al., 2013), KICH (n=65) (Davis et al., 2014), KIPR (n=274) (Cancer Genome Atlas Research Network, Linehan et al., 2016), LGG (n=509) (Cancer Genome Atlas Research N, Brat et al., 2015), LAML (n=197) (Cancer Genome Atlas Research Network, Ley TJ, et al., 2013), LUAD (n=562) (The Cancer Genome Atlas Research Network., Disease analysis working group., Collison et al., 2014), LUSC (n=469) (The Cancer Genome Atlas Research Network., Genome sequencing centres: Broad Institute. Et al., 2012; Campbell et al., 2012), LIHC (n=358) (Cancer Genome Atlas Research Network 2017), MESO (n=81) (Hmeljak et al., 2018), OV (n=409) (The Cancer Genome Atlas Research Network et al., 2011), PAAD (n=176) (Cancer Genome Atlas Research Network et al., 2017), PCPG (n=178) (Fishbein et al., 2017), PRAD (n=491) (Cancer Genome Atlas Research Network 2015), STAD (n=436), SARC (n=234) (Cancer Genome Atlas Research Network 2017), SKCM (n=440) (Cancer Genome Atlas Research Network 2015), THCA (n=485) (Cancer Genome Atlas Research Network 2014), TGCT (n=143) (Shen H. et al., 2018),

THYM (n=122) (Radovich et al., 2018), UCEC (n=515) (Levine et al., 2013), UCS (n=57) (Cherniack et al., 2017), and UVM (n=80) (Robertson et al., 2017) (Supplementary Table 1). From the TCGA compendium of 33 cancer types, 28 cancer types were mutual with CCLE and only 5 cancer types were unique to TCGA including KIRP, KICH, UCS, THYM, and PCPG. Unique cancer types were not combined with CCLE for cancer specific analyses; however, they were included in the pan-cancer analyses. Complete list of cancer type abbreviations from TCGA (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations).

Processing of CCLE Somatic Mutation Calls

All cancer cell lines were filtered for whole exome somatic mutation variants. The mutual cancer types with TCGA were annotated by TCGA cancer codes using the CCLE cell line annotations (https://portals.broadinstitute.org/ccle/data. File: Cell_lines_annotations_20181226.txt). From 1,461 annotated cell lines, we removed 329 cell lines containing WGS somatic mutations and 67 cell lines which had no annotation and were unable to be annotated. A group of 109 cell lines annotated as "CUP" was re-annotated using DepMap (https://depmap.org/portal/). After applying these metrics, only 1,065 cancer cell lines containing whole exome somatic mutation variants across 43 cancer types were included in eventual analyses. The 28 cancer types with cell lines mutual to TCGA cancer types include BRCA (n=28), BLCA (n=23), LAML (n=41), DLBC (n=54), OV (n=38), UCEC (n=19), CESC (n=16), KIRC (n=42), THCA (n=16), ESCA (n=28), STAD (n=24), COAD/READ (n=47), LIHC (n=17), PAAD (n=34), HNSC (n=57), LUAD (n=86), MESO (n=23), GBM (n=53), LGG (n=13), SARC (n=79), SKCM (n=61), LUSC (n=11), PRAD (n=7), ACC (n=1), TGCT (n=5), CHOL (n=4), and UVM (n=1) (Supplementary Table 1). Cancer types unique to CCLE include CLL (n=4), MB (n=5), NB (n=36), SCLC (n=30), ALL (n=61), LCML (n=15), MM (n=32), HL

(n=12), MNG (n=2), SCC (n=3), HCL (n=4), ATRT (n=12), NHL (n=10), MCC (n=2), and PENT (n=1) (Supplementary Table 1).

Whole Exome Somatic Mutations Analysis

SNVs in both TCGA PanCancer Atlas and CCLE somatic mutation calls were filtered to perform a unified pan-cancer analysis of somatic mutations in both compendiums and a series of cancer specific analyses. These analyses include somatic mutation variants that are specifically found in protein coding regions or CDS. The mutational effects included in analysis were missense, nonstop, nonsense, silent, inframe insertions, frameshift deletions, and frameshift insertions SNVs, all of which were mutual in both collections. The pan-cancer analysis of whole exome somatic mutations includes all tumor types from TCGA PanCancer Atlas (n=33) and CCLE (n=43) collection. We combined both TCGA PanCancer Atlas and CCLE somatic mutation landscapes into a single data set and used *OncodriveFML* to identify potential driver alterations across tumors. In addition, pan-cancer analyses were run specifically for TCGA and CCLE, respectively. Cancer specific analyses were run for TCGA and CCLE combined for the 28 mutual cancer types. In addition, cancer specific analyses were run for cancer types unique to TCGA (n=5) and CCLE (n=16).

Driver Calling Pipelines

OncodriveFML

O*ncodriveFML* analyzes the pattern of somatic mutations to identify signals of positive selection in both coding and non-coding genomic regions, thus predicting genes that have greater accumulation of mutations by higher functional impact (Mularoni et al., 2016). The predicted

functional impacts (FI) of mutations were scored using the CADD framework version 1.0 (Kircher et al., 2014) as described by Mularoni et al. for HG19 (GRCh37) reference genome. This framework provides a score for every possible nucleotide substitution in the human genome and can consequently be applied to coding and non-coding genomic elements. CADD scores can be downloaded in the Combined Annotation Dependent Depletion (CADD) website (https://cadd.gs.washington.edu/download). The mean of randomized FI scores for mutations was compared to permuted mutations within the same gene to calculate an empirical p-value. The resulting p-values are then adjusted using Benjamini-Hochberg multiple testing correction procedure. The results were calculated using CDS region coordinates from Gencode release 19 (HG19). HG19 CDS regions were downloaded from the Barcelona Biomedical Genomics Lab (BBGLAB) website (https://bitbucket.org/bbglab/oncodrivefml/downloads/). In addition, HG19 CDS regions can be obtained from the Encode website (http://www.gencodegenes.org/). As described by Mularoni et al., CDS regions were only considered for genes where both "gene_type" and "transcript_type" metadata were annotated as "protein_coding." When using HG19 from Encode to obtain CDS regions, it is required to remove any non-coding gene coordinates.

DriverPower

We aim to use *DriverPower* as an orthogonal method to validate potential novel driver genes predicted by *OncodriveFML*. *DriverPowe*r is a framework for identification of coding and non-coding driver mutations using mutational burden (MB) and FI scores. The current implementation of *DriverPower* can measure FI using any of four functional scoring schemes including CADD, DANN, EIGEN and LINSIGHT scores (Shuai et al., 2020). We aim to use

CADD scores and both somatic copy number and somatic mutation calls from TCGA and CCLE to run the analyses. Somatic copy number data will be downloaded from the cBioPortal for TCGA (https://www.cbioportal.org/datasets) and CCLE server (https://portals.broadinstitute.org/ccle/data) for the CCLE. We aim to apply the same metrics and analysis procedures as for *OncodriveFML*.

Mutex

To identify and analyze potential driver pathways/modules, we aim to use *Mutex*, a method that identifies potential driver pathways based on mutual exclusivity of alterations (Babur et al., 2015). *Mutex* scans groups of genes with a common downstream effect on a signaling network using a mutually exclusive test in which each gene in the group significantly contributes to the mutual exclusivity pattern as described by Babur et al. (Babur et al., 2015). We aim to use mutation and copy number profiles for TCGA PanCancer Atlas studies (https://www.cbioportal.org/datasets, Ref: TCGA, Cell 2018) and CCLE cancer cell lines (https://portals.broadinstitute.org/ccle/data) to identify potential mutually exclusion of mutations in these datasets and thus, driver associations and modules.

COSMIC CGC

The CGC is a catalog of driver genes whose mutations have been causally implicated in cancer and curated using tumor samples and cancer cell lines (Sondka et al., 2018). CGC describes in detail the effect of evidence-based, manually curated summaries of 719 cancer driver genes (version 86, August 2018) (https://cancer.sanger.ac.uk/census). The CGC

information was used to identify and compare annotated versus unannotated significant driver genes predicted by OncodriveFML analyses.

Lollipop/Mutational Needle Plots

Lollipop (Mutational needle) plots were produced using cBioPortal mutation mapper (Cerami et al., 2012; Gao et al., 2013). Data was formatted for specific driver genes predicted by OncodriveFML and uploaded to cBioPortal server (https://www.cbioportal.org/mutation_mapper). The GRCh37 reference genome was selected for the visualizations.

Statistical Analysis and Code Availability

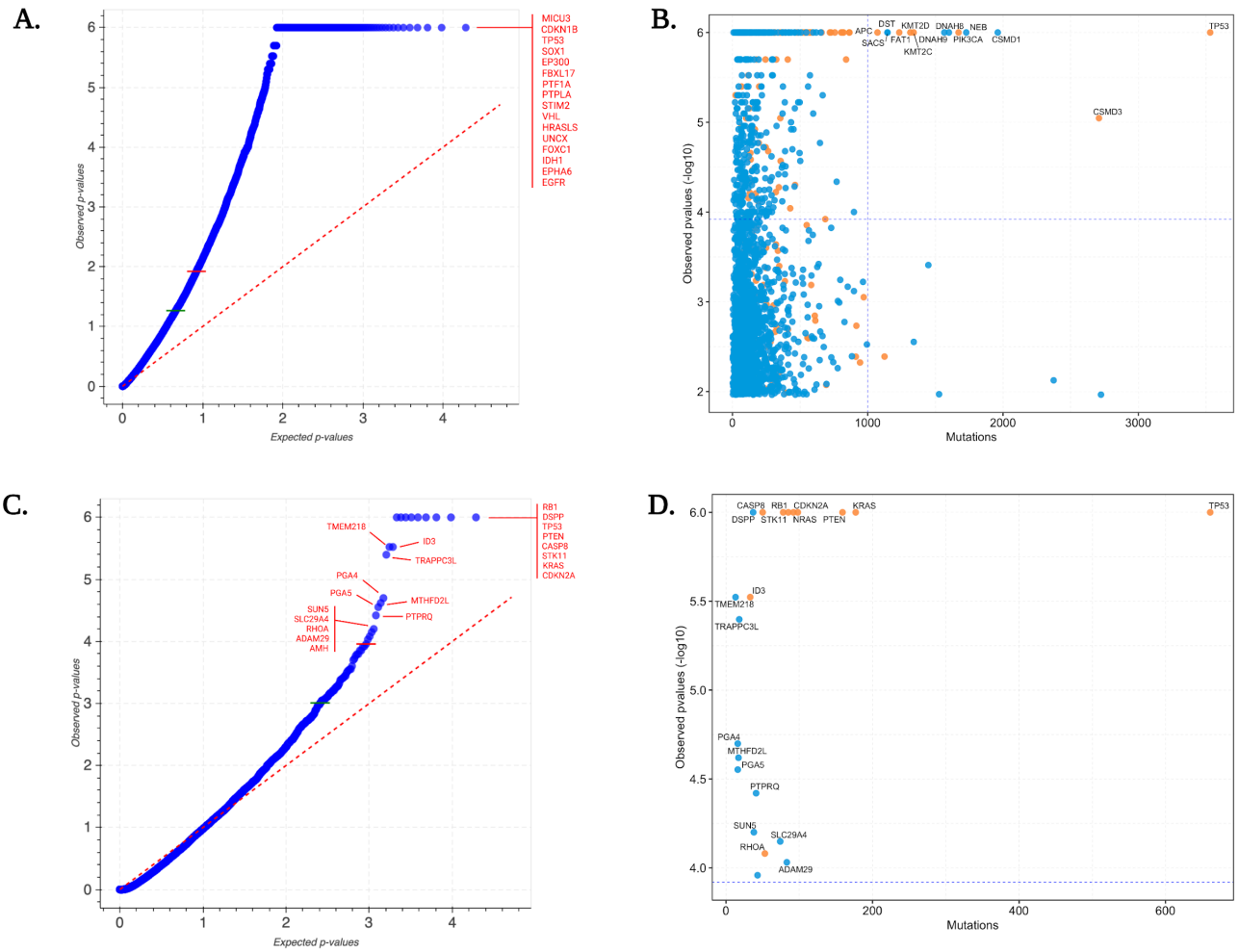
All data processing, data integration and analysis were performed in R/Rstudio. Statistical tests respective to *OncodriveFML* are noted above in respective sections. All data processing and statistical analyses code are available at https://github.com/caeareva/pcadmm/ and upon request from the author (caeareva@ucsc.edu).

Data Availability

The processed data for TCGA can be found at the Memorial Sloan Kettering Cancer Center cBioPortal (https://www.cbioportal.org/datasets, Ref: TCGA, Cell 2018) and CCLE at the Broad Institute CCLE server (https://portals.broadinstitute.org/ccle/data). In addition, TCGA mutations and processed data can be downloaded from the UCSC Xena Browser (https://xenabrowser.net/datapages/) by selecting cancer types that start with the term "GDC TCGA", NIH GDC (https://gdc.cancer.gov/about-data/publications/mc3-2017), and the Broad Institute FireBrowse Portal (https://gdac.broadinstitute.org/). CCLE data can also be downloaded

from the UCSC Xena Browser by selecting for the "Cancer Cell Line Encyclopedia (CCLE)" term and DepMap (https://depmap.org/portal/download/) by selecting CCLE_mutations.csv. Mutations data from the UCSC Xena Browser were annotated using HG38, however, they can be lifted from HG38 to HG19 using the UCSC Genome Browser liftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

Figure 4. TCGA and CCLE pan-cancer analyses of protein-coding genomic elements. A. QQ plot comparing the expected and observed distribution of functional mutational bias p-values of genes in TCGA pan-cancer analysis. Note: not all significant genes (red) are labeled in figure. **B.** Comparison of mutation count versus significance of TCGA driver genes identified by *OncodriveFML* in pan-cancer analysis. All genes compared had a q-value < 0.1. **C.** QQ plot comparing the expected and observed distribution of functional mutational bias p-values of genes in CCLE pan-cancer analysis. **D.** Comparison of mutation count versus significance of CCLE driver genes identified by *OncodriveFML* in pan-cancer analysis. All genes compared had a q-value < 0.1.



RESULTS

Specimens and Tumor Types

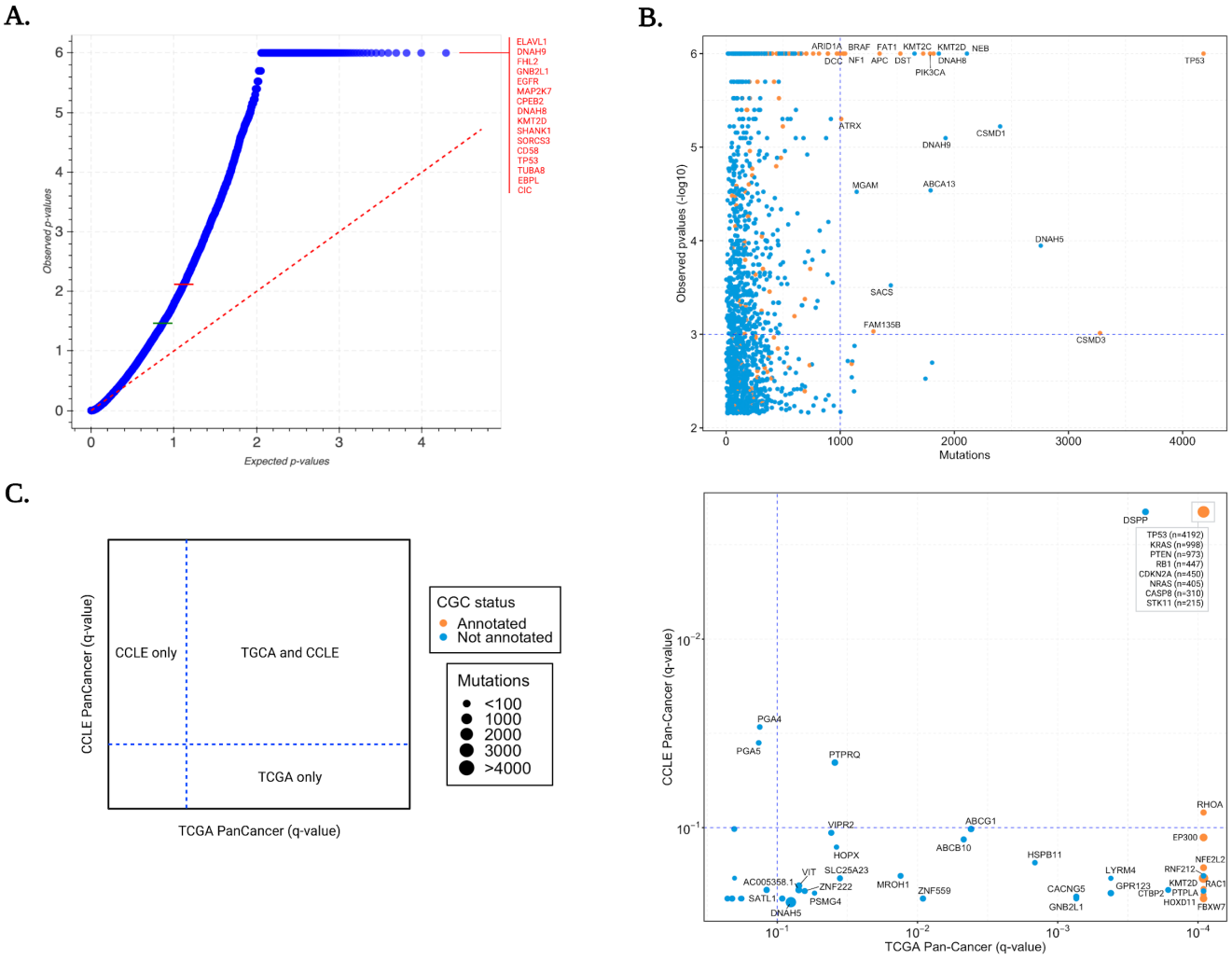
We evaluated all samples in the TCGA PanCancer Atlas collection (Hoadley et al., 2018) and CCLE (Ghandi *et al., 2019*) which eventually included whole-exome sequencing (WES) data for 10, 110 samples across 33 cancer types and 1,065 cancer cell lines across 43 cancer types (Fig 2A). From these compendiums, 28 tumor types were mutual in TCGA PanCancer Atlas and CCLE, 5 cancer types were unique to TCGA PanCancer Atlas, and 16 cancer types

were unique to CCLE (Fig. 3A). A total of 49 unique cancer types were included in the pan-cancer analysis from both TCGA and CCLE collections, counting COAD/READ as two separate cancer types.

To perform a unified analysis of driver alterations, the 49 unique cancer types were further stratified into 13 anatomical cohorts (Fig. 2A). Hematologic and lymphatic malignancies include acute myeloid leukemia (LAML), lymphoid neoplasm diffuse large B cell lymphoma (DLBC), thymoma (THYM), B-cell chronic lymphoblastic leukemia (CLL), T-cell acute lymphoblastic leukemia (ALL), chronic myelogenous leukemia (LCML), hairy cell leukemia (HCL), multiple myeloma [MM], Hodgkin's lymphoma [HL], and Non-Hodgkin's lymphoma [NHL]). Solid tumor types include gynecologic (ovarian [OV], uterine corpus endometrial carcinoma [UCEC], cervical squamous cell carcinoma and endo-cervical adenocarcinoma [CESC]), breast (breast invasive carcinoma [BRCA]), urologic (bladder urothelial carcinoma [BLCA], prostate adenocarcinoma [PRAD], testicular germ cell tumors [TGCT], kidney renal clear cell carcinoma [KIRC], kidney chromophobe [KICH], and kidney renal papillary cell carcinoma [KIRP]), endocrine (thyroid carcinoma [THCA] and adrenocortical carcinoma [ACC]), core gastrointestinal (esophagus carcinoma [ESCA], stomach carcinoma [STAD], colon adenocarcinoma [COAD], and rectum adenocarcinoma [READ]), developmental gastrointestinal (liver hepatocellular carcinoma [LIHC], pancreatic adenocarcinoma [PAAD], and cholangiocarcinoma [CHOL]), head and neck (head and neck squamous cell carcinoma [HNSC]), and thoracic (lung adenocarcinoma [LUAD], lung squamous cell carcinoma [LUSC], mesothelioma [MESO], and small cell lung cancer [SCLC]). Cancer types of the central nervous system (glioblastoma multiforme [GBM], brain lower-grade glioma [LGG], medulloblastoma [MB], neuroblastoma [NB], meningioma [MNG], and primitive neuroectodermal tumor [PENT])

and soft tissue (sarcoma [SARC], uterine carcinosarcoma [UCS], and atypical teratoid rhabdoid tumor [ATRT]). Cancer types from neural-crest-derived tissues (pheochromocytoma and paraganglia [PCPG]) and melanocytic tumors of the skin (skin cutaneous melanoma [SKCM], skin squamous cell carcinoma [SCC], and merkel cell carcinoma [MCC]) and eye (uveal melanoma [UVM]).

Figure 5. Pan-cancer analysis of TCGA and CCLE whole exome variants across 49 cancer types. A. Pan-cancer analysis results of TCGA-CCLE whole exome variants across 49 cancer types. Note: not all significant genes (red) are labeled in figure. **B.** Mutation distribution versus significance from pan-cancer analysis results. **C.** TCGA Pan-cancer versus CCLE Pan-cancer significance. **D.** Proportion of predicted significant (q-value < 0.1) driver genes annotated versus non-annotated in COSMIC CGC. **E.** Driver gene count of annotated and unannotated drivers in COSMIC CGC.


A.
B.
C.
D
E
Observed p-values
Expected p-values
ELAVL1
DNAH9
FHL2
GNB2L1
EGFR
MAP2K7
CPEB2
DNAH8
KMT2D
SHANK1
SORCS3
CD58
TP53
TUBA8
EBPL
CIC
Observed pvalues (-log10)
Mutations
ARID1A
BRAF
FAT1
KMT2C
KMT2D
NEB
DCC
NF1
APC
DST
DNAH8
PIK3CA
TP53
ATRX
CSMD1
DNAH9
MGAM
ABCA13
DNAH5
SACS
FAM135B
CSMD3
CCLE PanCancer (q-value)
TCGA PanCancer (q-value)
CCLE only
TGCA and CCLE
TCGA only
CGC status
Annotated
Not annotated
Mutations
<100
1000
2000
3000
>4000
CCLE Pan-Cancer (q-value)
TCGA Pan-Cancer (q-value)
TP53 (n=4192)
KRAS (n=998)
PTEN (n=973)
RB1 (n=447)
CDKN2A (n=450)
NRAS (n=405)
CASP8 (n=310)
STK11 (n=215)
DSPP
PGA4
PGA5
PTPRQ
VIPR2
ABCG1
ABCB10
HOPX
RHOA
EP300
HSPB11
NFE2L2
LYRM4
RNF212
GPR123
KMT2D
RAC1
CTBP2
PTPLA
HOXD11
FBXW7
CACNG5
GNB2L1
VIT
SLC25A23
AC005358.1
ZNF222
SATL1
PSMG4
DNAH5
MROH1
ZNF559
Cancer Type
Proportion of Predicted Driver Genes
CGC Status
Annotated
Unannotated
UCEC
COAD/READ
SKCM
LUAD
STAD
HSNC
LUSC
BLCA
BRCA
CESC
GBM
ALL
OV
LIHC
ESCA
SARC
DLBC
PAAD
KIRC
PRAD
LGG
LAML
KIRP
THCA
SCLC
MM
NB
LCML
MESO
UCS
ACC
NHL
MCC
HL
TGCT
CHOL
THYM
EHCH
ATRT
KICH
PCPG
UVM
SCC
CLL
MB
HCL
MNG
PENT
Number of predicted drivers
CGC Status
Condition
Annotated
Unannotated

Mutational Datasets

Mutational calls were produced by TCGA PanCancer Atlas contributing centers. We retrieved TCGA PanCancer Atlas WES somatic mutation calls for 10,124 tumor samples from cBioPortal (https://www.cbioportal.org/datasets). We removed all multiple nucleotide variants (MNVs), mutational bias, and non-coding genomic variants including splice sites, 5' untranslated regions (UTRs), and 3' UTRs that were included in the mutation data. A total of 10,106 samples were eventually included across 33 cancer types. The same metric was applied to 1570 cancer cell lines in the CCLE collection. CCLE mutation calls included whole-genome sequencing data for 329 cancer cells and 175 unclassified cell lines. After removing these groups, we ended with whole-exome somatic variants for 1065 cancer cells in our pan-cancer analysis.

The Pan-Cancer Analysis of Whole Cancer Exomes

The expansion of the CCLE to include sequencing data across cancer types represented an opportunity to undertake a pan-cancer analysis of driver mutation and modules in whole cancer exomes across tumor samples and cancer models. We implemented this analysis by combining both TCGA PanCancer Atlas and CCLE collections, and therefore increasing the number of samples per cancer type. Due to the low number of cancer cell lines including whole-genome sequencing data, we focus our endeavors in exome variants only.

We combined all processed somatic mutation variants from 49 cancer types from TCGA PanCancer Atlas and CCLE for uniform analysis. Urologic cancer types predominate in TCGA PanCancer anatomical cohorts with 1738 tumor samples from 6 cancer types including BLCA (n=409), PRAD (n=491), TGCT (n=143), KIRC (n=356), KICH (n=65), and KIRP (n=274). In the CCLE, the hematologic and lymphatic subtype predominates the sample size with somatic

variants for 229 cancer cell lines. Eye tissue was the cohort with the least number of samples and cell lines in both TCGA PanCancer Atlas (n=80) and CCLE (n=1) compendiums (Fig.2A).

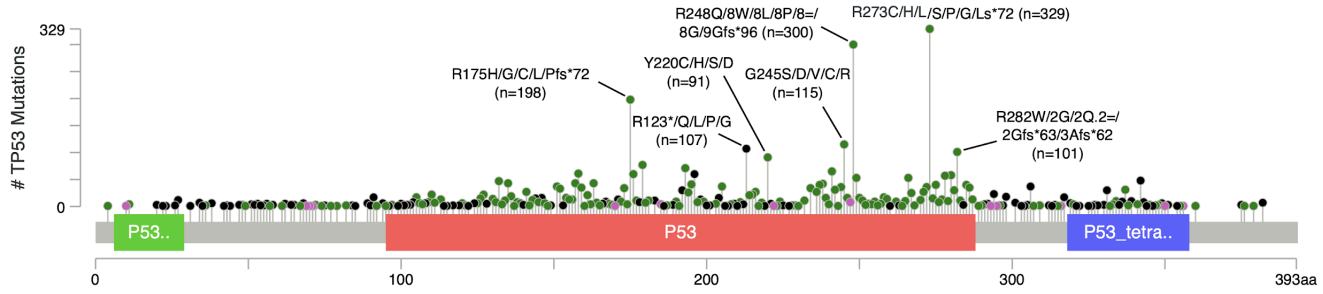
By merging TCGA PanCancer Atlas and CCLE, we build a landscape of more than 3 million (n=3,461,942) somatic mutation variants. This landscape includes mutations for all 49 cancer types across six mutational effects (Fig. 3B). The mutational landscape was made of 64% missense mutations (n=2,214,849), 28.6% silent mutations (n=989,533), 4.8% nonsense mutations (n=164,568), 2.6% SNV frameshift deletions (n=90,137), 0.1% nonstop mutations (n=2,558), and less than 0.1% for both frameshift insertion (n=302) and inframe insertion mutations (n=10) (Fig 3C).

By merging both TCGA and CCLE somatic variant calls by cancer type, we identified UCEC as the tumor type with the highest number of mutations followed by SKCM, COAD/READ, LUAD, STAD and HNSC (Fig. 3B). Unique tumor types to CCLE predominate the list of cancer types with the lowest number of mutations such as MNG, PENT, HCL, CLL, MB and SCC (Fig. 3B). This low mutation count might be associated with the low number of cell lines in those cancer types. Approximately 66% of all somatic variants in TCGA PanCancer Atlas and 56.9% of all somatic variants in CCLE collection had a missense effect (Fig. 3C). Approximately 26.1% of all mutations in TCGA PanCancer Atlas and 37.1% of all mutations in CCLE had a silent effect (Fig. 3C). Nonsense mutations formed 5.2% of all TCGA PanCancer Atlas somatic variants and 3.3% of all CCLE somatic variants (Fig. 3C). Frameshift deletions formed 2.6% of all TCGA PanCancer Atlas mutations and 2.6% of all CCLE mutations (Fig. 3C). Only 0.1% in both TCGA and CCLE collections were found to be nonstop mutations (Fig. 3C). The mutational effect landscapes of TCGA and CCLE show a high degree of similarity between TCGA tumor cancer genomes and CCLE cancer cell line genomes indicating an

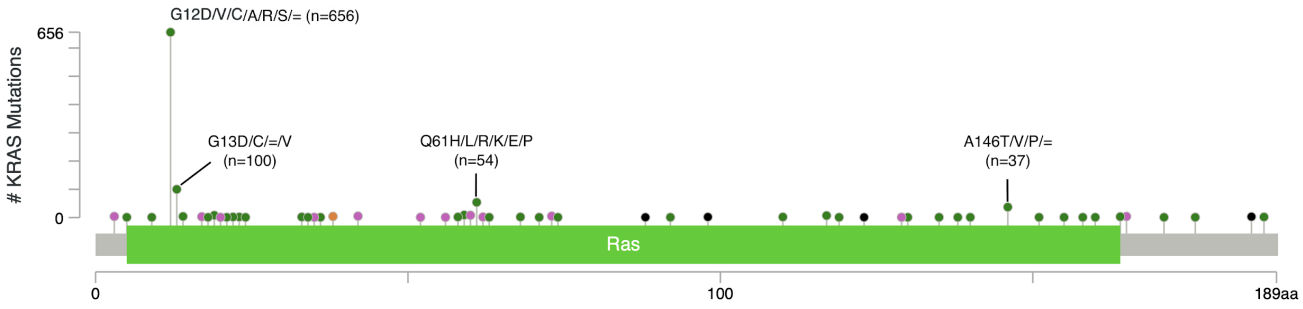
opportunity to undertake a TCGA PanCancer Atlas and CCLE combined pan-cancer analysis of driver mutations.

Figure 6. Lollipop/Mutational needle plots of *TP53*, *KRAS* and *PTEN*. A. Mutational needle plot showing the mutational landscape of *TP53* protein coding region across all cancer types. **B.** Mutational needle plot showing the mutational landscape of *KRAS* protein coding region across all cancer types **C.** Mutational needle plot showing the mutational landscape of *PTEN* protein coding region across all cancer types

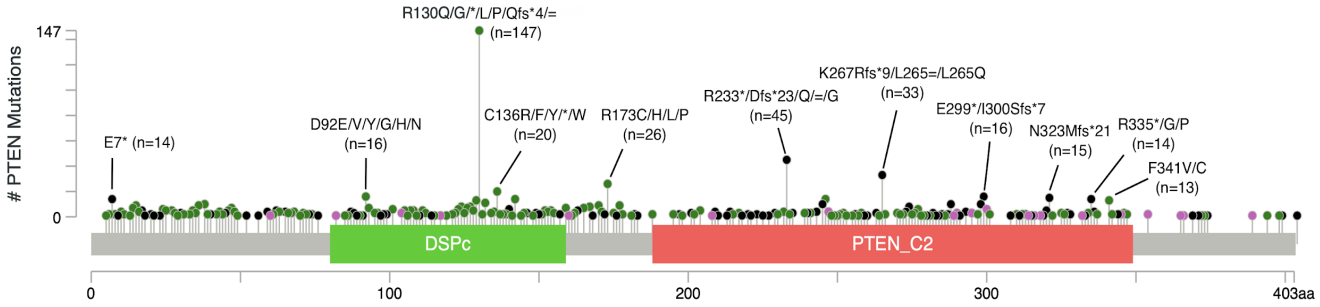
A. TP53 mutant (n=4192)



B. KRAS mutant (n=998)



C. PTEN mutant (n=973)



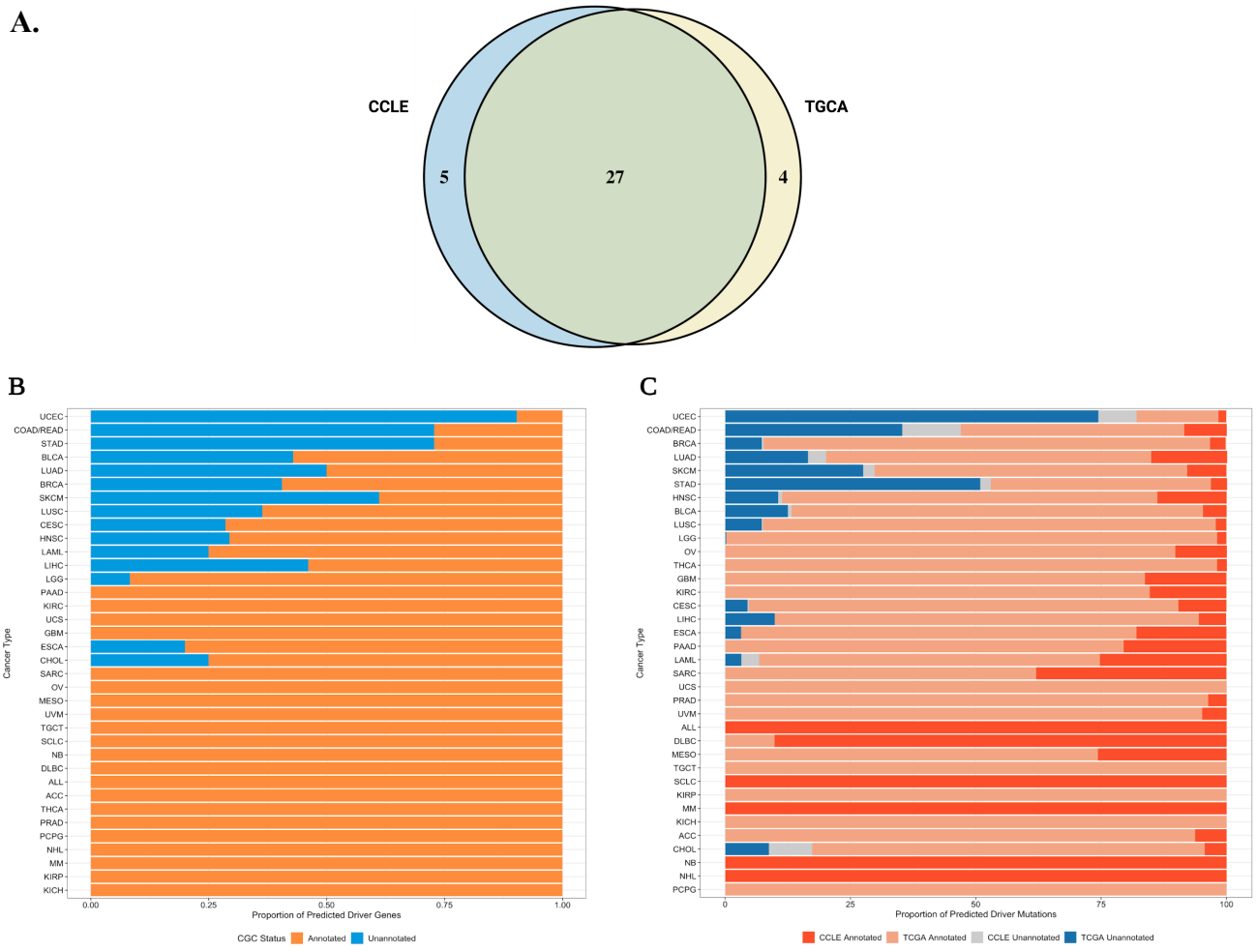
The Landscape of Cancer Driver Genes and Mutations

A compendium of whole exome driver mutations was characterized across 49 unique cancer types in TCGA PanCancer Atlas and CCLE using *OncodriveFML*. A total of 2,685,975 SNVs were processed for TCGA and 775,968 SNVs for CCLE. *OncodriveFML* was run three times for CDS genomic elements which includes TCGA pan-cancer variants only, CCLE pan-cancer variants only, and both TCGA and CCLE pan-cancer variants combined (Fig. 4, Fig 5). Taking this approach, we were able to identify the most significant driver genes in TCGA, CCLE, and both TCGA and CCLE pan-cancer compendiums. Our pan-cancer analysis approach predicted *ELAVL1* as the most significantly (observed p-value < 0.00001) altered driver gene (Fig. 5A). *ELAVL1* encodes the embryonic lethal abnormal vision like 1 (ELAVL1, also known as HuR) protein and plays a role regulating the alternative splicing of translation initiation 4E nuclear import factor 1 (Eif4enif1) which encodes the translation initiation factor 4E transporter (4E-T) and suppresses the expression of capped mRNAs (Chang et al., 2014). The knockout of endothelial-specific *ELAVL1* in mice has shown to exhibit reduced revascularization after hind limb ischemia and tumor angiogenesis resulting in attenuated blow flow and tumor growth

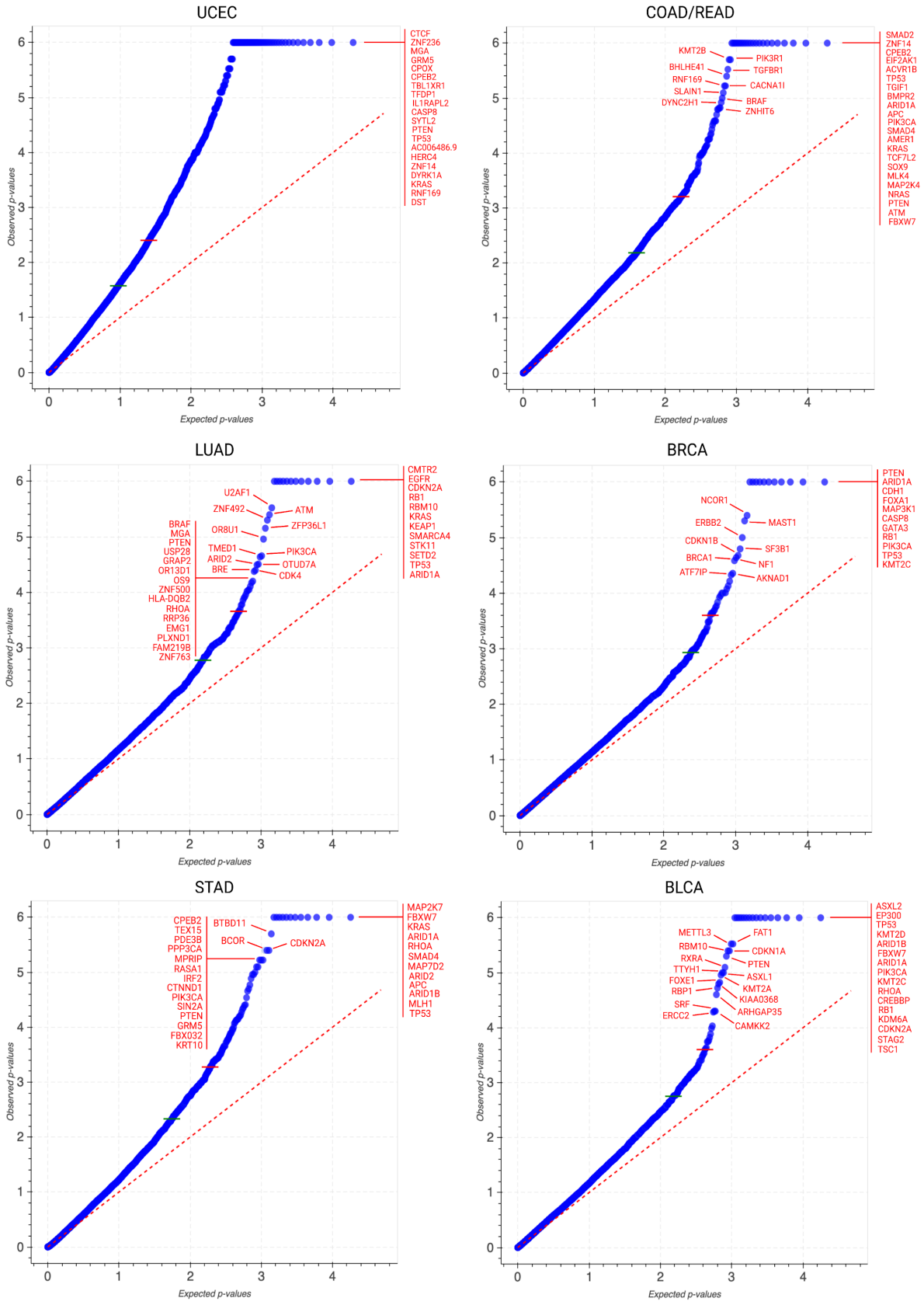
respectively (Chang et al., 2014). However, *ELAVL1* has not been previously characterized as a cancer driver gene to the best of our knowledge.

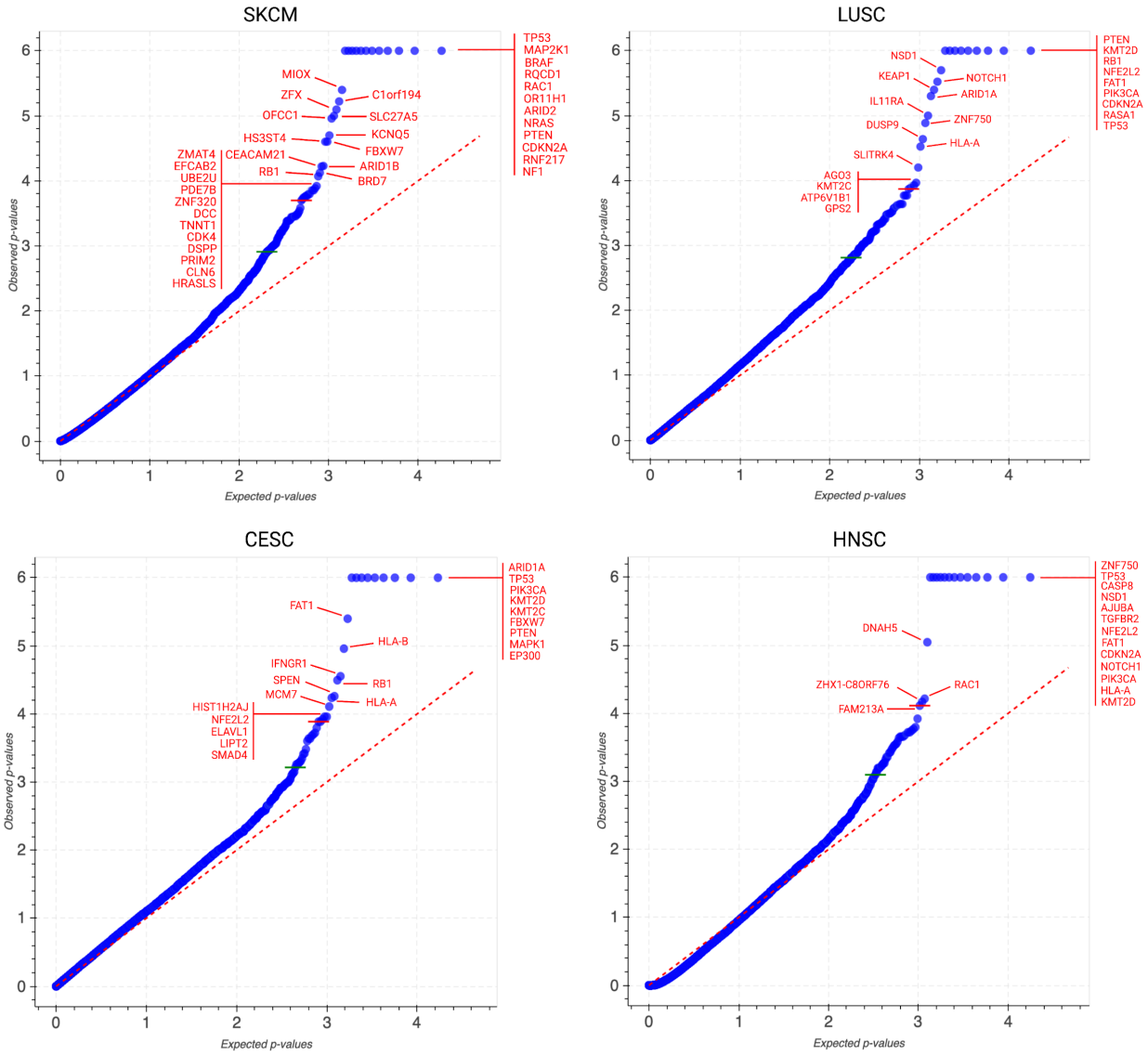
DNAH9 was predicted as the second most significantly (observed p-value < 0.00001) mutated driver gene followed by *FHL2* and *GNB2L1* in our TCGA and CCLE pan-cancer analysis results (Fig. 5A). *DNAH9* encodes dynein axonemal heavy chain 9 (DNAH9) protein which plays a role in cell motility and have been frequently associated with diverse kinds of malignant tumors (Zhu et al., 2019). However, *DNAH9* has not been previously described as a cancer driver gene to the best of our knowledge. *FHL2* encodes four-and-a-half LIM-domain protein 2 (FHL2), a transcriptional cofactor that regulates p21 gene expression that inhibits cell proliferation of soft surfaces therefore having an important role in cancer (Naotaka et al., 2016; Kleiber et al., 2007). *FHL2* is known to be deregulated in cancer, however, it has been found to be overexpressed in different tumor types, and its driver effect in a tissue-dependent manner is not well understood (Kleiber et al., 2007). *GNB2L1*, also known as *RACK1*, encodes guanine nucleotide-binding protein subunit beta-2-like 1 (GNB2L1) which has a significant role shuttling proteins, anchoring them at specific locations and stabilizing their activities (Adams et al., 2011). Studies have demonstrated that *GNB2L1* plays important roles in cancer progression and its expression is upregulated during angiogenesis in various cancer types, including lung cancer (Cox et al., 2003; Lee et al., 2017). Despite *DNAH9*, *FHL2,* and *GNB2L1* having been previously observed in cancer, they had not been characterized as potential driver genes and their mutational effects across tumors are not well understood. Many other potential driver genes which do not have COSMIC CGC annotations were predicted by our analysis including *SACS*, *MGAM*, *ABCA13*, *HSPB11*, *ZNF559*, and *MROH1* (Fig. 5B-C).

Figure 7. Predicted driver genes landscape of 11 most mutated cancer types. A. Venn diagram showing the number of mutual cancer types with at least one predicted driver. **B.** Fraction of predicted COSMIC CGC annotated and unannotated driver genes in cancer specific analyses. Plot includes all cancer types where at least one significant (q-value < 0.1) driver was predicted. **C.** Fraction of predicted driver mutations per cancer type in Figure B. **D.** QQ plots comparing the expected and observed distribution of functional mutational bias p-values of genes in TCGA-CCLE combined cancer types. Only tumors that had at list one significant driver gene identified by *OncodriveFML* were included. **E.** Mutation distribution of cancers in figure A. **F.** The q value of each driver tumor specific plotted against the q value in pan-cancer analysis for al cancer types in figure A.

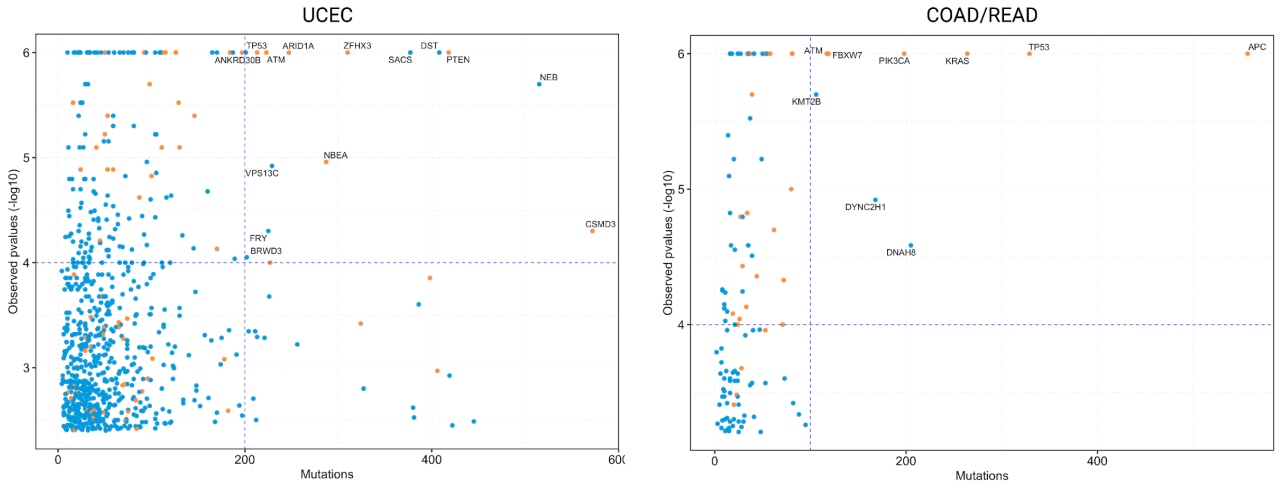


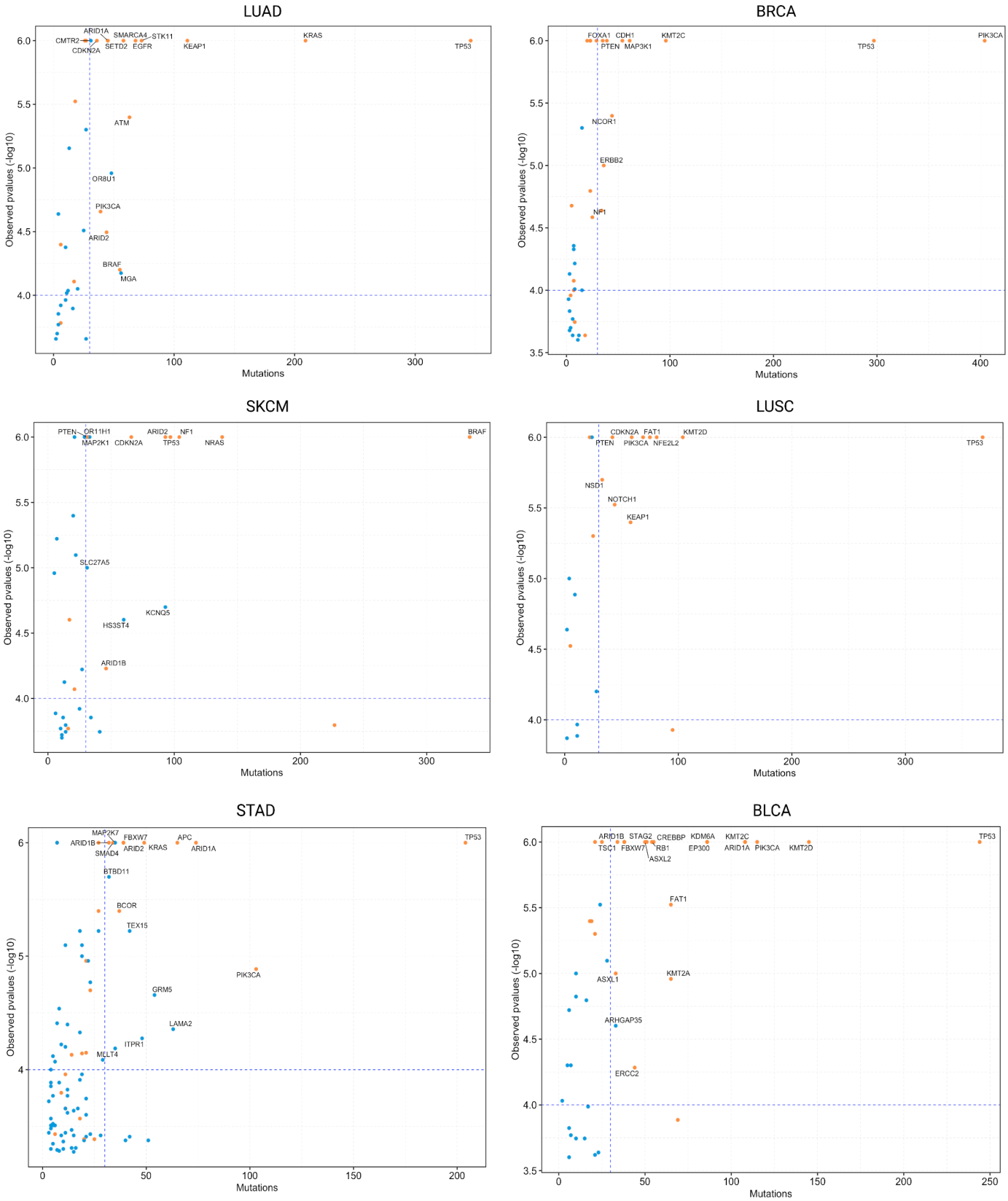
D.

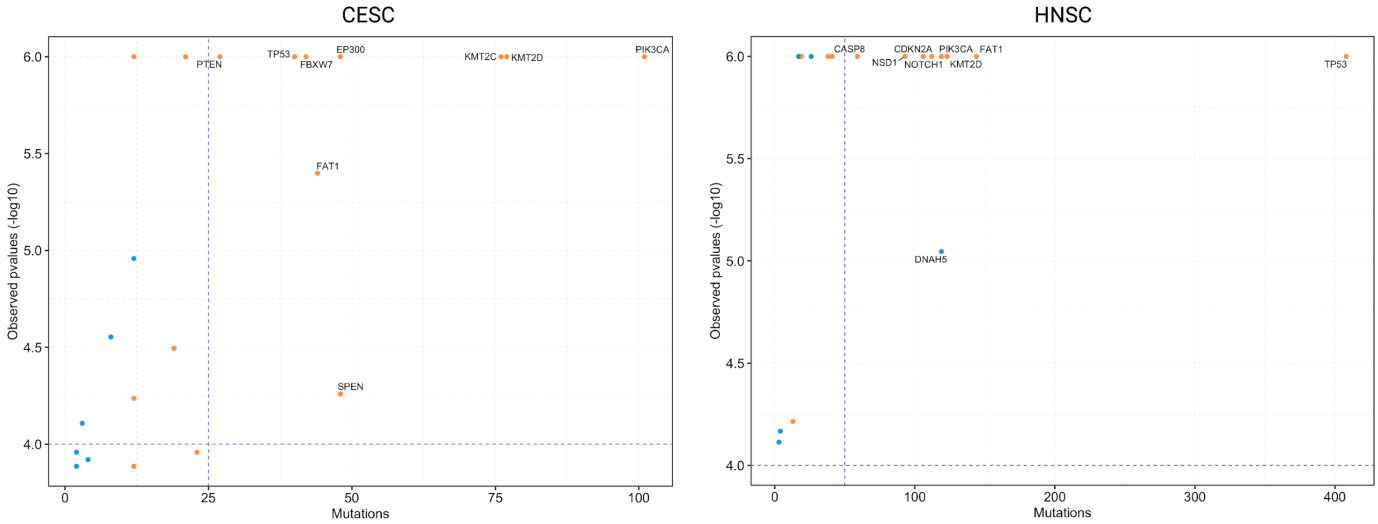

UCEC
CTCF
ZNF236
MGA
GRM5
CPOX
CPEB2
TBL1XR1
TFDP1
IL1RAPL2
CASP8
SYTL2
PTEN
TP53
AC006486.9
HERC4
ZNF14
DYRK1A
KRAS
RNF169
DST
COAD/READ
KMT2B
PIK3R1
BHLHE41
TGFBR1
RNF169
CACNA1I
SLAIN1
BRAF
DYNC2H1
ZNHIT6
SMAD2
ZNF14
CPEB2
EIF2AK1
ACVR1B
TP53
TGIF1
BMPR2
ARID1A
APC
PIK3CA
SMAD4
AMER1
KRAS
TCF7L2
SOX9
MLK4
MAP2K4
NRAS
PTEN
ATM
FBXW7
LUAD
U2AF1
ZNF492
ATM
ZFP36L1
OR8U1
TMED1
PIK3CA
ARID2
OTUD7A
BRE
CDK4
BRAF
MGA
PTEN
USP28
GRAP2
OR13D1
OS9
ZNF500
HLA-DQB2
RHOA
RRP36
EMG1
PLXND1
FAM219B
ZNF763
CMTR2
EGFR
CDKN2A
RB1
RBM10
KRAS
KEAP1
SMARCA4
STK11
SETD2
TP53
ARID1A
BRCA
NCOR1
MAST1
ERBB2
CDKN1B
SF3B1
BRCA1
NF1
ATF7IP
AKNAD1
PTEN
ARID1A
CDH1
FOXA1
MAP3K1
CASP8
GATA3
RB1
PIK3CA
TP53
KMT2C
STAD
CPEB2
TEX15
PDE3B
PPP3CA
MPRIP
RASA1
IRF2
CTNND1
PIK3CA
SIN2A
PTEN
GRM5
FBX032
KRT10
BTBD11
BCOR
CDKN2A
MAP2K7
FBXW7
KRAS
ARID1A
RHOA
SMAD4
MAP7D2
ARID2
APC
ARID1B
MLH1
TP53
BLCA
METTL3
FAT1
RBM10
CDKN1A
RXRA
PTEN
TTYH1
ASXL1
FOXE1
KMT2A
RBP1
KIAA0368
SRF
ARHGAP35
ERCC2
CAMKK2
ASXL2
EP300
TP53
KMT2D
ARID1B
FBXW7
ARID1A
PIK3CA
KMT2C
RHOA
CREBBP
RB1
KDM6A
CDKN2A
STAG2
TSC1
Observed p-values
Expected p-values


SKCM
TP53
MAP2K1
BRAF
RQCD1
RAC1
OR11H1
ARID2
NRAS
PTEN
CDKN2A
RNF217
NF1
MIOX
ZFX
OFCC1
HS3ST4
ZMAT4
EFCAB2
UBE2U
PDE7B
ZNF320
DCC
TNNT1
CDK4
DSPP
PRIM2
CLN6
HRASLS
CEACAM21
RB1
C1orf194
SLC27A5
KCNQ5
FBXW7
ARID1B
BRD7
Observed p-values
Expected p-values
LUSC
PTEN
KMT2D
RB1
NFE2L2
FAT1
PIK3CA
CDKN2A
RASA1
TP53
NSD1
KEAP1
IL11RA
DUSP9
SLITRK4
AGO3
KMT2C
ATP6V1B1
GPS2
NOTCH1
ARID1A
ZNF750
HLA-A
CESC
ARID1A
TP53
PIK3CA
KMT2D
KMT2C
FBXW7
PTEN
MAPK1
EP300
FAT1
IFNGR1
SPEN
MCM7
HIST1H2AJ
NFE2L2
ELAVL1
LIPT2
SMAD4
HLA-B
RB1
HLA-A
HNSC
ZNF750
TP53
CASP8
NSD1
AJUBA
TGFBR2
NFE2L2
FAT1
CDKN2A
NOTCH1
PIK3CA
HLA-A
KMT2D
DNAH5
ZHX1-C8ORF76
FAM213A
RAC1

E.


UCEC
TP53
ANKRD30B
ARID1A
ATM
ZFHX3
SACS
DST
PTEN
NEB
NBEA
VPS13C
CSMD3
FRY
BRWD3
Observed pvalues (-log10)
Mutations
COAD/READ
ATM
FBXW7
PIK3CA
KRAS
TP53
APC
KMT2B
DYNC2H1
DNAH8


LUAD
ARID1A
SMARCA4
STK11
CMTR2
SETD2
EGFR
KEAP1
KRAS
TP53
CDKN2A
ATM
OR8U1
PIK3CA
ARID2
BRAF
MGA
Observed pvalues (-log10)
Mutations
BRCA
FOXA1
CDH1
KMT2C
PTEN
MAP3K1
TP53
PIK3CA
NCOR1
ERBB2
NF1
SKCM
PTEN
OR11H1
ARID2
NF1
MAP2K1
CDKN2A
TP53
NRAS
BRAF
SLC27A5
KCNQ5
HS3ST4
ARID1B
LUSC
CDKN2A
FAT1
KMT2D
PTEN
PIK3CA
NFE2L2
TP53
NSD1
NOTCH1
KEAP1
STAD
MAP2K7
FBXW7
APC
ARID1B
ARID2
KRAS
ARID1A
SMAD4
TP53
BTBD11
BCOR
TEX15
PIK3CA
GRM5
LAMA2
ITPR1
MLLT4
BLCA
ARID1B
STAG2
CREBBP
KDM6A
KMT2C
TSC1
FBXW7
RB1
EP300
ARID1A
PIK3CA
KMT2D
ASXL2
TP53
FAT1
KMT2A
ASXL1
ARHGAP35
ERCC2



Superdrivers of Cancer

Our analysis successfully demonstrated the high frequency of cancer superdriver genes, drivers that can trigger the onset of cancer with little or no contribution from additional genetic alterations across tumor types (Grossmann et al., 2020). *TP53* was the most significantly mutated superdriver in our TCGA and CCLE pan-cancer analysis followed by *KRAS*, *PTEN*, *RB1*, *CDKN2A*, *NRAS*, *CASP8,* and *STK11* (q-value < 0.01) (Fig. 5B, Fig. 5C). *TP53* was found to be mutated in all 49 cancer types accumulating 4,192 protein coding somatic mutations (Fig. 5C). The most recurrent *TP53* alterations predicted in our analysis were *R23C/H/L/S/P/G/Ls*72* (n=329, 7.9% of all alterations in *TP53*), *R248Q/8W/8L/8P/8=/8G/8Gfs*96* (n=300, 7.1%), and *R175H/G/C/L/Pfs*72* (n=198, 4.72%) (Fig. 6A).

KRAS was found to be initially mutated in 69.3% (n=34) of all cancer types and was predicted to be a significant driver gene in 14.2% of all cancers (n=7) (Fig. 5C, Fig. 6B, Fig. 7B-D). *KRAS* was identified to host a total of 998 protein coding alterations in both TCGA and CCLE (Fig. 5C, Fig. 6B). Missense *KRAS G12D/V/C/A/R/S* was the alteration with highest

recurrence 65.7% (n=656) from all other alterations found in *KRAS* (Fig. 6A). Missense *KRAS G13D/C/=/V* was the second most recurrent mutation in *KRAS* forming 10% (n=100) of all other mutations found in the driver gene (Fig. 5C). The rest of *KRAS* mutations had less than 60 recurrences individually.

PTEN was among the highly significant superdrivers in both TCGA and CCLE pan-cancer analysis with 973 alterations (Fig. 5C). *PTEN* was initially mutated in 63.2% (n=31) of all cancer types and was predicted to be a significant driver gene in 26.5% (n=13) of all cancers (Fig. 5C). *PTEN R130Q/G/*/L/P/Qfs*4/=* was the most recurrent mutation forming 15.1% (n=147) of all mutations found in the protein coding region of the driver. *PTEN R233*/Dfs*23/Q/=/G* was the second most significant mutation in *PTEN* forming 4.6% (n=45) of all *PTEN* alterations across tumor types (Fig. 6C).

CDKN2A was initially mutated in 65.3% (n=32) of all cancer types and was predicted to be a significant driver in 16.3% (n=8) of all cancers with a mutational landscape of 450 alterations (Fig. 5C, Supplementary Fig. 1A). From all *CDKN2A* mutations, *CDKN2A R80*/Q,* * being a missense mutation and *Q* being a nonsense mutation, was the most recurrent alteration making 12.5% (n=57) of all protein coding region mutations (Supplementary fig. 1A). Missense *CDKN2A H83Y/D/R/=/L/P/Q* was the second most recurrent *CDKN2A* mutation making 6.2% of all mutations in the driver (Supplementary fig. 1A). The rest of mutations found in *CDKN2A* had less than 27 alterations (Supplementary fig. 1A).

RB1 was another highly significant superdriver initially mutated in 67.3% (n=33) of all cancers and predicted to be a significant driver gene in 22.4% (n=11) of them with a landscape of 447 protein coding mutations (Fig. 5C). The two most recurrent mutations found in *RB1* were *R552*/Q* (n=10) and *R320** (n=10) each one making 2.2% of all alterations found in the driver

(Supplementary Fig. 1A). *NRAS* was found to be initially mutated in 73.4% (n=36) of all cancers and predicted to be a significant driver gene in 10.2% (n=5) of them with a landscape of 405 mutations (Fig. 5C). The most recurrent mutation in *NRAS* was *Q61R/K/L/H/P/** making 64.7% (n=262) of all NRAS protein coding mutations (Supplementary Fig. 1A). The second most recurrent mutation was *NRAS G12D* forming only 10.9% (n=44) of all mutations.

CASP8 was found to be among the most significant driver genes. *CASP8* was initially mutated in 59.1% (n=29) of all cancer types and predicted a significant driver gene in 8.1% (n=4) of them with a landscape of 310 protein coding mutations (Fig. 5C). The two most recurrent mutations in *CASP8* were missense *668Q/** making 5.48% (n=17) and *R233W/R233Q/G234Dfs*4/R23P* making 4.5% (n=14) of all mutations, respectively (Supplementary Fig. 1A). The next two most recurrent mutations in *CASP8* were *R432** (n=11) and *CASP8 K473Nfs*/R471G/R471I* (n=9) making 3.5% and 2.9% of all mutations, respectively (Supplementary Fig. 1A).

STK11 was found to be the last most significant driver among the group of superdriver genes (Fig. 5C). *STK11* was mutated in 65.3% (n=32) of all cancer types and predicted to be a significant driver only in LUAD (2.0%, n=1) with a landscape of 215 protein coding mutations (Fig. 5C, Fig. 6A). *STK11 P281Rfs*6* was the most recurrent alteration making 4.1% (n=9) of all mutations in *STK11* (Supplementary Fig. 1A). The two second most recurrent alterations were *STK11 P281L/=* (n=7) and *STK11 S216F* (n=7) each one making 3.2% of all alterations, respectively (Supplementary Fig. 1A). The rest of mutations in *STK11* had less than 6 recurrences across all tumors.

These results show a cluster of highly significant cancer superdrivers. In this group, *TP53*, *KRAS*, *PTEN*, *RB1*, *CDKN2A*, *NRAS*, *CASP8* and *STK11* are shown to be the most

significant and some of the most mutated driver genes across all cancer types in both TCGA PanCancer Atlas and CCLE tumors. In addition, our pan-cancer analyses revealed the mutational landscapes of each of the drivers found in this group and their most recurrent alterations. This group of highly selective driver genes might be the dominant force driving the progression of malignant tumors and further analysis of their modules will provide better understanding of their systemic roles in cancer.

Pan-Cancer Analysis Reveals Potentially Novel Driver Genes

Our pan-cancer analysis revealed previously unannotated driver genes by combining TCGA PanCancer Atlas tumor samples and CCLE cancer cell lines somatic mutation variants (Fig. 5B-C, Fig. 6A). *DSPP* was predicted as a potential unannotated driver in the CGC and the most significant to be found in both CCLE and TCGA pan-cancer analyses (Fig. 5C; Fig 4C-D). *DSPP* encodes dentin sialophosphoprotein (DSPP) and plays an important role in dentin mineralization and has also been found to be upregulated in human cancers including head and neck squamous cell carcinoma (HNSC) (Gkouveris et al., 2018). In transiently transfected puromycin-free OSC2 cells, *DSPP* silencing down-regulated the mRNA expression levels of key ER stress regulators, including GRP78, SERCA2b, PERK, IRE1, ATF6 and MMP20, and it decreased cell viability and migration by enhancing apoptosis (Gkouveris et al., 2018). These results by Gkouveris et al. show a potential oncogenic activity of *DSPP;* however, further analyses and experimental approaches are needed to answer this question. *PTPRQ* is another TCGA and CCLE mutual and unannotated predicted driver gene in our analysis (Fig. 5C; Fig 4C-D). *PTPRQ* encodes the protein tyrosine phosphate receptor-like type Q (PTPRQ) which plays an important role in cellular proliferation and differentiation and has been shown to be important in the development of various cancer including sporadic colorectal cancer (CRC)

(Laczmanska et al. 2016). Laczmanska et al. demonstrated that *PTPRQ* expression was significantly (p=0.0293) higher in tissues presenting KRAS mutations and confirmed the contribution of *PTPRQ* in CRC development, therefore being a candidate oncogene (Laczmanska et al. 2016). Both *DSPP* and *PTPRQ* had been previously described to play key roles in various cancers; however, *both genes* had not been previously characterized as potential drivers to the best of our knowledge. However, *PTPRQ* has been proven to play an important role in the development of CRC (Laczmanska et al. 2016), therefore being a potential cancer driver gene.

In addition to the already described unannotated drivers, many others were found to be specific to TCGA (Fig. 4A) and CCLE (Fig. 4B), respectively. In TCGA pan-cancer results, we found *CSMD1* to be the most significant (q-value < 0.1) and altered potential unannotated driver. *CSMD1* encodes the human CUB and Sushi multiple domains 1 (CSMD1) membrane bound complement inhibitor suggested to be a tumor suppressor gene due to the fact that allelic loss of this region, including 8p23, characterizes various cancer types including breast cancer (Escudero-Esparza et al. 2016). Another potential uncharacterized and highly significant (q-value < 0.1) driver identified in TCGA pan-cancer analysis is *NEB* (Fig.4B). *NEB* encodes nebulin (NEB), a filamentous protein that is integral to the skeletal muscle thin filament (Labeit et a., 2011), and we could not find direct evidence in the science literature of a specific role of *NEB* in cancer. The *DNAH* genes family caught our attention among the potential unannotated drivers group in TCGA. We found *DNAH8* and *DNAH9* as highly significant potential drivers each one with over 1000 alterations (Fig.4B). *DNAH8* encodes dynein axonemal heavy chain 8 (DNAH8) and *DNAH9* encodes dynein axonemal heavy chain 9 (DNAH9) both important proteins for cell motility. In particular, *DNAH8* missense and frameshift deletion/insertion

mutations and *DNAH9* missense and silent mutations have been found to be highly sensitive to various chemotherapy compounds in gastric cancer compared to wild-type *DNAH8* and *DNAH9* ($P = 0.002$) (Zhu et al., 2019). However, *DNAH8* nor *DNAH9* has been previously reported as potential cancer drivers to the best of our knowledge.

Some of the unannotated and highly significant (q-value < 0.1) predicted driver genes in CCLE pan-cancer analysis includes *TMEM218*, *TRAPPC3L*, *PGA4*, *MTHFD2L*, and *PGA4* (Fig. 4D). Each of these potential drivers have less than 100 mutations in their protein coding regions, respectively. *TMEM218* encodes transmembrane protein 218 (TMEM218) whose function is not completely known nor understood, however, it has been associated with alternative splicing events in osteosarcoma (Green et al., 2020). We could not find any information in the literature stating the roles of *TRAPPC3L* in cancer. *PGA4* encodes pepsinogen A4 (PGA4) protein which functions in the digestion of dietary proteins and have been shown to be downregulated in cancer tissues and highly associated with higher survival in kidney renal clear cell carcinoma (KRCC) (Shen et al., 2020). *MTHFD2L* encodes mitochondrial methylenetetrahydrofolate dehydrogenase 2 like (MTHFD2L) protein which plays a role in the conversion of folate to formate in the mitochondrial pathway for 1-carbon metabolism and is expressed in highly proliferating cancers including non-small cell lung cancers (Tedeschi et al., 2015). *PGA5* encodes pepsinogen A5 (PGA5) protein and has been shown to have high expression in KRCC and kidney renal papillary cell carcinoma and decreased expression in stomach carcinoma, CHOL, COAD, UCEC, PRAD, BRCA, KICH, and THCA (Shen et al., 2020). In addition, PGA5 was shown to be involved in the K-RAS signaling pathway, bile acid metabolism, mitotic G2 M-phase and mTOR and DNA repair (Shen et al., 2020), and therefore may be a very important gene in cancer formation.

Our pan-cancer analysis approach predicted many novel cancer driver genes not previously described in the literature to the best of our knowledge. However, many of them have been observed to play key roles in the development of cancer. These results will provide an asset for future research endeavors and further assist in the discovery of the functional roles of these novel drivers in cancer.

The Landscape of Tumor Specific Driver Genes

We performed cancer specific analyses for 49 unique cancer types in TCGA and CCLE to identify cancer specific driver genes. We combined TCGA PanCancer Atlas and CCLE mutation calls for 28 mutual cancer types (Fig. 3A). In addition, 5 cancer types just contained somatic mutation calls from TCGA, and 16 cancer types contain mutation calls for CCLE only (Fig. 3A). We applied the same uniform filtering and processing metrics as for the pan-cancer analyses. After running *OncodriveFML* in all cancer types, we proceeded only with cancer types that had at least one significant driver gene predicted (q-value < 0.1).

A total of 37 cancer types had at least one significantly predicted driver gene (Fig. 7B-E). From this group, 28 cancer types were mutual in both TCGA and CCLE, 4 cancer types were unique to TCGA (KICH, KIRP, PCPG, and UCS), and 5 cancer types were unique to CCLE (ALL, NB, MM, SCLC, and NHL) (Fig. 7A). We determine the CGC status of every predicted driver in each cancer type. We found that 15 cancer types had a predicted amount of unannotated significant driver genes in the CGC (Fig. 7B). From this group, UCEC has the highest number of drivers unannotated by COSMIC CGC followed by COAD/READ, STAD, SKCM, LUAD and LICH (Fig. 7B). We labeled each predicted mutation from source origin, TCGA or CCLE, and driver gene CGC status, annotated or unannotated (Fig. 7C). Four cancer types including UCS, TGCT, KICH, and PCPG were found to have only TCGA mutations that are found in genes that

have been previously annotated as drivers in COSMIC CGC (Fig. 7C). Five cancer types including ALL, NB, MM, SCLC, and NHL had only CCLE mutations all of which were found to be annotated in the COSMIC CGC (Fig. 7C). UCEC was the cancer type with the highest number of significantly predicted driver genes (n=752) from which were 9.7% (n=73) found to be annotated and 90.3% (n=679) were not annotated in COSMIC CGC (Fig. 7B, Fig. 7D-E). COAD/READ were the second cancer type combined with the highest number of predicted drivers. COAD/READ had 113 significant predicted drivers from which 27.4% (n=31) were found to be annotated and 72.6% (n=82) were not found in COSMIC CGC (Fig. 7B, Fig. 7D-E). Our analysis predicted 92 significant drivers in STAD (Fig. 7B). From this group, we found that only 27.2% (n=25) were annotated and 72.8% (n=67) had not COSMIC CGC annotation (Fig. 7D-E).

BLCA and LUAD were among the group of cancer types with the highest number of predicted significant driver genes (Fig. 7B, Fig. 7D-E). Our cancer specific analysis predicted 42 significant driver genes for BLCA from which 57.1% (n=24) were found to be annotated and 42.9% (n=18) were not found in COSMIC CGC (Fig. 7D-E). LUAD were found to have a total of 38 significant driver genes (Fig. 7D). From this group, 50%(n=19) of all drivers were found to be annotated and 50% (n=19) were not found in COSMIC CGC (Fig. 7D-E).

BRCA and SKCM had a very similar driver landscape with 37 and 36 predicted driver genes, respectively (Fig. 7B-E). From all significant predicted drivers in BRCA, 59.5% (n=22) were found annotated and 40.5% (n=15) had no annotation in COMIC GCG (Fig. 7D-E). SKCM driver genes landscape was made of 38.9% (n=14) COSMIC CGC annotated and 61.1% (n=22) unannotated drivers (Fig. 7D-E).

LUSC, CESC and HNSC were among the top 10 cancer types with the highest number of significant drivers (Fig. 7B, Fig. 7D-E). These results might be associated with the high number of samples in these cancer types. Our cancer specific analysis predicted a landscape of 22 significant driver genes for LUSC (Fig. 7D-E). From this group, 63.3% (n=14) were found annotated and 36.4% (n=8) were found unannotated in COSMIC CGC (Fig. 7D-E). CESC had a landscape of 21 significant drivers (Fig. 7D) from which 71.2% (n=15) were annotated and only 28.6% (n=6) were unannotated in COSMIC CGC (Fig. 7D-E). HNSC was predicted to have a cancer landscape of 17 driver genes (Fig. 7D) from which 70.6% (n=12) were annotated and only 29.4% (n=5) were unannotated in COSMIC CGC (Fig. 7D-E).

The rest of cancer types from the group of 37 cancer types where at least one significant driver gene was predicted (Fig. 7B) contained less than 17 significant driver genes per cancer exome (Fig. 7D-E). This group includes LAML (n=16), LICH (n=13), LGG (n=12), PAAD (n=7), KIRC (n=6), UCS (n=5), GBM (n=5), ESCA (n=5), CHOL (n=4), SARC (n=3), OV (n=3), MESO (n=3), UVM (n=2), TGCT (n=2), SCLC (n=2), NB (n=2), DLBC (n=2), ALL (n=2), ACC (n=2), THCA (n=1), PRAD (n=1), PCPG (n=1), NHL (n=1), MM (n=1), KIRP (n=1), and KICH (n=1) (Supplementary Fig. 2A-M).

FUTURE WORK

This analysis of driver genes, mutations and modules across cancer exomes will provide a compendium for future research and clinical endeavors. Here, we applied *OncodriveFML* to identify potential driver genes and mutations across cancer types in TCGA PanCancer Atlas and CCLE. Further, our current analyses focus in using orthogonal methods to validate potential driver genes and mutations discovered in this study. To do so, we aim to apply *DriverPower*, a framework for identification of coding and non-coding driver mutations using MB and FI scores

(Shuai et al., 2020). By incorporating MB and FI scores, *DriverPower* predicts potential driver mutations more accurately compared to other driver mutation calling pipelines as described by Shuai et al. and its performance had the highest F1 score, highest precision and recall in protein coding elements (precision = 0.84; recall = 0.79; F1 = 0.81)

Eventually, we aim to identify driver modules using *Mutex,* a method that identifies potential driver pathways based on mutual exclusivity of alterations (Babur et al., 2015). The *Mutex* analysis will allow us to identify potential novel driver modules/pathways which had not been previously characterized using different approaches. These results will help us understand better the systemic role of driver genes, mutations and modules across malignant tumors.

CONCLUSION AND FUTURE PERSPECTIVES

A fundamental understanding of cancer driver genes, mutations and modules is necessary to develop early detection advances, more efficient treatments, and potentially, a universal cure for cancer. Cancer driver discovery is challenging due to the limitations of tumor sample sizes and complexity to define the borders of biological pathways. In this paper, we report an approach to discover novel driver genes, mutations and modules that play significant roles in tumorigenesis and tumor malignancy.

By combining the TCGA and CCLE cancer samples, we created a mutational landscape over 3 millions of mutations across 49 cancer types. This approach allowed us to identify very well-known and studied driver genes such as *TP53*, *KRAS*, *PTEN*, *RB1*, and *CDKN2A* across cancer types. We also revealed a landscape of unannotated driver genes across tumor types that might have the potential to be drug targets across tumors. Some of the most significant unannotated driver genes include *DSPP*, *PTPRQ*, *ELAVL1*, *FHL2*, *GNB2L1*, *DNAH8* and

DNAH9. Many of these unannotated drivers have been previously observed to play key roles in cancer. For example, *DSSP* which has been shown to downregulate expression levels of key ER stress regulators (Gkouveris et al., 2018) and *PTPRQ* in cell proliferation and differentiation leading to tumor development (Laczmanska et al. 2016).

These results yield insights into the nature of cancer and altered biological processes that lead to development of solid human neoplasms and hematopoietic and lymphoid malignancies. We further aim to use orthogonal computational methods to validate potential novel driver genes discovered in this study to provide a framework and model to study cancer driver mutations and modules using both tumor samples and cancer cell lines. In addition, we aim to identify potential novel driver modules/pathways. Future efforts could include the development of computational and experimental methods to predict the functional outcomes of driver mutations, driver pathways and combinations of these. When concluded, this study will provide a comprehensive landscape of driver genes, mutations and modules across cancer types that will yield significant insights for future research and clinical efforts against cancer malignancies. Extending these endeavors to demonstrate how newly discovered driver genes impact tumor samples and cancer types will bring us closer to fundamentally understanding the mechanistic and systemic role of driver mutations.

REFERENCES

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4; 144(5): 646-74. doi: 10.1016/j.cell.2011.02.013.
2. Warburg O. The metabolism of carcinoma cells. The Journal of Cancer Research. 1925; 9(1): 148–163.
3. Warburg O. On the origin of cancer cells. Science. 1956 Feb 24;123(3191):309-14. doi: 10.1126/science.123.3191.309.
4. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science. 2009 May 22;324(5930):1029-33. doi: 10.1126/science.1160809.
5. Liberti MV, Locasale JW. The Warburg Effect: How Does it Benefit Cancer Cells? Trends Biochem Sci. 2016 Mar; 41(3): 211-218. doi: 10.1016/j.tibs.2015.12.001.
6. Watson, J. D. & Crick, F. H. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* 171, 737–738 (1953). URL: https://doi.org/10.1038/171737a0.
7. International Human Genome Sequencing Consortium., Whitehead Institute for Biomedical Research, Center for Genome Research:., Lander, E. *et al.* Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001). URL: https://doi.org/10.1038/35057062.
8. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, et al. The sequence of the human genome. Science. 2001 Feb 16; 291 (5507): 1304-51. doi: 10.1126/science.1058040. Erratum in: Science 2001 Jun 5; 292 (5523): 1838.
9. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945 (2004). URL: https://doi.org/10.1038/nature03001.
10. Hood, L., Galas, D. The digital code of DNA. *Nature* 421, 444–448 (2003). https://doi.org/10.1038/nature01410
11. Gibbs, R.A. The Human Genome Project changed everything. *Nat Rev Genet* 21, 575–576 (2020). URL: https://doi.org/10.1038/s41576-020-0275-3.
12. The Cancer Genome Atlas Research Network., Genome Characterization Center., Chang, K. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45, 1113–1120 (2013). URL: https://doi.org/10.1038/ng.2764
13. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018 Apr 5; 173(2): 291-304.e6. doi: 10.1016/j.cell.2018.03.022.
14. Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, et al. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell. 2018 Apr 5;173(2):305-320.e10. doi: 10.1016/j.cell.2018.03.033.

15. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, Samaras D, Shroyer KR, Zhao T, Batiste R, Van Arnam J; Cancer Genome Atlas Research Network, Shmulevich I, Rao AUK, Lazar AJ, Sharma A, Thorsson V. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. Cell Rep. 2018 Apr 3;23(1):181-193.e7. doi: 10.1016/j.celrep.2018.03.086.
16. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kamińska B, Huelsken J, Omberg L, Gevaert O, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. Cell. 2018 Apr 5; 173 (2): 338-354.e15. doi: 10.1016/j.cell.2018.03.034.
17. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL., et al. The Immune Landscape of Cancer. Immunity. 2018 Apr 17; 48(4): 812-830.e14. doi: 10.1016/j.immuni.2018.03.023.
18. The Cancer Genome Atlas Pan-Cancer analysis project Cancer Genome Atlas Research Network; Genome Characterization Center, Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, Ally A, Balasundaram M, Birol I, Butterfield YS, Chu A, Chuah E, Chun HJ, et. al. Nature Genetics 45, 1113–1120 (2013) doi:10.1038/ng.2764
19. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium., Campbell, P.J., Getz, G. *et al.* Pan-cancer analysis of whole genomes. *Nature* 578, 82–93 (2020). URL: https://doi.org/10.1038/s41586-020-1969-6.
20. Ghandi, M., Huang, F.W., Jané-Valbuena, J. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 569, 503–508 (2019). URL: https://doi.org/10.1038/s41586-019-1186-3.
21. Li, Y., Roberts, N.D., Wala, J.A. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* 578, 112–121 (2020). URL: https://doi.org/10.1038/s41586-019-1913-9.
22. PCAWG Transcriptome Core Group., Calabrese, C., Davidson, N.R. *et al.* Genomic basis for RNA alterations in cancer. *Nature* 578, 129–136 (2020). URL: https://doi.org/10.1038/s41586-020-1970-0.
23. Rheinbay, E., Nielsen, M.M., Abascal, F. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* 578, 102–111 (2020). URL: https://doi.org/10.1038/s41586-020-1965-x.
24. Cortés-Ciriano, I., Lee, J.JK., Xi, R. *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* 52, 331–341 (2020). URL: https://doi.org/10.1038/s41588-019-0576-7.
25. Gerstung, M., Jolly, C., Leshchiner, I. *et al.* The evolutionary history of 2,658 cancers. *Nature* 578, 122–128 (2020). URL: https://doi.org/10.1038/s41586-019-1907-7
26. Li, S.C., Tachiki, L.M.L., Kabeer, M.H. *et al.* Cancer genomic research at the crossroads: realizing the changing genetic landscape as intratumoral spatial and temporal heterogeneity becomes a confounding factor. *Cancer Cell Int* 14, 115 (2014). URL: https://doi.org/10.1186/s12935-014-0115-7.

27. Kadota K, Sima CS, Arcila ME, Hedvat C, Kris MG, Jones DR, Adusumilli PS, Travis WD. KRAS Mutation Is a Significant Prognostic Factor in Early-stage Lung Adenocarcinoma. Am J Surg Pathol. 2016 Dec; 40(12): 1579-1590. doi: 10.1097/PAS.0000000000000744.
28. Wodarz D, Newell AC, Komarova NL. Passenger mutations can accelerate tumour suppressor gene inactivation in cancer evolution. J R Soc Interface. 2018 Jun;15(143):20170967. doi: 10.1098/rsif.2017.0967.
29. Waks, Z., Weissbrod, O., Carmeli, B. *et al.* Driver gene classification reveals a substantial overrepresentation of tumor suppressors among very large chromatin-regulating proteins. *Sci Rep* 6, 38988 (2016). URL: https://doi.org/10.1038/srep38988
30. Krug, U., Ganser, A. & Koeffler, H. Tumor suppressor genes in normal and malignant hematopoiesis. *Oncogene* 21, 3475–3495 (2002). URL: https://doi.org/10.1038/sj.onc.1205322
31. Shen, L., Shi, Q. & Wang, W. Double agents: genes with both oncogenic and tumor-suppressor functions. *Oncogenesis* 7, 25 (2018). URL: https://doi.org/10.1038/s41389-018-0034-x
32. Cline MJ. The role of proto-oncogenes in human cancer: implications for diagnosis and treatment. Int J Radiat Oncol Biol Phys. 1987 Sep;13(9):1297-301. doi: 10.1016/0360-3016(87)90219-7.
33. Campbell PJ. Cliques and Schisms of Cancer Genes. Cancer Cell. 2017 Aug 14; 32 (2): 129-130. doi: 10.1016/j.ccell.2017.07.009.
34. Zhang, J., Wu, LY., Zhang, XS. *et al.* Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics* 15, 271 (2014). URL: https://doi.org/10.1186/1471-2105-15-271
35. Ulz, P., Heitzer, E. & Speicher, M. Co-occurrence of *MYC* amplification and *TP53* mutations in human cancer. *Nat Genet* 48, 104–106 (2016). URL: https://doi.org/10.1038/ng.3468
36. Chen, H., Liu, H. & Qing, G. Targeting oncogenic Myc as a strategy for cancer treatment. *Sig Transduct Target Ther* 3, 5 (2018). URL: https://doi.org/10.1038/s41392-018-0008-7
37. Aubrey, Brandon J et al. Tumor-Suppressor Functions of the TP53 Pathway. *Cold Spring Harbor perspectives in medicine* vol. 6,5 a026062. 2 May. 2016, doi:10.1101/cshperspect.a026062
38. Hermeking H, Eick D. Mediation of c-Myc-induced apoptosis by p53. Science. 1994 Sep 30;265(5181):2091-3. doi: 10.1126/science.8091232.
39. Cisowski, J., & Bergo, M. O. (2017). What makes oncogenes mutually exclusive?. *Small GTPases*, *8*(3), 187–192. URL: https://doi.org/10.1080/21541248.2016.1212689
40. Silverbush D., Cristea S., Yanovich-Arad G., Geiger T., Beerenwinkel N., Sharan R. Simultaneous integration of multi-omics data improves the identification of cancer driver modules. Cell Syst., 8 (5) (2019), pp. 456-466.

41. B. Vogelstein, N. Papadopoulos, V.E. Velculescu, S. Zhou, L.A. Diaz, K.W. Kinzler Cancer genome landscapes. Science, 339 (2013), pp. 1546-1558.
42. Mateo, L., Duran-Frigola, M., Gris-Oliver, A. *et al.* Personalized cancer therapy prioritization based on driver alteration co-occurrence patterns. *Genome Med* 12, 78 (2020). URL: https://doi.org/10.1186/s13073-020-00774-x
43. Remy E, Rebouissou S, Chaouiya C, Zinovyev A, Radvanyi F, Calzone L. A Modeling Approach to Explain Mutually Exclusive and Co-Occurring Genetic Alterations in Bladder Tumorigenesis. Cancer Res. 2015 Oct 1; 75(19): 4042-52. doi: 10.1158/0008-5472.CAN-15-0602.
44. Mularoni, L., Sabarinathan, R., Deu-Pons, J. *et al.* OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol* 17, 128 (2016). URL: https://doi.org/10.1186/s13059-016-0994-0
45. Babur, Ö., Gönen, M., Aksoy, B.A. *et al.* Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol* 16, 45 (2015). URL: https://doi.org/10.1186/s13059-015-0612-6
46. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 Jul 11;499(7457):214-218. doi: 10.1038/nature12213.
47. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41. doi: 10.1186/gb-2011-12-4-r41. Epub 2011 Apr 28.
48. Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, et al. Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. Cancer Cell. 2016 May 9; 29(5): 723-736. doi: 10.1016/j.ccell.2016.04.002.
49. Huang, F. W. et al. TERT promoter mutations and monoallelic activation of TERT in cancer. *Oncogenesis* 4, e176 (2015).
50. Lek, M., Karczewski, K., Minikel, E. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291 (2016). URL: https://doi.org/10.1038/nature19057
51. The Cancer Genome Atlas Network., Genome sequencing centres: Washington University in St Louis., Koboldt, D. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70 (2012). URL: https://doi.org/10.1038/nature11412
52. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015 Oct 8;163(2):506-19. doi: 10.1016/j.cell.2015.09.033.
53. The Cancer Genome Atlas Network., Genome Sequencing Center Baylor College of Medicine., Muzny, D. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337 (2012). URL: https://doi.org/10.1038/nature11252

54. The Cancer Genome Atlas Research Network., Analysis Working Group: Dana-Farber Cancer Institute., Bass, A. *et al.* Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202–209 (2014). URL: https://doi.org/10.1038/nature13480
55. Farshidfar F, Zheng S, Gingras MC, Newton Y, Shih J, Robertson AG, et al. Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. Cell Rep. 2017 Mar 14;18(11):2780-2794. doi: 10.1016/j.celrep.2017.02.033.
56. The Cancer Genome Atlas Research Network., Albert Einstein College of Medicine., Burk, R. *et al.* Integrated genomic and molecular characterization of cervical cancer. *Nature* 543, 378–384 (2017). URL: https://doi.org/10.1038/nature21386
57. The Cancer Genome Atlas Research Network., Analysis Working Group: Asan University., Kim, J. *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* 541, 169–175 (2017). URL: https://doi.org/10.1038/nature20805
58. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, et al. The somatic genomic landscape of glioblastoma. Cell. 2013 Oct 10;155(2):462-77. doi: 10.1016/j.cell.2013.09.034.
59. The Cancer Genome Atlas Network., Genome sequencing centre: Broad Institute., Lawrence, M. *et al.* Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 517, 576–582 (2015). URL: https://doi.org/10.1038/nature14129
60. The Cancer Genome Atlas Research Network., Analysis working group: Baylor College of Medicine., Creighton, C. *et al.* Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499, 43–49 (2013). https://doi.org/10.1038/nature12222
61. Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, Buhay C, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell. 2014 Sep 8;26(3):319-330. doi: 10.1016/j.ccr.2014.07.014.
62. Cancer Genome Atlas Research Network, Linehan WM, Spellman PT, Ricketts CJ, Creighton CJ, Fei SS, Davis C, Wheeler DA, et al. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. N Engl J Med. 2016 Jan 14;374(2):135-45. doi: 10.1056/NEJMoa1505917.
63. Cancer Genome Atlas Research N, Brat DJ, Verhaak RG, Aldape KD, Yung WK, Salama SR, Cooper LA, Rheinbay E, Miller CR, Vitucci M, Morozova O, Robertson AG, Noushmehr H, et al. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. N Engl J Med. 2015; 372:2481–2498.
64. Cancer Genome Atlas Research Network, Ley TJ, Miller C, Ding L, Raphael BJ, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med. 2013 May 30;368(22):2059-74. doi: 10.1056/NEJMoa1301689. Epub 2013 May 1. Erratum in: N Engl J Med. 2013 Jul 4;369(1):98.
65. The Cancer Genome Atlas Research Network., Disease analysis working group., Collisson, E. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511, 543–550 (2014). URL: https://doi.org/10.1038/nature13385
66. The Cancer Genome Atlas Research Network., Genome sequencing centres: Broad Institute., Hammerman, P. *et al.* Comprehensive genomic characterization of squamous

cell lung cancers. *Nature* 489, 519–525 (2012). URL: https://doi.org/10.1038/nature11404
67. Campbell, J., Alexandrov, A., Kim, J. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat Genet* 48, 607–616 (2016). URL: https://doi.org/10.1038/ng.3564
68. Cancer Genome Atlas Research Network. Electronic address: elizabeth.demicco@sinaihealthsystem.ca; Cancer Genome Atlas Research Network. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. Cell. 2017 Nov 2;171(4):950-965.e28. doi: 10.1016/j.cell.2017.10.014.
69. Hmeljak J, Sanchez-Vega F, Hoadley KA, Shih J, Stewart C, Heiman D, et al. Integrative Molecular Characterization of Malignant Pleural Mesothelioma. Cancer Discov. 2018 Dec;8(12):1548-1565. doi: 10.1158/2159-8290.CD-18-0804.
70. The Cancer Genome Atlas Research Network., (Participants are arranged by area of contribution and then by institution.)., Disease working group and tissue source sites. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615 (2011). URL: https://doi.org/10.1038/nature10166
71. Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu; Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell. 2017 Aug 14;32(2):185-203.e13. doi: 10.1016/j.ccell.2017.07.007.
72. Fishbein L, Leshchiner I, Walter V, Danilova L, Robertson AG, Johnson AR, et al. Cancer Genome Atlas Research Network, Pacak K, Nathanson KL, Wilkerson MD. Comprehensive Molecular Characterization of Pheochromocytoma and Paraganglioma. Cancer Cell. 2017 Feb 13;31(2):181-193. doi: 10.1016/j.ccell.2017.01.001.
73. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015 Nov 5;163(4):1011-25. doi: 10.1016/j.cell.2015.10.025.
74. Cancer Genome Atlas Research Network. Electronic address: wheeler@bcm.edu; Cancer Genome Atlas Research Network. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017 Jun 15;169(7):1327-1341.e23. doi: 10.1016/j.cell.2017.05.046.
75. Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. Cell. 2015 Jun 18;161(7):1681-96. doi: 10.1016/j.cell.2015.05.044.
76. The Cancer Genome Atlas Research Network., Analysis working group: The University of Texas MD Anderson Cancer Center., Weinstein, J. *et al.* Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* 507, 315–322 (2014). URL: https://doi.org/10.1038/nature12965
77. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. Cell. 2014 Oct 23;159(3):676-90. doi: 10.1016/j.cell.2014.09.050.

78. Shen H, Shih J, Hollern DP, Wang L, Bowlby R, Tickoo SK, Thorsson V, et al. Integrated Molecular Characterization of Testicular Germ Cell Tumors. Cell Rep. 2018 Jun 12;23(11):3392-3406. doi: 10.1016/j.celrep.2018.05.039.
79. Radovich M, Pickering CR, Felau I, Ha G, Zhang H, Jo H, Hoadley KA, Anur P, et al. The Integrated Genomic Landscape of Thymic Epithelial Tumors. Cancer Cell. 2018 Feb 12;33(2):244-258.e10. doi: 10.1016/j.ccell.2018.01.003.
80. Levine, D., The Cancer Genome Atlas Research Network., Genome sequencing centres: Broad Institute. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* 497, 67–73 (2013). URL: https://doi.org/10.1038/nature12113
81. Cherniack AD, Shen H, Walter V, Stewart C, Murray BA, Bowlby R, et al. Cancer Genome Atlas Research Network, Weinstein JN, Zhang J, Akbani R, Levine DA. Integrated Molecular Characterization of Uterine Carcinosarcoma. Cancer Cell. 2017 Mar 13;31(3):411-423. doi: 10.1016/j.ccell.2017.02.010.
82. Robertson AG, Shih J, Yau C, Gibb EA, Oba J, Mungall KL, Hess JM, et al. Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. Cancer Cell. 2017 Aug 14;32(2):204-220.e15. doi: 10.1016/j.ccell.2017.07.003.
83. Kircher, M., Witten, D., Jain, P. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46, 310–315 (2014). URL: https://doi.org/10.1038/ng.2892
84. Shuai, S., Abascal, F., Amin, S.B. *et al.* Combined burden and functional impact tests for cancer driver discovery using DriverPower. *Nat Commun* 11, 734 (2020). URL: https://doi.org/10.1038/s41467-019-13929-1
85. Sondka, Z., Bamford, S., Cole, C.G. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* 18, 696–705 (2018). URL: https://doi.org/10.1038/s41568-018-0060-1
86. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012 May;2(5):401-4. doi: 10.1158/2159-8290.CD-12-0095.
87. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013 Apr 2;6(269):pl1. doi: 10.1126/scisignal.2004088.
88. Chang SH, Elemento O, Zhang J, Zhuang ZW, Simons M, Hla T. ELAVL1 regulates alternative splicing of eIF4E transporter to promote postnatal angiogenesis. Proc Natl Acad Sci U S A. 2014 Dec 23;111(51):18309-14.doi: 10.1073/pnas.1412172111.
89. Zhu, C., Yang, Q., Xu, J. *et al.* Somatic mutation of DNAH genes implicated higher chemotherapy response rate in gastric adenocarcinoma patients. *J Transl Med* 17, 109 (2019). URL: https://doi.org/10.1186/s12967-019-1867-6

90. Naotaka Nakazawa, Aneesh R. Sathe, G. V. Shivashankar, Michael P. Sheetz. Matrix mechanics controls FHL2 movement to the nucleus to activate p21 expression. PNAS November 1, 2016 113 (44) E6813-E6822. https://doi.org/10.1073/pnas.1608210113
91. Kleiber K, Strebhardt K, Martin BT. The biological relevance of FHL2 in tumour cells and its role as a putative cancer target. Anticancer Res. 2007 Jan-Feb;27(1A):55-61.
92. Adams, D.R., Ron, D. & Kiely, P.A. RACK1, A multifaceted scaffolding protein: Structure and function. *Cell Commun Signal* 9, 22 (2011). URL: https://doi.org/10.1186/1478-811X-9-22
93. Cox EA, Bennin D, Doan AT, O'Toole T, Huttenlocher A. RACK1 regulates integrin-mediated adhesion, protrusion, and chemotactic cell migration via its Src-binding site. Mol Biol Cell. 2003 Feb;14(2):658-69. doi: 10.1091/mbc.e02-03-0142.
94. Lee WK, Lee SY, Choi JE, Seok Y, Lee EB, Lee HC, Kang HG, Yoo SS, Lee MH, Cho S, Jheon S, Kim YC, Oh IJ, Na KJ, Jung CY, Park CK, Kim MH, Lee MK, Park JY. Development of a prognosis-prediction model incorporating genetic polymorphism with pathologic stage in stage I non-small cell lung cancer: A multicenter study. Thorac Cancer. 2017 May; 8(3): 251-259. doi: 10.1111/1759-7714.12434.
95. Grossmann, P., Cristea, S. & Beerenwinkel, N. Clonal evolution driven by superdriver mutations. *BMC Evol Biol* 20, 89 (2020). URL: https://doi.org/10.1186/s12862-020-01647-y
96. Gkouveris I, Nikitakis NG, Aseervatham J, Ogbureke KUE (2018). The tumorigenic role of DSPP and its potential regulation of the unfolded protein response and ER stress in oral Cancer cells. *International Journal of Oncology* 53:1743–1751.
97. Laczmanska I, Karpinski P, Gil J, Laczmanski L, Bebenek M, Sasiadek MM. High PTPRQ Expression and Its Relationship to Expression of PTPRZ1 and the Presence of KRAS Mutations in Colorectal Cancer Tissues. Anticancer Res. 2016 Feb; 36(2): 677-81.
98. Escudero-Esparza, A., Bartoschek, M., Gialeli, C., Okroj, M., Owen, S., Jirström, K., Orimo, A., Jiang, W. G., Pietras, K., & Blom, A. M. (2016). Complement inhibitor CSMD1 acts as tumor suppressor in human breast cancer. *Oncotarget*, *7*(47), 76920–76933. URL: https://doi.org/10.18632/oncotarget.12729
99. Labeit, S., Ottenheijm, C. A., & Granzier, H. (2011). Nebulin, a major player in muscle health and disease. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, *25*(3), 822–829. https://doi.org/10.1096/fj.10-157412
100. Green, D., Eyre, H., Singh, A. *et al.* Targeting the MAPK7/MMP9 axis for metastasis in primary bone cancer. *Oncogene* 39, 5553–5569 (2020). URL: https://doi.org/10.1038/s41388-020-1379-0
101. Shen, S, Li, H, Liu, J, Sun, L, Yuan, Y. The panoramic picture of pepsinogen gene family with pan-cancer. *Cancer Med*. 2020; 9: 9064– 9080. URL: https://doi.org/10.1002/cam4.3489
102. Tedeschi PM, Vazquez A, Kerrigan JE, Bertino JR. Mitochondrial Methylenetetrahydrofolate Dehydrogenase (MTHFD2) Overexpression Is Associated

with Tumor Cell Proliferation and Is a Novel Target for Drug Development. Mol Cancer Res. 2015 Oct;13(10):1361-6. doi: 10.1158/1541-7786.MCR-15-0117.

103. Shen, L., Shi, Q. & Wang, W. Double agents: genes with both oncogenic and tumor-suppressor functions. *Oncogenesis* 7, 25 (2018). URL: https://doi.org/10.1038/s41389-018-0034-x.

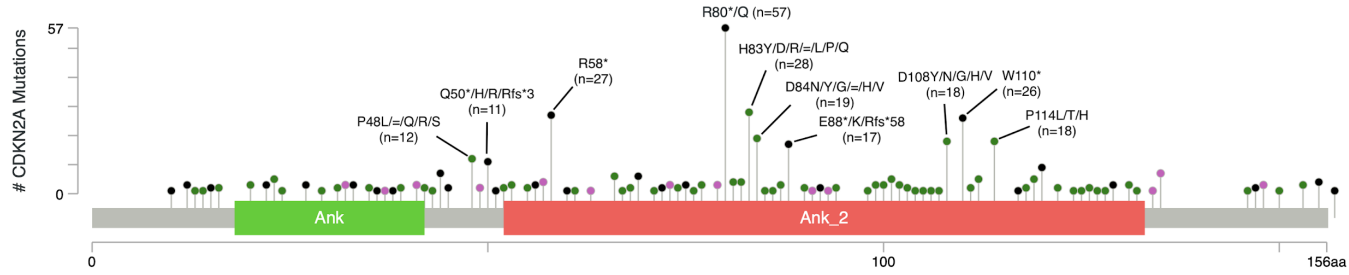
SUPPLEMENTARY INFORMATION

Supplementary table 1. Data sets information including tumor type descriptions, tumor abbreviations, number of samples, reference genome, and institute source. Table available at https://github.com/caeareva/pcadmm/blob/de7da87a424863b29b3ff2b3f5886dd99f2b27e0/Drivers%20supplemental%20table%201.xlsx.

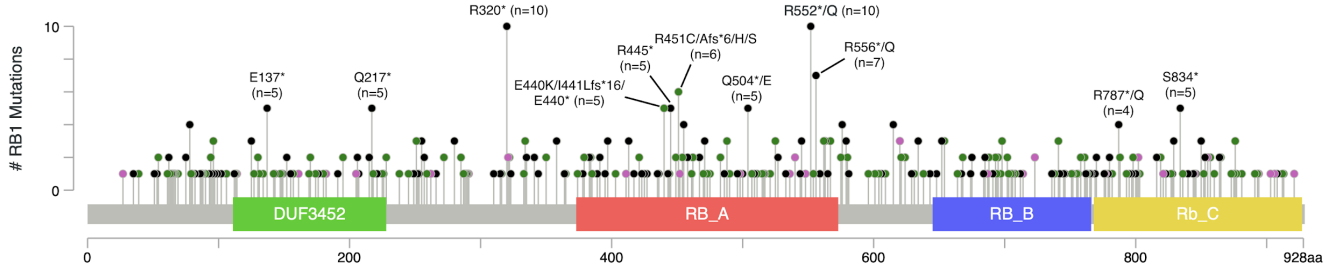
Supplementary figure 1. **Mutation distribution in most mutated driver genes. A.** Needle plots for *CDKN2A, RB1*, *NRAS*, *CASP8*, and *STK11*, predicted among the most significant drivers in our TCGA and CCLE pan-cancer analysis using *oncodriveFML*.

A.

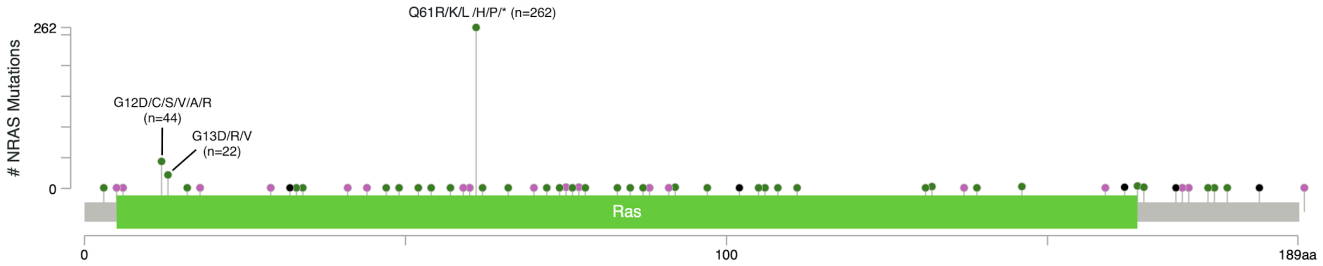
CDKN2A mutations (n= 450)



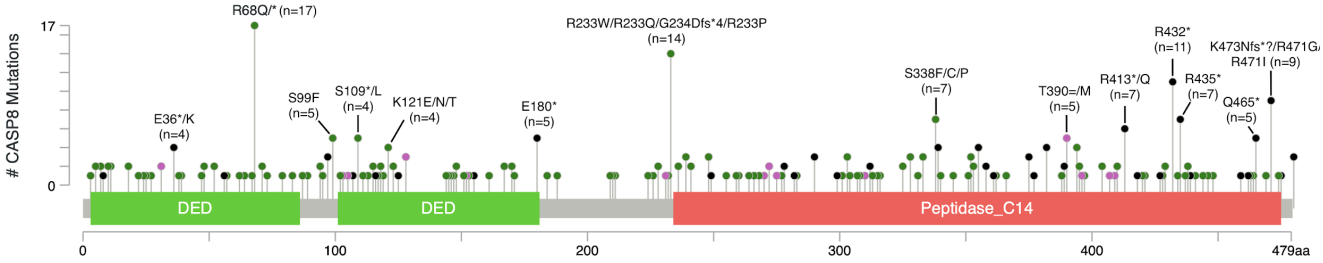
RB1 mutations (n=447)



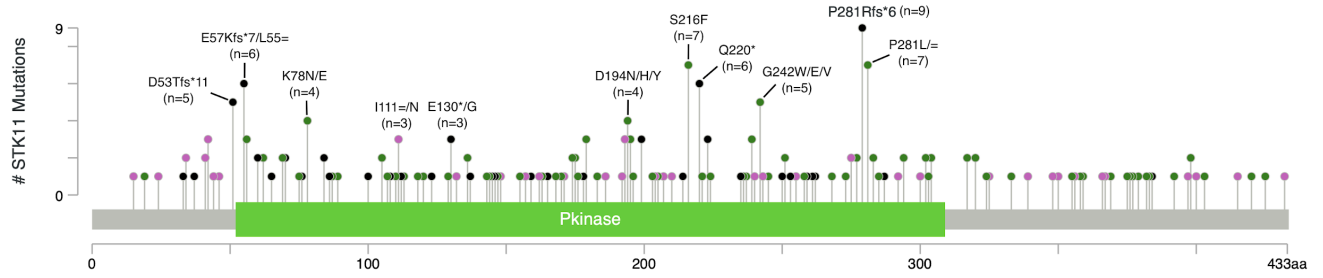
NRAS mutations (n=405)



CASP8 mutations (n=310)



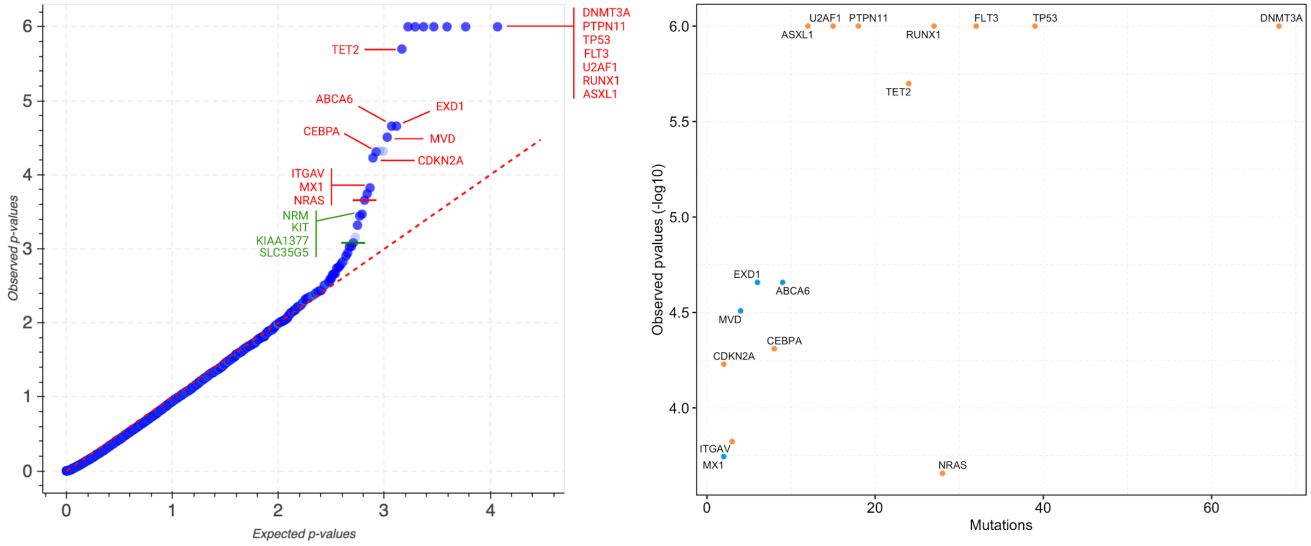
STK11 mutations (n=215)



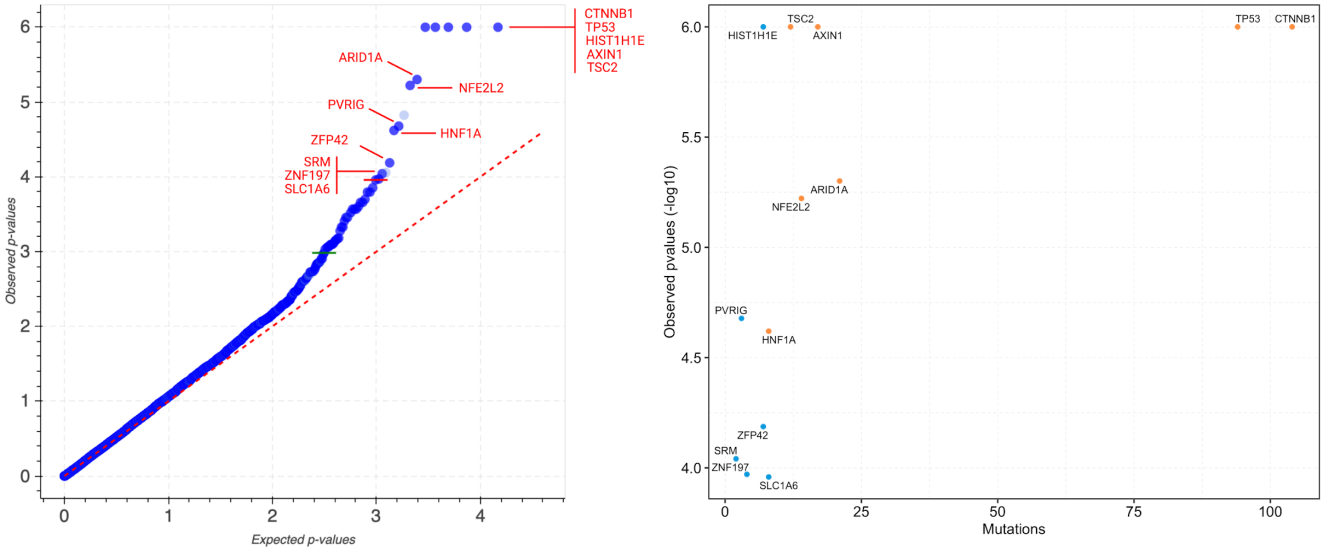
Supplementary figure 2. **Predicted driver genes landscape of 26 least mutated cancer types with at least one significant driver gene. A.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel) showing the mutation distribution of cancer types versus observed p-value for **LAML**. Only tumors that had at list one significant driver gene identified by *OncodriveFML* were included. Red color indicates genes with q-value < 0.1 and green color indicates genes with q-value < 0.25. **B.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel) showing the mutation distribution of cancer types versus observed p-value for **LIHC**. **C.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel) showing the mutation distribution of cancer types versus observed p-value for **LGG**. **D.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel) showing the mutation distribution of cancer types versus observed p-value for PAAD. **E.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel) showing the mutation distribution of cancer types versus observed p-value for KIRC. **F.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel) showing the mutation distribution of cancer types versus observed p-value for UCS. **G.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel) showing the mutation distribution of cancer types versus observed p-value for GBM. **H.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel)

showing the mutation distribution of cancer types versus observed p-value for ESCA. **I.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel) showing the mutation distribution of cancer types versus observed p-value for CHOL. **J.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel) showing the mutation distribution of cancer types versus observed p-value for SARC. **K.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel) showing the mutation distribution of cancer types versus observed p-value for OV. **L.** QQ plot (left panel) comparing the expected and observed distribution of functional mutational bias p-values of genes, and scatter plot (right panel) showing the mutation distribution of cancer types versus observed p-value for MESO. **M.** QQ plots for rest of cancer types containing less than 3 significant (q-value < 0.1) driver genes including UVM (n=2), TGCT (n2), SCLC (n=2), NB (n=2), DLBC (n=2), ALL (n=2), ACC (n=2), THCA (n=1), PRAD (n=1), PCPG (n=1), MM (n=1), NHL (n=1), KIRP (n=1), and KICH (n=1).

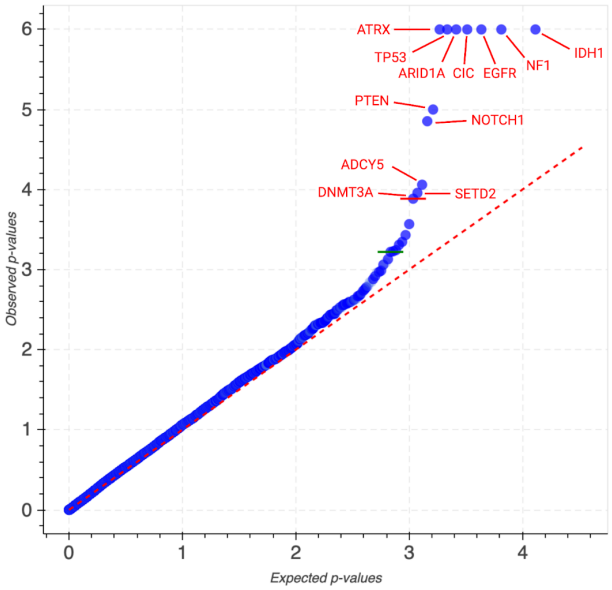
A.

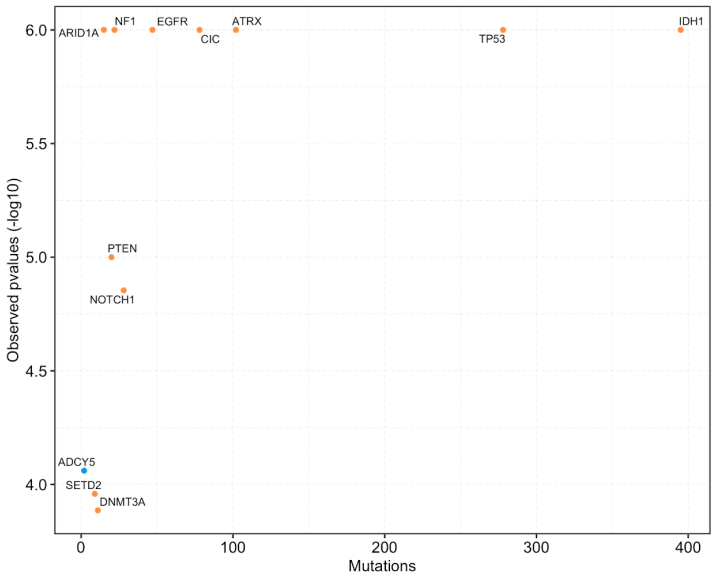

DNMT3A
PTPN11
TP53
FLT3
U2AF1
RUNX1
ASXL1
TET2
ABCA6
EXD1
CEBPA
MVD
CDKN2A
ITGAV
MX1
NRAS
NRM
KIT
KIAA1377
SLC35G5
Observed p-values
Expected p-values
Observed pvalues (-log10)
Mutations

B.

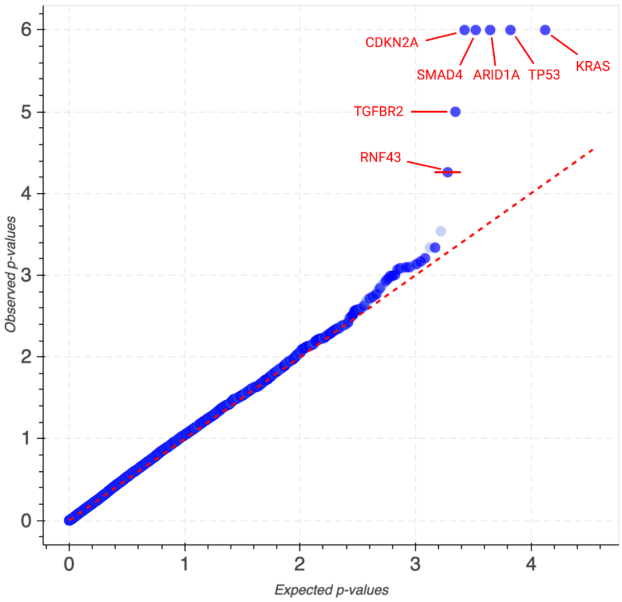

CTNNB1
TP53
HIST1H1E
AXIN1
TSC2
ARID1A
NFE2L2
PVRIG
HNF1A
ZFP42
SRM
ZNF197
SLC1A6
Observed p-values
Expected p-values
Observed pvalues (-log10)
Mutations

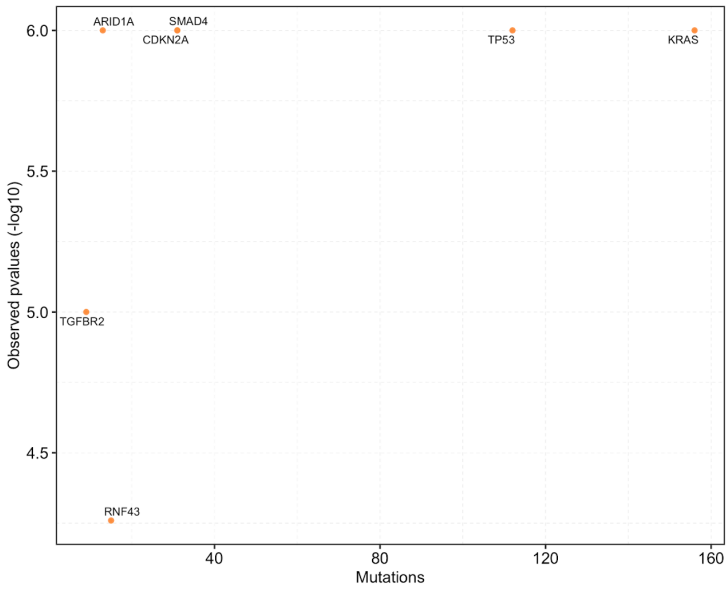
C.


ATRX
TP53
ARID1A
CIC
EGFR
NF1
IDH1
PTEN
NOTCH1
ADCY5
DNMT3A
SETD2
Observed p-values
Expected p-values

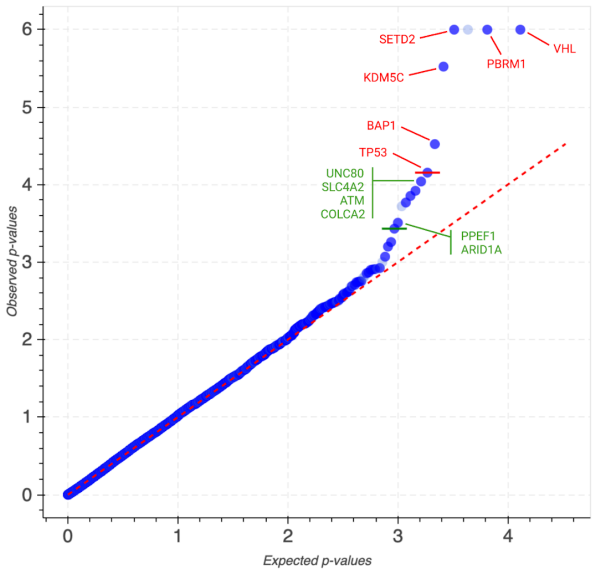

ARID1A
NF1
EGFR
CIC
ATRX
TP53
IDH1
PTEN
NOTCH1
ADCY5
SETD2
DNMT3A
Observed pvalues (-log10)
Mutations

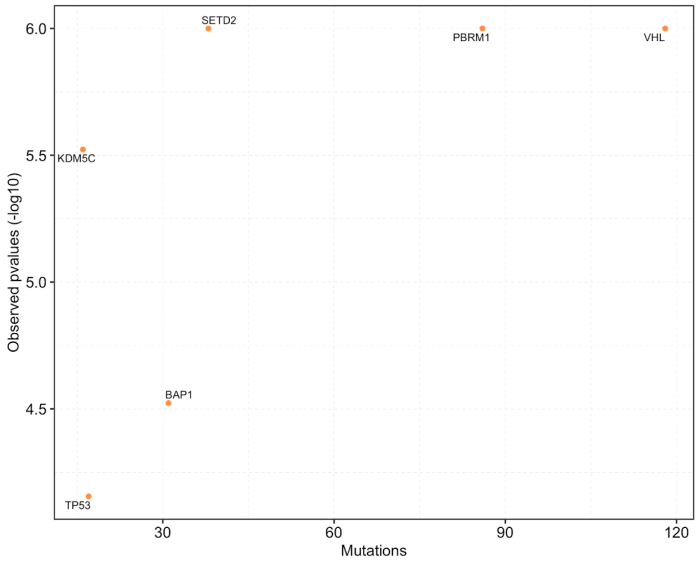
D.


CDKN2A
SMAD4
ARID1A
TP53
KRAS
TGFBR2
RNF43
Observed p-values
Expected p-values

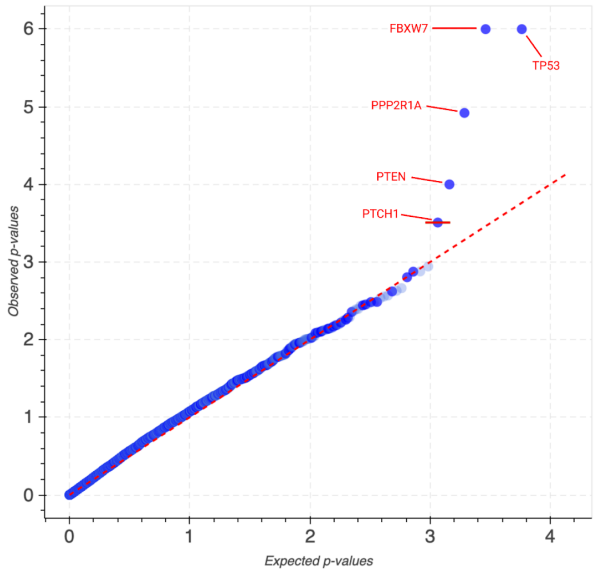

ARID1A
SMAD4
CDKN2A
TP53
KRAS
TGFBR2
RNF43
Observed pvalues (-log10)
Mutations

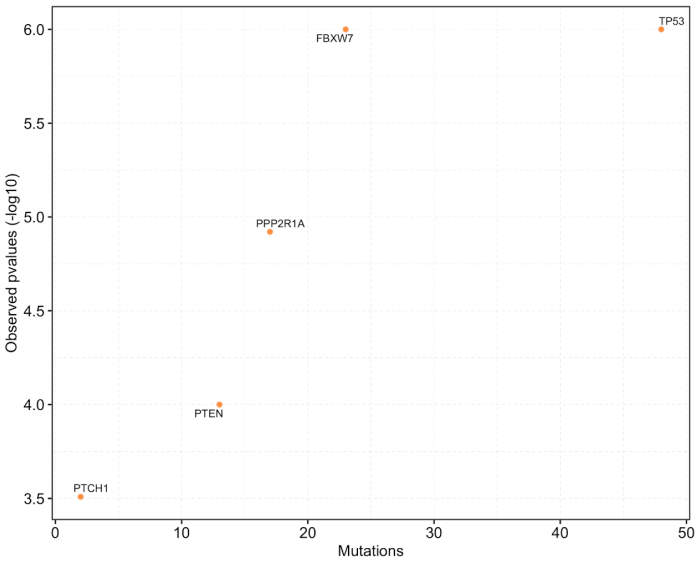
E.


SETD2
VHL
PBRM1
KDM5C
BAP1
TP53
UNC80
SLC4A2
ATM
COLCA2
PPEF1
ARID1A
Observed p-values
Expected p-values

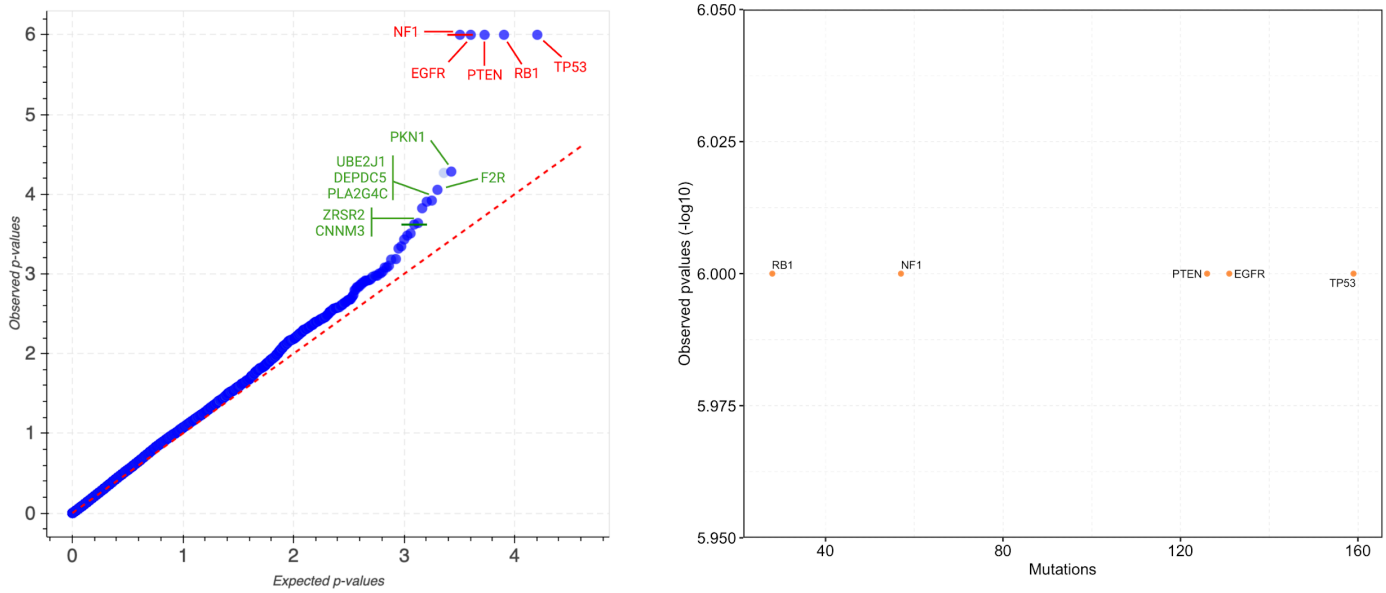

SETD2
PBRM1
VHL
KDM5C
BAP1
TP53
Observed pvalues (-log10)
Mutations

F.

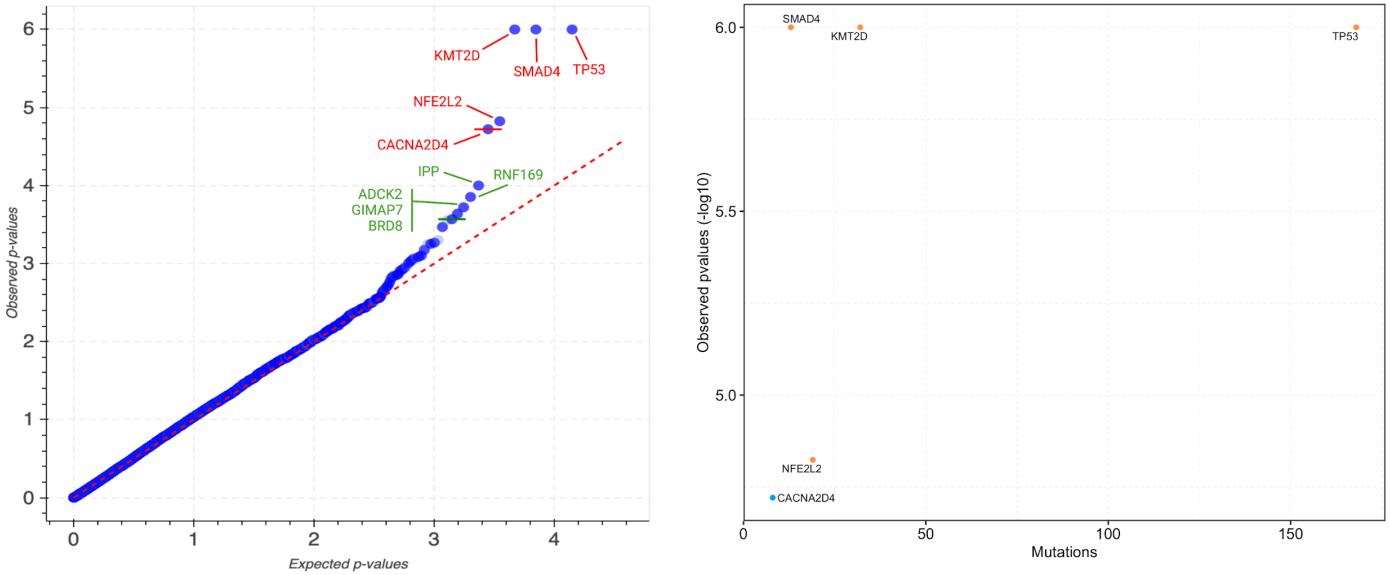

FBXW7
TP53
PPP2R1A
PTEN
PTCH1
Observed p-values
Expected p-values


TP53
FBXW7
PPP2R1A
PTEN
PTCH1
Observed pvalues (-log10)
Mutations

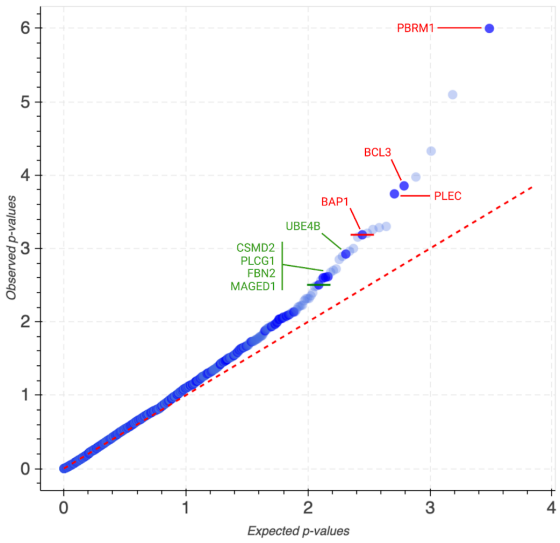
G.

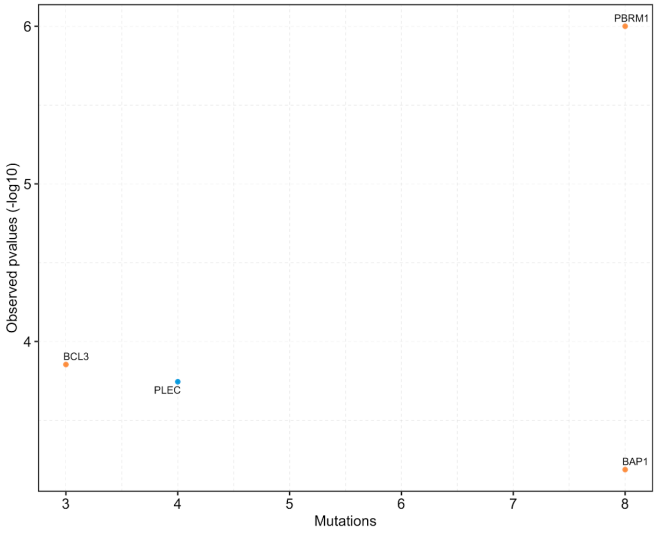

NF1
EGFR
PTEN
RB1
TP53
PKN1
UBE2J1
DEPDC5
PLA2G4C
F2R
ZRSR2
CNNM3
Observed p-values
Expected p-values
Observed pvalues (-log10)
Mutations
RB1
NF1
PTEN
EGFR
TP53

H.

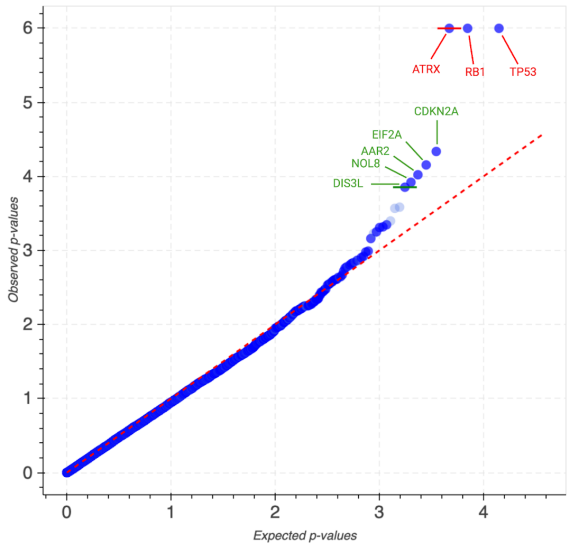

KMT2D
SMAD4
TP53
NFE2L2
CACNA2D4
IPP
RNF169
ADCK2
GIMAP7
BRD8
Observed p-values
Expected p-values
Observed pvalues (-log10)
Mutations
SMAD4
KMT2D
TP53
NFE2L2
CACNA2D4

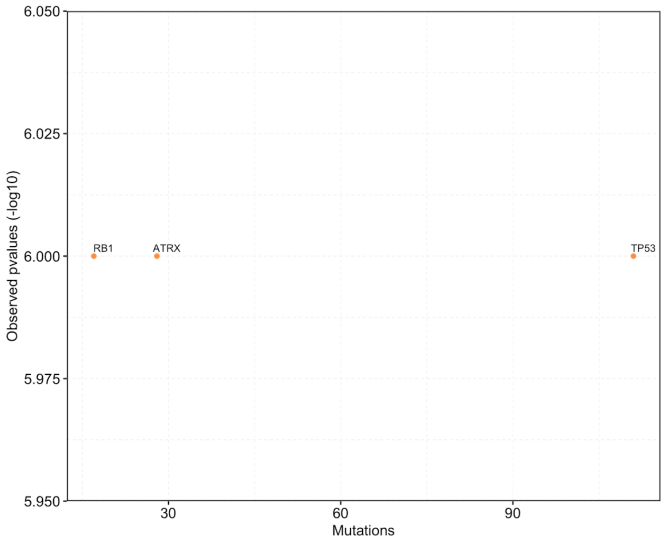
I.


PBRM1
BCL3
PLEC
BAP1
UBE4B
CSMD2
PLCG1
FBN2
MAGED1
Observed p-values
Expected p-values

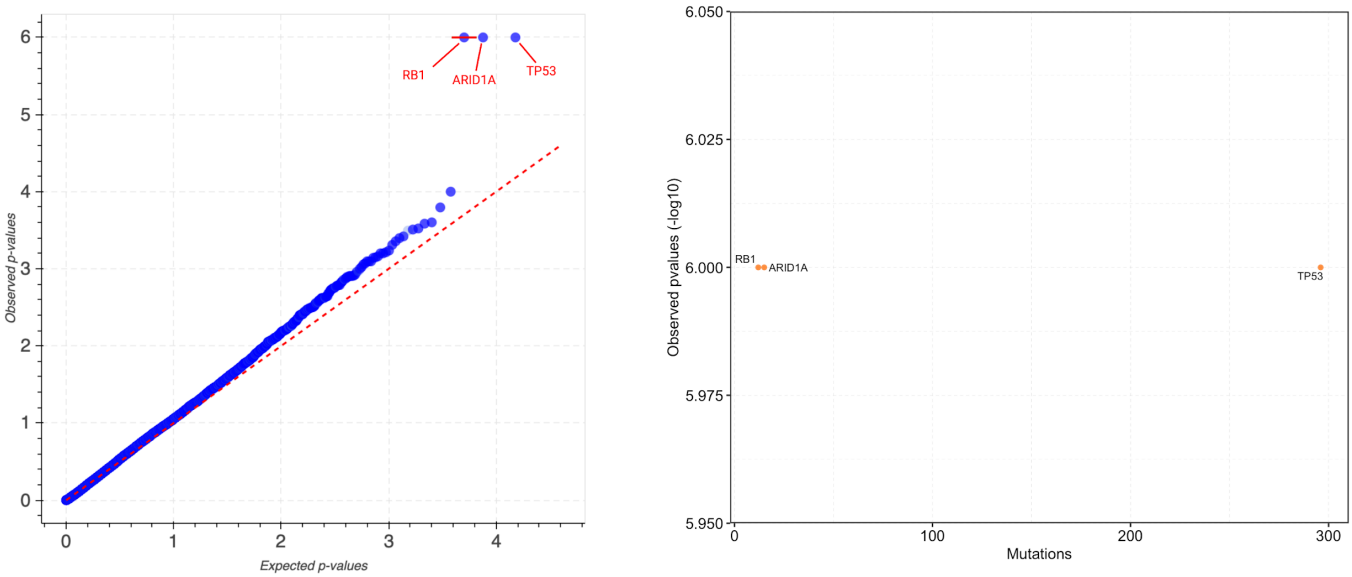

PBRM1
BCL3
PLEC
BAP1
Observed pvalues (-log10)
Mutations

J.

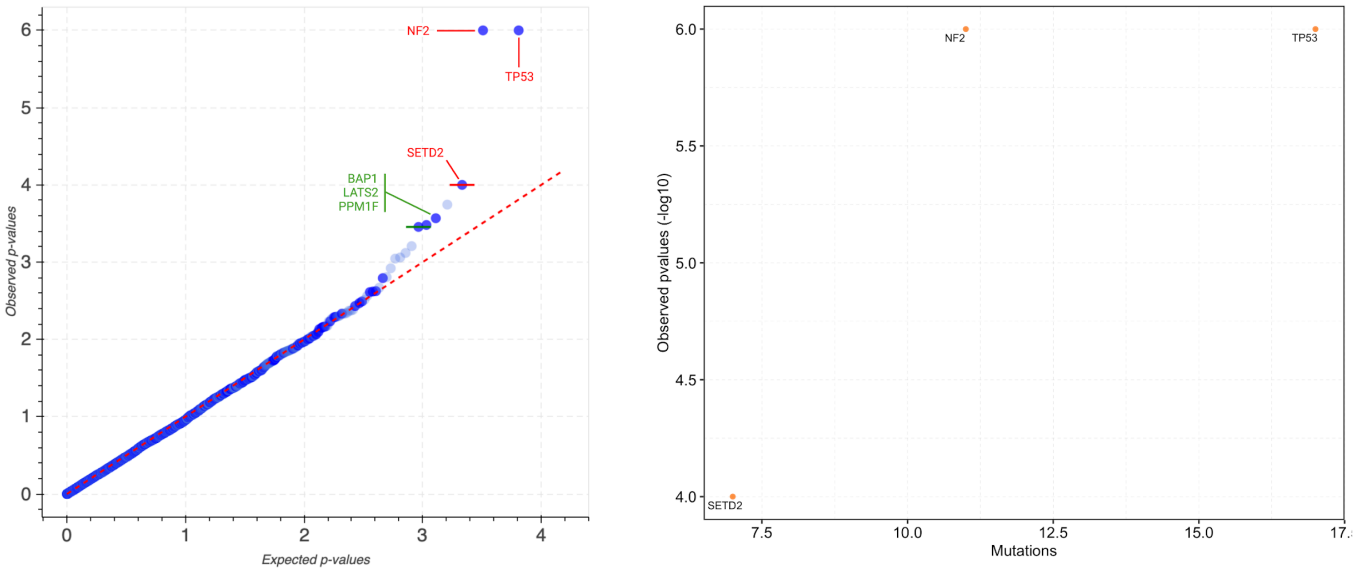

ATRX
RB1
TP53
CDKN2A
EIF2A
AAR2
NOL8
DIS3L
Observed p-values
Expected p-values


RB1
ATRX
TP53
Observed pvalues (-log10)
Mutations

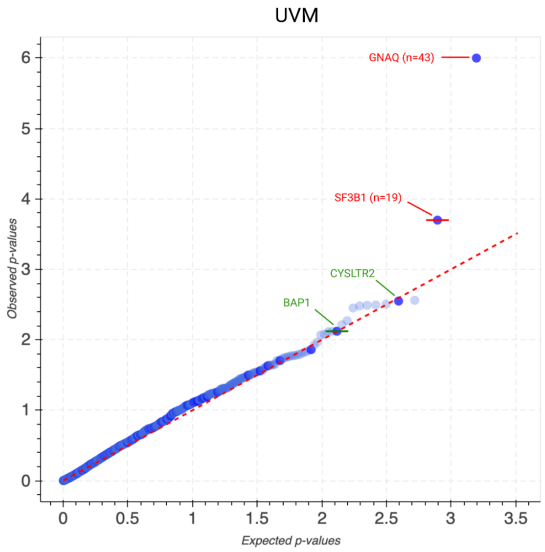
K.

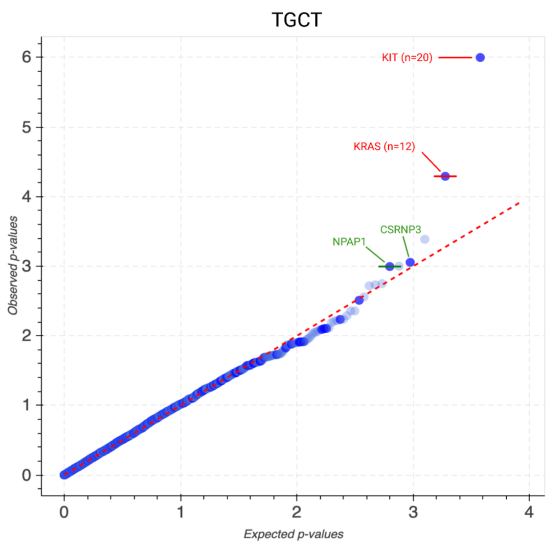

RB1
ARID1A
TP53
Observed p-values
Expected p-values
Observed pvalues (-log10)
Mutations

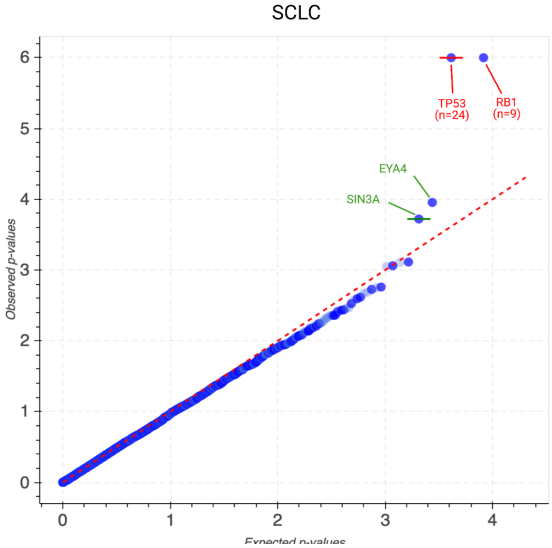
L.

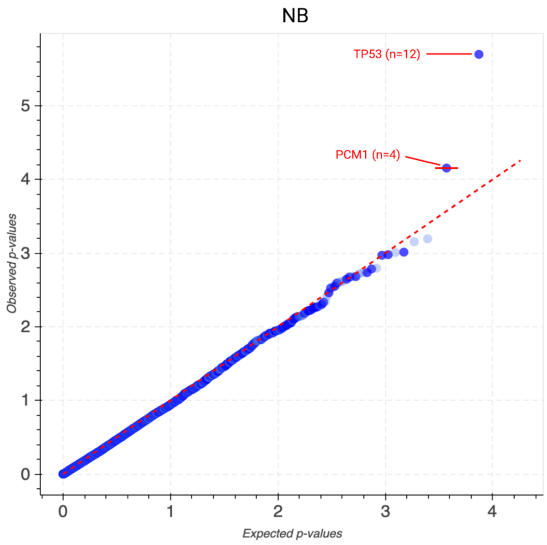

NF2
TP53
SETD2
BAP1
LATS2
PPM1F
Observed p-values
Expected p-values
Observed pvalues (-log10)
Mutations

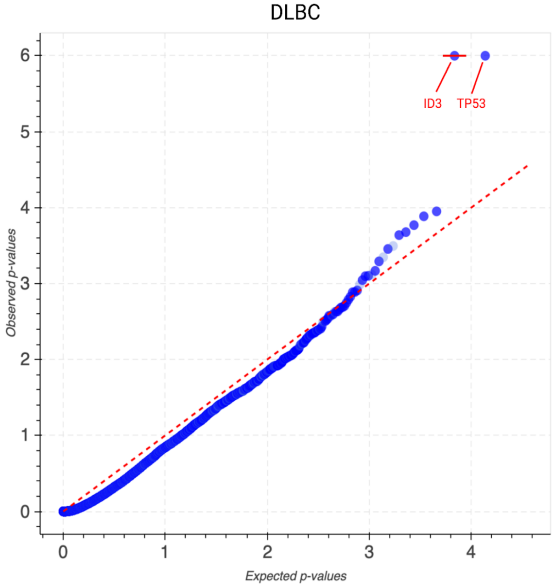
M.

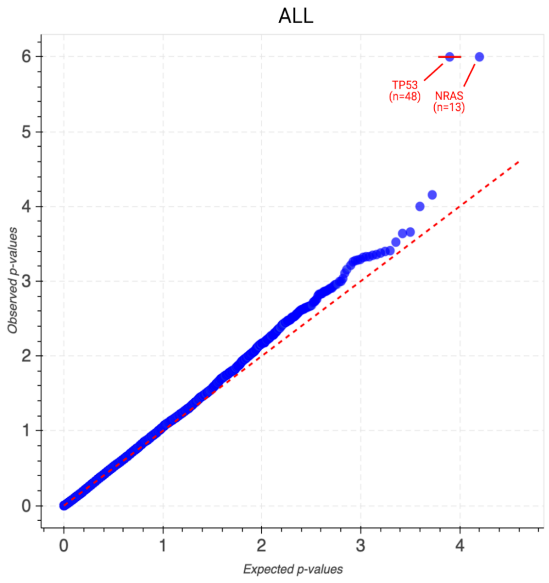

UVM
GNAQ (n=43)
SF3B1 (n=19)
CYSLTR2
BAP1
Observed p-values
Expected p-values

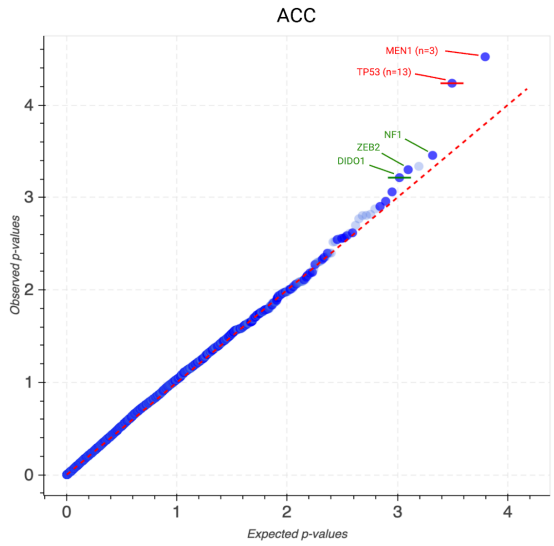

TGCT
KIT (n=20)
KRAS (n=12)
CSRNP3
NPAP1
Observed p-values
Expected p-values

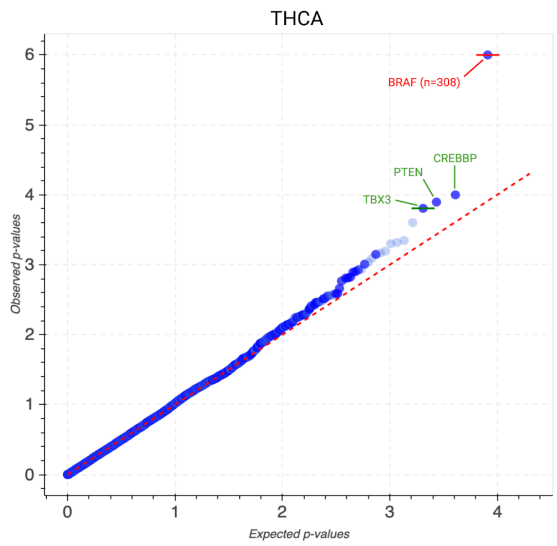

SCLC
TP53 (n=24)
RB1 (n=9)
EYA4
SIN3A
Observed p-values

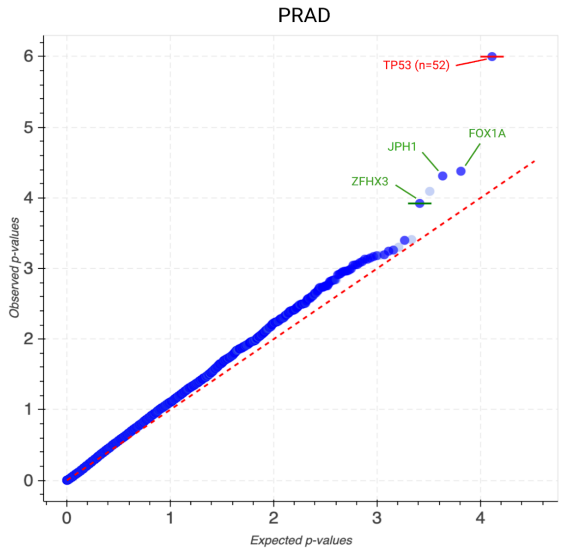

NB
TP53 (n=12)
PCM1 (n=4)
Observed p-values
Expected p-values

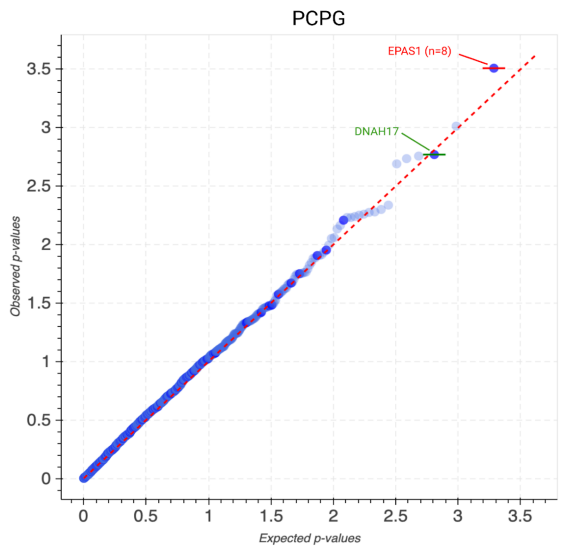

DLBC
ID3
TP53
Observed p-values
Expected p-values

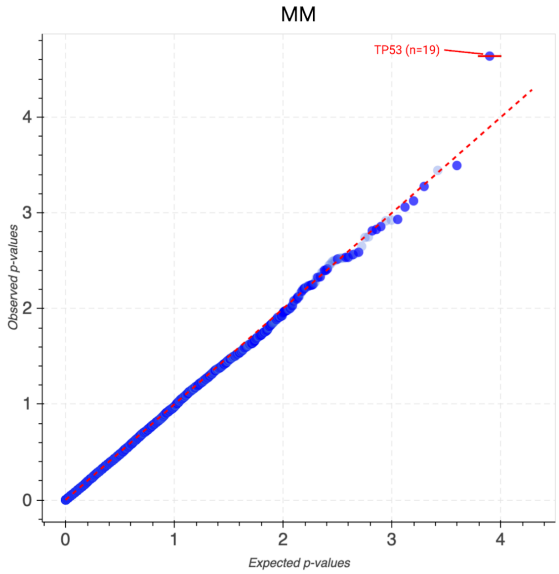

ALL
TP53 (n=48)
NRAS (n=13)
Observed p-values
Expected p-values

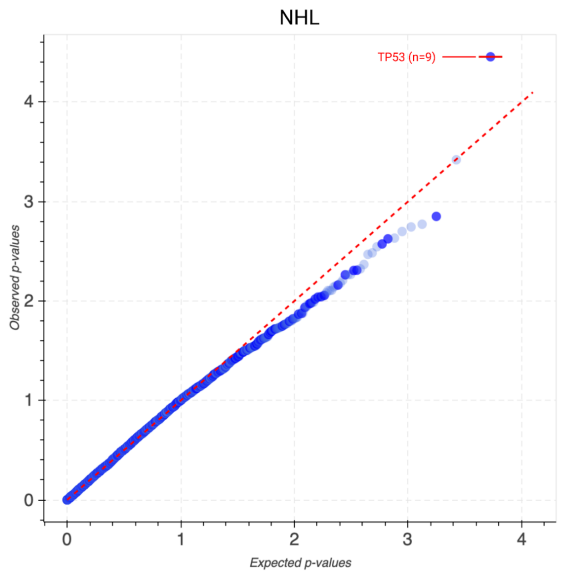

ACC
MEN1 (n=3)
TP53 (n=13)
NF1
ZEB2
DIDO1
Observed p-values
Expected p-values

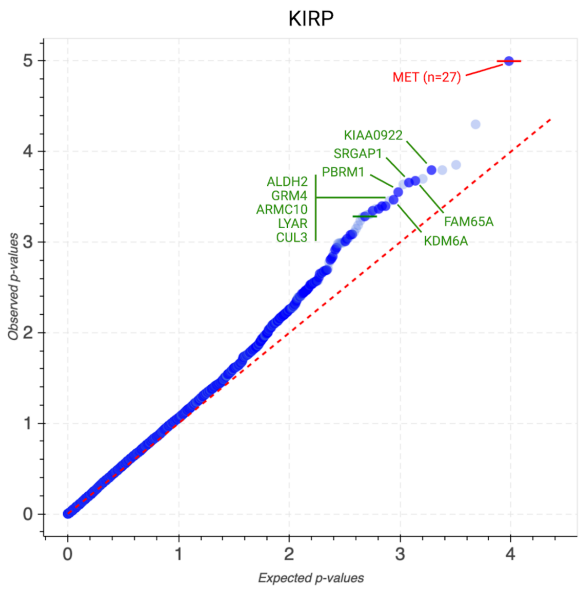

THCA
BRAF (n=308)
CREBBP
PTEN
TBX3
Observed p-values
Expected p-values

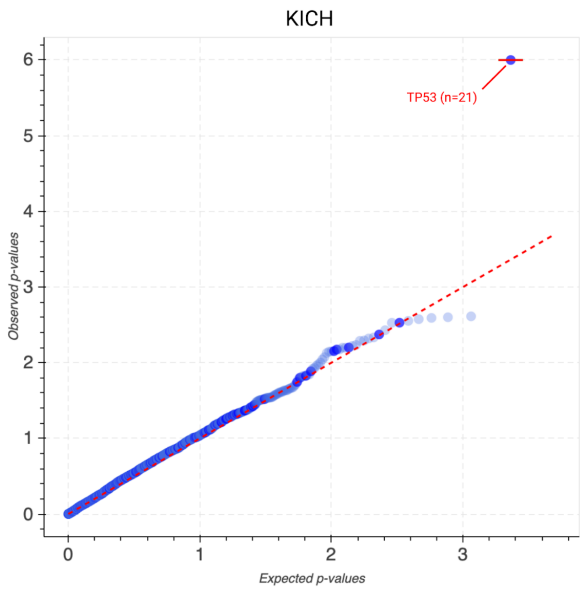

PRAD
TP53 (n=52)
FOX1A
JPH1
ZFHX3
Observed p-values
Expected p-values


PCPG
EPAS1 (n=8)
DNAH17
Observed p-values
Expected p-values


MM
TP53 (n=19)
Observed p-values
Expected p-values


NHL
TP53 (n=9)
Observed p-values
Expected p-values


KIRP
MET (n=27)
KIAA0922
SRGAP1
PBRM1
ALDH2
GRM4
ARMC10
LYAR
CUL3
FAM65A
KDM6A
Observed p-values
Expected p-values
0
1
2
3
4
5


KICH
TP53 (n=21)
Observed p-values
Expected p-values
0
1
2
3
4
5
6