Analiza povezanosti somatskih varijanti tumora s mutacijma i mutacijskim potpisima

Tomljanović, Ingrid

UNIVERSITY OF ZAGREB

FACULTY OF SCIENCE

DEPARTMENT OF BIOLOGY

INGRID TOMLJANOVIĆ

Association analysis of somatic tumor variants with mutations and mutational signatures

Analiza povezanosti somatskih varijanti tumora s mutacijama i mutacijskim potpisima

Graduation Thesis

Zagreb, 2019.

This graduation thesis was conducted at The Institute for Research in Biomedicine, Barcelona, Spain, under the supervision of ICREA Research Professor Fran Supek, PhD, and co-supervision of Asst. Prof. Dubravko Pavoković, PhD. The thesis was submitted for evaluation to the Department of Biology at the Faculty of Science, University of Zagreb with the aim of obtaining the title of Master in Molecular Biology.

ACKNOWLEDGEMENTS

I thank my thesis supervisor, dr. sc. Fran Supek, for his scientific insight, wise advice and understanding.

I thank dr. sc. Dubravko Pavoković for his kindness, support and mentorship over the past five years. Thank you for encouraging me to grow since the moment I entered University and for being the strong person you are.

I thank Vlatka Marjan and Sanjica Mihaljević for their good-heartedness and for helping me and numerous other biology students get through college.

I thank my mother, grandfather and grandmother for supporting me with their immense love in every way they possibly could.

I thank Zorana, Alka, Lejla and Henna for being true friends, as well as Jurica and Nevenka for becoming a part of my life in Barcelona.

I thank Hrvoje for being my best friend.

BASIC DOCUMENTATION CARD

University of Zagreb

Faculty of Science

Department of Biology

Graduation Thesis

Association analysis of somatic tumor variants with mutations and mutational signatures

Ingrid Tomljanović

Rooseveltov trg 6, 10000 Zagreb, Croatia

Throughout the lifetime of a patient, the genome of a cancer cell gradually accumulates somatic mutations due to a variety of endogenous and exogenous mutational processes. The genomic fingerprints of those processes can be deconvoluted from tumor somatic mutation data and are termed mutational signatures. The goal of this master thesis was to assess the relationships of a comprehensive set of somatic mutations belonging to a list of 721 selected genes with the 30 mutational signatures and their proxies across 33 different cancer types. Two-sided nonparametric testing was employed on both pan-cancer and per-cancer type levels to individually test recurrent somatic mutations and to perform burden testing of recurrently mutated genes based on the TCGA MC3 dataset comprising ~10,000 samples. Analyses were stratified by sample hypermutation status and predicted benign or pathogenic mutation character. Results were submitted to genomic control and Benjamini-Hochberg multiple testing correction procedures and filtered using absolute effect sizes. Individual mutation testing identified 62 significant associations, while burden testing identified 162 significant associations on the per-cancer type level across data subsets. The majority of identified associations were novel findings and constitute potentially valuable targets for future research into the mutational processes operative in cancer cells.

(66 pages, 4 figures, 14 tables, 66 references, original in: English)

Thesis deposited in the Central Biological Library.

Key words: mutational signatures, TCGA, somatic mutations, driver genes, COSMIC, burden testing, genomic control

Supervisor / Co-supervisor: ICREA Research Professor Fran Supek, PhD, Asst. Prof. Dubravko Pavoković, PhD

Reviewers: Asst. Prof. Dubravko Pavoković, PhD, Assoc. Prof. Antun Alegro, PhD, Asst. Prof. Marin Ježić, PhD

Thesis accepted: February 28th, 2019

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odsjek

Diplomski rad

Analiza povezanosti somatskih varijanti tumora s mutacijama i mutacijskim potpisima

Ingrid Tomljanović

Rooseveltov trg 6, 10000 Zagreb, Croatia

Tijekom života pacijenta, genom stanice raka postupno nakuplja somatske mutacije uslijed djelovanja raznih endogenih i egzogenih mutacijskih procesa. Genomske otiske ovih procesa moguće je razlučiti iz podataka o somatskim mutacijama tumora te ih nazivamo mutacijskim potpisima. Cilj ovog diplomskog rada bio je proučiti povezanost opsežnog skupa somatskih mutacija koje pripadaju 721 odabranom genu sa 30 mutacijskih potpisa i zamjenskih varijabli na 33 različita tipa raka. Dvostrano neparametarsko testiranje primijenjeno je na razini svih tipova raka te pojedinih tipova raka u svrhu testiranja pojedinih ponavljajućih somatskih mutacija te testiranja tereta mutacija u ponavljajuće mutiranim genima na osnovi podatkovnog seta TCGA MC3 koji sadrži ~10,000 uzoraka. Analize su podijeljene po statusu hipermutiranosti uzoraka te predviđenom dobroćudnom ili zloćudnom tipu mutacija. Rezultati su u svrhu ispravka višestrukog testiranja hipoteza podvrgnuti metodi genomske kontrole te Benjamini-Hochbergovoj metodi i filtrirani po apsolutnoj vrijednosti veličine učinka. Testiranjem pojedinih mutacija otkrivene su 62 značajne poveznice, dok su testiranjem tereta mutacija otkrivene 162 značajne poveznice na razini pojedinih tipova raka i svim podskupovima podataka. Većina otkrivenih poveznica predstavljaju nova otkrića i moguće vrijedne ciljeve budućih istraživanja mutacijskih procesa aktivnih u stanicama raka.

(66 stranica, 4 slike, 14 tablica, 66 literaturnih navoda, jezik izvornika: engleski)

Rad je pohranjen u Središnjoj biološkoj knjižnici.

Ključne riječi: mutacijski potpisi, TCGA, somatske mutacije, pogonski geni, COSMIC, testiranje tereta, genomska kontrola

Voditelj / Suvoditelj: ICREA Research Professor Fran Supek, PhD, Asst. Prof. Dubravko Pavoković, PhD

Ocjenitelji: doc. dr. sc. Dubravko Pavoković, izv. prof. dr. sc. Antun Alegro, doc. dr. sc. Marin Ježić

Rad prihvaćen: 28. veljače, 2019.

LIST OF ABBREVIATIONS

BH: Benjamini-Hochberg

CGC: Cancer Gene Census

COSMIC: Catalogue of Somatic Mutations in Cancer

DDR: DNA damage repair

FDR: false discovery rate

GC: genomic control

GDC: Genomic Data Commons

ICGC: International Cancer Genome Consortium

MAF: Mutation Annotation Format

MC3: Multi-Center Mutation-Calling in Multiple Cancers

NGS: next-generation sequencing

NER: nucleotide-excision repair

NMF: non-negative matrix factorization

PCAWG: Pan-Cancer Analysis of Whole Genomes

SNP: single-nucleotide polymorphism

TCGA: The Cancer Genome Atlas

WES: whole-exome sequencing

WGA: whole-genome amplified

WGS: whole-genome sequencing

1. INTRODUCTION

1.1. CANCER GENOMICS

Cancer is a disease of the genome characterized by progressive acquisition of mutations in individual cells, which are subsequently subjected to the effects of natural selection (Vogelstein et al., 2013). As a rapidly advancing field, cancer genomics is focused on conducting comprehensive analyses of genomic data from both tumor and normal tissue in order to provide insights into cancer pathogenesis and guide efforts towards the establishment of novel treatment options (Garraway and Lander, 2013).

Until recently, systematic studies of cancer at the genomic level were inaccessible to researchers due to the technological and financial constraints related to high-throughput sequencing of whole exomes and whole genomes at an appropriate depth of coverage. However, the development of next generation sequencing (NGS) technologies in the past decade has resulted in vast amounts of publicly available DNA sequence information from large patient cohorts and a variety of different cancer types. NGS allows for detection of all classes of somatic alterations in a cancer genome, resulting in a mutational catalogue composed of point mutations, insertions, deletions, large genomic rearrangements and copy number changes. In addition, epigenomic, transcriptomic and germline variant data has been made available through use of NGS platforms. Collectively, the generated data has already provided detailed, numerous and clinically actionable insights into the oncogenic processes operative in human cells (Stratton, Campbell and Futreal, 2009).

Our understanding of cancer biology has been and continues to be strongly propelled by research findings stemming from collective initiatives such as the The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). TCGA was a large-scale project that started with a pilot in 2006 and officially took place from 2009 to 2015. It was launched by the US National Institutes of Health with a mission to aggregate information on alterations present in all tumor types. Around 10 million mutations from over 10,000 tumor-normal exome pairs involving 33 cancer types were provided to the scientific community through TCGA efforts (The future of cancer genomics, Nature Medicine Editorial, 2015).

TCGA data has been mined for multiple uses, including studies of tumor subtypes, pan-cancer characteristics and therapeutic resistance mechanisms. TCGA is currently conducting the largest set of cross-cancer type analyses thus far, integrated in form of the PanCancer Atlas project with an aim of rigorously characterizing the molecular features of cancers including cell-of-origin patterns, oncogenic processes and signaling pathways (Nawy, 2018). Current knowledge of these features is organized in a series of in-depth PanCan Atlas publications, with a flagship PanCan paper establishing that tumor classification based on nearly any data type will result in a clustering by cell-of-origin patterns, primarily tissue type, histology or anatomic origin (Hoadley et al., 2018). This highlights the importance of stratifying data by cancer (tissue) type in performing large-scale cancer studies. The basis of PanCancer Atlas publications is the MC3 dataset, a harmonized set of 3.5 million somatic variants from TCGA exome data that was produced by the Multi-Center Mutation-Calling in Multiple Cancers (MC3) network based on the output of seven different mutation-calling algorithms (Ellrott et al., 2018).

The ICGC was the second major initiative of this kind and was launched in 2008 as an international effort to conduct broad studies of 50 different cancer types and catalogue their genomic abnormalities. More than 25,000 cancer genomes were analyzed at the genomic, epigenomic and transcriptomic levels and assessed against clinical features in order to discover oncogenic mutations, mutational signatures and tumor subtypes, thus enabling development of new research lines and new cancer therapies. The ICGC is dedicated to a set of clearly defined goals, namely the coordinated generation and rapid dissemination of standardized high-quality data under the aegis of bioethical principles (Hudson et al., 2010).

Thus far, sequence data from over 50,000 cancers has been produced worldwide by leading entreprises such as the ICGC and TCGA, as well as many other smaller-scale projects, with whole exome sequencing (WES) being the prevalent platform for data accumulation and whole genome sequencing (WGS) becoming the main method of choice for future studies (Nakagawa and Fujita, 2018). The reported mutations have been manually curated in the Catalogue of Somatic Mutations in Cancer (COSMIC), the world's largest database of cancer somatic mutation information (Tate et al., 2018). The latest release of COSMIC, v86 from August 2018,

consists of 6 million coding mutations collected from 1.4 million tumour samples and curated from 26,000 research publications (https://cancer.sanger.ac.uk).

1.2. SOMATIC MUTATIONS IN CANCER

Throughout the lifetime of a patient, the genome of a cancer cell gradually accumulates changes from the diploid genome of its biologically normal ancestral cell, which is itself a descendant of the fertilized egg alongside all other cells in the patient's body (Figure 1). These genomic changes were named somatic mutations in order to differentiate them from germline mutations that get passed on from parents to offspring. Somatic mutations arise from a variety of intrinsic and extrinsic mutagens and can occur while the ancestral cell is still phenotypically normal as well as after clonal expansion begins. The consequences of a mutation on oncogenesis, i.e. on the selective advantages it may confer to a cell, allow us to distinguish between two basic types of somatic mutations: "drivers" and "passengers" (Stratton, Campbell and Futreal, 2009).

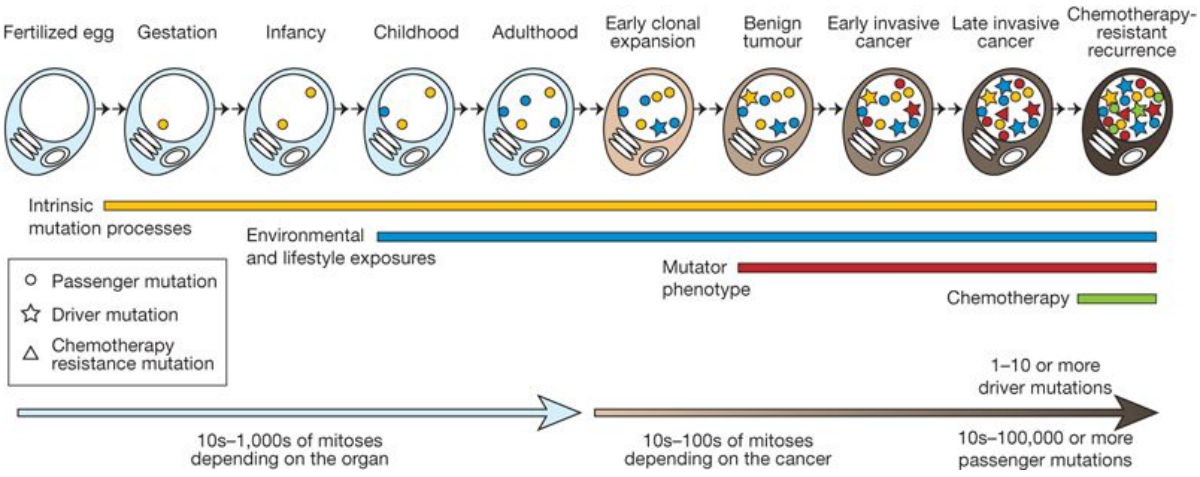

Figure 1. Cancer cell descendance from the fertilized egg through a lineage of mitotic cell divisions depicted with timing of mutation acquisition and the contributing mutagenic processes. (From Stratton, Campbell and Futreal, 2009).

Driver mutations are causally implicated in oncogenesis and positively selected for in the tumor microenvironment (Garraway and Lander, 2013). They confer hallmark

cancer traits to cells harbouring them, including growth advantage, ability for surrounding tissue invasion, metastasis, neoangiogenesis and evasion of apoptosis (Hanahan and Weinberg, 2000). By definition, drivers occur in a subset of genes collectively termed “cancer genes”. These genes and the mutations in them are therefore of paramount therapeutic significance, making their thorough identification pertinent for novel biomarker and drug target discovery (Stratton, Campbell and Futreal, 2009).

Cancer genes are distinguished by a higher mutation frequency than the background mutation rate. However, driver genes that are mutated at low frequencies also exist, which prompted the ICGC to set the threshold for identifying genes mutated in >3% of tumors per subtype as one of its main goals (Hudson et al., 2010). Over 1% of human genes have been causally implicated in cancer and they are listed in the Cancer Gene Census (CGC) within COSMIC (Futreal et al., 2004). The latest Census contains 719 genes divided into two tiers based on evidence supporting their oncogenic status (Tate et al., 2018). Tier 1 genes have extensively documented roles in oncogenesis and substantial evidence in support, while Tier 2 genes have strong literature links to such roles, but with less supporting evidence. The Census also describes the genes’ contribution to pathogenesis, the types of mutations altering their function in cancer, and cancer types in which they are frequently mutated.

Cancer driver genes encode proteins with a broad range of functions, including signaling, cell metabolism, RNA splicing and numerous others. In addition, genomics studies have provided proof that epigenomic changes such as dysregulated DNA methylation and chromatin modification have roles in driving tumorigenesis (Garraway and Lander, 2013). Recently, a comprehensive identification of oncogenic driver mutations and genes was conducted on 9423 TCGA exomes spanning 33 cancer types (Bailey et al., 2018). By merging and manually curating the output of 26 different computational tools, a total of 299 driver genes and >3400 driver mutations were identified and partially validated using cell lines.

Besides drivers, the majority of mutations present in cancer genomes are passenger mutations. Passengers are approximately randomly distributed and accumulate without being positively selected for over the course of tumorigenesis. They do not contribute to cancer development by conferring growth advantage and inciting clonal expansion, but they are present in the clonal cell population in case they

arose spontaneously during cell divisions prior to the acquisition of a driver (Stratton, Campbell and Futreal, 2009).

Upon exposure to mutagens such as UV radiation, numbers of certain mutation types increase dramatically. The resulting set of mutations therefore exhibits distinct patterns, or signatures, known to be associated with the corresponding mutagens (Vogelstein and Kinzler, 1992). Therefore, along with drivers, the numerous passenger mutations can reveal valuable information on the mutational processes that were operative throughout cancer development.

1.3. MUTATIONAL SIGNATURES

The earliest studies that revealed links between patterns of mutations and their underlying causal mutagens were conducted long before the advent of NGS using DNA sequencing with chain-terminating inhibitors (Sanger, Nicklen and Coulson, 1977). As the throughput of this method was relatively limited, mutational patterns were first examined in a single gene of choice, *TP53*, due to its high frequency of mutation across cancer types (Lehman et al., 1994). The reported patterns, or spectra, of *TP53* mutations were then compared and associated to spectra generated experimentally through exposures to various carcinogens.

Studies such as these were appropriate only for cancers where one mutational process, such as exposure to UV radiation, generates the majority of observed mutations and their resulting pattern, thus allowing for the establishment of etiological associations. In most cancer types, however, a combination of multiple active mutational processes is responsible for the mutational pattern observed in a sample. Reporting such mixed mutational patterns, even in later studies using whole exome sequence data, was inadequately informative to elucidate the causal mutational processes (Pleasance et al., 2009, Alexandrov and Stratton, 2014).

Mutational processes in a cancer cell may be operative at different times and intensities during the lifetime of a patient, and thus have varying contributions to the final tumor mutation set. This results in a mixed mutational pattern that does not resemble any of the individual processes. A central achievement in decomposing such

patterns was presented as a theoretical model and computational framework for deriving individual mutational signatures from somatic mutations belonging to a set of cancer samples. Signatures of mutational processes were modeled as a blind source separation problem wherein original signals must be deconvoluted from a mix of superimposed signals, and non-negative matrix factorization (NMF) was shown to be an effective method for identifying the signatures due to the intrinsic nonnegativity of mutations. The output of the algorithm consists of a minimal set of mutational signatures and the proportion or number of mutations contributed by each signature to each of the cancer samples (Alexandrov et al., 2013a).

The method was initially applied to WGS data from 21 breast cancer patients, followed by a seminal study that analyzed 5 million mutations from 7,042 cancer samples to identify a set of >20 mutational signatures present in cancer (Nik-Zainal et al., 2012, Alexandrov et al., 2013b). The set is curated in COSMIC and has currently grown to 30 validated signatures based on >12,000 samples from 40 cancer types (Figure 2). New studies are expanding the repertoire of mutational signatures even further by increasing sample size, complementing predominant WES data with WGS data, and incorporating less studied mutation classes into their analyses (Alexandrov et al., 2018).

Mutational signatures are the fingerprints of exogenous and endogenous mutational processes, including the intrinsic infidelity of the DNA replication machinery, various mutagen exposures, enzymatic DNA modification and defective DNA repair (Alexandrov et al., 2013). Each signature profile is displayed according to a 96-substitution classification defined using 6 possible base substitution types (C>A, C>G, C>T, T>A, T>C, and T>G), all of which are referred to by the pyrimidine of the mutated Watson–Crick base pair. The number of mutation types is expanded using information on the sequence context of each mutation, namely by incorporating bases immediately 5' and 3' to the mutated base, which results in 96 possibilities (6 types of substitution * 4 types of 5' base * 4 types of 3' base). This classification is advantageous for discerning signatures causing the same types of substitutions in alternate sequence contexts. Mutational signatures are reported as the relative proportions of each of the 96 mutation types generated by a signature and based on the trinucleotide frequencies of the reference human genome (Figure 1). The mutational catalogue (or spectrum) of

a cancer genome is the distribution of mutations over each of the 96 possible types, e.g. the number of C > T mutations at CCA, where the mutated based is underlined.

Figure 2. 30 mutational signatures (https://cancer.sanger.ac.uk/cosmic/signatures).

The mutational processes causing some mutational signatures are known and annotated in COSMIC. At least 11 mutational signatures are thought to be generated by endogenous processes such as DNA editing by enzymes from the *AID/APOBEC* family of cytidine deaminases (Signatures 2 and 13), spontaneous deamination of 5-

methylcytosines (Signature 1), defective DNA mismatch repair (Signatures 6, 15, 20 and 26) and defective double-strand break repair (Signature 3). Signature 10 is associated with mutations in the proofreading domain of polymerase ε. Signature 5 is thought to arise from a clock-like process, similarly as Signature 1 (Alexandrov et al., 2015). Known links between exogenous processes and mutational signatures include the following: tobacco smoking (Signature 4), UV light (Signature 7), alkylating agents (Signature 11), aristolochic acid (Signature 22), aflatoxin (Signature 24), tobacco chewing (Signature 29). However, the associated etiologies of Signatures 8, 12, 14, 16, 17, 18, 19, 21, 23, 25, 27, 28 and 30 are still unknown. A number of variant association studies was published so far with the aim of providing evidence of known, speculated and novel etiological associations (Petljak and Alexandrov, 2016).

1.4. PREVIOUS RESEARCH ON VARIANT ASSOCIATIONS WITH MUTATIONAL SIGNATURES

1.4.1. GERMLINE VARIANT ASSOCIATIONS

While somatic mutation is the main process driving cancer development, many germline genetic variants have also been implicated in cancer susceptibility. Some have been linked with mutational signatures, including pathogenic germline *BRCA1/2* variants, which are associated with Signature 3, and a common germline *APOBEC3A-APOBEC3B* deletion allele which confers increased breast cancer susceptibility in carriers by elevating the levels of *APOBEC* signatures through a yet unknown mechanism (Lu et al., 2015, Nik-Zainal et al., 2016, Nik-Zainal et al., 2014).

In order for germline *BRCA1/2* variants to modulate Signature 3 levels, bi-allelic inactivation is required through a second, somatic alteration event. The focus of a newly published study was to establish whether this mechanism is a ubiquitous requirement for germline mutations to exert their effect on somatic phenoypes. Paired tumor and normal TCGA sequence data was assessed along with other molecular information to perform an exome-wide search for genes affected by bi-allelic alteration, as well as for their associations with different somatic phenotypes including microsatellite instability and mutational signatures. In the case of mutational

signatures, associations were tested on a set of DNA damage repair genes and the results revealed only one significant association due to the small number of samples carrying bi-allelic germline-somatic alterations (Buckley et al., 2018).

A preprint of the most extensive analysis of germline variant associations with mutational signatures was published by the Pan-Cancer Analysis of Whole Genomes (PCAWG) Collaboration in 2017. The data comprised 88 million germline variants from 2,642 whole genomes belonging to ICGC and TCGA patients and spanned 39 cancer types. Notably, the analysis revealed elevated Signature 1 levels are associated with damaging germline variants in the *MBD4* DNA glycosylase gene. In addition, common germline variants were implicated in reducing *APOBEC* signature levels (Waszak et al., 2017).

1.4.2. SOMATIC MUTATION ASSOCIATIONS

It is known that somatic mutations modulate the activity of mutagenic processes and consequently influence the levels of their respective signatures in a cancer sample. For example, early research on this topic revealed that tumors harboring somatic mutations in the *ERCC2* gene have an increased level of Signature 5* (resembling Signature 5), thus associating this signature with the nucleotide-excision repair pathway (Kim et al., 2016).

In 2018, three broad association analyses of somatic mutations with mutational signatures were published. In a PanCancer Atlas study of DNA damage repair (DDR) deficiency across cancer types, which focused on 276 curated DDR genes, signatures levels were assessed according to three somatic alteration types. This included point mutations, deletions and epigenetic silencing from a subset of 48 driver genes which were evaluated for associations with signatures (Knijnenburg et al., 2018). The analysis revealed novel and reproduced known associations, e.g. those of *POLE* mutations with Signature 10.

The two other recent association studies were rather similar and focused on driver mutations with potential to affect signature levels. Firstly, in a paper that assessed the dual effect of mutation and selection on the set of drivers present in a

cancer cell, differences in signature levels were evaluated with respect to 53 chosen driver mutations using nonparametric testing (Temko et al., 2018). The study produced 43 significant positive associations, predominantly with signatures of endogenous and unknown etiologies. The second study used logistic regression to identify 39 significant associations, both negative and positive, between 50 driver mutations and 30 known mutational signatures (Poulos et al., 2018).

Although extensive, these studies have limited their scope to selected gene classes or relatively small sets of driver mutations. This is justifiable insofar as it restricts the number of hypotheses tested and increases statistical power. However, the oncogenic effect of somatic mutations is not distributed in a binary manner between two traditional classes, passengers and drivers, but rather lies on a continuum which encompasses major drivers, those of intermediate effect and 'mini-drivers' (Castro-Giner, Ratcliffe and Tomlinson, 2015, Vogelstein et al., 2013). In addition, it was recently shown that rare, recurrent mutations in DDR genes have the potential to alter protein stability (Knijnenburg et al., 2018). Therefore, there is still considerable space left for investigating signature associations with somatic mutations recurring at lower frequencies in the data, as well as with those occurring in yet unexplored gene classes.

2. RESEARCH AIM

Based on the knowledge that certain drivers are specific only to some cancer types or simply appear less frequently than the canonical drivers investigated in previous studies, there may be valuable insight to gain through analyzing the effect of a comprehensive set of pan-cancer and subtype-specific recurrent somatic mutations on the modulation of mutational signature levels. Aside from drivers, which by definition have a considerable impact on cells, recurrent somatic mutations of putatively benign character have hitherto been excluded from signature association studies even though they could influence the underlying mutational processes or be a consequence of those processes. In addition, there is prospect for both hypothesis generation and etiological findings in associating mutational signatures with the somatic mutation status of gene from classes yet unexplored in this context, but known to be involved in oncogenesis, such as chromatin remodeler genes.

Therefore, the motivation behind this thesis was to employ mutational signature analyses to broaden the current understanding on how somatic mutations influence mutagenic processes. Considering all the above, the specific aim of this thesis was to conduct an extensive association study of 30 COSMIC mutational signatures with tumor somatic mutations belonging to a curated list of 721 genes. This included both pan-cancer and per-cancer-type association testing of up to 721 individual benign or pathogenic somatic mutations and up to 711 genes recurrently mutated in ~10,000 samples across 33 cancer types in the TCGA MC3 dataset.

3. MATERIALS AND METHODS

3.1. DATA ACQUISITION AND PREPARATION

3.1.1. SOMATIC MUTATION DATA

All somatic mutations were obtained from TCGA through the Genomic Data Commons (GDC) Portal (https://gdc.cancer.gov/about-data/publications/mc3-2017, Grossman et al., 2016). Input for all analyses was the publicly available MC3 file (mc3.v0.2.8.PUBLIC.maf.gz). The MC3 dataset represents the complete set of TCGA somatic mutation calls and was published on March 28, 2018 as the result of the Multi-Center Mutation Calling in Multiple Cancers project (Ellrott et al., 2018). The MC3 project aimed to create a comprehensive and harmonized dataset that would enable robust pan-cancer and cross-tumor type analyses. An ensemble of seven different mutation-calling algorithms (MuTect, MuSE, VarScan2, Radia, Pindel, Somatic Sniper, Indelocator) was applied to exome sequencing data from over 10,000 patients across 33 different cancer types to generate a dataset of 3.6 million somatic mutations. Extensive efforts were made to remove low-quality and duplicated samples and include the best tumor-normal sequencing pairs for each patient (Ellrott et al., 2018). The list of 33 cancer types present in the data and used in all analyses is presented in Table 1 along with their abbreviations.

The initial dataset prior to filtering comprised 3,600,963 mutations from 10,224 samples contained in a MAF (Mutation Annotation Format) file. File specifications can be found at https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/. Mutation coordinates were given for the GRCh37 reference genome build. The data contained mutations from only one tumor per patient and there were no duplicate mutations present. Filtering was applied similarly to Knijnenburg et al., 2018, keeping only mutations labelled as “PASS”. This label in MC3 indicates high-quality mutations which are not potential artifacts and have been called by two or more mutation-calling algorithms. Mutations from whole-genome amplified (WGA) samples were rescued if not marked as “PASS” due to the fact that many early OV and LAML samples in TCGA were sequenced using WGA DNA. Only mutations with the variant type single-nucleotide polymorphism (SNP) were kept for downstream analyses.

Mutations were further filtered based on their type and location. Mutations in introns, 3' or 5' UTRs or UTR flanking regions, as well as all insertions, deletions and synonymous mutations were removed, keeping only missense, splice site, nonsense, nonstop and translation start site single-nucleotide mutations. Previous similar studies regularly excluded missense mutations predicted to be benign and focused their analyses on pathogenic mutations that are more likely to exert a cellular effect (Temko et al., 2018, Knijnenburg et al., 2018, Kim et al., 2016). Given the possibility that putatively benign mutations may also correlate with mutational process activity as its consequence, these mutations were included in the analysis but examined separately from pathogenic mutations in all subsequent steps (see further details in section 3.1.3.). Similarly, somatic mutations reported as common in ExAC database of germline variants could have been excluded due to their probable benign character, but were instead investigated separately from others (Lek et al., 2016). Mutation-level filtering reduced the overall dataset to a final of 1,899,524 mutations from 10,109 samples. The TCGA sample IDs were uploaded to the GDC Data Portal in order to obtain the corresponding cancer type for each sample. The disease data was added to the filtered MC3 dataset along with a unique identifier that was created for all mutations.

The filtered MC3 dataset was further divided into two sets of mutations, belonging either to hypermutated or to non-hypermutated samples. A list of hypermutated sample IDs was obtained from Bailey et al., 2018. The list was then expanded to include samples with >1900 mutations, corresponding to the 99th percentile of the mutation burden distribution in the filtered MC3 data. Using 38 Mb as the estimate of exome size, the mutation rate for these hypermutated samples was estimated at 50 mutations per Mb. The final hypermutator MC3 dataset comprised 917,319 mutations from 355 samples. In previous similar studies, hypermutated samples were removed from analyses to avoid high mutation burden as a source of confounding. Here, analyses were simply stratified based on hypermutation and conducted on both sets of samples. The remaining 9754 samples formed the non-hypermutator MC3 dataset, which was the main discovery dataset used in this study.

Table 1. TCGA cancer types and their abbreviations.

Abbreviation	Cancer type
ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical and endocervical cancers
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UVM	Uveal Melanoma

3.1.2. GENE LIST PREPARATION

A list of genes of interest was compiled based on multiple sources and only mutations belonging to these genes were considered for subsequent association analyses in order limit the number of hypotheses tested.

Firstly, 299 driver genes discovered in a recent PanCan Atlas study on MC3 data were obtained from Bailey et al., 2018. The gene list was then expanded to span all major DNA damage repair pathways by adding a curated set of 276 DDR genes from Knijnenburg et al., 2018.

Two recent studies provided evidence that polymerized actin has roles in the nucleus and may be essential in cellular responses to DNA damage (Caridi et al., 2018, Schrank et al., 2018). Thus, several genes coding for actin, WAS and Arp2/3 complex proteins were added to the gene list to investigate their potential associations with DNA double-strand break repair (Signature 3) activity.

The dysregulation of chromatin remodelers and the resulting abnormal gene expression patterns have been ascertained as mechanisms driving oncogenesis (Kumar et al., 2016). Therefore, an extensive set of chromatin remodelers was added to the gene list. Specifically, chromatin remodeling genes were downloaded from EpiFactors, a database of human epigenetic factors and complexes (Medvedeva et al., 2015, http://epifactors.autosome.ru/genes), as well as from the curated collection provided at http://www.dnarepairgenes.com/chromremodgenes.html. In addition, Gene Ontology AmiGO 2 annotation was used to extract genes involved in chromatin remodeling (http://amigo.geneontology.org/amigo/term/GO:0006338). The unique gene names from these three sources were verified to be HGNC approved symbols using the tool provided at https://www.genenames.org/cgi-bin/symbol_checker and corrected where necessary to produce a final set of 224 chromatin remodeler genes. Genes for which there were no mutations present in the MC3 dataset were removed from the gene list, resulting in a total of 721 genes (Supplementary Table 1).

3.1.3. MATRIX GENERATION

Two basic types of matrices were generated for association testing in a manner similar to Knijnenburg et al., 2018. The SNP-level matrix contained binary calls for the presence of each somatic mutation in each sample, with sample IDs as rows and mutation IDs as columns. The gene-level matrix contained binary calls for the presence of any somatic mutation within the given gene for each sample.

Matrices were generated separately for the three main datasets: the non-hypermutator MC3 dataset, the hypermutator MC3 dataset and the MC3 dataset comprising all samples. Firstly, each dataset was divided into two subsets, one containing benign mutations and another with pathogenic mutations. This was done in order to enable separate investigation of these mutation types in all downstream analyses due to their predicted opposing cellular effects.

The separation was based on combining mutation annotations from two functional prediction algorithms, SIFT and PolyPhen (Kumar, Henikoff and Ng, 2009, Adzhubei, Jordan and Sunyaev, 2013). The missense mutations marked as “benign” by PolyPhen or “tolerated” by SIFT were defined as benign mutations and those remaining were considered pathogenic. In addition, mutations labelled as “common in ExAC” were considered as benign. For each of the three main datasets, all mutations common in ExAC were extracted first, after which the benign missense mutations were also removed. These two groups together formed the benign mutation subset, while those remaining formed the pathogenicsubset, resulting in a total of six mutation sets. The two matrix types were generated for each mutation set using different approaches, producing 12 binary matrices in total.

Only those mutations recurring at least five times in either of the three main datasets were selected as a first step in SNP-level matrix generation due to the requirement of having at least 5 mutated samples for later statistical analyses. Of these recurrent mutations, only those belonging to genes in the predefined gene list were further selected to reduce the number of hypotheses tested and increase statistical power. For each mutation, binary calls across samples were recorded. The matrix was then divided into two smaller matrices with mutations belonging either to the benign or pathogenic subset.

Gene-level matrices were created for genes recurrently mutated in each given pathogenic or benign mutation subset. Firstly, mutations were filtered using the predefined gene list. For each subset, the number of samples having at least one mutation present in a given gene was counted and only the genes mutated in at least five samples were included in the matrix. Similarly to previous studies, the matrix was created in a manner which assumes pooling of pathogenic (or benign) somatic

mutations per gene, resembling the approaches used in rare germline variant burden testing (Kim et al., 2016, Wagner, 2013).

3.1.4. MUTATIONAL SIGNATURE DATA

Data on the activity of mutational processes was obtained in the form of relative contributions of each of the 30 validated COSMIC mutational signatures for all samples used in this analysis. Mutational signature data was downloaded from mSignatureDB, a recently created database containing the 30 COSMIC signature contributions for ~16,000 individual tumors across 33 TCGA and 40 ICGC cancer projects (Huang et al., 2017, http://tardis.cgu.edu.tw/msignaturedb/).

3.1.5. PROXY VARIABLE PREPARATION

Proxy variables for certain mutational signatures of known etiology were created in order to conduct a preliminary round of association testing on a reduced number of hypotheses compared to testing with all 30 signatures. Signature proxies were calculated for Signature 1 (aging), Signature 4 (smoking), Signature 7 (UV light), Signatures 2 and 13 (APOBEC activity) and Signature 15 (defective mismatch repair). The proxies correspond to percentages of certain mutation types from the 96-substitution classification or their mutual combinations, and are described in Table 2 using IUPAC notation. Proxies 2, 3 and 4 were normalized, continuity-corrected by adding 0.5 to their respective values and $\log_2$ transformed into novel variables. These variables were included in testing in addition to the non-transformed proxies.

Table 2. Proxy variable descriptions, formulas and matching mutational signatures.

Proxy	Description	Formula	Mutational signature
1	C>T mutation % at CpG sites	NCG > T %	Signature 1
2	C>A mutation %	NCN > A %	Signature 4
3	C>T mutation % at pyrimidine dimers	YCN > T %	Signature 7
4	C>T or G mutation %	(TCW > T + TCW > G) %	Signature 2, 13
5	C>T mutation % at GpC sites	GCN > T %	Signature 15

3.2. STATISTICAL ANALYSES

3.2.1. ASSOCIATION TESTING SETUP

Associations between somatic mutations and mutational signature levels were assessed using the nonparametric Wilcoxon-Mann-Whitney test as previously described in Temko et al., 2018. Specifically, it was tested whether the relative contribution of each signature in samples bearing a mutation was statistically different from those without the mutation. All tests were two-sided.

The sign of the Hodges-Lehmann estimator was used to indicate if the direction of association was positive or negative, while the absolute value of the estimator was taken as a measure of effect size for each of the hypotheses tested. In nonparametric testing, the Hodges-Lehmann estimator is used to provide a point estimate of the difference between values in two sets of samples. For *m* and *n* values belonging to set A and set B, respectively, the difference between the values in *m***n* pairs can be obtained and the Hodges-Lehmann estimator is defined as their median. Therefore, the estimator does not represent the difference in medians or means, but rather the median of the differences between values from set A and set B (Hodges and Lehmann, 1963).

All analyses were performed with R version 3.5.2 and statistical assessment was implemented using the 'parallel' package (R Core Team, 2018). An equivalent approach was used in carrying out analyses with mutational signature proxy variables. A total of 48 separate association analyses were performed based on the following principle: somatic mutations recurring at least five times in their corresponding dataset were tested individually for associations with signatures, while mutations recurring less frequently were pooled together at the level of their respective genes which were the functional units ultimately tested for signature associations.

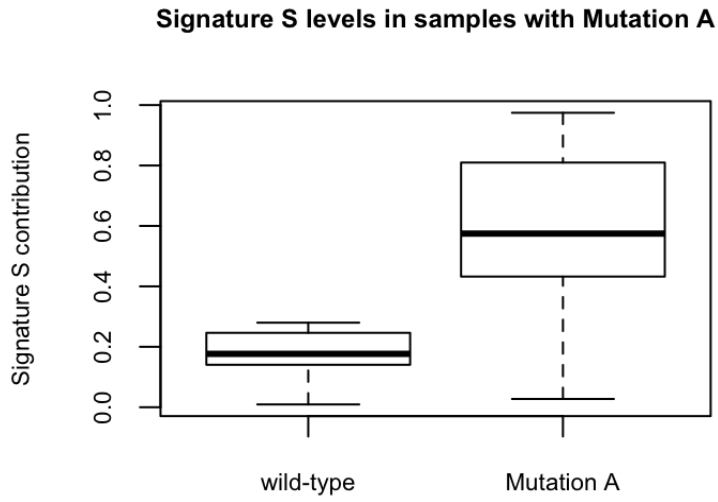

Figure 3. Example comparison of Signature S levels in cancer samples with and without the presence of Mutation A. Reproduced from Temko et al., 2018.

3.2.2. INDIVIDUAL MUTATION TESTING

The input for individual mutation testing were six SNP-level matrices containing binary calls for either benign or pathogenic mutations belonging to each of the three main datasets: the non-hypermutator MC3 dataset, the hypermutator MC3 dataset and the complete dataset comprising all samples. Details of each matrix can be found in Table 3. Association testing was conducted with mutational signatures and proxy variables for all six matrices on two separate levels, resulting in 24 analyses.

On the pan-cancer level, mutational signature contributions for samples across all 33 distinct cancer types were compared with respect to the presence or absence of each analyzed mutation. Therefore, the total set of individually tested mutations, distributed into pathogenic and benign subsets, was initially chosen based on pan-cancer level recurrence (frequency in the main datasets).

In addition to pan-cancer testing, mutations were assessed within each cancer type in order to empower detection of subtype-specific associations by removing tissue type as a source of confounding which existed in the pan-cancer analyses. On the per-cancer type level, somatic mutations from each matrix were considered as possible tissue-specific drivers based on recurrence in at least five samples of a given cancer type according to the methodology outlined by Temko et al., 2018. Only those

mutations were subjected to association testing in the 33 cancer types or additional groups of similar cancers. Groups were formed by combining the samples from i) COAD and READ, ii) LGG and GBM and iii) LUSC and HNSC. Two additional groups were formed to facilitate discovery of *APOBEC*-related (Signature 2 and 13) associations: LUSC, HNSC and BLCA were grouped together due to the high levels of Signature 13 in these cancers, while LUAD and BRCA were grouped based on high Signature 2 levels.

Table 3. Details of input matrices used in individual mutation association testing.

Dataset	Sample count	Total recurrent mutations	Mutation count (benign matrix)	Mutation count (pathogenic matrix)
non-hypermutator	9754	486	56	430
hypermutator	355	148	36	112
complete	10109	721	125	596

3.2.3. GENE-LEVEL (BURDEN) TESTING

By definition, driver mutations appear often enough in tumor samples to allow their individual oncogenic roles and properties to be characterized fairly well. However, the driver genes harboring these mutations also contain numerous other rarely recurring somatic mutations which can either be inconsequential passengers or have variable impact on normal gene function in a manner more similar to established drivers. Therefore, valuable information on the oncogenic significance of a gene may be acquired by studying such mutations. Their infrequent nature necessitates an approach that pools data across the affected gene, rather than focusing on a particular nucleotide as the recurrently mutated unit being examined. This involves aggregating samples that have rare somatic mutations in a particular gene and assessing their molecular phenotypes, such as mutational signatures, against those of samples with the wild-type gene. It is important to avoid pooling together samples that harbor pathogenic mutations with those harboring protective mutations of opposing effect direction, as well as to separate samples that hold mutations of no functional influence. This is difficult because gene function is altered only by a small subset of the rarely recurring somatic mutations, while the majority are passengers that appear

infrequently because they are not selected for conferring any advantages to the tumor cell. Approaches which use various methodologies to aggregate rare mutations and their effects into a pooled unit for the purpose of association testing are termed “burden” tests and rely heavily on computational predictions of the degree and type of effect exerted by each variant (Wagner, 2013). Burden tests were developed for rare germline variant analyses within the scope of genome-wide association studies (Asimit and Zeggini, 2010).

A burden testing approach was applied to cancer somatic mutations in this study. Burden testing included all mutations regardless of their recurrence frequency in the data, with their respective genes being subsequently tested if mutated in five or more samples. Mutations were divided into pathogenic and benign subsets prior to being pooled per gene in order to avoid cancelling out opposite signals. The burden testing input consisted of six matrices described in Table 4 and all analyses were stratified and performed as previously explained for individual mutation testing.

Table 4. Details of input matrices used in gene-level association testing.

Dataset	Sample count	Sample count (benign subset)	Gene count (benign matrix)	Sample count (pathogenic subset)	Gene count (pathogenic matrix)
non-hypermutator	9754	9708	682	9709	686
hypermutator	355	355	676	355	680
complete	10109	10063	706	10064	711

3.2.4. MULTIPLE HYPOTHESIS TESTING CORRECTION

The nature of genome- and exome-wide association studies assumes that a large number of statistical tests are performed for each dependent variable. This leads to the multiple hypothesis testing problem, where the number of hypotheses reaching significance due to chance alone grows unacceptably large as number of tests increases. Numerous methods have been developed and used in genomics to alleviate the consequences of large numbers of false positives (Type I errors) produced in multiple hypothesis testing (Goeman and Solari, 2014).

One such method is the Benjamini-Hochberg (BH) procedure which aims to control the false discovery rate (FDR), defined as the proportion of Type I errors among all significant hypotheses (Benjamini and Hochberg, 1995). An FDR-adjusted p-value is termed a q-value and is a measure of significance in terms of the FDR. The q-value of a hypothesis test is the expected proportion of false positives incurred when using that value as the significance threshold (Storey and Tibshirani, 2003). This procedure was used to correct for multiple hypothesis testing in previous mutational signature association studies. Depending on the study, an FDR < 0.1, 0.05 or 0.01 was used as the significance threshold (Kim et al., 2016, Temko et al., 2018, Knijnenburg et al., 2018). In this study, the BH procedure was also applied and associations with an FDR q-value < 0.01 were considered significant based on a more conservative approach outlined in Knijnenburg et al., 2018.

Due to the presence of underlying confounders, test statistics in association studies can be inflated (and inversely, p-values can be deflated) relative to the expectation of no association. This leads to an increase in false positives among the significant results, which can then be controlled by applying multiple testing correction methods. Genomic control (GC) is a method developed as an alternative to the conservative Bonferroni correction that was traditionally used for this purpose in genome-wide association studies. It controls for sources of bias such as population heterogeneity and cryptic relatedness in order to avoid spurious associations (Devlin and Roeder, 1999).

Given a set of test scores or p-values, the extent of confounding can be quantified in the form of the genomic inflation factor lambda (λ), which may be simultaneously estimated from the set and used to correct the constituent values. λ is assumed to be constant for all markers (SNPs or genes) being tested as it is a consequence of genome-wide bias and is derived based on the premise that most markers being tested are not associated with the dependent variable. These are termed null markers and they are used to calculate λ, which is assumed to be the factor by which their p-values deviate from an expected uniform distribution. Given that this deviation is not due to association with the dependent variable, the total value of λ is attributed to the effect of underlying confounding factors. All p-values can thereby be divided by this factor in order to account for systematic bias, albeit at a cost in power

to detect true associations (Devlin and Roeder, 1999, Devlin, Bacanu and Roeder, 2004).

There are several ways to calculate λ, such as by dividing the mean or median of the observed test statistic distribution by the expected median of the chi-squared distribution with one degree of freedom (0.4549). In this study, λ was calculated by taking the $-\log_{10}$ of a set of sorted p-values and fitting a linear model based on uniformly distributed p-values to the bottom half of values in the resulting vector (Tsepilov et al., 2013). The obtained slope coefficient represented λ and was used to divide all values of the vector if the coefficient was >1. Where λ was <1, it was assumed there was no confounding present and hence there was no need to apply the correction. Values divided by λ were converted back to p-values and subjected to BH correction to control the FDR at 1%. λ was calculated separately for each dependent variable and only if the set of markers available to estimate it was greater than 200. The threshold of 200 initial markers (p-values) needed to calculate λ was chosen conservatively based on the results from Marchini et al., 2004. If the number of genes or mutations being tested was < 200 or λ was <1, only the BH correction procedure was applied to a given set of p-values.

4. RESULTS

Due to the scope of the study, only the output of analyses conducted in association with the 30 mutational signatures on the per-cancer type level is presented, while selected remaining results are part of the Supplementary Material (SuppM). Of the three main datasets used in this analysis, results are presented for the non-hypermutator and hypermutator datasets. Significant associations obtained for the complete dataset comprising all MC3 samples are part of the SuppM. Signatures have been summarized by the origin of their etiology in Table 5.

Table 5. COSMIC mutational signatures and their putative etiologies.

Etiology type	**Etiology**	**Signature**
endogenous	5-methylcytosine deamination	Signature 1
	APOBEC	Signature 2
	DNA double-strand break repair	Signature 3
	DNA mismatch repair	Signature 15
	APOBEC	Signature 13
	DNA mismatch repair	Signature 20
	DNA mismatch repair	Signature 26
	DNA mismatch repair	Signature 6
	error-prone polymerase η	Signature 9
	POLE mutations	Signature 10
exogenous	Smoking	Signature 4
	Ultraviolet light	Signature 7
	Temozolomide	Signature 11
	Aristolochic acid	Signature 22
	Aflatoxin	Signature 24
	Tobacco chewing	Signature 29
unknown		Signature 5
		Signature 8
		Signature 12
		Signature 14
		Signature 16
		Signature 17
		Signature 18
		Signature 19
		Signature 21
		Signature 23
		Signature 25
		Signature 27
		Signature 28
		Signature 30

4.1. PER-CANCER TYPE TESTING

In each of the tables containing per-cancer type testing results, signature associations are presented grouped by cancer type and sorted per group from largest to smallest by absolute effect size. Due to the large number of significant results obtained even after both BH and GC methods were applied to correct for multiple hypothesis testing, significant associations were additionally filtered according to absolute effect size and only this subset of associations is reported in the tables. Effect size filtering was based on a threshold value chosen by inspection of density plots for each set of absolute effect size values (Figure 4). Thresholds were chosen to include associations with the largest Hodges-Lehmann estimator values corresponding to the long tail of each distribution.

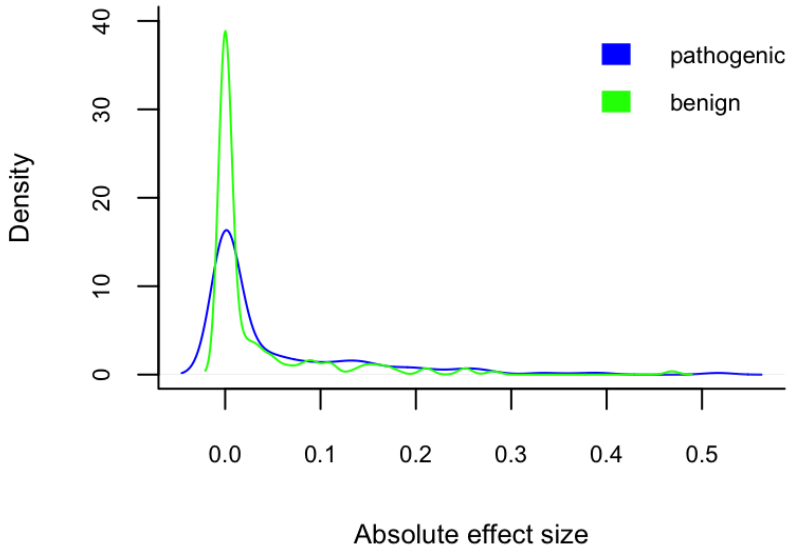

Figure 4. Density distribution of significant association absolute effect sizes shown for burden testing of pathogenic and benign mutation subsets of the non-hypermutator dataset. N(pathogenic) = 282, N(benign) = 154. Chosen thresholds were 0.05 (pathogenic) and 0.02 (benign), with 81/282 and 41/154 associations having effect sizes greater than the threshold value. Associations were filtered and reported using the same approach for all other analyses and datasets in this study. Different thresholds were applied for each analysis due to differences in effect sizes attributable to associations with benign versus pathogenic mutations, as well as associations resulting from individual mutation testing versus burden testing.

Where negative associations were present, indicating that samples with a mutation or mutated gene have significantly lower levels of a particular signature, their reciprocal associations were inspected (Poulos et al., 2018). If present, reciprocals comprised all significant associations of that mutation or gene with other signatures in the given cancer type, but in the opposite (positive) direction. The absolute effect size of the negative association was then compared to each of the reciprocals. Regardless of initial direction, the association(s) for which the absolute effect size was larger was considered to more likely represent the causal relationship among them and was retained, while those with small effect sizes were taken to represent its opposite-direction consequences and excluded from further consideration. In other words, an increase or decrease in the levels of one signature automatically affects the proportions of other signatures in a sample. As a consequence, this will sometimes cause a given predictor to appear as associated with those changes as well, albeit with a smaller effect size. Reciprocals of this kind are simple side-effects and should therefore not be examined.

4.1.1. BURDEN TESTING

For each of the four burden analyses (A-D) reported, results were summarized in the form of tables with a common format. The “Disease” column contained TCGA cancer type abbreviations (Table 1) or abbreviations for additional cancer groups described previously. It should be noted that the LUSC_HNSC_BLCA group was abbreviated to “LHB” in the tables. The “Gene” column contained HGNC approved gene symbols, while the “Signature” column contained response variables from the list of 30 COSMIC mutational signatures (Table 5). For each association, raw p-values and FDR q-values were given in columns titled “p_value” and “q_value”, respectively. It should be noted that, in cases where λ was > 1 and GC had to be applied, the p-values corrected by λ were reported, along with their corresponding FDR q-values. The direction of association was contained in the “direction” column, with value -1 marking positive associations (increased signature levels in samples with the mutated gene version) and 1 marking negative associations due to the sign of the Hodges-Lehmann estimator attributed to each association type. Absolute value of the estimator (absolute effect size) was contained in the “effect size” column.

4.1.1.1. Pathogenic mutations

A) Non-hypermutator dataset pathogenic mutation burden testing

This analysis resulted in a total of 77 significant associations (Table 6). Twenty-four associations involved signatures of exogenous etiology, 49 involved signatures of endogenous etiology and four involved signatures with no known etiology. There were ten negative associations, four of which had reciprocal associations of smaller effect sizes that were removed from the final subset of significant associations reported. Interestingly, nine out of ten negative associations were related to Signature 1 levels, while one was related to signature 22 levels. Based on the set of pathogenic mutations, recurrently mutated genes were present and tested in all cancer types. Eleven associations appeared in more than one cancer type and they are summarized in Table 7.

Table 6. Significant mutational signature associations obtained by burden testing of pathogenic mutations from samples in the non-hypermutator MC3 dataset.

Disease	Gene	Signature	p_value	q_value	direction	effect size
BRCA	ERBB3	Signature 2	3.77E-04	8.73E-03	-1	2.85E-01
BRCA	SETD2	Signature 2	5.21E-04	8.73E-03	-1	2.72E-01
BRCA	SETD2	Signature 13	2.66E-05	1.58E-03	-1	2.55E-01
BRCA	TET2	Signature 13	1.38E-04	5.48E-03	-1	2.55E-01
BRCA	UBR5	Signature 2	1.84E-04	7.31E-03	-1	1.24E-01
BRCA	FOXA1	Signature 2	5.09E-05	3.03E-03	-1	1.18E-01
BRCA	MYH9	Signature 2	4.46E-04	8.73E-03	-1	1.11E-01
BRCA	NF1	Signature 2	5.87E-04	8.73E-03	-1	1.00E-01
CESC	KMT2D	Signature 1	6.39E-06	4.73E-04	1	1.53E-01
CESC	PIK3CA	Signature 2	1.26E-05	9.34E-04	-1	9.55E-02
COAD	CREBBP	Signature 6	9.69E-05	8.43E-03	-1	1.38E-01
GBM	APC	Signature 3	4.83E-05	1.63E-03	-1	5.17E-01
GBM	IDH1	Signature 1	2.10E-07	8.60E-06	1	3.95E-01
GBM	HUWE1	Signature 25	4.52E-06	1.85E-04	-1	6.51E-02
HNSC	CUL1	Signature 2	1.57E-04	7.51E-03	-1	1.84E-01
HNSC	HRAS	Signature 1	1.12E-04	8.03E-03	-1	1.62E-01
HNSC	NSD1	Signature 1	2.59E-07	3.71E-05	1	1.50E-01
HNSC	HLA-B	Signature 2	7.06E-05	5.05E-03	-1	1.17E-01
HNSC	APOB	Signature 4	9.95E-05	7.12E-03	-1	7.39E-02
HNSC	CASP8	Signature 2	2.93E-05	4.19E-03	-1	7.08E-02
HNSC	HUWE1	Signature 2	2.35E-04	8.40E-03	-1	6.72E-02
KIRP	FGFR3	Signature 2	1.98E-04	4.35E-03	-1	9.38E-02
LGG	EGFR	Signature 1	2.70E-05	2.16E-04	-1	2.25E-01
LGG	IDH1	Signature 1	6.32E-07	7.58E-06	1	1.80E-01

LGG	TP53	Signature 1	1.41E-07	3.38E-06	1	1.31E-01
LGG_GBM	APC	Signature 3	6.32E-06	4.55E-04	-1	5.17E-01
LGG_GBM	MACF1	Signature 1	5.83E-04	6.99E-03	1	3.64E-01
LGG_GBM	IDH1	Signature 1	4.38E-23	3.16E-21	1	2.14E-01
LGG_GBM	EGFR	Signature 1	1.41E-07	3.40E-06	-1	1.58E-01
LGG_GBM	ATRX	Signature 1	3.43E-06	6.17E-05	1	1.48E-01
LGG_GBM	TP53	Signature 1	1.38E-08	4.96E-07	1	1.25E-01
LGG_GBM	PTEN	Signature 1	1.03E-04	1.48E-03	-1	1.19E-01
LGG_GBM	SMARCA2	Signature 4	1.02E-06	4.92E-05	-1	7.09E-02
LGG_GBM	ZFHX3	Signature 4	1.37E-06	4.92E-05	-1	6.42E-02
LGG_GBM	APC	Signature 9	1.95E-06	7.03E-05	-1	6.03E-02
LHB	STAG2	Signature 13	2.17E-05	3.35E-03	-1	1.74E-01
LHB	MACF1	Signature 13	9.56E-06	3.35E-03	-1	1.32E-01
LHB	ERBB2	Signature 13	8.89E-05	8.21E-03	-1	1.25E-01
LHB	ERBB2	Signature 2	1.75E-05	2.02E-03	-1	1.25E-01
LHB	STAG2	Signature 2	6.73E-05	6.22E-03	-1	1.11E-01
LHB	ARID1A	Signature 13	1.96E-05	3.35E-03	-1	1.00E-01
LHB	MACF1	Signature 2	1.09E-05	2.02E-03	-1	9.62E-02
LHB	ARID1A	Signature 2	1.61E-05	2.02E-03	-1	8.15E-02
LHB	PIK3CA	Signature 13	5.11E-05	5.90E-03	-1	5.92E-02
LHB	PIK3CA	Signature 2	1.38E-06	6.38E-04	-1	5.78E-02
LUAD	ATF7IP	Signature 4	7.74E-06	2.96E-04	-1	3.36E-01
LUAD	GABRA6	Signature 4	7.02E-04	7.45E-03	-1	2.73E-01
LUAD	PLCB4	Signature 4	3.75E-05	1.02E-03	-1	2.66E-01
LUAD	PTPRD	Signature 4	9.61E-09	6.12E-07	-1	2.51E-01
LUAD	APOB	Signature 4	5.13E-10	4.90E-08	-1	2.51E-01
LUAD	PTPRC	Signature 4	2.23E-04	4.26E-03	-1	2.50E-01
LUAD	HGF	Signature 4	6.73E-04	7.45E-03	-1	2.33E-01
LUAD	ZFHX3	Signature 4	5.80E-04	7.15E-03	-1	2.25E-01
LUAD	NF1	Signature 4	1.51E-05	4.82E-04	-1	2.08E-01
LUAD	ERBB4	Signature 4	2.20E-04	4.26E-03	-1	2.01E-01
LUAD	KRAS	Signature 4	3.45E-10	4.90E-08	-1	2.00E-01
LUAD	SETBP1	Signature 4	5.98E-04	7.15E-03	-1	1.99E-01
LUAD	COL5A1	Signature 4	4.82E-04	7.15E-03	-1	1.94E-01
LUAD	HERC2	Signature 4	4.00E-04	6.94E-03	-1	1.61E-01
LUAD	SPTA1	Signature 4	6.68E-06	2.96E-04	-1	1.54E-01
LUAD	KMT2C	Signature 4	5.99E-04	7.15E-03	-1	1.47E-01
LUAD	DMD	Signature 4	5.59E-04	7.15E-03	-1	1.39E-01
LUAD	TP53	Signature 4	5.79E-05	1.38E-03	-1	9.09E-02
LUAD_BRCA	PIK3CA	Signature 2	5.21E-10	1.78E-07	-1	5.01E-02
LUSC_HNSC	SPTA1	Signature 4	9.44E-06	1.59E-03	-1	1.86E-01
LUSC_HNSC	HLA-B	Signature 2	4.32E-05	4.85E-03	-1	9.22E-02
LUSC_HNSC	CASP8	Signature 13	1.84E-05	6.20E-03	-1	7.08E-02
LUSC_HNSC	CASP8	Signature 2	1.49E-05	2.52E-03	-1	5.94E-02
PAAD	TP53	Signature 1	5.43E-05	4.34E-04	-1	1.77E-01
READ_COAD	LATS2	Signature 15	7.35E-05	8.74E-03	-1	1.47E-01
READ_COAD	CREBBP	Signature 6	1.08E-05	1.28E-03	-1	1.31E-01
READ_COAD	BRCA2	Signature 8	3.50E-08	4.17E-06	-1	5.43E-02
STAD	TLR4	Signature 17	6.64E-05	5.84E-03	-1	3.88E-01
STAD	CHD6	Signature 6	7.33E-05	6.45E-03	-1	3.28E-01
THCA	EIF1AX	Signature 18	4.93E-04	2.47E-03	-1	7.38E-02
UVM	GNA11	Signature 22	1.23E-11	6.17E-11	-1	1.41E-01
UVM	GNAQ	Signature 22	9.19E-11	2.30E-10	1	1.35E-01

Table 7. Associations significant in multiple cancers within analysis A).

Gene	Signature	Disease
APC	Signature 3	GBM, LGG_GBM
APOB	Signature 4	HNSC, LUAD
CASP8	Signature 2	HNSC, LUSC_HNSC
CREBBP	Signature 6	COAD, READ_COAD
EGFR	Signature 1	LGG, LGG_GBM
HLA-B	Signature 2	HNSC, LUSC_HNSC
IDH1	Signature 1	GBM, LGG, LGG_GBM
PIK3CA	Signature 2	CESC, LUAD_BRCA, LHB
SPTA1	Signature 4	LUAD, LUSC_HNSC
TP53	Signature 1	LGG, LGG_GBM
ZFHX3	Signature 4	LGG_GBM, LUAD

B) Hypermutator dataset pathogenic mutation burden testing

This analysis resulted in a total of five significant associations (Table 8). Three associations involved signatures of exogenous etiology, one involved signatures of endogenous etiology and four involved signatures with no known etiology. There were no negative associations in the final subset of significant associations selected by effect size. It should be noted that, in this particular analysis, effect sizes were smaller in comparison to all other analyses. The effect size threshold was chosen accordingly and was lowered as far as possible to the value of 0.005 based on the effect size filtering methodology described previously. Even with such a low threshold, most of the initial significant associations were filtered out. Based on the set of pathogenic mutations, there were no recurrently mutated genes to test for in ACC, CHOL, ESCA, GBM, HNSC, KICH, LAML, LGG LIHC, OV, PAAD, PRAD, SARC, THYM and UCS. None of the associations appeared in more than one cancer type.

Table 8. Significant mutational signature associations obtained by burden testing of pathogenic mutations from samples in the hypermutator MC3 dataset.

Disease	Gene	Signature	p_value	q_value	direction	effect size
COAD	MED12	Signature 21	2.79E-05	6.19E-03	-1	2.40E-02
COAD	AR	Signature 28	4.08E-05	3.02E-03	-1	1.98E-02
COAD	ATR	Signature 28	4.08E-05	3.02E-03	-1	1.98E-02
COAD	PPP4R4	Signature 28	4.08E-05	3.02E-03	-1	1.98E-02
SKCM	PARP1	Signature 2	2.87E-06	3.42E-04	-1	3.29E-02

In this analysis, only two cancer types, COAD and UCEC, had values of $\lambda > 1$ for certain signatures and GC was applied in these cases. However, it should be noted that the well-known association of *POLE* gene with Signature 10 is not among the significant burden testing results presented in Table 8. This is due to the fact that λ with an abnormally large value was present for Signature 10 in COAD (λ=4.52) and for several other signatures in COAD and UCEC, causing this and numerous other associations to be lost after applying GC prior to BH correction. For this reason, it was acknowledged that λ may overcorrect in certain settings and additional results for COAD and UCEC (corrected using only the BH procedure where GC was previously applied) are presented in Table 9. Results were filtered by effect size as described previously. Table 9 comprises 35 associations, of which 31 are positive and four are negative. The *POLE* association with Signature 10, regarded as a proof-of-method, is now present among them.

Table 9. Significant mutational signature associations obtained by burden testing of pathogenic mutations from COAD and UCEC samples in the hypermutator MC3 dataset. Results are shown only for signatures where λ was > 1, but only the BH procedure was applied to avoid overcorrecting by GC.

Disease	Gene	Signature	p_value	q_value	direction	effect size
COAD	TBL1XR1	Signature 10	2.82E-06	2.77E-04	-1	6.99E-01
COAD	MECOM	Signature 10	9.74E-10	5.83E-07	-1	6.99E-01
COAD	ACVR2A	Signature 10	5.14E-04	9.51E-03	-1	6.99E-01
COAD	ATR	Signature 10	7.15E-08	1.41E-05	-1	6.91E-01
COAD	RMI1	Signature 10	5.39E-06	4.17E-04	-1	6.91E-01
COAD	RFC1	Signature 10	1.69E-04	3.80E-03	-1	6.91E-01
COAD	ASXL1	Signature 10	2.32E-05	9.61E-04	-1	6.85E-01
COAD	PPP4R1	Signature 10	8.23E-06	5.27E-04	-1	6.85E-01
COAD	BRCA2	Signature 10	3.20E-09	8.39E-07	-1	6.85E-01
COAD	REV3L	Signature 10	1.82E-07	2.38E-05	-1	6.85E-01
COAD	POLE	Signature 10	1.48E-09	5.83E-07	-1	6.80E-01

COAD	AMER1	Signature 10	3.43E-07	3.85E-05	-1	6.75E-01
COAD	JAK2	Signature 10	4.78E-05	1.50E-03	-1	6.75E-01
COAD	SHPRH	Signature 10	4.78E-05	1.50E-03	-1	6.75E-01
COAD	HLTF	Signature 10	8.07E-05	2.19E-03	-1	6.75E-01
COAD	PPP4R4	Signature 10	4.78E-05	1.50E-03	-1	6.75E-01
COAD	PDGFRA	Signature 10	5.56E-04	9.71E-03	-1	6.75E-01
COAD	DNA2	Signature 10	2.28E-05	9.61E-04	-1	6.75E-01
COAD	CUL5	Signature 10	1.24E-05	6.52E-04	-1	6.75E-01
COAD	TAF1	Signature 10	1.24E-05	6.52E-04	-1	6.75E-01
COAD	ASXL2	Signature 10	1.87E-05	8.67E-04	-1	6.75E-01
COAD	BRIP1	Signature 10	5.84E-06	4.17E-04	-1	6.75E-01
COAD	ERCC4	Signature 10	1.45E-04	3.40E-03	-1	6.75E-01
COAD	TP63	Signature 10	8.07E-05	2.19E-03	-1	6.64E-01
COAD	RASA1	Signature 10	1.87E-05	8.67E-04	-1	6.64E-01
COAD	TLR4	Signature 10	6.78E-05	1.98E-03	-1	6.64E-01
COAD	RET	Signature 10	6.78E-05	1.98E-03	-1	6.64E-01
COAD	CHD9	Signature 10	1.62E-07	2.38E-05	-1	6.64E-01
COAD	ATAD2	Signature 10	4.17E-04	7.99E-03	-1	6.64E-01
COAD	PDS5B	Signature 10	1.47E-04	3.40E-03	-1	6.64E-01
COAD	PARPBP	Signature 10	5.56E-04	9.71E-03	-1	6.64E-01
UCEC	POLD1	Signature 10	2.86E-05	1.10E-03	1	6.57E-01
UCEC	EP400	Signature 10	3.03E-04	6.12E-03	1	6.03E-01
UCEC	BAP1	Signature 10	2.11E-04	4.48E-03	1	5.98E-01
UCEC	NOTCH1	Signature 10	1.76E-04	3.85E-03	1	5.77E-01

4.1.1.2. Benign mutations

C) Non-hypermutator dataset benign mutation burden testing

This analysis resulted in a total of 42 significant associations (Table 10). Twelve associations involved signatures of exogenous etiology, 19 involved signatures of endogenous etiology and 11 involved signatures with no known etiology. There were two negative associations, both with Signature 1, and they had no reciprocal associations in the final subset of significant associations selected by effect size. Based on the set of benign mutations, there were no recurrently mutated genes to test for in ACC, CHOL, KICH, LAML, MESO, TGCT, THYM and UVM. Two associations appeared in more than one cancer type. Association of *BRD4* with Signature 25 appeared in two cancer groups, GBM and LGG_GBM. Association of *SPTA1* with Signature 4 appeared in two cancer groups, LUAD and LUSC_HNSC. *SPTA1* was also associated with Signature 13 in BRCA, Signature 7 in SKCM and Signature 17 in STAD.

Table 10. Significant mutational signature associations obtained by burden testing of benign mutations from samples in the non-hypermutator MC3 dataset.

Disease	Gene	Signature	p_value	q_value	direction	effect size
BRCA	SMARCC2	Signature 13	1.19E-04	8.32E-03	-1	2.54E-01
BRCA	HUWE1	Signature 1	1.49E-05	1.04E-03	1	1.80E-01
BRCA	AMER1	Signature 13	3.81E-04	8.88E-03	-1	1.45E-01
BRCA	SPTA1	Signature 13	2.61E-04	8.88E-03	-1	8.84E-02
BRCA	TAF1	Signature 10	8.73E-05	6.11E-03	-1	4.96E-02
BRCA	NOTCH2	Signature 11	1.83E-06	1.28E-04	-1	3.78E-02
COAD	ALK	Signature 20	3.32E-10	1.30E-08	-1	2.82E-01
COAD	ALK	Signature 26	2.47E-11	9.63E-10	-1	7.82E-02
COAD	CHD6	Signature 21	1.68E-09	6.55E-08	-1	5.53E-02
COAD	ALK	Signature 14	1.64E-11	6.40E-10	-1	3.57E-02
GBM	BRD4	Signature 25	2.06E-04	3.49E-03	-1	6.50E-02
KIRP	KMT2C	Signature 9	6.40E-04	2.56E-03	-1	2.90E-02
LGG_GBM	FOXA1	Signature 4	4.36E-04	9.16E-03	-1	9.04E-02
LGG_GBM	BRD4	Signature 25	1.22E-07	5.11E-06	-1	6.50E-02
LGG_GBM	AMER1	Signature 30	7.54E-05	3.17E-03	-1	4.51E-02
LGG_GBM	CHD8	Signature 24	1.60E-04	5.41E-03	-1	3.01E-02
LUAD	SPTA1	Signature 4	6.34E-08	9.06E-06	-1	2.13E-01
LUAD	FLT3	Signature 18	4.35E-05	6.23E-03	-1	4.97E-02
LUAD	FANCI	Signature 25	2.81E-06	4.02E-04	-1	2.60E-02
LUSC_HNSC	SPTA1	Signature 4	8.69E-06	2.27E-03	-1	1.68E-01
OV	APOB	Signature 13	2.78E-04	5.83E-03	-1	3.22E-02
READ_COAD	RIF1	Signature 1	1.61E-04	9.51E-03	1	4.69E-01
READ_COAD	MGA	Signature 26	5.09E-11	1.50E-09	-1	4.79E-02
SKCM	RAI1	Signature 7	1.90E-04	5.43E-03	-1	1.68E-01
SKCM	CARD11	Signature 7	1.64E-04	5.43E-03	-1	1.57E-01
SKCM	ERBB4	Signature 7	1.87E-04	5.43E-03	-1	1.34E-01
SKCM	MECOM	Signature 7	1.33E-04	5.43E-03	-1	1.11E-01
SKCM	SPTA1	Signature 7	2.26E-04	5.52E-03	-1	1.07E-01
SKCM	APOB	Signature 7	1.54E-05	1.67E-03	-1	1.05E-01
SKCM	NRAS	Signature 7	1.95E-05	1.67E-03	-1	9.48E-02
STAD	ERBB4	Signature 17	6.70E-06	1.47E-04	-1	2.09E-01
STAD	SPTA1	Signature 17	4.79E-06	1.47E-04	-1	1.46E-01
STAD	RAD54L2	Signature 26	6.41E-08	2.82E-06	-1	3.40E-02
UCEC	FAT1	Signature 6	1.43E-05	2.66E-03	-1	2.49E-01
UCEC	BAP1	Signature 20	2.88E-05	1.79E-03	-1	1.55E-01
UCEC	KANSL1	Signature 20	3.44E-06	6.39E-04	-1	1.15E-01
UCEC	WHSC1L1	Signature 20	1.32E-04	4.91E-03	-1	8.92E-02
UCEC	SMARCAD1	Signature 26	3.65E-08	6.79E-06	-1	7.78E-02
UCEC	TP53	Signature 13	1.87E-05	3.49E-03	-1	4.09E-02
UCEC	CHD8	Signature 12	5.04E-08	9.37E-06	-1	3.03E-02
UCEC	PSIP1	Signature 12	1.49E-05	1.39E-03	-1	2.59E-02
UCEC	PRKDC	Signature 20	7.44E-05	3.46E-03	-1	2.07E-02

D) Hypermutator dataset benign mutation burden testing

Due to the inextricable amount of confounding resulting from the hypermutated nature of these samples and the pooling of mutations predicted to be benign, neither causality nor consequentiality could be inferred well from this analysis and it was therefore regarded as unnecessary.

4.1.2. INDIVIDUAL MUTATION TESTING

For each of the four individual mutation analyses reported (E-H), results were summarized in the form of tables with a common format as described previously for burden testing, but with two additional columns present. The “Mutation” column contained a unique identifier created for each somatic mutation in the form of “Chr_Pos_Ref_Alt” (chromosome, position in GRCh37, reference allele, alternative allele). The “HGVSp” column was added to enable easier interpretation of individual mutation associations and contained the mutations written according to HGVS nomenclature.

4.1.2.1. Pathogenic mutations

E) Non-hypermutator dataset pathogenic mutation individual testing

This analysis resulted in a total of 47 significant associations (Table 11). Eleven associations involved signatures of exogenous etiology, 33 involved signatures of endogenous etiology and three involved signatures with no known etiology. There were seven negative associations, five of which had reciprocal associations that were removed from the final subset of significant associations selected by effect size. There were no recurrent pathogenic mutations to test for in CHOL, DLBC, KICH, KIRC, KIRP, MESO and SARC.

Four associations appeared in more than one cancer type. Associations of *PIK3CA* p.E542K with Signature 2 appeared in BRCA, LUAD_BRCA, LUSC_HNSC and LHB, while *PIK3CA* p.E545K associations with Signature 2 appeared in BRCA, CESC, LUAD_BRCA and LHB.

Table 11. Significant mutational signature associations obtained by individual testing of pathogenic mutations from samples in the non-hypermutator MC3 dataset.

Disease	Gene	Mutation	HGVSp	Signature	p_value	q_value	direction	effect size
BLCA	ERCC2	19_45867687_T_C	p.N238S	Signature 5	6.41E-12	1.15E-10	-1	3.28E-01
BRCA	TP53	17_7578263_G_A	p.R196*	Signature 3	8.46E-04	7.19E-03	-1	2.24E-01
BRCA	PIK3CA	3_178936082_G_A	p.E542K	Signature 2	9.87E-06	8.39E-05	-1	6.67E-02
BRCA	AKT1	14_105246551_C_T	p.E17K	Signature 7	9.80E-05	1.67E-03	-1	5.51E-02
BRCA	PIK3CA	3_178936091_G_A	p.E545K	Signature 2	4.31E-06	7.33E-05	-1	5.28E-02
CESC	PIK3CA	3_178936091_G_A	p.E545K	Signature 2	1.12E-06	4.47E-06	-1	1.27E-01
GBM	IDH1	2_209113112_C_T	p.R132H	Signature 1	1.08E-06	1.52E-05	1	3.84E-01
HNSC	TP53	17_7578190_T_C	p.Y220C	Signature 16	1.08E-04	2.92E-03	-1	1.98E-01
LGG	IDH1	2_209113113_G_A	p.R132C	Signature 15	1.40E-04	1.19E-03	-1	1.05E-01
LGG	TP53	17_7577121_G_A	p.R273C	Signature 15	1.51E-06	2.57E-05	-1	6.18E-02
LGG_GBM	IDH1	2_209113113_G_A	p.R132C	Signature 1	2.79E-04	2.88E-03	1	2.59E-01
LGG_GBM	TP53	17_7577121_G_A	p.R273C	Signature 1	4.40E-06	6.83E-05	1	1.90E-01
LGG_GBM	IDH1	2_209113112_C_T	p.R132H	Signature 1	1.06E-14	3.29E-13	1	1.68E-01
LGG_GBM	ATRX	X_76909629_G_A	p.R1426*	Signature 15	1.16E-03	9.00E-03	-1	1.20E-01
LHB	TP53	17_7577099_C_G	p.R280T	Signature 13	1.10E-05	3.20E-04	-1	2.03E-01
LHB	RXRA	9_137328351_C_T	p.S427F	Signature 2	2.11E-04	3.67E-03	-1	1.97E-01
LHB	TP53	17_7577085_C_T	p.E285K	Signature 13	2.59E-05	4.51E-04	-1	1.95E-01
LHB	ERBB2	17_37868208_C_T	p.S310F	Signature 2	1.00E-07	4.37E-06	-1	1.77E-01
LHB	ERBB2	17_37868208_C_T	p.S310F	Signature 13	1.70E-06	1.48E-04	-1	1.65E-01
LHB	KDM6A	X_44922802_C_T	p.Q555*	Signature 2	5.45E-04	5.92E-03	-1	1.53E-01
LHB	TP53	17_7577099_C_G	p.R280T	Signature 2	3.98E-04	4.95E-03	-1	1.38E-01
LHB	TP53	17_7577085_C_T	p.E285K	Signature 2	3.44E-04	4.95E-03	-1	1.35E-01
LHB	FGFR3	4_1803568_C_G	p.S249C	Signature 2	4.35E-06	9.46E-05	-1	9.58E-02
LHB	FGFR3	4_1803568_C_G	p.S249C	Signature 13	1.98E-05	4.30E-04	-1	9.53E-02
LHB	PIK3CA	3_178936091_G_A	p.E545K	Signature 2	5.08E-08	4.37E-06	-1	8.01E-02
LHB	PIK3CA	3_178936091_G_A	p.E545K	Signature 13	6.63E-06	2.88E-04	-1	7.40E-02
LHB	PIK3CA	3_178936082_G_A	p.E542K	Signature 2	2.38E-06	6.91E-05	-1	7.19E-02
LHB	PIK3CA	3_178936082_G_A	p.E542K	Signature 13	2.63E-04	3.81E-03	-1	6.78E-02
LIHC	TP53	17_7577534_C_A	p.R249S	Signature 24	1.77E-06	1.24E-05	-1	2.92E-01
LUAD	KRAS	12_25398284_C_A	p.G12V	Signature 4	6.13E-04	3.06E-03	-1	1.53E-01
LUAD	KRAS	12_25398285_C_A	p.G12C	Signature 4	3.25E-04	3.06E-03	-1	1.43E-01
LUAD	EGFR	7_55259515_T_G	p.L858R	Signature 1	2.77E-04	2.77E-03	-1	1.23E-01
LUAD_BRCA	PIK3CA	3_178936082_G_A	p.E542K	Signature 2	1.08E-07	1.77E-06	-1	6.91E-02
LUAD_BRCA	PIK3CA	3_178936091_G_A	p.E545K	Signature 2	5.75E-08	1.77E-06	-1	5.63E-02
LUSC	PIK3CA	3_178936082_G_A	p.E542K	Signature 4	1.50E-04	2.69E-03	1	2.31E-01
LUSC_HNSC	MAPK1	22_22127164_C_T	p.E322K	Signature 2	3.85E-04	6.93E-03	-1	1.26E-01
LUSC_HNSC	PIK3CA	3_178936082_G_A	p.E542K	Signature 2	8.58E-06	4.63E-04	-1	5.88E-02
LUSC_HNSC	TP53	17_7578190_T_C	p.Y220C	Signature 16	1.35E-04	7.27E-03	-1	5.20E-02
OV	TP53	17_7577538_C_T	p.R248Q	Signature 6	6.82E-04	6.13E-03	-1	1.22E-01
SKCM	KIT	4_55594221_A_G	p.K642E	Signature 7	9.51E-04	3.99E-03	1	6.00E-01
SKCM	PPP6C	9_127912080_G_A	p.R301C	Signature 7	1.91E-03	5.74E-03	-1	1.54E-01
SKCM	BRAF	7_140453137_C_T	p.V600M	Signature 7	2.84E-06	3.40E-05	-1	1.47E-01

SKCM	NRAS	1_115256529_T_C	p.Q61R	Signature 7	9.97E-04	3.99E-03	-1	8.59E-02
UCEC	PIK3CA	3_178916726_G_A	p.R38H	Signature 20	5.91E-08	3.55E-06	-1	2.08E-01
UVM	SF3B1	2_198267483_C_T	p.R625H	Signature 1	1.94E-03	9.70E-03	-1	3.63E-01
UVM	GNA11	19_3118942_A_T	p.Q209L	Signature 22	1.67E-11	8.34E-11	-1	1.41E-01
UVM	GNAQ	9_80409488_T_G	p.Q209P	Signature 22	2.39E-06	5.98E-06	1	9.83E-02

F) Hypermutator dataset pathogenic mutation individual testing

This analysis resulted in a total of 5 significant associations (Table 12). No associations involved signatures of exogenous etiology, four involved signatures of endogenous etiology and one involved signatures with no known etiology. Initially, there were two negative associations which were both removed while their reciprocal associations were kept as the true associations due to their greater effect size. One of these is the well-known association of *POLE* p.P286R with Signature 10, which is normally used to validate the methodology of signature association studies and confirms the quality of approaches used in this study. This also proves the rationale behind choosing true associations among reciprocal associations based on the effect size criterion that was previously described. Recurrent pathogenic mutations were present and tested in BLCA, COAD, SKCM, STAD and UCEC. No associations appeared in multiple cancer types.

Table 12. Significant mutational signature associations obtained by individual testing of pathogenic mutations from samples in the hypermutator MC3 dataset.

Disease	Gene	Mutation	HGVSp	Signature	p_value	q_value	direction	effect size
COAD	APC	5_112174631_C_T	p.R1114*	Signature 10	1.16E-06	5.79E-06	-1	7.85E-01
COAD	BRAF	7_140453136_A_T	p.V600E	Signature 6	2.87E-03	9.11E-03	-1	1.52E-01
UCEC	POLE	12_133253184_G_C	p.P286R	Signature 10	9.25E-08	2.96E-06	-1	7.44E-01
UCEC	NF1	17_29677227_C_T	p.R2450*	Signature 10	3.42E-04	5.47E-03	-1	6.32E-01
UCEC	PTEN	10_89692940_C_T	p.R142W	Signature 18	3.69E-05	9.19E-04	-1	2.37E-02

4.1.2.2. Benign mutations

G) Non-hypermutator dataset benign mutation individual testing

This analysis resulted in a total of seven significant associations (Table 13). Zero associations involved signatures of exogenous etiology, five involved signatures of endogenous etiology and two involved signatures with no known etiology. There was one negative association which had no reciprocal associations in the final subset of significant associations selected by effect size. There were no recurrent benign mutations to test for in ACC, CHOL, DLBC, KICH, KIRC, KIRP, LAML, MESO, PCPG, PRAD, SARC, TGCT, THCA, THYM, UCS and UVM. One association appeared in more than one cancer type: *NFE2L2* p.R34P was associated with Signature 18 in LUSC and LUSC_HNSC.

Table 13. Significant mutational signature associations obtained by individual testing of benign mutations from samples in the non-hypermutator MC3 dataset.

Disease	Gene	Mutation	HGVSp	Signature	p_value	q_value	direction	effect size
BRCA	PIK3CA	3_178938934_G_A	p.E726K	Signature 1	6.31E-05	5.05E-04	1	2.02E-01
LGG_GBM	PIK3CA	3_178952085_A_G	p.H1047R	Signature 6	5.48E-05	2.74E-04	-1	1.45E-01
LUAD	KRAS	12_25398284_C_T	p.G12D	Signature 1	2.62E-03	7.86E-03	-1	9.77E-02
LUSC	NFE2L2	2_178098944_C_G	p.R34P	Signature 18	5.66E-04	1.13E-03	-1	2.98E-02
LUSC_HNSC	TP53	17_7574018_G_A	p.R337C	Signature 15	1.41E-03	7.05E-03	-1	3.21E-02
LUSC_HNSC	NFE2L2	2_178098944_C_G	p.R34P	Signature 18	1.67E-03	8.34E-03	-1	2.98E-02
UCEC	TP53	17_7578406_C_T	p.R175H	Signature 13	2.28E-04	3.43E-03	-1	6.25E-02

H) Hypermutator dataset benign mutation individual testing

This analysis resulted in one significant association whose reciprocal associations were excluded (Table 14). It involved *PIK3CA* p.R88Q and Signature 10, which is of endogenous etiology. Recurrent benign mutations were present and tested in COAD, SKCM, STAD and UCEC. This association serves as an example of inconsistency among the output of different computational tools used to predict the functional consequences of mutations. During data preparation and filtering, *PIK3CA* p.R88Q was removed from the pathogenic mutation subset due the fact that it was annotated as tolerated by SIFT (score 0.06), while PolyPhen predicted it to be probably damaging (score 0.998). Upon subsequent inspection in COSMIC, this mutation was confirmed to be pathogenic and the association should therefore be regarded as an addition to results presented in Table 12. Thus, there were no significant signature

associations in this analysis, which is in agreement with anticipation for mutations predicted to be inconsequential (benign).

Table 14. Significant mutational signature associations obtained by individual testing of benign mutations from samples in the hypermutator MC3 dataset.

Disease	Gene	Mutation	HGVSp	Signature	p_value	q_value	direction	effect size
COAD	PIK3CA	3_178916876_G_A	p.R88Q	Signature 10	1.29E-06	2.59E-06	-1	6.75E-01

5. DISCUSSION

In this study, relationships between mutational signatures and a comprehensive set of somatic mutations were assessed across 33 cancer types. Separate analyses involved individual and burden testing of somatic mutations and were stratified based on sample and mutation characteristics. Individual mutation testing per-cancer type identified a total of 53 significant positive associations and nine significant negative associations across data subsets. Burden testing per-cancer type identified 146 significant positive associations and 16 significant negative associations across data subsets. Unless stated otherwise, the associations discussed further in the text were positive.

In order to validate the methodology used in this study, previously established associations were examined first. Recurrent mutations in the *POLE* exonuclease domain, such as p.P286R and p.V411L, alter the activity of error-prone polymerase epsilon and were proposed to be the underlying cause of Signature 10 (Alexandrov et al., 2013b, Kane and Shcherbakova, 2014). An association between *POLE* p.P286R and Signature 10 (q-value=2.96E-06, effect size=7.44E-01) was established in UCEC, as was done previously by Poulos et al., 2018, and Knijnenburg et al., 2018. In addition, burden testing based on hypermutated COAD samples confirmed that pathogenic mutations in *POLE* associate with Signature 10 (q-value=5.83E-07, effect size=6.80E-01), as well as revealed Signature 10 associations of similar effect size and q-value for pathogenic mutations in *MECOM*, *ATR*, *CHD9*, *BRCA2* and *REV3L*.

It should be noted that the association of pathogenic *POLE* mutations with Signature 10 was initially lost among significant results due to overcorrection by a large value of λ during genomic control. As shown by this association, which is normally used as a proof-of-method, genomic control can be too conservative for multiple testing correction in certain settings. While this study aimed to be rigorous and use a very conservative approach by applying both GC and subsequent BH correction to restrict the FDR at 1%, it was noted that values of λ were often larger than anticipated which may have caused a number of valid and biologically interesting results to be deemed statistically insignificant. This characteristic of λ was also described by others. Namely, it was demonstrated that the extent of genomic inflation is overestimated by the

inflation factor even if a moderate proportion of true associations is present (van Iterson, van Zwet and Heijmans, 2017). This is in line with expectations from this study, where the set of genes used in association testing was carefully selected based on their established relevance in oncogenesis and is anticipated to produce a set of results which is moderately or even strongly enriched in true associations with oncogenic processes.

A particularly strong association with Signature 10 was also present for *PIK3CA* p.R88Q in COAD (q-value=2.59E-06, effect size=6.75E-01), as identified recently in colorectal cancer and UCEC by Temko et al., 2018, and Poulos et al., 2018. Another strong link was established in COAD for *APC* p.R1114* and Signature 10 (q-value=5.79E-06, effect size=7.85E-01). This association had the largest effect size among all significant results in this study, surpassing even that of *POLE* p.P286R and Signature 10. This result was reported for colorectal cancer by Temko et al., 2018, where it was suggested that Signature 10 activity causes truncating driver lesions in the *APC* gene. Burden testing of pathogenic mutations additionally revealed the *APC* gene is linked to Signature 3 levels in GBM (q-value=1.63E-03, effect size=5.17E-01) and LGG_GBM.

A novel association of notable significance was found for *NF1* p.R2450* and Signature 10 (q-value=5.47E-03, effect size=6.32E-01). This mutation was not examined by previous studies of this kind. Whereas p.R2450* may be causal due to the fact that it is a truncating mutation affecting a tumor suppressor gene, the underlying mechanistic basis could be difficult to unravel due to the complexity of *NF1* signalling (Rad and Tee, 2016). In contrast, Signature 10 activity may cause this characteristic driver lesion in *NF1*, similar to what was previously posited for truncating *APC* lesions by Temko et al., 2018.

A well-known association of *ERCC2* mutations and Signature 5 from Kim et al., 2016, was also among the significant results. Specifically, it linked *ERCC2* p.N238S with Signature 5 in BLCA (q-value=1.15E-10, effect size=3.28E-01), which has not been previously reported. *ERCC2* mutations pointed towards deficiencies in the nucleotide-excision repair (NER) pathway as the underlying etiology of Signature 5 (Kim et al., 2016). The NER pathway is used to remove, among others, lesions caused by cisplatin. The p.N238S mutation is located in a highly conserved helicase motif, thus

disrupting protein function, and was identified in complete responders to cisplatin (Van Allen et al., 2014). Its strong association with Signature 5 and large accompanying effect size may delineate it as a priority candidate predictor of cisplatin sensitivity. Indeed, the translational relevance of *ERCC2* missense mutations, including p.N238S, was very recently recognized by demonstrating their direct roles in conferring NER deficiency and driving cisplatin response (Li et al., 2019).

The association of *BRAF* p.V600E and Signature 6 was reproduced in COAD (q-value=9.11E-03, effect size=1.52E-01), as reported for colorectal cancer by Poulos et al., 2018, and Temko et al., 2018. A known association involving *BRAF* p.V600M and Signature 7 (q-value=3.40E-05, effect size=1.47E-01) was also detected in SKCM (Poulos et al, 2018). The characteristic aflatoxin-induced *TP53* p.R249S mutation was linked to aflatoxin Signature 24 (q-value=1.24E-05, effect size=2.92E-01) in LIHC samples, as reported in Temko et al., 2018 and serves as a valuable positive control. In the set of pooled LGG and GBM samples, *ATRX* p.R1426* showed a correlation with Signature 15 levels (q-value=9.00E-03, effect size=1.20E-01). This mutation was previously correlated with Signature 14 in LGG by Temko et al., 2018.

The recurrent *IDH1* p.R132H mutation has been negatively correlated with Signature 1 levels in GBM and LGG_GBM samples in this study (GBM: q-value=1.52E-05, effect size=3.84E-01), as well as in previous studies, where it was also positively associated with Signature 5 and 6 levels in brain cancers (Poulos et al, 2018, Temko et al., 2018). Pathogenic mutation burden testing supported *IDH1* association with Signature 1 in non-hypermutated brain cancers. Previous studies did not find associations relating to *IDH1* p.R132C, which was negatively correlated with Signature 1 in LGG_GBM in this study, but also positively correlated with Signature 25 in LGG (q-value=1.19E-03, effect size=1.05E-01).

PIK3CA mutations were linked to a variety of different signatures. *PIK3CA* p.H1047R was previously linked to Signatures 6, 15, 20, 21 and 26 in colorectal cancer and STAD by Poulos et al., 2018 and Temko et al., 2018. A link with Signature 6 was also established in this study based on LGG_GBM samples (q-value=2.74E-04, effect size=1.45E-01). Other *PIK3CA* mutations and *ERBB2* p.S310F were previously linked to APOBEC-related signatures in multiple cancer types (Poulos et al., 2018 and Temko et al., 2018). The majority of those findings were obtained in this study, along with a

novel negative association involving *PIK3CA* p.E542K and Signature 4 in LUSC (q-value=2.69E-03, effect size=2.31E-01), which should be studied further with respect to patient smoking status.

Signature 4 levels were found to be increased with respect to *KRAS* mutation status in LUAD based on pathogenic mutation burden testing (q-value=4.90E-08, effect size=2.00E-01). This was supported by *KRAS* p.G12V (q-value=3.06E-03, effect size=1.53E-01) and p.G12C (q-value=3.06E-03, effect size=1.43E-01) Signature 4 associations in LUAD, the latter of which was also reported by Temko et al., 2018. Pathogenic mutation burden testing identified a series of Signature 4 associations of noteworthy effect size involving *ATF7IP*, *COL5A1*, *DMD*, *ERBB4*, *GABRA6*, *HERC2*, *HGF*, *KMT2C*, *NF1*, *PLCB4*, *PTPRC*, *PTPRD*, *SETBP1*, *TP53* and *ZFHX3*. These associations could be investigated further with respect to patient smoking history to clarify whether mutation burden in these genes is increased as a consequence of exposure to tobacco carcinogens.

Several previously unreported associations may be of interest for future investigation based on their noted small q-values along with substantial effect sizes. These include the associations of Signature 2 with pathogenic *ERBB3* and *SETD2* mutations in BRCA (q-values=8.73E-03 for both, respective effect sizes=2.85E-01 and 2.72E-01). In addition, pathogenic *SETD2* and *TET2* mutations were linked to Signature 13 levels in BRCA (respective q-values=1.58E-03 and 5.48E-03, effect sizes=2.55E-01 for both). In non-hypermutated STAD samples, pathogenic mutations in *TLR4* were linked to Signature 17, while mutations in *CHD6* were linked to Signature 6 (q-values=5.84E-03 and 6.45E-03, effect sizes=3.88E-01 and 3.28E-01, respectively). *EGFR* mutations were found to be associated with Signature 1 levels in LGG and LGG_GBM (q-values=2.16E-04 and 3.40E-06, effect sizes=2.25E-01 and 1.58E-01). Interestingly, the association of *EGFR* and Signature 1 appeared in another cancer type, LUAD, where *EGFR* p.L858R was identified as the significantly associated mutation (q-value=2.77E-03, effect size=1.23E-01).

Most associations described thus far were positive, indicating an increase in the levels of a given signature for samples harboring an alteration. However, additional negative associations of interest were present among significant results. Pathogenic mutation burden testing in hypermutated UCEC samples identified *POLD1*, *EP400*,

BAP1 and *NOTCH1* as negatively associated with Signature 10 levels with substantial effect sizes of ~0.6. Since UCEC is known to comprise of genetically heterogeneous subtypes, these associations should be further inspected by stratifying samples accordingly (Berger et al., 2018).

A novel negative Signature 1 association with *KMT2D* was identified in CESC (q-value=4.73E-04, effect size=1.53E-01). A recent study identified *KMT2D* as a methylation driver gene due to a strong association with genome-wide methylation changes, including hypermethylation in STAD and hypomethylation in BLCA (Youn et al., 2018). Another unreported negative Signature 1 association was identified for *NSD1* mutations in HNSC (q-value=3.71E-05, effect size=1.50E-01). In addition, significant associations with hypomethylation were reported by others for *NSD1* mutations in HNSC (Saghafinia et al., 2018). Considering that the etiology of Signature 1 involves the spontaneous deamination of 5-methylcytosines and that recent studies found global changes in 5-methylcytosine levels are related to *KMT2D* and *NSD1*, the associations reported here could be studied further in the context of methylation levels in order to infer potential causality.

KIT p.K642E was associated with substantially lower Signature 7 levels in SKCM samples (q-value=3.99E-03, effect size=6.00E-01). Previously, this mutation was positively correlated with Signatures 1 and 5 in melanoma in a study by Temko et al., 2018, where a one-sided test was applied due to which the Signature 7 association was overlooked, while other studies did not assess this mutation. *KIT* p.K642E most often does not overlap with other melanoma drivers (e.g. *NRAS*, *NF1* or *BRAF* mutations). This association is probably due to confounding from underlying stratification in SKCM samples, where the *KIT* mutation may define a subgroup of melanomas with lower Signature 7 levels. A recent genomic study supporting this explanation demonstrated that *KIT* is a driver specific to acral and mucosal melanomas and suggested that the principal mechanisms driving oncogenesis in these subtypes were not attributable to UV light exposure (Hayward et al., 2017).

The study by Hayward et al. also suggested that non-UV oncogenic mechanisms were shared by acral and mucosal melanoma subtypes and uveal melanoma, where they identified *GNAQ* and *SF3B1* as common driver genes. Interestingly, the three significant associations obtained by testing individual

pathogenic mutations in UVM involved *GNA11*, *GNAQ* and *SF3B1*. *SF3B1* p.R625H was associated with a substantial increase in Signature 1 levels (q-value=9.70E-03, effect size=3.63E-01), comparable to the one found for *IDH1* mutations. Mutually exclusive activating *GNA11* and *GNAQ* mutations, which appear in 85% of uveal melanomas and are thought to initiate tumorigenesis, both showed associations with Signature 22 (aristolochic acid). Specifically, *GNA11* p.Q209L was positively associated (q-value= 8.34E-11, effect size=1.41E-01), while *GNAQ* p.Q209P was negatively associated with Signature 22 levels (q-value=5.98E-06, effect size= 9.83E-02). These two results were also reflected in pathogenic mutation burden testing. The *GNA11* p.Q209L mutation (A>T) corresponds to the major mutational channel characterizing Signature 22 (T>A, referred to by the pyrimidine of the Watson-Crick pair). Interestingly, non-cutaneous melanomas account for a higher proportion of melanomas in Asians than Europeans (Chi et al., 2011) and the consumption of aristolochic acids is known to be prevalent in Asia. Although speculative, the potential link between aristolochic acid and non-cutaneous melanoma oncogenesis could be investigated further, firstly by verifying the presence of Signature 22 in UVM samples and then by stratifying samples based on the presence of mutually exclusive *GNAQ* and *GNA11* mutations.

Genes that were significantly associated with signatures based on benign mutation burden testing should not be interpreted as causal factors affecting signature levels. Rather, it could be hypothesized that the increased benign mutation load in these genes is a consequence of increased activity of corresponding mutational processes. The association of *SPTA1* with Signature 4 appeared in 2 cancer groups, LUAD and LUSC_HNSC, both in benign and pathogenic burden testing, which serves against this association being causal and supports the notion that the mutation burden in this gene is a consequence of higher Signature 4 activity.

Considering the correlative nature of association analyses, the existence of alternate possible explanations for these results, particularly where causality was suggested, should be acknowledged. In order to clarify whether a mutation that positively associates with a given signature is a consequence of the mutational process activity underlying that signature or its potential cause, additional analyses can be leveraged. For example, Temko et al. used the correspondence between a mutational channel enriched in a given signature and the channel of the associated driver mutation

as supporting evidence of the driver mutation being caused by a certain mutational process. Generally, the results of association studies should be interpreted with caution due to a number of factors that can affect the analysis and lead to false positive or negative results. These include the quality of data processing and sample stratification, discrepancies between tools used to annotate individual mutation effects, as well as dependance on tools used in assigning signature contributions to individual samples which are known to be affected by the number of mutations present in a sample (Alexandrov et al., 2013a).

The analyses presented in this study could be expanded in several future directions. Firstly, additional genomic alteration types including copy number alterations, small insertions and deletions, as well as methylation status could be incorporated in the analysis along with somatic mutations or be examined separately from somatic mutations using the same methodology (Knijnenburg et al., 2018). Recording the presence of different types of deleterious alterations for a given gene in a single binary hit matrix could increase the number of samples where this gene can be tested for association with signatures and lead to an increase in power to detect true findings. Separate alteration types, however, may have biologically distinct effects on mutational processes and pooling them all together for a given gene may lead to confounding. In addition to incorporating additional alterations, the repertoire of mutational signatures being tested can be expanded to include the newly identified 49 single base substitution, 11 doublet base substitution, four clustered base substitution and 17 insertion and deletion mutational signatures (Alexandrov et al., 2018).

Beside expanding the scope of the analysis, it could also be refined in several ways to provide clearer mechanistic insight, especially for selected results of interest. Samples within a given cancer type could be stratified based on the newest established molecular subtypes (Berger et al., 2018, Campbell et al., 2018, Ricketts et al., 2018, Liu et al., 2018). Where relevant, such as in associations with aging- and smoking-related signatures, stratification can be further performed based on clinical information. Power could be increased by grouping genes into molecular pathways or grouping similar molecular subtypes of different cancers. Specificity could be enhanced by classifying mutations as benign or pathogenic using a consensus-based approach that considers predictions of more than two functional prediction algorithms, as well as with manual curation of the resulting mutation sets. Lastly, a number of different burden

testing approaches have been developed so far that could be employed in this setting and yield varying results (Wagner 2013).

Caution was taken to avoid biasing the analyses conducted within this study and the applied methodology was shown to be successful in identifying several important proof-of-concept results along with a number of recent findings published by others in the past year. In addition to these, a large number of associations identified in this study were previously unknown. Based on the validity of the approaches used, some of these novel associations could provide important oncogenic insights upon future investigation.

6. CONCLUSION

This study statistically assessed the relationships of a comprehensive tumor somatic mutation set from a list of 721 genes with the 30 COSMIC mutational signatures and their proxies across ~10,000 samples of 33 different cancer types. Two-sided nonparametric testing on both pan-cancer and per-cancer type levels was employed to individually test recurrent somatic mutations, as well as to perform burden testing of recurrently mutated genes based on the TCGA MC3 dataset. Analyses were stratified based on sample hypermutation status and predicted benign or pathogenic character of mutations. Results were submitted to genomic control and Benjamini-Hochberg multiple testing correction procedures and subsequently filtered using absolute effect sizes. Based on the number of genes and mutations included in association testing, the scope of this study was an order of magnitude greater in comparison to previous similar studies. Individual mutation testing identified a total of 62 significant associations, while burden testing identified 162 significant associations on the per-cancer type level across data subsets. Among these, well-known associations including *POLE* and Signature 10 were identified, serving as a validation of the methodology that was used. The majority of identified associations were novel findings and constitute potentially valuable targets for future research into the mutational processes operative in cancer cells.

7. REFERENCES

Adzhubei, I., Jordan, D. and Sunyaev, S. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, 76(1), pp.7.20.1-7.20.41.

Alexandrov, L., Nik-Zainal, S., Wedge, D., Campbell, P. and Stratton, M. (2013a). Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*, 3(1), pp.246-259.

Alexandrov, L. et al. (2013b). Signatures of mutational processes in human cancer. *Nature*, 500(7463), pp.415-421.

Alexandrov, L. and Stratton, M. (2014). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current Opinion in Genetics & Development*, 24, pp.52-60.

Alexandrov, L., Jones, P., Wedge, D., Sale, J., Campbell, P., Nik-Zainal, S. and Stratton, M. (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47(12), pp.1402-1407.

Alexandrov, L., Kim, J., Haradhvala, N., Huang, M., Ng, A., Boot, A., Covington, K., Gordenin, D., Bergstrom, E., Lopez-Bigas, N., Klimczak, L., McPherson, J., Morganella, S., Sabarinathan, R., Wheeler, D., Mustonen, V., Getz, G., Rozen, S. and Stratton, M. (2018). The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv*. [online] Available at: https://doi.org/10.1101/322859 [Accessed 17 Feb. 2019].

Asimit, J. and Zeggini, E. (2010). Rare Variant Association Analysis Methods for Complex Traits. *Annual Review of Genetics*, 44(1), pp.293-308.

Bailey, M. et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, 173(2), pp.371-385.e18.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), pp.289-300.

Berger, A. et al. (2018). A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell*, 33(4), pp.690-705.e9.

Buckley, A., Ideker, T., Carter, H., Harismendy, O. and Schork, N. (2018). Exome-wide analysis of bi-allelic alterations identifies a Lynch phenotype in The Cancer Genome Atlas. *Genome Medicine*, 10(1).

Campbell, J. et al. (2018). Genomic, Pathway Network, and Immunologic Features Distinguishing Squamous Carcinomas. *Cell Reports*, 23(1), pp.194-212.e6.

Caridi, C., D'Agostino, C., Ryu, T., Zapotoczny, G., Delabaere, L., Li, X., Khodaverdian, V., Amaral, N., Lin, E., Rau, A. and Chiolo, I. (2018). Nuclear F-actin and myosins drive relocalization of heterochromatic breaks. *Nature*, 559(7712), pp.54-60.

Castro-Giner, F., Ratcliffe, P. and Tomlinson, I. (2015). The mini-driver model of polygenic cancer evolution. *Nature Reviews Cancer*, 15(11), pp.680-685.

Devlin, B. and Roeder, K. (1999). Genomic Control for Association Studies. *Biometrics*, 55(4), pp.997-1004.

Chi, Z., Li, S., Sheng, X., Si, L., Cui, C., Han, M. and Guo, J. (2011). Clinical presentation, histology, and prognoses of malignant melanoma in ethnic Chinese: A study of 522 consecutive cases. *BMC Cancer*, 11(1).

Devlin, B., Bacanu, S. and Roeder, K. (2004). Genomic Control to the extreme. *Nature Genetics*, 36(11), pp.1129-1130.

Ellrott, K. et al. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Systems*, 6(3), pp.271-281.e7.

Futreal, P., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M. (2004). A census of human cancer genes. *Nature Reviews Cancer*, 4(3), pp.177-183.

Garraway, L. and Lander, E. (2013). Lessons from the Cancer Genome. *Cell*, 153(1), pp.17-37.

Goeman, J. and Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11), pp.1946-1978.

Grossman, R., Heath, A., Ferretti, V., Varmus, H., Lowy, D., Kibbe, W. and Staudt, L. (2016). Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, 375(12), pp.1109-1112.

Hanahan, D. and Weinberg, R. (2000). The Hallmarks of Cancer. *Cell*, 100(1), pp.57-70.

Hayward, N., Wilmott, J., Waddell, N., Johansson, P., Field, M., Nones, K., Patch, A., Kakavand, H., Alexandrov, L., Burke, H., Jakrot, V., Kazakoff, S., Holmes, O., Leonard, C., Sabarinathan, R., Mularoni, L., Wood, S., Xu, Q., Waddell, N., Tembe, V., Pupo, G., De Paoli-Iseppi, R., Vilain, R., Shang, P., Lau, L., Dagg, R., Schramm, S., Pritchard, A., Dutton-Regester, K., Newell, F., Fitzgerald, A., Shang, C., Grimmond, S., Pickett, H., Yang, J., Stretch, J., Behren, A., Kefford, R., Hersey, P., Long, G., Cebon, J., Shackleton, M., Spillane, A., Saw, R., López-Bigas, N., Pearson, J., Thompson, J., Scolyer, R. and Mann, G. (2017). Whole-genome landscapes of major melanoma subtypes. *Nature*, 545(7653), pp.175-180.

Hoadley, K. et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2), pp.291-304.e6.

Hodges, J. and Lehmann, E. (1963). Estimates of Location Based on Rank Tests. *The Annals of Mathematical Statistics*, 34(2), pp.598-611.

Huang, P., Chiu, L., Lee, C., Yeh, Y., Huang, K., Chiu, C. and Tang, P. (2017). mSignatureDB: a database for deciphering mutational signatures in human cancers. *Nucleic Acids Research*, 46(D1), pp.D964-D970.

Hudson (Chairperson), T. et al. (2010). International network of cancer genome projects. *Nature*, 464(7291), pp.993-998.

Kane, D. and Shcherbakova, P. (2014). A Common Cancer-Associated DNA Polymerase Mutation Causes an Exceptionally Strong Mutator Phenotype, Indicating Fidelity Defects Distinct from Loss of Proofreading. *Cancer Research*, 74(7), pp.1895-1901.

Kim, J., Mouw, K., Polak, P., Braunstein, L., Kamburov, A., Tiao, G., Kwiatkowski, D., Rosenberg, J., Van Allen, E., D'Andrea, A. and Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature Genetics*, 48(6), pp.600-606.

Knijnenburg, T., et al. (2018). Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Reports*, 23(1), pp.239-254.e6.

Kumar, P., Henikoff, S. and Ng, P. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7), pp.1073-1081.

Kumar, R., Li, D., Müller, S. and Knapp, S. (2016). Epigenomic regulation of oncogenesis by chromatin remodeling. *Oncogene*, 35(34), pp.4423-4436.

Lehman, T., Greenblatt, M., Bennett, W. and Harris, C. (1994). Mutational Spectrum of the P53 Tumor Suppressor Gene: Clues to Cancer Etiology and Molecular Pathogenesis. *Drug Metabolism Reviews*, 26(1-2), pp.221-235.

Lek, M. et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), pp.285-291.

Li, Q., Damish, A., Frazier, Z., Liu, D., Reznichenko, E., Kamburov, A., Bell, A., Zhao, H., Jordan, E., Gao, S., Ma, J., Abbosh, P., Bellmunt, J., Plimack, E., Lazaro, J., Solit, D., Bajorin, D., Rosenberg, J., D'Andrea, A., Riaz, N., Van Allen, E., Iyer, G. and Mouw, K. (2019). ERCC2 Helicase Domain Mutations Confer Nucleotide Excision Repair Deficiency and Drive Cisplatin Sensitivity in Muscle-Invasive Bladder Cancer. *Clinical Cancer Research*, 25(3), pp.977-988.

Liu, Y. et al. (2018). Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell*, 33(4), pp.721-735.e8.

Lu, C. et al. (2015). Patterns and functional implications of rare germline variants across 12 cancer types. *Nature Communications*, 6(1).

Marchini, J., Cardon, L., Phillips, M. and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5), pp.512-517.

Medvedeva, Y., Lennartsson, A., Ehsani, R., Kulakovskiy, I., Vorontsov, I., Panahandeh, P., Khimulya, G., Kasukawa, T. and Drabløs, F. (2015). EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database*, 2015, p.bav067.

Nakagawa, H. and Fujita, M. (2018). Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Science*, 109(3), pp.513-522.

Nawy, T. (2018). A pan-cancer atlas. *Nature Methods*, 15(6), pp.407-407.

Nik-Zainal, S. et al. (2012). Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, 149(5), pp.979-993.

Nik-Zainal, S. et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605), pp.47-54.

Nik-Zainal, S., Wedge, D., Alexandrov, L., Petljak, M., Butler, A., Bolli, N., Davies, H., Knappskog, S., Martin, S., Papaemmanuil, E., Ramakrishna, M., Shlien, A., Simonic, I., Xue, Y., Tyler-Smith, C., Campbell, P. and Stratton, M. (2014). Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nature Genetics*, 46(5), pp.487-491.

Petljak, M. and Alexandrov, L. (2016). Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis*, 37(6), pp.531-540.

Pleasance, E. et al. (2009). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278), pp.191-196.

Poulos, R., Wong, Y., Ryan, R., Pang, H. and Wong, J. (2018). Analysis of 7,815 cancer exomes reveals associations between mutational processes and somatic driver mutations. *PLOS Genetics*, 14(11), p.e1007779.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rad, E. and Tee, A. (2016). Neurofibromatosis type 1: Fundamental insights into cell signalling and cancer. *Seminars in Cell & Developmental Biology*, 52, pp.39-46.

Ricketts, C. et al. (2018). The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Reports*, 23(1), pp.313-326.e5.

Saghafinia, S., Mina, M., Riggi, N., Hanahan, D. and Ciriello, G. (2018). Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors. *Cell Reports*, 25(4), pp.1066-1080.e8.

Sanger, F., Nicklen, S. and Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), pp.5463-5467.

Schrank, B., Aparicio, T., Li, Y., Chang, W., Chait, B., Gundersen, G., Gottesman, M. and Gautier, J. (2018). Nuclear ARP2/3 drives DNA break clustering for homology-directed repair. *Nature*, 559(7712), pp.61-66.

Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16), pp.9440-9445.

Stratton, M., Campbell, P. and Futreal, P. (2009). The cancer genome. *Nature*, 458(7239), pp.719-724.

Tate, J., Bamford, S., Jubb, H., Sondka, Z., Beare, D., Bindal, N., Boutselakis, H., Cole, C., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S., Kok, C., Noble, K., Ponting, L., Ramshaw, C., Rye, C., Speedy, H., Stefancsik, R., Thompson, S., Wang, S., Ward, S., Campbell, P. and Forbes, S. (2018). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1), pp.D941-D947.

Temko, D., Tomlinson, I., Severini, S., Schuster-Böckler, B. and Graham, T. (2018). The effects of mutational processes and selection on driver mutations across cancer types. *Nature Communications*, 9(1).

The future of cancer genomics. (2015). *Nature Medicine*, [online] 21(2), pp.99-99. Available at: https://doi.org/10.1038/nm.3801 [Accessed 17 Feb. 2019].

Tsepilov, Y., Ried, J., Strauch, K., Grallert, H., van Duijn, C., Axenovich, T. and Aulchenko, Y. (2013). Development and Application of Genomic Control Methods for Genome-Wide Association Studies Using Non-Additive Models. *PLoS ONE*, 8(12), p.e81431.

Van Allen, E., Mouw, K., Kim, P., Iyer, G., Wagle, N., Al-Ahmadie, H., Zhu, C., Ostrovnaya, I., Kryukov, G., O'Connor, K., Sfakianos, J., Garcia-Grossman, I., Kim, J., Guancial, E., Bambury, R., Bahl, S., Gupta, N., Farlow, D., Qu, A., Signoretti, S., Barletta, J., Reuter, V., Boehm, J., Lawrence, M., Getz, G., Kantoff, P., Bochner, B., Choueiri, T., Bajorin, D., Solit, D., Gabriel, S., D'Andrea, A., Garraway, L. and Rosenberg, J. (2014). Somatic ERCC2 Mutations Correlate with Cisplatin Sensitivity in Muscle-Invasive Urothelial Carcinoma. *Cancer Discovery*, 4(10), pp.1140-1153.

van Iterson, M., van Zwet, E. and Heijmans, B. (2017). Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology*, 18(1).

Vogelstein, B. and Kinzler, K. (1992). Carcinogens leave fingerprints. *Nature*, 355(6357), pp.209-210.

Vogelstein, B., Papadopoulos, N., Velculescu, V., Zhou, S., Diaz, L. and Kinzler, K. (2013). Cancer Genome Landscapes. *Science*, 339(6127), pp.1546-1558.

Wagner, M. (2013). Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. *Pharmacogenomics*, 14(4), pp.413-424.

Waszak, S. et al. (2017). Germline determinants of the somatic mutation landscape in 2,642 cancer genomes.

Youn, A., Kim, K., Rabadan, R., Tycko, B., Shen, Y. and Wang, S. (2018). A pan-cancer analysis of driver gene mutations, DNA methylation and gene expressions reveals that chromatin remodeling is a major mechanism inducing global changes in cancer epigenomes. *BMC Medical Genomics*, 11(1).

LINKS TO ONLINE RESOURCES

https://cancer.sanger.ac.uk

https://cancer.sanger.ac.uk/cosmic/signatures

https://gdc.cancer.gov/about-data/publications/mc3-2017

mc3.v0.2.8.PUBLIC.maf.gz

https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/

http://epifactors.autosome.ru/genes

http://www.dnarepairgenes.com/chromremodgenes.html
http://amigo.geneontology.org/amigo/term/GO:0006338

https://www.genenames.org/cgi-bin/symbol_checker

http://tardis.cgu.edu.tw/msignaturedb/

8. SUPPLEMENTARY MATERIAL

8.1. CURRICULUM VITAE

PERSONAL INFORMATION

E-mail: tomljanovic.ingrid@gmail.com

Date and place of birth: 29/10/1994, Zagreb, Croatia

Languages (CEFR): Croatian (C2), English (C2), French (B1/B2)

EDUCATION

MSc in Molecular Biology (2016 - 2019)

University of Zagreb, Faculty of Science | Zagreb, Croatia

GPA: 4.733/5 (thesis grade N/A)

MSc thesis internship (2018)

Institute for Research in Biomedicine (IRB Barcelona) | Barcelona, Spain

BSc in Molecular Biology, cum laude (2013 - 2016)

University of Zagreb, Faculty of Science | Zagreb, Croatia

GPA: 4.571/5 (195 ECTS credits)

ERASMUS+ study abroad (2016)

Universitat de Vic | Vic, Spain

High school degree (2009 - 2011, 2012 - 2013)

Classical High School of Zagreb | Zagreb, Croatia

ASSIST Scholar (2011 - 2012)

The White Mountain School | Bethlehem, New Hampshire

8.2. SUPPLEMENTARY RESULTS

Supplementary Table 1. HGNC symbols of 721 genes used in this study. The list of gene symbols was distributed into multiple columns (V1-V9) for ease of representation.

V1	V2	V3	V4	V5	V6	V7	V8	V9
APLF	DCLRE1C	POLE4	TDP2	NASP	TOP1MT	AXIN2	HLA-A	PTPRD
APTX	DDB1	POLG	ENDOV	NCOA2	SP140L	B2M	HLA-B	RAC1
ASCC3	DDB2	POLH	SPRTN	NFRKB	BAZ1A	BAP1	HRAS	RAD21
DNTT	DMC1	POLI	RNF4	PHF10	GATA3	BCL2	HUWE1	RAF1
LIG1	DNA2	POLK	SMARCA4	JADE1	PAX7	BCL2L11	IDH2	RARA
LIG3	DUT	POLN	IDH1	PHF19	RERE	BCOR	IL6ST	RASA1
LIG4	EID3	POLQ	SOX4	PIWIL4	SP140	BRAF	IL7R	RBM10
MRE11A	EME1	PPP4C	WEE1	PSIP1	MSL3	BRD7	INPPL1	RET
NBN	EME2	PPP4R1	RAD9B	RAD54L2	FOXP3	BTG2	IRF2	RHEB
NHEJ1	ERCC1	PPP4R2	AEN	RB1	HDAC2	CACNA1A	IRF6	RHOA
PARG	ERCC2	PPP4R4	PLK3	RSF1	MYC	CARD11	JAK1	RHOB
PARP1	ERCC3	PRPF19	EXO5	RUVBL1	KAT2B	CASP8	JAK2	RIT1
PARP3	ERCC4	RAD1	CDC5L	RUVBL2	TOP1	CBFB	JAK3	RNF111
PARPBP	ERCC5	RAD17	BCAS2	SETD6	RBBP4	CBWD3	KANSL1	RNF43
PNKP	ERCC6	RAD18	PLRG1	SMARCA1	CBX3	CCND1	KDM5C	RPL22
POLB	ERCC8	RAD23A	YWHAB	SMARCA5	HDAC4	CD70	KEL	RPL5
POLL	FAM175A	RAD23B	YWHAG	SMARCAL1	NPM1	CD79B	KIF1A	RPS6KA3
POLM	FAN1	RAD51	YWHAE	SMARCC2	ANP32B	CDH1	KIT	RQCD1
PRKDC	FANCA	RAD51B	CDC25A	SMARCD1	SP110	CDK12	KLF5	RRAS2
RAD50	FANCB	RAD51C	CDC25B	SMARCD2	HMGXB4	CDK4	KMT2A	RUNX1
RNF168	FANCC	RAD51D	CDC25C	SMARCD3	UBTF	CDKN1A	KMT2B	RXRA
RNF8	FANCD2	RAD52	BABAM1	SMARCE1	PADI4	CDKN1B	KMT2C	SCAF4
TP53BP1	FANCE	RAD54B	BRCC3	SRCAP	MYB	CDKN2A	KMT2D	SETBP1
XRCC1	FANCF	RAD54L	TTK	TFPT	SCMH1	CDKN2C	KRAS	SETD2
XRCC2	FANCG	RAD9A	SMARCC1	TOP2A	ESR1	CEBPA	KRT222	SF1
XRCC3	FANCI	RBBP8	SWI5	TOP2B	NPM3	CIC	LATS1	SF3B1
XRCC4	FANCL	RBX1	MORF4L1	TP73	TADA2A	CNBD1	LATS2	SIN3A
XRCC5	FANCM	RDM1	RNF169	VPS72	DAXX	COL5A1	LEMD2	SMAD2
XRCC6	FEN1	RECQL	HERC2	YY1	ANP32D	CREB3L3	LZTR1	SMAD4
UBE2A	GADD45A	RECQL4	ACTB	ZHX1	TAF6L	CREBBP	MACF1	SMC1A
EXO1	GADD45G	RECQL5	ACTL6A	ZNHIT1	NUDT5	CSDE1	MAP2K1	SMC3
HMGB1	GEN1	REV1	ACTL6B	ZRANB3	ZBTB1	CTNNB1	MAP2K4	SOS1
MLH1	GTF2H1	REV3L	ACTR3B	ADNP	TOX	CTNND1	MAP3K1	SOX17
MLH3	GTF2H2	RIF1	ACTR5	ANP32A	KAT2A	CYLD	MAP3K4	SPOP
MSH2	GTF2H3	RMI1	ACTR6	CDC6	SMARCA2	CYSLTR2	MAPK1	SPTA1
MSH3	GTF2H4	RMI2	ARID1A	CUL1	MORF4L2	DACH1	MAX	SPTAN1
MSH6	GTF2H5	RNMT	ARID2	CUL2	SUPT6H	DAZAP1	MECOM	SRSF2
PCNA	H2AFX	RRM2B	ATAD2	DNTTIP2	HMGB4	DDX3X	MED12	STAG2
PMS1	HELQ	RTEL1	BAHD1	EPC2	SOX9	DHX9	MEN1	STK11
PMS2	HES1	SETMAR	BAZ2A	FOXA1	BRDT	DIAPH2	MET	TAF1
POLD1	HFM1	SHFM1	BRD4	GFI1	PBRM1	DICER1	MGA	TBL1XR1
POLD2	HLTF	SHPRH	BPTF	HDGF	TADA2B	DMD	MTOR	TBX3
POLD3	HMGB2	SLX1A	BRMS1	HMG20A	SUPT4H1	DNMT3A	MUC6	TCF12
POLD4	HUS1	SLX1B	CHAF1B	HMGN1	TFAM	EEF1A1	MYCN	TCF7L2
RFC1	INO80	SLX4	CHD1	HMGN2	TOX4	EEF2	MYD88	TET2
RFC2	KAT5	SMARCAD1	CHD1L	HMGN3	CHRAC1	EGFR	MYH9	TGFBR2
RFC3	MAD2L2	SMC5	CHD2	HMGN4	SUV39H2	EGR3	NCOR1	TGIF1
RFC4	MBD4	SMC6	CHD3	HMGN5	ACTR8	EIF1AX	NF1	THRAP3
RFC5	MDC1	SMUG1	CHD4	HP1BP3	PIH1D1	ELF3	NF2	TLR4
RPA1	MGMT	SPO11	CHD5	JDP2	ANP32E	EP300	NIPBL	TMSB4X

RPA2	MMS19	STRA13	CHD6	KEAP1	TOX2	EPAS1	NOTCH1	TNFAIP3
RPA3	MNAT1	SWSAP1	CHD7	MAPKAPK3	HMGB3	EPHA2	NOTCH2	TRAF3
RPA4	MPG	TCEA1	CHD8	MAZ	HDAC1	EPHA3	NRAS	TSC1
ALKBH1	MPLKIP	TCEB1	CHD9	MLLT1	NPM2	ERBB2	NSD1	TSC2
ALKBH2	MRPL40	TCEB2	CTBP1	NFYB	KDM6B	ERBB3	NUP133	TXNIP
ALKBH3	MUS81	TCEB3	CTCF	NOC2L	KDM6A	ERBB4	NUP93	U2AF1
APEX1	MUTYH	TDG	CTCFL	NPAS2	MYSM1	EZH2	PAX5	UNCX
APEX2	NABP2	TDP1	CXXC1	PPM1G	BAZ1B	FAM46D	PCBP1	USP9X
APITD1	NEIL1	TELO2	DEK	RAI1	TP63	FAT1	PDGFRA	VHL
ATM	NEIL2	TOP3A	DMAP1	RCC1	SMYD1	FBXW7	PDS5B	WHSC1
ATR	NEIL3	TOP3B	DPF1	SAFB	NFE2L2	FGFR1	PGR	WT1
ATRIP	NFATC2IP	TOPBP1	DPF2	SATB1	ABL1	FGFR2	PHF6	XPO1
ATRX	NSMCE1	TP53	DPF3	SATB2	ACVR1	FGFR3	PIK3CA	ZBTB20
BARD1	NSMCE2	TREX1	EP400	SFPQ	ACVR1B	FLNA	PIK3CB	ZBTB7B
BLM	NSMCE4A	TREX2	GADD45B	SP1	ACVR2A	FLT3	PIK3CG	ZC3H12A
BRCA1	NTHL1	TYMS	HCFC1	SP100	AJUBA	FOXA2	PIK3R1	ZCCHC12
BRCA2	NUDT1	UBE2B	HELLS	SS18L1	AKT1	FOXQ1	PIK3R2	ZFHX3
BRE	NUDT15	UBE2N	HMG20B	SS18L2	ALB	FUBP1	PIM1	ZFP36L1
BRIP1	NUDT18	UBE2T	IKZF1	SSRP1	ALK	GABRA6	PLCB4	ZFP36L2
CCNH	RRM1	UBE2V2	ING3	TLE1	AMER1	GNA11	PLCG1	ZMYM2
CDK7	RRM2	UIMC1	INO80B	TNP1	APC	GNA13	PLXNB2	ZMYM3
CETN2	OGG1	UNG	INO80C	TNP2	APOB	GNAQ	POLRMT	ZNF133
CHAF1A	PALB2	USP1	INO80D	TONSL	AR	GNAS	PPM1D	ZNF750
CHEK1	PARP2	UVSSA	INO80E	UBR5	ARAF	GPS2	PPP2R1A	
CHEK2	PARP4	WDR48	LRWD1	VDR	ARHGAP35	GRIN2D	PPP6C	
CLK2	PAXIP1	WRN	MBD5	WHSC1L1	ARID5B	GTF2I	PRKAR1A	
CUL3	PER1	XAB2	MBD6	YAF2	ASXL1	H3F3A	PTCH1	
CUL4A	POLA1	XPA	MTA1	ZNF541	ASXL2	H3F3C	PTMA	
CUL5	POLE	XPC	MTA3	SMARCB1	ATF7IP	HGF	PTPDC1	
DCLRE1A	POLE2	ZSWIM7	MYBBP1A	TOX3	ATXN3	HIST1H1C	PTPN11	
DCLRE1B	POLE3	PTEN	MYO1C	HDAC5	AXIN1	HIST1H1E	PTPRC	

Supplementary Table 2. Significant associations obtained per-cancer type by burden testing of pathogenic mutations from samples in the complete dataset comprising all MC3 samples. This analysis resulted in a total of 78 significant associations. 5 associations involved signatures of exogenous etiology, 70 involved signatures of endogenous etiology and 3 involved signatures with no known etiology.

Disease	Gene	Signature	p_value	q_value	direction	effect size
BLCA	ERCC2	Signature 13	2.00E-05	5.11E-03	1	1.56E-01
BRCA	SETD2	Signature 13	1.26E-05	1.09E-03	-1	2.59E-01
BRCA	TET2	Signature 13	7.05E-05	4.07E-03	-1	2.43E-01
BRCA	ERCC6	Signature 26	4.02E-05	2.32E-03	-1	1.60E-01
BRCA	FOXA1	Signature 2	5.06E-06	4.38E-04	-1	1.37E-01
CESC	KMT2B	Signature 1	8.83E-05	3.71E-03	1	2.10E-01
CESC	MACF1	Signature 1	3.13E-04	7.88E-03	1	2.03E-01
CESC	KMT2D	Signature 1	8.11E-07	1.02E-04	1	1.57E-01
CESC	HUWE1	Signature 1	1.83E-04	5.75E-03	1	1.49E-01
CESC	KMT2C	Signature 1	3.80E-05	2.39E-03	1	1.22E-01
CESC	PIK3CA	Signature 2	2.32E-05	2.93E-03	-1	8.91E-02
COAD	RECQL	Signature 20	6.11E-10	1.89E-07	-1	1.69E-01
GBM	APC	Signature 1	2.58E-04	4.82E-03	1	6.35E-01
GBM	MACF1	Signature 1	1.31E-04	3.68E-03	1	5.87E-01

GBM	IDH1	Signature 1	2.53E-07	1.42E-05	1	3.93E-01
GBM	SMARCA2	Signature 4	1.09E-04	6.10E-03	-1	7.09E-02
GBM	HUWE1	Signature 25	3.40E-05	1.90E-03	-1	6.50E-02
GBM	PTPRD	Signature 10	2.86E-05	1.60E-03	-1	5.43E-02
HNSC	NSD1	Signature 1	1.57E-07	2.57E-05	1	1.51E-01
HNSC	APOB	Signature 1	6.00E-05	4.92E-03	1	1.49E-01
KIRP	FGFR3	Signature 2	1.98E-04	4.35E-03	-1	9.38E-02
LGG	EGFR	Signature 1	2.60E-05	2.42E-04	-1	2.26E-01
LGG	IDH1	Signature 1	1.44E-06	2.01E-05	1	1.76E-01
LGG	TP53	Signature 1	9.88E-08	2.77E-06	1	1.33E-01
LGG_GBM	APC	Signature 1	1.29E-04	2.69E-03	1	5.10E-01
LGG_GBM	MACF1	Signature 1	2.47E-05	6.17E-04	1	4.03E-01
LGG_GBM	IDH1	Signature 1	2.08E-22	2.60E-20	1	2.13E-01
LGG_GBM	ATRX	Signature 1	6.04E-07	2.50E-05	1	1.59E-01
LGG_GBM	EGFR	Signature 1	8.00E-07	2.50E-05	-1	1.49E-01
LGG_GBM	TP53	Signature 1	4.28E-09	2.68E-07	1	1.30E-01
LGG_GBM	PTEN	Signature 1	3.52E-04	6.28E-03	-1	1.09E-01
LGG_GBM	BAZ1B	Signature 10	4.73E-11	5.91E-09	-1	1.08E-01
LGG_GBM	ZMYM2	Signature 10	1.01E-08	4.22E-07	-1	1.08E-01
LGG_GBM	BLM	Signature 10	2.10E-05	1.64E-04	-1	1.08E-01
LGG_GBM	CHD3	Signature 10	2.10E-05	1.64E-04	-1	1.08E-01
LGG_GBM	CHD6	Signature 10	2.10E-05	1.64E-04	-1	1.08E-01
LGG_GBM	EP300	Signature 10	2.10E-05	1.64E-04	-1	1.08E-01
LGG_GBM	MET	Signature 10	2.10E-05	1.64E-04	-1	1.08E-01
LGG_GBM	SUPT6H	Signature 10	2.10E-05	1.64E-04	-1	1.08E-01
LGG_GBM	TCF7L2	Signature 10	2.10E-05	1.64E-04	-1	1.08E-01
LGG_GBM	UBR5	Signature 10	2.10E-05	1.64E-04	-1	1.08E-01
LGG_GBM	CXXC1	Signature 10	4.19E-09	2.62E-07	-1	1.08E-01
LGG_GBM	KAT5	Signature 10	3.74E-05	2.63E-04	-1	5.43E-02
LGG_GBM	PTPRD	Signature 10	3.15E-06	6.56E-05	-1	5.43E-02
LGG_GBM	EPAS1	Signature 10	2.51E-07	7.84E-06	-1	5.12E-02
LHB	ERBB2	Signature 2	4.58E-06	1.15E-03	-1	1.39E-01
LHB	MACF1	Signature 2	2.05E-05	3.43E-03	-1	8.84E-02
LHB	ARID1A	Signature 2	2.83E-05	3.56E-03	-1	8.43E-02
LHB	PIK3CA	Signature 13	1.72E-05	8.65E-03	-1	6.78E-02
LHB	PIK3CA	Signature 2	1.08E-07	5.45E-05	-1	6.61E-02
LUAD_BRCA	FOXA1	Signature 2	8.51E-06	1.68E-03	-1	1.86E-01
LUAD_BRCA	PIK3CA	Signature 2	2.91E-11	1.15E-08	-1	5.15E-02
LUSC_HNSC	SPTA1	Signature 4	2.00E-07	7.48E-05	-1	2.07E-01
LUSC_HNSC	HLA-B	Signature 2	4.63E-05	8.66E-03	-1	1.00E-01
LUSC_HNSC	BLM	Signature 11	2.36E-05	4.42E-03	-1	5.66E-02
PAAD	TP53	Signature 1	3.86E-05	4.24E-04	-1	1.81E-01
READ	ATR	Signature 10	5.84E-04	3.28E-03	-1	8.12E-01
READ	FANCM	Signature 10	5.84E-04	3.28E-03	-1	8.12E-01
READ	NOTCH2	Signature 10	5.84E-04	3.28E-03	-1	8.12E-01
READ	PIK3R1	Signature 10	5.84E-04	3.28E-03	-1	8.12E-01
READ	POLE	Signature 10	5.84E-04	3.28E-03	-1	8.12E-01
READ	SETBP1	Signature 10	5.84E-04	3.28E-03	-1	8.12E-01
READ	ARID2	Signature 10	2.18E-03	8.43E-03	-1	8.12E-01
READ	CHD9	Signature 10	1.88E-03	8.41E-03	-1	8.12E-01
READ	PTPRC	Signature 10	2.17E-04	3.28E-03	-1	7.77E-01
READ	SETD2	Signature 10	2.53E-04	3.28E-03	-1	7.77E-01
READ	CDK12	Signature 10	2.08E-03	8.43E-03	-1	7.77E-01
READ	HGF	Signature 10	2.77E-03	9.14E-03	-1	7.77E-01
READ	KANSL1	Signature 10	2.52E-03	9.12E-03	-1	7.77E-01
READ	CHD6	Signature 10	1.50E-04	3.28E-03	-1	5.56E-02
READ	NF1	Signature 10	1.57E-05	9.12E-04	-1	5.56E-02
READ_COAD	BRAF	Signature 6	1.36E-06	5.37E-04	-1	2.90E-01

STAD	WHSC1L1	Signature 26	4.43E-09	1.24E-06	-1	7.52E-02
STAD	ERCC4	Signature 26	8.15E-06	5.70E-04	-1	6.05E-02
STAD	WHSC1L1	Signature 21	5.96E-07	1.67E-04	-1	5.98E-02
THCA	EIF1AX	Signature 18	4.93E-04	2.47E-03	-1	7.38E-02
UVM	GNA11	Signature 22	1.23E-11	6.17E-11	-1	1.41E-01
UVM	GNAQ	Signature 22	9.19E-11	2.30E-10	1	1.35E-01

Supplementary Table 3. Significant associations obtained per-cancer type by burden testing of benign mutations from samples in the complete dataset comprising all MC3 samples. This analysis resulted in a total of 50 significant associations. 12 associations involved signatures of exogenous etiology, 32 involved signatures of endogenous etiology and 6 involved signatures with no known etiology.

Disease	Gene	Signature	p_value	q_value	direction	effect size
BRCA	RAD54B	Signature 13	1.92E-05	2.24E-03	-1	2.71E-01
BRCA	SMARCC2	Signature 13	1.37E-04	5.36E-03	-1	2.51E-01
BRCA	ATRX	Signature 13	2.41E-04	7.06E-03	-1	2.12E-01
BRCA	BLM	Signature 13	4.22E-04	9.88E-03	-1	1.93E-01
BRCA	HUWE1	Signature 1	3.15E-06	3.68E-04	1	1.82E-01
BRCA	NF1	Signature 1	1.34E-04	7.85E-03	1	1.77E-01
BRCA	FANCI	Signature 13	5.09E-04	9.93E-03	-1	1.53E-01
BRCA	SPTA1	Signature 1	2.23E-04	8.69E-03	1	1.25E-01
BRCA	SPTA1	Signature 13	7.97E-05	4.66E-03	-1	9.04E-02
BRCA	GATA3	Signature 30	3.95E-05	4.62E-03	-1	6.41E-02
BRCA	TAF1	Signature 10	2.16E-05	2.53E-03	-1	4.96E-02
CESC	WHSC1L1	Signature 1	1.64E-04	8.00E-03	1	2.24E-01
CESC	KMT2C	Signature 1	2.42E-04	8.00E-03	1	1.81E-01
CESC	CHD3	Signature 28	1.79E-12	1.18E-10	-1	2.86E-02
COAD	MUS81	Signature 21	3.99E-06	1.02E-03	-1	5.72E-02
GBM	RTEL1	Signature 10	4.99E-05	9.93E-04	-1	2.07E-02
KIRP	KMT2C	Signature 9	6.40E-04	2.56E-03	-1	2.90E-02
LGG_GBM	JAK2	Signature 10	2.05E-05	1.85E-04	-1	1.08E-01
LGG_GBM	NIPBL	Signature 10	2.05E-05	1.85E-04	-1	1.08E-01
LGG_GBM	SETD2	Signature 10	2.05E-05	1.85E-04	-1	1.08E-01
LGG_GBM	MDC1	Signature 10	4.29E-09	3.48E-07	-1	1.08E-01
LGG_GBM	CDC5L	Signature 10	2.43E-05	1.97E-04	-1	8.04E-02
LGG_GBM	NSD1	Signature 10	3.44E-05	2.32E-04	-1	5.70E-02
LGG_GBM	TELO2	Signature 10	3.01E-05	2.21E-04	-1	5.59E-02
LGG_GBM	ATR	Signature 10	3.19E-06	5.17E-05	-1	5.10E-02
LGG_GBM	SETD2	Signature 14	4.18E-12	3.39E-10	-1	3.94E-02
LGG_GBM	RTEL1	Signature 10	4.77E-05	2.97E-04	-1	2.07E-02
LUAD	CTCFL	Signature 4	2.89E-05	1.31E-03	-1	3.10E-01
LUAD	APC	Signature 4	4.97E-04	8.18E-03	-1	3.00E-01
LUAD	HFM1	Signature 4	3.03E-04	6.10E-03	-1	2.61E-01
LUAD	PDGFRA	Signature 4	2.69E-06	1.62E-04	-1	2.32E-01
LUAD	SPTA1	Signature 4	1.97E-10	3.56E-08	-1	2.31E-01
LUAD	HGF	Signature 4	8.94E-05	3.24E-03	-1	2.27E-01
LUAD	ERBB4	Signature 4	1.94E-04	5.18E-03	-1	2.22E-01
LUAD	POLQ	Signature 4	4.29E-04	7.77E-03	-1	2.20E-01
LUAD	APOB	Signature 4	9.29E-07	8.41E-05	-1	2.01E-01
LUAD	PTPRD	Signature 4	2.00E-04	5.18E-03	-1	1.72E-01
LUAD	DMD	Signature 4	2.80E-04	6.10E-03	-1	1.72E-01

LUAD	FANCI	Signature 25	1.96E-06	3.54E-04	-1	2.59E-02
LUSC_HNSC	SPTA1	Signature 4	1.26E-05	3.73E-03	-1	1.31E-01
OV	APOB	Signature 13	2.63E-04	6.58E-03	-1	3.22E-02
PRAD	SPTA1	Signature 15	9.05E-04	6.34E-03	-1	8.04E-02
READ	ATRX	Signature 10	4.75E-06	9.03E-05	-1	8.14E-01
READ	FANCM	Signature 10	5.84E-04	3.70E-03	-1	8.12E-01
READ	RIF1	Signature 10	4.64E-04	3.70E-03	-1	7.77E-01
READ	RIF1	Signature 1	1.74E-04	3.30E-03	1	5.95E-01
READ	BRCA2	Signature 1	6.75E-04	6.41E-03	1	5.71E-01
READ_COAD	SFPQ	Signature 21	1.15E-05	3.81E-03	-1	9.14E-02
READ_COAD	TNFAIP3	Signature 9	4.81E-07	1.59E-04	-1	3.78E-02
STAD	CDH1	Signature 26	2.44E-05	5.97E-03	-1	6.21E-02

Supplementary Table 4. Significant associations obtained per-cancer type by individual testing of pathogenic mutations from samples in the complete dataset comprising all MC3 samples. This analysis resulted in a total of 62 significant associations. 2 associations involved signatures of exogenous etiology, 59 involved signatures of endogenous etiology and 1 involved signatures with no known etiology.

Disease	Gene	Mutation	HGVSp	Signature	p_value	q_value	direction	effect size
BLCA	ERCC2	19_45867687_T_C	p.N238S	Signature 5	1.80E-10	3.61E-09	-1	3.28E-01
COAD	BRAF	7_140453136_A_T	p.V600E	Signature 6	4.87E-16	1.80E-14	-1	4.22E-01
GBM	IDH1	2_209113112_C_T	p.R132H	Signature 1	1.27E-06	1.78E-05	1	3.82E-01
LGG_GBM	IDH1	2_209113113_G_A	p.R132C	Signature 1	3.10E-04	3.31E-03	1	2.58E-01
LIHC	TP53	17_7577534_C_A	p.R249S	Signature 24	1.70E-06	1.19E-05	-1	2.92E-01
READ_COAD	PTEN	10_89692905_G_A	p.R130Q	Signature 10	2.37E-06	8.03E-05	-1	7.85E-01
READ_COAD	BRAF	7_140453136_A_T	p.V600E	Signature 6	3.42E-17	1.75E-15	-1	4.25E-01
SKCM	KIT	4_55594221_A_G	p.K642E	Signature 7	6.90E-04	6.21E-03	1	6.25E-01
STAD	FBXW7	4_153249385_G_A	p.R465C	Signature 6	1.33E-05	1.86E-04	-1	4.73E-01
STAD	KRAS	12_25398281_C_T	p.G13D	Signature 6	6.71E-05	4.70E-04	-1	4.36E-01
UCEC	PAX7	1_18962743_C_T	p.S155L	Signature 10	5.72E-06	3.52E-05	-1	8.81E-01
UCEC	TOX	8_59750747_G_A	p.R273C	Signature 10	9.00E-06	4.16E-05	-1	8.80E-01
UCEC	POLE	12_133253184_G_C	p.P286R	Signature 10	5.91E-21	6.55E-19	-1	8.77E-01
UCEC	PIK3CA	3_178952018_A_G	p.T1025A	Signature 10	6.58E-06	3.54E-05	-1	8.77E-01
UCEC	LATS1	6_150023019_G_A	p.R82*	Signature 10	6.46E-06	3.54E-05	-1	8.68E-01
UCEC	NF1	17_29677227_C_T	p.R2450*	Signature 10	2.04E-11	7.54E-10	-1	8.57E-01
UCEC	PTEN	10_89624245_G_T	p.E7*	Signature 10	2.50E-10	6.92E-09	-1	8.53E-01
UCEC	SMC3	10_112337617_C_T	p.R99*	Signature 10	6.69E-06	3.54E-05	-1	8.53E-01
UCEC	SMAD2	18_45375016_G_A	p.S276L	Signature 10	1.07E-03	2.98E-03	-1	8.49E-01
UCEC	DHX9	1_182852658_C_T	p.R1050*	Signature 10	8.40E-06	4.05E-05	-1	8.39E-01
UCEC	TP53	17_7578212_G_A	p.R213*	Signature 10	1.71E-04	5.28E-04	-1	8.37E-01
UCEC	PBRM1	3_52643768_G_A	p.R710*	Signature 10	1.11E-05	4.91E-05	-1	8.31E-01
UCEC	MGA	15_42041074_C_T	p.R1818*	Signature 10	9.41E-04	2.68E-03	-1	8.21E-01
UCEC	ATRX	X_76938406_C_T	p.R781Q	Signature 10	1.53E-05	5.85E-05	-1	8.17E-01
UCEC	ARHGAP35	19_47424921_C_T	p.R997*	Signature 10	5.15E-08	7.01E-07	-1	8.09E-01
UCEC	XPO1	2_61719472_C_T	p.E571K	Signature 10	1.18E-05	5.04E-05	-1	8.05E-01
UCEC	ARID1A	1_27106354_C_T	p.R1989*	Signature 10	1.72E-18	9.54E-17	-1	7.85E-01
UCEC	APC	5_112178000_C_T	p.R2237*	Signature 10	1.33E-05	5.29E-05	-1	7.69E-01
UCEC	CASP8	2_202137487_G_T	p.E239*	Signature 10	1.84E-05	6.58E-05	-1	7.51E-01
UCEC	APC	5_112177901_C_T	p.R2204*	Signature 10	2.07E-05	7.16E-05	-1	7.43E-01
UCEC	FBXW7	4_153244185_G_A	p.R658*	Signature 10	2.59E-06	2.05E-05	-1	7.38E-01

UCEC	FUBP1	1_78428511_G_A	p.R430C	Signature 10	1.76E-06	1.63E-05	-1	7.38E-01
UCEC	CHD4	12_6692411_C_A	p.R1338I	Signature 10	1.81E-05	6.58E-05	-1	7.23E-01
UCEC	APC	5_112175639_C_T	p.R1450*	Signature 10	3.21E-05	1.05E-04	-1	7.11E-01
UCEC	PTEN	10_89720744_G_T	p.E299*	Signature 10	3.37E-08	5.34E-07	-1	6.67E-01
UCEC	NF1	17_29576111_C_T	p.R1362*	Signature 10	3.11E-06	2.16E-05	-1	6.44E-01
UCEC	PTEN	10_89692940_C_T	p.R142W	Signature 10	5.06E-06	3.31E-05	-1	6.29E-01
UCEC	POLE	12_133250289_C_A	p.V411L	Signature 10	5.68E-08	7.01E-07	-1	4.02E-01
UCEC	PAX7	1_18962743_C_T	p.S155L	Signature 1	7.22E-04	8.73E-03	1	2.97E-01
UCEC	PBRM1	3_52643768_G_A	p.R710*	Signature 1	8.99E-04	8.73E-03	1	2.96E-01
UCEC	APC	5_112178000_C_T	p.R2237*	Signature 1	9.09E-04	8.73E-03	1	2.95E-01
UCEC	LATS1	6_150023019_G_A	p.R82*	Signature 1	9.59E-04	8.73E-03	1	2.89E-01
UCEC	SMC3	10_112337617_C_T	p.R99*	Signature 1	9.59E-04	8.73E-03	1	2.89E-01
UCEC	ATRX	X_76938406_C_T	p.R781Q	Signature 1	1.02E-03	8.73E-03	1	2.89E-01
UCEC	NF1	17_29677227_C_T	p.R2450*	Signature 1	2.67E-06	9.87E-05	1	2.89E-01
UCEC	PIK3CA	3_178952018_A_G	p.T1025A	Signature 1	1.11E-03	8.73E-03	1	2.88E-01
UCEC	POLE	12_133253184_G_C	p.P286R	Signature 1	5.56E-10	4.97E-08	1	2.82E-01
UCEC	ARID1A	1_27106354_C_T	p.R1989*	Signature 1	8.96E-10	4.97E-08	1	2.66E-01
UCEC	PTEN	10_89624245_G_T	p.E7*	Signature 1	6.70E-05	1.86E-03	1	2.62E-01
UCEC	PTEN	10_89692940_C_T	p.R142W	Signature 1	1.09E-04	2.41E-03	1	2.59E-01
UVM	SF3B1	2_198267483_C_T	p.R625H	Signature 1	1.94E-03	9.70E-03	-1	3.63E-01