UNIVERSITAT DE BARCELONA

Variability in domain-scale regional mutation density across tissues and individuals

Marina Salvadores Ferreiro

Variability in domain-scale regional mutation density across tissues and individuals

Marina Salvadores Ferreiro

Doctoral Thesis

2023

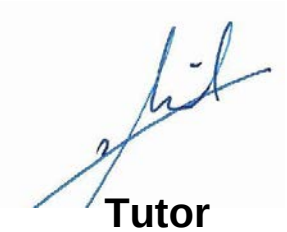
Variability in domain-scale regional mutation density across tissues and individuals

---

Author
Marina Salvadores Ferreiro

Director
Fran Supek

Co-director
Travis H Stracker



Tutor
Josep Francesc Abril Ferrando

Genome Data Science Lab
Institute for Research in Biomedicine

Programa de Doctorat en Biomedicina
Facultad de Biología - **Universitat de Barcelona**

Memòria presentada per ...Marina Salvadores Ferreiro...
per optar al grau de doctor/a per la Universitat de Barcelona

A mi familia, la
que es y la que elegí.

Abstract

Mutations exhibit an uneven distribution across the human genome, with a regional mutation density (RMD) influenced by chromatin landscapes and replication timing. While there is a good understanding of this variation, considered in aggregate across individuals, the existence of inter-individual differences in local and regional mutational risk remains elusive. In this thesis, we aim to systematically characterize the variability in RMD across different somatic tissues and individuals, use the RMD patterns for tumor type classification, and use strand biases in mutations to detect origins of replication.

First, we developed a statistical method based on non-negative matrix factorization to discern mutation risk heterogeneity across megabase-sized chromosomal domains, utilizing over 4,000 tumor whole-genome sequences (WGS) to derive RMD profiles. From this analysis, we identified 13 genome-wide mutation risk redistribution trends, or “RMD signatures”. Ten of these were tissue-specific, ranging from patterns unique to one individual tissue to those shared between tissues with similar physiology. These tissue-specific patterns may help identify the cell-of-origin and subtype of a tumor. Additionally, we identified three global, tissue-independent RMD signatures, which were universally found across many cancer types.

Of these, the RMDflat signature, which implies a more uniform distribution of mutations throughout the genome, was previously identified in MMR-deficient tumors, serving as a validation of the method. Beyond MMR deficiency, we found HR deficiency to be responsible for this uniform distribution. The remaining two global signatures, which we named RMDglobal1 and RMDglobal2, are novel. The RMDglobal1 preferentially affects facultative heterochromatin, is enriched in B1 subcompartment and subtelomeric regions. This RMD mutation risk signature strongly reflects recurrent programs in regional plasticity in replication timing (RT), heterochromatin and DNA accessibility, which we observed across tumors, tissues and cultured cells, and is robustly linked with a higher expression of cell proliferation genes. Consistently, this RMD redistribution pattern is associated with altered cell cycle control via loss of the *RB1* tumor suppressor gene. The RMDglobal2 signature was associated with loss-of-function of the *TP53* pathway, affecting the redistribution of mutation rates within late RT regions. These global signatures divert the local mutation supply towards certain chromosomal domains, impacting 26%-75% of cancer driver genes and potentially altering the trajectory of cancer evolution.

Second, we applied machine learning classifiers, trained on the RMD profiles of the tumors to predict 18 tumor types. These classifiers exhibited a striking accuracy (median AUC of 0.99), further highlighting their diagnostic potential for cancers of unknown primary; we also propose value of RMD for subtyping tumors. When compared with driver mutations, these passenger mutation features including RMD patterns and trinucleotide mutational signatures, exhibited a superior 92% accuracy, underscoring the value of WGS for tumor diagnostics. A further refinement of these mutation-based predictors that incorporated transcriptomic and epigenomic profiles enabled the prediction of tissue-of-origin in 600 cell lines. Upon evaluation, 60% of the

cell lines closely matched their known tumor type of origin, while the rest were poorer matches, and 6% were suspected of being annotated with the incorrect tissue.

Lastly, we developed a computational method, MuSAS, to identify replication origins based on strand asymmetry in mutation densities of various mutational signatures. By comparing the strand bias profiles switch points across approximately 300 of tumors, we discerned tissue-specific replication origin usage patterns in at least seven tissues, with a prominent signal of brain-specific origins. These tissue-specific replication origins correlate with earlier replication timing in the matching tissue and have a higher prevalence of nearby tissue-specific genes. These origins also display enrichment of chromatin accessibility and enhancer and promoter histone marks in the matched tissues. Altogether, this suggests a mechanistic link between tissue-specific gene/chromatin activation and tissue-specific origin usage.

In summary, this thesis presents a comprehensive study of the regional distribution of somatic mutations, and their variability both across tissues and individuals. Our studies help elucidate mechanisms behind this variability, quantify their impact on cancer evolution, as well as evaluate its applications to detect cancer type.

Resumen

Las mutaciones se acumulan de manera no uniforme a lo largo del genoma humano. Esta distribución está influenciada por la estructura de la cromatina y el tiempo de replicación. Aunque esta variabilidad está caracterizada en el conjunto de individuos, la existencia de diferencias interindividuales en el riesgo mutacional local y regional aún no ha sido elucidada. En esta tesis, nuestro objetivo es caracterizar sistemáticamente la variabilidad en la distribución de las mutaciones entre diferentes tejidos e individuos, usar estos patrones de mutación para clasificar los tipos de cáncer y para detectar orígenes de replicación.

Primero, desarrollamos un método estadístico basado en la factorización de matrices no negativas para identificar patrones de redistribución de mutaciones en el dominios cromosómicos de una megabase, utilizando más de 4,000 secuencias del genoma completas para calcular perfiles de densidad de mutaciones. En este análisis, identificamos 13 patrones que describen redistribuciones del riesgo de mutación en el genoma. Diez de estas son específicas para un tejido o grupo de tejidos con una fisiología similar. Estos patrones específicos de tejido pueden servir para identificar la célula de origen y el subtipo de un tumor. Además, identificamos tres patrones de redistribución de mutaciones que son globales e independientes del tejido, ya que están presentes en la mayoría de los tipos de cáncer.

El primer patrón, RMDflat, causa una distribución más uniforme de mutaciones a lo largo del genoma y fue previamente identificada en tumores deficientes en MMR, sirviendo como validación del método. Aparte de la deficiencia en MMR, en este estudio encontramos que la deficiencia en HR también causa esta distribución uniforme. Los otros dos patrones de distribución globales son novedosos y los llamamos RMDglobal1 y RMDglobal2. RMDglobal1 afecta preferentemente a la heterocromatina facultativa y está enriquecida en regiones subteloméricas. El patrón de RMDglobal1 está correlacionado con plasticidad en el tiempo de replicación, la heterocromatina y la accesibilidad al ADN, observada en tumores, tejidos y células cultivadas, asociada con una mayor expresión de genes de proliferación celular. Consistentemente, este patrón está asociado con la alteración del ciclo celular a través de la pérdida del gen supresor de tumores RB1. RMDglobal2 está asociado con la pérdida de función de gen TP53, afectando la redistribución de mutaciones en las regiones con tiempo de replicación tardío. Estos tres patrones de redistribución globales pueden cambiar el suministro local de mutaciones en ciertas regiones genómicas, impactando en el 26% al 75% de los genes impulsores del cáncer y potencialmente alterando la trayectoria de la evolución del cáncer.

En segundo lugar, aplicamos clasificadores de aprendizaje automático, entrenados en los perfiles de densidad de mutaciones en ventanas de 1 megabase de los tumores, para predecir 18 tipos de cáncer. Estos clasificadores mostraron una alta precisión (AUC mediana de 0.99), destacando su potencial diagnóstico para cánceres de origen desconocido; también proponemos que pueden servir para identificar el subtipo de cáncer. Al comparar con mutaciones ‘driver’ (oncogénicas), las mutaciones pasajeras (no oncogénicas), incluyendo nuestros patrones de densidades, mostraron una precisión superior del 92%, subrayando el valor de secuenciar el genoma

completo para el diagnóstico. Además, incorporando perfiles transcriptómicos y epigenómicos a estos clasificadores pudimos predecir el tejido de origen en 600 líneas celulares. Donde vimos que el 60% de las líneas celulares coincidían estrechamente con su tipo de tumor de origen conocido, mientras que el resto coincidían peor, en particular sospechamos que el 6% de las líneas celulares están anotados con el tejido incorrecto.

Por último, desarrollamos un método computacional, MuSAS, para identificar orígenes de replicación basado en la asimetría de las mutaciones entre las dos hebras del ADN. Al comparar los orígenes de replicación identificados en aproximadamente 300 tumores, encontramos siete patrones de uso de origen de replicación que son específicos de un tejido, con una señal prominente de orígenes de replicación específicos del cerebro. Estos orígenes de replicación específicos de un tejido, se correlacionan con un tiempo de replicación temprano en el tejido correspondiente y tienen una mayor prevalencia de genes específicos de ese tejido. Estos orígenes también muestran un enriquecimiento de la accesibilidad de la cromatina y marcas específicas de histonas en 'enhancers' y 'promoters' en los tejidos correspondientes. En conjunto, esto sugiere un vínculo entre la activación génica y de la cromatina con el uso de orígenes de replicación específicos para un tejido.

En resumen, esta tesis presenta un estudio exhaustivo de la distribución regional de las mutaciones y su variabilidad en diferentes tejidos e individuos. Nuestros estudios ayudan a aclarar los mecanismos detrás de esta variabilidad, cuantificar su impacto en la evolución del cáncer, así como evaluar sus aplicaciones para detectar el tipo de cáncer.

List of content

Chapter 1
Introduction

1.1 Mutagenesis: the process of generating mutations

1.1.1 What is a mutation and types of mutations

A mutation is defined as any change in the DNA sequence of a cell. Mutations can be caused by endogenous (e.g. a mistake during DNA replication) or exogenous (exposure to DNA damaging agents) sources (1). Mutations can affect a single nucleotide or involve larger segments of DNA. Point mutations or single nucleotide variants (SNVs) occur when only a single nucleotide base is changed (substitution). When there is an addition or removal of several nucleotide bases, they are insertions or deletions (indels) respectively (2).

Besides, there are other types of genetic alterations, such as structural variants (SVs). Among SVs, Copy Number Variants (CNVs) involve a change in the number of copies of a particular DNA segment of variable length. In addition, there are duplications when a region of the DNA is replicated and inserted; inversions when a region of the DNA is reversed within the genome; and translocations when a region of the genome is exchanged between chromosomes (3).

Mutations can be separated into germline and somatic mutations. Germline mutations happen in the reproductive cells (egg or sperm) and they become incorporated into the DNA of every cell in the offspring. When the mutation occurs in other types of cells, they are not inherited and are called somatic mutations (1). Examining distributions of each type of mutation, we can study different biological processes related to genome maintenance.

In the process of mutagenesis, first there is a DNA damage which could be endogenous or exogenous that modifies the DNA sequence. Second, this change in the DNA sequence should escape the DNA repair pathways. Once the region with the DNA lesion is replicated, the mutation becomes fixed in the genome.

1.1.2 Causes of somatic DNA mutations

Somatic mutations can be caused by endogenous or exogenous sources. The cause is endogenous when it arises from an internal errors in DNA maintenance process, for instance, a mistake during DNA replication. The cause is exogenous when there is an external agent affecting DNA integrity, usually by damaging DNA, such as UV light (4).

Endogenous DNA damaging processes

Replication errors

When a human cell divides, its complete DNA sequence is copied with high fidelity. However, during this DNA replication, errors can occur and if they escape the proofreading, they result in mutations. Even though most of these errors are repaired by the DNA mismatch match repair

pathway, mutations accumulate at a frequency of 10^{-6} to 10^{-8} per cell per generation (5,6) in human.

Spontaneous base deamination

In the DNA sometimes there is a spontaneous deamination of the cytosine (C), adenine (A), guanine (G) or 5-methylcytosine (5mC) that becomes an uracil (U), hypoxanthine, xanthine or thymine (T) respectively (4). This phenomenon occurs more frequently in single-stranded rather than double-stranded DNA (4). These altered nucleobases result in mutations in the next round of replication.

Oxidative DNA damage

In the cells, there are naturally reactive oxygen species (ROS) because they are a byproduct of the mitochondrial oxidative metabolism. In addition, they are generated in response to xenobiotics, cytokines, and bacterial invasion (7). In excess, these ROS can cause base lesions, including 8-oxoG and thymine glycol, and DNA strand breaks (4). The ROS can also oxidate the free nucleotide pool; when a free 8-oxoG gets misincorporated into DNA it generates A>C mutations.

DNA methylation

Per cell and per day, the DNA methyltransferases can spontaneously produce ~40000 methylated residues (4). These methylated bases can be repaired by direct reversion of the methylation (e.g. MGMT protein) or through the base excision repair (BER) pathway. Methylated DNA bases left unrepaired are a major source of spontaneous DNA damage (4).

APOBEC mutagenesis

The APOBEC3 family are cytosine deaminases that defend the cell against viruses and retrotransposons by damaging their genetic material via deamination. However, in human cells, the APOBEC3 enzymes can be activated and create U:G mismatches via the deamination of cytosines in single-stranded DNA. If left unrepaired, this leads to the accumulation of mutations (8). APOBEC3 mutagenesis has been extensively described in several cancer types (8). However, it is a process that can also occur in normal cells in an episodic manner throughout the human lifespan (9).

Exogenous DNA damaging agents

Ionizing Radiation

Ionizing radiation (alpha, beta, gamma, neutrons, and X-rays) is present in the environment (cosmic radiation, radioactive minerals and radon gas) and exposures also result from medical

devices (diagnostics and cancer therapy). The ionizing radiation can damage the DNA directly or indirectly via an increase of ROS. It causes base lesions and single-strand breaks (4), with cumulative effects.

Ultraviolet (UV) Radiation

UV radiation emanates from the sun and damages the DNA mainly by causing covalent linkages between two adjacent pyrimidines. The two main photoproducts are cyclobutane pyrimidine dimers and pyrimidine (6-4) pyrimidone photoproducts. These dimers distort the helix and can be repaired by the Nucleotide Excision Repair pathway or be bypassed by the translesion synthesis polymerases, therefore contributing to the accumulation of mutations (4). UV radiation is the main cause of skin cancers in humans (10).

Chemical agents

Alkylating agents are components in day-to-day exposures, such as tobacco smoke or chemotherapeutic agents. They add alkyl groups to the DNA generating adducts that, if left unrepaired, lead to a mutation (4). Aromatic amines are produced from tobacco smoke, fuel burning, pesticides, etc. They are converted into a carcinogenic alkylating agent that attacks the C8 position of guanines. These lesions lead to base substitutions and frameshift mutations (4). In addition, there are other chemical agents, such as Polycyclic aromatic hydrocarbons (PAHs) from air pollution, natural toxins (e.g. aflatoxin, aristolochic acid) and occupational hazards that can alter the DNA and cause mutations.

1.1.3 DNA repair pathways

The cells have multiple DNA repair mechanisms to counteract the DNA damaging processes described above. Different DNA damages activate different pathways of repair. The aim of these repair mechanisms is to maintain the integrity of the genome by preventing the accumulation of mutations (4,11), and to resolve DNA damage rapidly such that it does not prevent transcription and replication.

Base Excision Repair (BER)

BER is the primary mechanism that repairs base lesions arising from oxidation, deamination and alkylation damage. These are adducts that cause a small distortion on the structure of the DNA helix (12,13). BER performs the repair through two general pathways: the short-patch and long-patch. The short-patch repairs one single nucleotide while the long-patch produces a repair tract of at least two nucleotides (12,13). BER starts when a DNA glycosylase recognizes the base with the lesion. BER removes the damaged base leaving an abasic site (also called AP site i.e. when there is no base but DNA backbone is intact) which is then processed by short-patch or long-patch repair. There are several distinct DNA glycosylases and each of them recognizes certain specific lesions (12,13).

Nucleotide Excision Repair (NER)

NER is a mechanism that repairs bulky lesions that are damages that cause a distortion to the DNA helix (e.g. UV photolesions, adducts from PAHs). There are two main pathways of NER: global genome NER (GG-NER) and transcription-coupled NER (TC-NER) (4).

In GG-NER, there are damage sensors that scan the genome to identify the structural disruption of the DNA helix due to the lesion. Once the lesion is recognized, the rest of NER-associated proteins are recruited. These proteins cut the damaged strand upstream and downstream of the lesion, synthesize a repair patch copying the opposite undamaged strand and ligates it into the genome to restore the DNA (14).

The TC-NER is initiated when a lesion stalls the transcription machinery. This recruits TC-NER-specific proteins and forms a complex that moves backwards the transcription machinery, exposing the lesion base. Additional NER-associated proteins are recruited and the subsequent repair process is as in GG-NER (14).

DNA Mismatch Repair (MMR)

DNA MMR mechanism is a highly conserved post-replicative process that ensures the DNA replication fidelity. The MMR substrates are base-base mismatches and the insertion-deletion mispairs that arise during replication and recombination. MMR is also implicated in microsatellite stability, DNA-damage signaling, apoptosis and somatic hypermutation (15,16). MMR has two main components: MutS which recognizes the lesions and MutL which performs the processing and excision. When left unrepaired, the mismatch generates a single nucleotide substitution mutation and the insertion-deletion mispair generates an indel mutation (15).

Translesion Synthesis (TLS)

Translesion synthesis is performed by highly conserved TLS polymerases that can replicate opposite and pass aberrant DNA lesions. The TLS polymerases have lower fidelity than DNA polymerases, are error-prone and are a major source of mutagenesis (17). If the wrong nucleotide is incorporated by the TLS, it becomes a mutation in the next round of replication. The major function of TLS is to overcome replication blocks if the normal DNA replication is blocked by DNA damage or by structured DNA in the template strand (18).

Repair of ds DNA Breaks: Homologous repair (HR) and Non-homologous end joining (NHEJ)

DNA breaks can be single-strand breaks (SSBs) or double-strand breaks (DSBs). SSBs are often caused by oxidative damage or erroneous activity of the DNA topoisomerase 1 enzyme. The main mechanism that repairs SSB is the single-strand breaks repair (SSBR). When the SSB is not repaired, the DNA replication collapses and the transcription is stalled, which can lead to activation of apoptosis (4).

DSBs are very toxic and are induced by several chemical and physical DNA damaging agents. There are two major pathways involved in the repair of the DSBs: HR and NHEJ (4). When there is a DSB, it is detected by chromatin modification and triggers a cascade of events leading to the recruitment of 53BP1 (NHEJ factor) or BRCA1 (HR factor) proteins. In the NHEJ pathway, the 53BP1 recruits the NHEJ components to the break site, activates the checkpoint signaling and facilitates the synapsis of the two ends. The HR pathway uses DNA strand invasion and template-directed DNA repair synthesis (4).

1.1.4 Accumulation of somatic mutations

The final fixed mutations that a cell harbors during its life span arise from a combination of exposure to DNA damage (endogenous or exogenous) and the DNA pathways that can repair them. A summary of the different DNA damaging agents, the DNA damage they cause and the DNA pathways involved in their repair is shown in Table 1.

DNA damaging agents ↓	Toxins Alkylating agents Base deamination Replication errors ↓	Oxidative damage electrophiles ↓	Ionizing radiation UV radiation Crosslinking agents Aromatic compounds ↓
Damage DNA ↓	Mismatches Uracil Abasic sites Adducts ↓	Lesions Single strand break Double strand break ↓	Bulky lesions Intra and interstrand crosslink Single strand break Double strand break ↓
DNA repair pathways	Mismatch repair Base excision repair	Base excision repair Single strand break repair Homologous repair Non-homologous end joining	Nucleotide excision repair Single strand break repair Homologous repair Non-homologous end joining Translesion synthesis

Table 1. Summary of the DNA damaging agents, separated by the damage they cause in the DNA and the repair pathway involved in their repair. Adapted from Fig4 in (4).

The accumulation of somatic mutations is a process that occurs during the whole life span of a normal cell. This accumulation of mutations over time can lead to cancer and contribute to aging (19). Initially, the mutagenic processes were studied in tumor samples as a means to understand cancer development. In addition, characterizing the somatic mutations in normal cell populations was technically changed due to the lack of clonality. Only recently, different approaches have been developed to detect mutations in normal tissues such as sequencing localized clonal populations (e.g. colorectal crypts) (20).

In healthy cells, the mutation rates differ between cell types from the same individual. In *Moore et al.* study, the intestine and appendix show the highest mutation rates (43 to 73 mutations per year) compare to gastric glands (20-32 muts/year), pancreatic acini (8-23 muts/year) and bile ductules (5-21 muts/year). In the testis, the mutation rates were much lower (1.83–2.53 muts/year) than in any of the somatic cell types analyzed (20). The mutational processes ubiquitously active in normal tissues were the deamination of 5-methylcytosine and another endogenous process with unclear mechanisms called SBS5. In addition, other mutational processes were found in a subset of cell types, such as mutations due to oxidative damage, chemotherapy agents, UV-light exposure and APOBEC mutagenesis (20).

In normal conditions, the cells accumulate somatic mutations. The process of tumorigenesis starts when one (or more) of these mutations confer a fitness advantage to the cell. This abnormal cell starts to grow in a clonal expansion and forms the tumor mass (19,20). In contrast with normal cells, the tumors tend to have a mutator phenotype and may accumulate a much higher number of mutations (1000 to 20,000 somatic point mutations) during tumor growth and clonal expansion (20). These somatic mutations that are accumulated during cancer development can be used to study different biological mechanisms, including the identification of cancer driver mutations or the characterization of mutational processes. During the past decade the cancer genomes have been extensively used as models to study the process of mutagenesis.

1.2. Studies of somatic mutations in human tumors

1.2.1 Basic principles of cancer

Cancer is a disease that starts from one of the body's cells that grows uncontrollably forming a tumor mass and spreads to other parts of the body (21). Cancer is considered a collection of diseases because each cancer type has its own characteristics in terms of cell of origin, risk factors, driver gene repertoire, symptoms and mortality, and effective treatment. According to the World Health Organization and the International Agency for Research on Cancer, cancer is one of the main causes of death worldwide, leading to approximately 10 million deaths per year (22).

Cancer development (or tumorigenesis) in humans is a multistep process. Each step is caused by mutations (or in some cases, epigenetic alterations) that progressively lead to the transformation of a normal cell into a malignant tumor (23). One classification proposed there are 8 biological capabilities, known as the hallmarks of cancer, that enable the cells to become tumorigenic and ultimately malignant (23,24). These hallmarks comprise: sustaining proliferative signaling, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis, resisting apoptosis, deregulating cellular energetics and avoiding immune destruction (24).

These hallmarks describe the biological capabilities that the tumor cells need to acquire in order to proliferate and spread. The acquisition of these capabilities usually happens when a relevant gene is genetically altered (21,24). There are two types of relevant genes involved in cancer development: tumor suppressor genes and oncogenes. Tumor suppressor genes (TSG) are genes involved in controlling cell growth and division. When the tumor suppressor gene is mutated (causing its inactivation) the cell starts to divide uncontrollably and may lose the ability to apoptose. On the other hand, the oncogenes are genes involved in normal cell growth and division. When the oncogene suffers a gain-of-function mutation in a specific location (hotspot) it causes its activation, similarly allowing the cells to grow and survive when they should not (21). TSG and oncogenes are also often affected by copy number alterations.

The process of tumorigenesis is a clonal expansion that follows the Darwinian evolution principles (25). The cells that receive a mutation with positive fitness effect (within the microenvironment provided by the tissue) will become selected and amplified compared with the other cells (25,26). Eventually, a cell may acquire a set of sufficiently advantageous mutations that confers it with the hallmarks of cancer capabilities and allows it to form the tumor mass. This tumor can proliferate autonomously and metastasize to other tissues (25,26).

The principal driving force behind the tumorigenesis are the various types of mutations (epigenetic changes notwithstanding). According to their consequence for cancer development the mutations can be classified into driver or passenger mutations. Driver mutations are a small proportion and are those that have a functional consequence on the gene, confer a growth advantage and are positively selected during tumorigenesis (25). In contrast, the rest of mutations (the majority) are passenger mutations. They do not confer a growth advantage and are not selected, but they were present in an ancestor of the cancer cell when it acquired one of the driver mutations (25). Therefore, they can serve as a track record or the mutational process that a cell has undergone.

The last step of tumorigenesis is the invasion and spread to other tissues. When the cancer is where it is first formed, it is called primary cancer. Once the cancer is spread to other parts of the body, it is called metastatic cancer. The metastatic cancers still present the molecular features of the primary cancer despite being in a different location (21).

1.2.2 Basic principles of next generation sequencing and available datasets

Next generation sequencing (NGS) is a collection of high-throughput technologies which have revolutionized genomic research, by enabling the sequencing of DNA and RNA samples in a rapid and cost-effective manner. In the last decade, through a collective effort thousands of cancer patients have been sequenced around the world. Among the biggest projects we should mention The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC) (27), the Pan-cancer Analysis of Whole Genomes (PCAWG) (28) and the Hartwig Medical Foundation (HMF) (29) among others. Both whole-genome sequencing (WGS) and whole-exome sequencing

(WES) of somatic mutations for thousands of patients have become available for the scientific community.

As aforementioned, the mutations can be separated into germline and somatic mutations. When a sample (e.g. a tumor biopsy) is sequenced, we can observe genetic variants that are a mixture of germline and somatic mutations. Usually, a blood sample from the same patient is also taken and DNA sequenced to identify the mutations that are from the germline (i.e. inherited). Once we remove the mutations present both in the tumor and in the blood (germline), this generates a set consisting only of somatic mutations (30). Each type of mutation can be used to study different biological processes. For instance, germline variants are often studied through genome wide association studies (GWAS) to find genetic markers linked to a disease or trait (31).

For somatic mutations in human cancers, the first studies focused on determining which genes are responsible for the initiation and progression of cancer (32,33). For the emerging tumor-normal samples sequenced, researchers applied mathematical approaches to determine which genes accumulate mutations more frequently than expected by random chance (32,33). Nowadays, hundreds of cancer genes have been identified using different methods (34).

After the initial efforts to identify the driver genes (32,33), the field rapidly expanded to additionally analyze the passenger mutations. Understanding the patterns and frequencies of these passenger mutations provides valuable insights into the mutational processes, DNA repair deficiencies, chromatin organization and evolutionary dynamics that shape the tumor genome.

1.2.3 Mutational signatures

One of the first passenger mutation patterns studied was the differences in the trinucleotide mutational frequency. These patterns are called mutational signatures, and they provide information about the underlying biological processes and can help identify the factors responsible for the mutagenesis. The methodology to extract mutational signatures from cancer genomes was introduced by Alexandrov *et al.* in 2013 (35) and has been subsequently extended (36,37).

The standard methodology to calculate mutational signatures consists of applying a non-negative matrix factorization (NMF) algorithm into a matrix of counts (or relative abundance) of mutations in each trinucleotide context (5' and 3' neighboring nucleotides) (37). In addition to trinucleotide context, pentanucleotide and heptanucleotide can also be used. The NMF is a dimensionality reduction method which is able to decompose the information from the trinucleotide mutation patterns into the more interpretable mutational signatures, which are combinations of trinucleotide frequencies with various weights. The NMF output provide two reduced matrices which correspond to: (1) the profile of the mutational signatures, which trinucleotide context is more enriched for that specific signatures; and (2) the exposures to the mutational signatures, which samples has a higher contribution of that specific signature into its mutation burden.

The emergence of the mutational signatures methods was transformative, since it helped to identify the existence of several novel, commonly occuring mutational patterns across various

human somatic cell types. The underlying mechanistic processes of many of these mutational signatures were elucidated, such as the patterns for known mutagenic factors (UV light or tobacco) (35). As well as the signatures caused by DNA repair defects (36). Likewise, it was demonstrated that localized hypermutation (clustered mutations) is generated by endogenous mutagens, such as the cytosine deaminase APOBEC family (8,38,39) and also by error-prone DNA polymerases (40,41).

Since these initial studies, the mutational signatures were improved several times by including more patients. In addition, the mutational signatures, which were initially obtained with single nucleotide variants only, have been extended to be derived from indels and structural variants, generating additional sets of mutational signatures (37,42). Largely, the NMF algorithm has been widely adopted in the field of cancer genomics and mutational signature analysis, however related methods were applied to cell line models (43) and to model organisms (44).

1.2.4 Relative mutation rates

The study of the distribution of the passenger somatic mutations through human chromosomes has revealed that they are distributed unequally across the genome, with some regions acquiring substantially more mutations than others (45). The regional variability in the mutation rates has been observed at different resolutions, from single nucleotides to megabase-sized domains, and different mechanisms seem to be implicated (35,45–48).

Single nucleotide scale (1–10 bp)

As aforementioned, the variability at the resolution of a single nucleotide is highly dependent on the trinucleotide, pentanucleotide and heptanucleotide context (5' and 3' neighboring nucleotides). The relative abundance of mutations in these contexts depend on the mutagenic processes that the individual has been exposed to. The mutational patterns generated by the changes in the mutability of trinucleotides across individuals were studied and cataloged as mutational signatures (35) [see section 1.2.3].

Sub-gene scale ($10–10^3$ bp)

The variability at the sub-gene scale has been linked with different features. For instance, the mutation rates have a peak at the binding sites of the CTCF protein across cancer types (46) and at the binding sites for the ETS family of transcription factors in skin cancer (49). Similarly, the DNA structures known as "hairpins" are preferentially targeted by APOBEC3A and therefore enriched in mutations (50). At a similar scale, the nucleosome occupancy has been associated with altered patterns in the somatic and germline mutation rates (47).

Gene scale ($10^3–10^5$ bp)

In the variability at the gene scale a well known example is the asymmetry between the transcribed and non-transcribed DNA strand in the mutation rates for some mutational processes (35). Besides, a reduction in the mutation rates in genes that have high transcription levels has

been detected (51). A mechanism that might explain this is the association between H3K36me3 (mark associated with transcription elongation) and an increase in DNA repair deficiency (52).

Domain scale (10^5 –10^6 bp)

At the domain scale high mutation rates has been correlated with different features such as late replication timing regions (48); high density of repressive chromatin marks; lower accessibility of DNA as estimated via the density of DNAse hypersensitive sites (53); and a lower density of transcribed genes (54).

In this thesis, we focus on understanding the variability in mutation rates at the domain scale and its variability across tissues and across individuals.

1.3. Regional mutation density at the megabase scale

1.3.1 Determinants of the general regional mutation density at the megabase scale

At the megabase scale, the mutation densities in the human are not uniformly distributed. If we split the genome in 1-Mb windows and compare the somatic mutation density across windows we found increases and decreases up to five-fold between different windows (54). The profile of mutation densities in each of these 1-Mb windows is called hereafter Regional Mutation Density (RMD) spectrum (Fig 1).

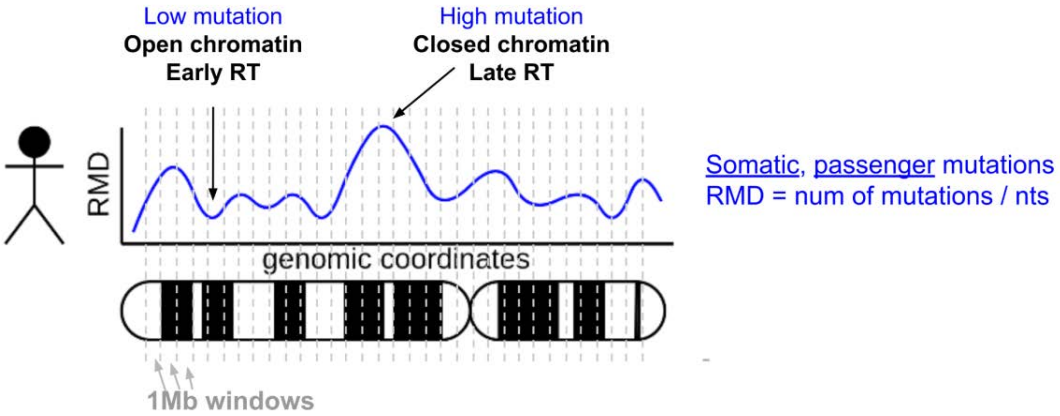


Fig 1. Representation of the regional mutation density (RMD) profile for one human tumor in one chromosome.

Initial studies examine these regional mutation density variations in relation to the distribution of sequence features, gene expression and epigenetic marks throughout the genome. In 2011, Hodgkinson *et al.* showed that there is significant variation in the somatic mutation density at the 1-Mb scale within three cancer genomes and observed correlation between them (55). They also showed that the density of mutations is correlated to the GC content, replication timing, distance to telomere and centromere, gene density and nucleosome occupancy (55). In 2012, Schuster-Böckler *et al.* evaluated 46 different genetic and epigenetic features and showed that at the megabase scale the RMD is correlated with many features of chromatin organization (highest correlation with H3K9me3) (56). Concurrently, two additional studies showed the correlation between the RMD and replication timing and high-order nuclear organization (57,58).

In 2015, two comprehensive studies examined the correlation between the RMD with chromatin organization and replication separating by tissue-of-origin. Polak *et al*. checked the distribution of RMD in comparison with cell-type-specific epigenomic features, showing that chromatin accessibility, modification and replication timing, explains 86% of the variance in the RMD along cancer genomes. The best determinants are the epigenomic features of the most likely cell type of origin for each tumor, rather than the epigenomic features of the matched cancer cell lines (54). Simultaneously, Supek and Lehner showed that the RMD is stable across cancer types, with differences related to changes in tissue-specific replication timing and gene expression. However, they observed that the RMD profile of mutations arising after the inactivation of DNA MMR are no longer enriched in late replicating heterochromatin relative to early replicating euchromatin. Therefore, they propose that differential DNA repair is the primary cause of the large-scale regional mutation rate variation across the human genome and not differential mutation supply (48). In addition, Akdemir *et al.* (59) showed that the somatic mutational load changes in cancer genomes vary with three-dimensional chromatin structure and co-localized with topologically-associating-domain boundaries.

The current working hypothesis is that the uneven distribution in mutation densities along the genome is caused by the differential activity of DNA MMR and NER, which preferentially prevents mutations in early-replicating, euchromatic regions (45,60). There are several pieces of evidence supporting this: (i) in Supek *et al.* when the MMR mechanism is impaired in a sample, its mutational landscape becomes more “flat” (Fig 2) and the earlier the MMR fails in the history of a tumor, the flatter its regional mutation rate landscape is with respect to replication timing (48); (2) in experiments, this “flat” distribution of mutations has also been observed in human cell lines and organoids where the MMR activity was abolished (61,62). In addition, the global genome NER was shown to target early replicating regions and when switched off the mutational distribution becomes “flatter” (60). Altogether, domain-scale variability is strongly affected by how the global DNA repair mechanisms prioritizes the repair of mutations in early-replicating euchromatin regions.

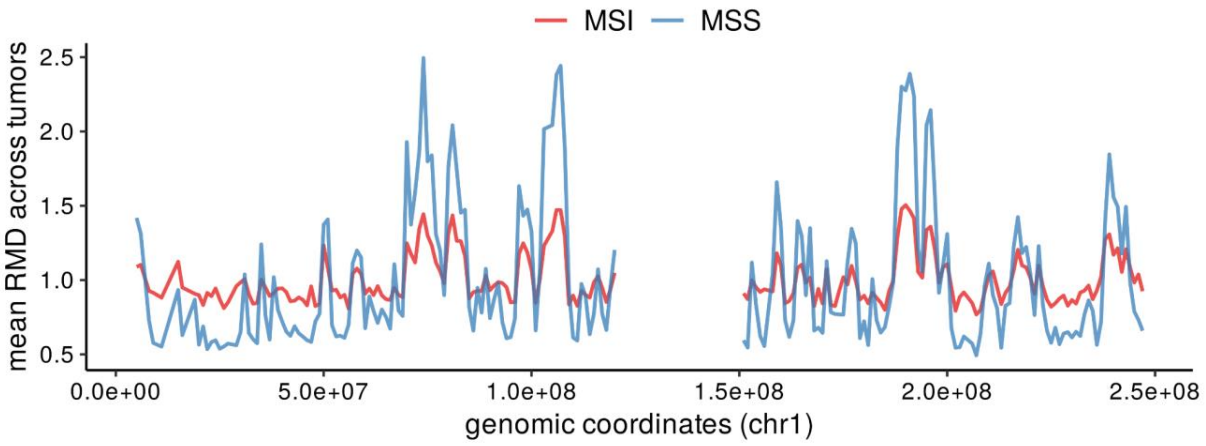


Fig 2. Mean regional mutation density across tumors with DNA MMR failure (MSI tumors) versus with the DNA MMR mechanism working (MSS tumors). The MSI samples suffer a relative reduction of mutations in late replicating heterochromatic regions and a relative increase of mutations in early replicating euchromatic regions. Figure inspired from (48) and reproduced with data from Chapter 3.

1.3.2 Variability in regional mutation density across tissues

Even though there exists a common, canonical regional mutation rate landscape shared across all human cells, there are regional mutation density patterns superimposed that are tissue-specific. This is consistent with the fact that each cell type has somewhat different RT programs (63) and chromatin landscapes, and changes towards late replication and towards closed chromatin correlate with higher mutation rates. As explained above, it was shown that the regional mutational profile of one cancer type (e.g. melanoma) best matches the replication timing profile (48) and chromatin landscape (54) of the same tissue type (e.g. melanocytes). This explains why the mutation rates profile varies across tissues.

In the work presented in this thesis and by others, the tissue-specificity of the regional mutation density was used to predict the tissue of origin based on the tumor genome sequence. Overall, determining the cancer type and molecular subtype of a tumor has important clinical implications because it provides context for understanding biological mechanisms and predicting therapy response. However, in approximately 3% of metastatic cancers the standard diagnostic procedure cannot identify the site of origin (64). While there exist algorithms to classify such ‘cancers of unknown primary’ (CUP) based on gene expression (65,66) or DNA methylation (67), genome sequencing provides an opportunity to obtain independent predictions, and might be better suited for low-purity scenarios like liquid biopsies (sampling circulating tumor DNA or cells from blood). Therefore, in this thesis, we defined and used a Regional Mutation Density (RMD) feature to predict the cancer-type-of-origin and compare it with other features obtained from sequencing data (see Chapter 4).

1.3.3 Variability in regional mutation density across individuals

In addition to the variability in regional mutation rates across cancer types, in this thesis we hypothesized whether there is also variability at the domain scale across individuals. We hypothesized that there might be factors, such as exposure to mutagens and/or DNA repair failures and/or cancer gene mutations, that can cause domain-scale variability across individuals. As mentioned above, the “flat” mutational landscape of microsatellite instability (MSI) tumors (tumors with DNA MMR failure) is an example of this type of variability. The described MMR loss implies an increase of mutations in the early-replicating, gene-rich domains, generating a redistribution of mutation risks (observed as variable mutation counts). We consider this redistribution to be sample specific, because only a few percent of the tumor samples will be MSI and therefore be affected by this. It is important to note that the increase of mutations will be more evident in early-replicating, gene-rich domains, revealing that this kind of redistribution can have important functional consequences for those tumors. In addition, we anticipate several mechanisms to be able to affect the mutational landscape independently and in different ways. In particular, this concerns features that are spread genome-wide such as DNA replication timing, DNA accessibility and distribution of various histone marks.

1.3.4 Variability in regional mutation density stratifying by mutational signatures

The uneven distribution in mutation densities is overall shaped by the differential activity of MMR and/or NER in early versus late replicating regions (45,60). However, the DNA damage distribution caused by different endogenous or exogenous factors might vary due to their association with RT and/or DNA repair.

To study this, Yaacov *et al.* systematically analyzed the mutational landscape of 2787 tumors separately for early and late replicating regions (68). They observed that some signatures were atypical in that they were associated with early RT regions such as SBS16, SBS5 and SBS2; others exhibited a more stereotyped pattern of enrichment in late RT regions such as SBS8 and SBS4; and others did not show associations such as SBS3. In a second study, Yaacov *et al.* tested the associations between mutational signatures in non-cancerous tissues and RT (69). They also compared the RT bias in 4 matching non-cancerous and cancerous tissues. While for most signatures the RT bias was consistent in normal and cancer samples, they found SBS1 late RT bias to be lost in cancer (69). Similarly, Tomkova *et al.* (70) evaluated the mutational signature distribution taking into account the replication timing and the strand asymmetry. They found several signatures associated with late RT such as SBS17, 8 and 4 (as in Yaacov *et al.*), however they don't find any signatures significantly associated with early RT (70).

These studies were performed in regions that are always early or late replicating across several tissues. However, Caballero *et al.* performed an analysis of the mutational signatures with their matched replication timing profile in 5 cell types (71). They demonstrated that the association between mutation rates and RT is heterogeneous among cell types and that this extends to the mutational signatures, which showed inconsistent replication timing bias between cell types. This

underscores that there likely exist multiple complex mechanisms that govern the details of regional mutation risk distribution across individuals.

1.4. Variability in genome organization landscapes

Based on their role in modulating the regional mutation risk at the domain scale, we described here the main genome-wide chromatin organization landscapes, and their known mechanisms of variability between cell types or individuals.

1.4.1 Replication timing landscape

Replication timing (RT) refers to the temporal order in which the DNA sequence is copied during the S phase of the cell cycle. This order of copying has links with maintaining genome stability and with regulating cellular processes. Initial studies using cytogenetic techniques showed that chromosomes are organized into territories (72). With the appearance of new genome-wide methodologies, RT was measured at high resolution in different cell types. This led to the identification of replication domains (RDs), which are units of roughly 400 to 800 kb composed of multiple adjacent replicons that replicate coordinately (72,73). Other methods, such as chromatin conformation capture, which maps chromatin interactions, showed that the genome is partitioned into distinct A and B compartments that exhibit early and late RT, respectively. They also confirmed the presence of stable units of chromosome structure named topologically associating domains (TADs) that correspond to RDs (72). At the domain scale, RT measures revealed multi-megabase segments of chromosomes, named constant timing regions (CTRs), which replicate via the coordinated activation of adjacent replicons at certain times during the S phase. These CTRs are delimited by large timing transition regions (TTRs) which RT varies (74).

RT is associated with different genomic features such as GC content, gene density and transcriptional activity (74). In general, RT is highly stable and the RT profiles from different samples correlate strongly. However, there is variability in RT in some domains across different types of biological samples. The most prominent case is cell type-specific differences, in which each cell type presents a particular RT program which differs in certain regions. These cell type-specific differences reflect diverse functional requirements and gene expression programs (73,75). In addition, RT variability also occurs during development. When the cells differentiate, the variation in RT can define the establishment of cell identity (73).

Other sources of RT variability can stem from genetic or epigenetic factors. Genetic variations, such as single nucleotide polymorphisms (SNPs) or somatic mutations, can affect the timing of DNA replication; conceivably, they might alter the binding sites of replication factors. For instance, genetic polymorphisms (called rtQTLs) have been described to act in *cis* upon megabase-scale DNA segments and associated with the local DNA replication timing (75,76). Additionally, the loss of the genes *RIF1* and *MCM10* has been found to disrupt replication timing via RIF1 KO experiments (77) and in a comparison across a panel of 184 cell lines (71) respectively. Similarly,

epigenetic modifications, such as DNA methylation and histone modifications, can impact RT variability by regulating chromatin accessibility and/or replication initiation (mammalian origins tend to be located at accessible DNA (78)). Finally, other external factors or environmental conditions can influence the RT variability such as cellular stress or exposure to DNA-damaging agents.

The process of carcinogenesis can also produce changes in RT. Ryba *et al.* reported that approximately 18% of the genome show RT variability in leukemias compared to non-cancerous lymphoblastoids (79). However, *Du et al.* found that the RT is largely conserved in cancer cells, however noting a change in 5.7% of the genome between normal and cancer cells in the prostate. Additionally, when they cluster RT profiles from different available cell lines, they observed that the cell types are divided in 3 major clusters: the normal cells, the cancer cell lines (except Hela S3) and the EBV-transformed normal lymphoblasts (80). Altogether, this shows that indeed there are some differences in RT profiles of cancer versus healthy cells.

1.4.2 DNA accessibility and chromatin landscape

DNA accessibility reflects the degree to which a specific DNA segment is accessible for interactions with other molecules, such as transcription factors, enzymes such as DNA or RNA polymerases, or regulatory proteins. In a cell, the DNA is tightly packed around histone proteins forming the chromatin. However, for some cellular processes such as gene transcription, DNA replication or DNA repair the specific regions of the DNA needs to be open (or accessible, normally corresponding to nucleosome-free regions) (81). The chromatin landscape refers to the overall pattern of chromatin organization and epigenetic modifications across the entire genome of a cell.

The DNA accessibility is influenced by different factors: (i) the chromatin landscape that can be by default more open or closed in a specific region (e.g. active genes are usually located in more open chromatin regions); (ii) the DNA cytosine methylation that, when occurring in promoter regions, leads to gene silencing therefore it can associate with the accessibility of that DNA region; and (iii) the histone covalent modifications (“marks”) that can modify the structure of the chromatin and affect the accessibility of the DNA region (e.g. histone acetylation usually increase DNA accessibility) (81). There are methods, such as DNase-seq and ATAC-seq, to experimentally measure chromatin accessibility by identifying regions of the genome that are more accessible to DNase enzymes or transposases, respectively (82). The chromatin accessibility of a large number of diverse samples has become available through the ENCODE (83), and for hundreds of tumors via the TCGA consortium (84).

In the global ENCODE study of accessible DNA, they measured DNase I hypersensitive sites (DHSs) that are regulatory DNA markers, creating a consensus high-resolution map of 3.6 million DHSs across 438 human cells (83). DHS patterns are shared across both individual biosamples and groups of biosamples, however they found several cell- and tissue-specific regulatory programs (83). In the TCGA study using ATAC-Seq, they determined the chromatin accessibility landscape in 410 TCGA tumors across 23 cancer types (84). Similarly, they found several chromatin accessibility profiles that are cancer- and tissue-specific DNA regulatory elements.

They additionally classified the tumors into subtypes that showed different prognostic outcomes (84). Both studies evidence the presence of a strong variability in chromatin accessibility across tissue or cell samples, particularly across different cell types.

1.4.3 Histone marks

The chromatin marks are chemical modifications that occur on the histone proteins around which DNA is wrapped to form chromatin. These marks are also called histone modifications or epigenetic marks (85) (the latter term would also encompass DNA cytosine methylation). They play a crucial role in regulating gene expression and chromatin structure by influencing how genes are activated or silenced, without altering the DNA. The presence or absence of specific chromatin marks can impact various cellular processes, including DNA replication, transcription and DNA repair (85).

There are several histone modifications with different features and functions. H3K9me3 is a marker of constitutive heterochromatin (highly compacted chromatin, usually nuclear lamina-associated) while H3K27me3 is a marker of facultative heterochromatin (repressed regions and promoters). It marks bivalent genes ready for activation or inactivation during development and during differentiation of stem cells (85). H3K9ac and H3K27ac are markers of transcription and active chromatin mostly placed at promoters and enhancers. H3K4me1 is a marker of enhancers while H3K4me3 marks promoters (85). A comprehensive dataset for many histone marks across hundreds of mammalian samples has been generated and released by ENCODE and Roadmap Epigenomics consortia (86).

In addition, using computational methods for segmentation, locally distinct and stable patterns of histone modifications and other chromatin-associated features were described, and named chromatin states (87); a commonly employed method is ChromHMM. Each state is associated with a functional outcome, such as gene activation, gene repression, or “poised” states that allow rapid activation or repression in response to cellular signals (87).

The loci bearing certain histone marks can span a range of lengths, between several hundred nucleotides to multiple kilobases. Therefore, the variability across tissues and individuals for the histone marks landscape has been studied at lower scales than the megabase scale we focus on. However, there are some studies that show that epigenetic deregulation in cancer can span large domains of long-range epigenetic silencing (88) and activation (89). These regions showed coordinated gene expression, histone modification, DNA methylation changes and disruption of topologically associated domains (TADs) between several kilobases to megabases (80,90).

1.4.4 DNA methylation

DNA methylation is an epigenetic modification that consists of adding a methyl group to the cytosine base of DNA. This covalent modification in mammals happens at CpG dinucleotides, where a cytosine (C) is followed by a guanine (G) in the DNA sequence (in the brain there is non CpG methylation (91)). During cell division, DNA methylation patterns are faithfully transmitted

to daughter cells, contributing to epigenetic memory and maintaining cell identity across generations. DNA methylation is involved in key cellular processes such as gene regulation, differentiation and genome stability (92–94).

There are regions of the genome, called CpG islands (CGIs) that contain short (approximately 1 kb) CpG-rich regions while the rest of the genome is depleted for CpGs. These CpG islands are often found near the promoter regions of the gene. Methylation of CpG islands in gene promoters is strongly associated with gene silencing, as it can inhibit the binding of some transcription factors and other regulatory proteins necessary for gene expression (92,93). In the gene bodies, DNA methylation also occurs even though it is less frequent. Its methylation is not associated with repression but rather with increased expression, and its functional role is still not fully established (92).

DNA methylation is involved in differentiation and development. The DNA methylation patterns change as cells transition from stem to differentiated cell types causing the methylation or demethylation of promoters of specific genes and facilitating the establishment of cell-specific gene expression programs (92,93). The disruption of the DNA methylation can have significant consequences and is associated with cancer and developmental disorders (94). In addition, DNA hypomethylation was reported to occur within lamina-associated, late-replicating regions termed partially methylated domains (PMDs) and to be associated with the number of mitotic cell divisions (age); the effect does not seem particular to cancer (95).

The variability in DNA methylation across tissues and individuals has been widely studied. Between tissues, it was shown that there exist methylation ‘grand canyons’ which differ substantially between cell types (96). At the inter-individual level, the difference in DNA methylation has been studied between healthy and cancer patients. For instance, regions of focal DNA hypermethylation and long-range hypomethylation were described in colorectal cancer (97), breast cancer (98), ovarian cancer (99) and others (89).

1.5. Origins of replication

1.5.1 Overview of the origins of replication mechanism

During cell division, the replication of the DNA initiates from specific sites known as origins of replication (henceforth, origins). In eukaryotes, the origins are determined in two subsequent steps. The initial phase, known as ‘licensing’, consists of the formation of the pre-replication complex (pre-RC). For this, first the Origin Recognition Complex (ORC) recognizes and binds to the origins. Next, the ORC recruits other proteins that load the MMC complex (with helicase activity). The helicase creates a replication bubble allowing the polymerases to access the DNA (100–102).

Following this, the initiation of DNA synthesis commences, known as origin ‘firing’. At each origin, the DNA synthesis proceeds bidirectionally. Since the DNA synthesis always advances in the 5’

to 3’ direction, one strand (called the leading strand) is synthesized continuously by polymerase ε while the other strand (called the lagging strand) is synthesized discontinuously through short RNA-primed DNA segments by polymerase δ. This dual-step approach to origin usage, where the two steps are separate in time, is vital to ensure that re-replication doesn't occur in a single cell cycle. Notably, in each cell cycle only a subset of all licensed origins are activated (100–103).

Mammalian cells, due to their large genomes, require thousands of replication forks that start from origins to fully duplicate their DNA. Their replication initiation mechanism is highly flexible, with numerous origins activating at varied frequencies based on the cell lineage and/or cell population (102,104). According to their frequency of usage, the origins of replication can be separated into 3 categories: (i) constitutive origins, very few origins that are activated almost all the time; (ii) flexible origins, the majority of the origins that are used in an stochastic manner in each cell cycle; and (iii) dormant origins, which are used only when adjacent origins fail or under stress conditions (100,102).

Two main theories exist about how replication origins are chosen. The first suggests there's a specific decision point in the G1 phase deciding which origins to activate. The other theory posits that origins are activated at random, with their efficiency rising throughout the S phase (102). Various factors, such as nuclear position, epigenetic cues, and the presence of MCM complexes, impact the efficacy of these origins (102). However, many details remain unclear, and more research is needed to fully understand origin selection.

1.5.2 Methods to detect origins of replication

In yeast, the origins are primarily dictated by specific DNA sequences, known as the ARS (100,101). In contrast, discerning these replication origins in humans poses a greater challenge because a simple motif analogous to the ARS does not appear to exist, and moreover. While it is not straightforward to accurately identify these origins from DNA sequence in humans, a modest degree of prediction is attainable. This predictive capability arises from the amalgamation of specific sequence features enriched in origins, including for instance the presence of CpG islands and G-quadruplex structure (G4) forming DNA motifs (105). In addition, the epigenomic landscape also contains correlates of origin usage. Features such as regions of open chromatin, areas of DNA hypomethylation, and the presence of the H2A.Z histone variant, are statistically associated with loci of replication origins (106,107).

In the last decades, the use of next generation sequencing-based techniques enabled the identification of thousands of replication origin loci in the human genome:

SNS-seq method is based on the purification and quantification of short nascent strands (SNS) of DNA specific to replication origins. This method detects between 40000 to 110000 origins per sample and presents a very high resolution in pinpointing the exact location of the origins (105).

Bubble-seq method is based on the sequencing of the DNA replication bubble, which is formed when two replication forks diverge from a single origin. The DNA is fragmented and the circular

replication bubbles are trapped in agarose gel. This method detects more than 100000 replication origins (108).

Ini-seq method is based on the biochemical synchronization of the initiation events in a cell-free system. The newly replicated DNA is labeled and immuno-precipitated. This method detects around 50000 origins (109).

OK-seq method is based on sequencing purified Okazaki fragments. Using the Okazaki fragment to determine the fork polarity allows to determine the initiation and termination sites. This method detects between 5000 to 10000 origins, and the resolution is lower (110,111).

Other methods. The location of the origins can also be interfered with repli-seq data (106), using computational methods based on the nucleotide skew bias (112) or from Chip-seq data of proteins associated with replication origins such as ORC1, ORC2 and MCM7 (113–115).

Although several experimental technologies are used to identify DNA replication origins, their results can vary greatly, particularly in terms of the number of origins identified, and moreover the overlap in the set of origin loci is not always high. These variations are frequently linked to technical aspects of the assays, including sensitivity, potential for false positives, or resolution challenges. While this likely explains a part of the discrepancy, we hypothesize that additionally the differences might reflect genuine differences between specific cell types and/or individuals in origin usage.

Despite previous reports indicating that the replication initiation mechanism in mammals is highly flexible, with individual examples of origins activating at different frequencies based on cell type or cell group (102,104), a comprehensive study of the variability in origin usage across tissue types remains lacking. In this thesis, we developed a method based on genomic data analysis to pinpoint origins using mutation strand asymmetry and explore the differences in origin usage among distinct types of human somatic tissues. It uses the mutational strand asymmetry of certain mutational signatures at origins of replication to locate the origins. This approach was first suggested by Shinbrot *et al.* (103) and has been further implemented by Jaksik et al. (106), who used the asymmetry of POLE mutated samples to locate origins of replication.

Chapter 2
Objectives

Recent studies in cancer genomes have characterized the regional mutation density (RMD) profiles along the genome, which presented a strong tissue-specific correlation with RT and heterochromatin landscapes. However, the existence of inter-individual variability in the regional mutation density independently of the tissue has not been explored.

In this thesis, we aim to systematically characterize the patterns of somatic regional mutation density variability between tissues and individuals. We aim to detect and quantify the inter-individual variability in the RMD profiles of human tumors, to identify their underlying mechanism, and to understand their causes and consequences. Additionally, we aim to use the tissue-specificity properties of the RMD profile to predict the cancer-type of origin. Finally, we intend to use regional strand asymmetry patterns in mutations to identify the location of origins of replication and study their variability across tissues.

The specific objectives of this thesis are:

1. To study and characterize the inter-individual variability in domain-scale regional mutation density profiles in human tumors.
    a. To develop a method to systematically extract the patterns (signatures) of regional mutation density variability both across tissues and individuals in an unsupervised manner.
    b. To apply systematic statistical analyses to uncover the mechanisms underlying the inter-individual variability signatures by characterizing the regions which suffer the redistribution of mutation rates
    c. To perform association testing between candidate driver events (CNA and mutations) and mutation rate redistribution
    d. To explore the possible consequences of these redistribution of mutations on the mutation supply towards regions harboring cancer genes
2. To develop and apply machine learning classifiers to infer the tissue-of-origin in human tumors using the RMD profiles. To compare the accuracy of the RMD models against, and in combination with other features extracted from DNA sequencing data, such as the trinucleotide composition or driver mutations
3. To extend these cancer type classifiers to predict the tumor-type characteristics of cell lines, in order to identify the cell lines that are more faithful models for cancers. We will extend our models to use other features such as gene expression and methylation. Associations of cell type to drug responses will be studied.
4. To develop a method to study regional mutation rates in DNA strand-specific manner, thereby identifying origins of replication based on the mutational asymmetry. Using these newly-detected origins, we aim to investigate whether there is tissue or inter-individual specific variability in replication origin usage in humans.

IRB BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE

Fran Supek
Group Leader

To whom it may concern,

As requested for the purposes of depositing the PhD thesis for ms. Marina Salvadores Ferreiro, this is a report of the published articles related to her thesis, describing their impact.

First-author papers of Marina Salvadores that form part of the thesis:

Cell cycle gene alterations associate with a redistribution of mutation risk across chromosomal domains in human cancers. **M Salvadores**, F Supek (2023) *Nature Cancer*, in press. (impact factor 2021 **IF_{2021} = 23.2**)

This is a global study of mutation risk heterogeneity across large-scale chromosomal domains in human, describing systematic analyses on thousands of cancer whole-genome sequences. We classified changes in such mutation risk particular to different somatic tissues, which can be applied for cancer subtyping. Next, by use of a custom methodology for mutation pattern recognition, we identified variation in megabase-scale mutation risk between individuals, which was observed across almost all tissues. These changes in mutation risk associate with disruptions in cell cycle control genes *RB1* and *TP53*, and moreover coincide with domains showing chromatin remodelling and replication time changes. Thus, alterations in cell cycling cause a redistribution of mutation risk across human chromatin.

The PhD candidate ms. Marina Salvadores has performed all statistical analyses in this very extensive study, and had major contributions to conceptualization, interpretation, and writing of the manuscript.

Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. **M Salvadores**, F Fuster-Tormo, F Supek (2020) *Science Advances* 6 (27), eaba1862 (**IF_{2021} = 15.0**)

Cancer cell lines are widely used experimental models of tumors, however it is often unclear whether they represent the tumors well with respect to their genomes, transcriptomes and epigenomes. We performed a large-scale analysis that bridges tumor DNA methylation and mRNA expression levels with the cell line data, and further validated these analysis by analysis of mutation patterns in cancer cell line genomes. This revealed convergent evidence that, remarkably, tens of cell lines are likely to originate from a different tissue than originally thought, revealing prevalent mislabeling. Additionally we demonstrate the use of the method for tumor subtyping. Correcting these tumor type labels when analysing drug screening data improves power to find associations between driver mutations and cancer drug sensitivity, thus improving the ability to identify new potential treatments.

Ms. Salvadores has performed all statistical analyses in this study (save for bionformatics processing of DNA sequences), and had major contributions to conceptualization, interpretation, and writing of the manuscript.

Passenger mutations accurately classify human tumors. **M Salvadores**, D Mas-Ponte, F Supek. *PLoS Computational Biology* (2019) 15 (4), e1006953 (**IF_{2021} = 4.8**)

Parc Científic de Barcelona
Baldiri Reixac, 10
E-08028 Barcelona
Spain
www.irbbarcelona.org

Fran Supek
Group Leader

INSTITUTE FOR RESEARCH IN BIOMEDICINE

We studied the utility of mutation patterns in tumor genomes for cancer typing and subtyping, and in particular the applications discovering the tissue-of-origin for the metastatic cancers of unknown primary. We contrasted driver mutation patterns to the passenger mutations, the latter represented by megabase domain-scale mutation density and, independently, by trinucleotide mutation signatures. The passenger mutation signatures are vastly more powerful in classifying cancer types and subtypes than the driver changes, underscoring the power of generating WGS data (rather than exome-only or panel sequencing data) in cancer diagnostics and monitoring.

Ms. Salvadores has performed nearly all statistical analyses in this study, and had major contributions to conceptualization, interpretation, and writing of the manuscript.

Middle-author papers with substantial contributions of Marina Salvadores:

Prevalence, causes and impact of *TP53*-loss phenocopying events in human tumors. B Fito-Lopez, **M Salvadores**, MM Alvarez, F Supek (2023) *BMC biology* 21 (1), 92 **(IF_{2021} = 7.4)**

Ms. Salvadores had considerable contributions to analyses and interpretation of the results, providing guidance for a junior colleague (first author on the study), and contributed writing of the manuscript.

Loss of the abasic site sensor HMCES is synthetic lethal with the activity of the APOBEC3A cytosine deaminase in cancer cells. J Biayna, I Garcia-Cao, MM Álvarez, **M Salvadores**, J Espinosa-Carrasco, ... F Supek*, T Stracker* (2021) *PLoS Biology* 19 (3), e3001176 (**IF_{2021} = 9.6**)

Ms. Salvadores had substantial contributions by performing analyses of large-scale genetic screening data across various cell lines, thus providing important support for the experimental data from CRISPR screens in the study.

Mutational signatures are markers of drug sensitivity of cancer cells. J Levatić, **M Salvadores**, F Fuster-Tormo, F Supek (2022) *Nature communications* 13 (1), 2926. (**IF_{2021} = 17.7**)

Ms. Salvadores had substantial contributions to developing methods and performing analyses that link genetic markers to drug resistance profiles, and replicating the associations between data sets to assess robustness, thus providing important support for the conclusions of the study.

Sincerely,

Fran Supek

In Barcelona 26/09/2023.

Travis Stracker

In Bethesda 091023

Parc Científic de Barcelona
Baldiri Reixac, 10
E-08028 Barcelona
Spain
www.irbbarcelona.org

Chapter 3

Cell cycle gene alterations associate with a redistribution of mutation risk across chromosomal domains in human cancers

---

The following chapter has been selected from the publication:

M Salvadores and F Supek (2023). Cell cycle gene alterations associate with a redistribution of mutation risk across chromosomal domains in human cancers. Nature Cancer (in press).

This manuscript has been accepted for publication at Nature Cancer.

Cell cycle gene alterations associate with a redistribution of mutation risk across chromosomal domains in human cancers

Marina Salvadores [1], Fran Supek [1,2] *

[1] Genome Data Science, Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, 08028 Barcelona, Spain.
[2] Catalan Institution for Research and Advanced Studies (ICREA), 08010 Barcelona, Spain.
* correspondence to: fran.supek@irbbarcelona.org

Abstract

Somatic mutations in human cells have a highly heterogeneous genomic distribution, with increased burden in late-replication time (RT), heterochromatic domains of chromosomes. This regional mutation density (RMD) landscape is known to vary between cancer types, in association with tissue-specific RT or chromatin organization. We hypothesized that regional mutation rates additionally vary between individual tumors in a manner independent of cell type, and that recurrent alterations in chromatin organization may underlie this. Here, we identified three RMD signatures that describe mutation risk redistribution across megabase-sized domains in genomes of >4000 tumors; these 3 somatic mutation landscapes were universally observed across various tissue types. First, we identified a mutation redistribution RMD signature preferentially affecting facultative heterochromatin, malleable RT and Hi-C domains, and enriched in the B1 subcompartment and in subtelomeric regions. This RMD mutation risk signature strongly reflects recurrent programs in plasticity in DNA replication time, and in heterochromatin and accessible DNA remodeling, observed across many tumors, tissues and cultured cells, robustly linked with a higher expression of cell proliferation genes. Consistently, occurrence of this mutation redistribution pattern in cancers is associated with altered cell cycle control *via* loss of activity of the *RB1* tumor suppressor gene. Second, another independant global RMD signature was associated with loss-of-function of the *TP53* pathway, mainly affecting the redistribution of mutation rates within late RT regions. Third, we quantify a known "flat" RMD pattern associated with DNA repair failures, additionally linking it to homologous recombination deficiencies. These three mutation risk redistribution signatures can change the local mutation supply towards 26%-75% cancer driver genes, potentially diverting the course of cancer evolution.. Our study highlights that somatic mutation rates at the domain scale can be variable across tumors in a manner independent of tissue-of-origin, but associated with cell proliferation and alterations of cell cycle control genes, which trigger the local remodeling of heterochromatin, spatial chromatin contacts and the RT program in many human cancers.

Introduction

Mutation rates in somatic cells are highly heterogeneous across megabase-scale segments, with higher mutation rates in late DNA replication time (RT), inactive, heterochromatic DNA. This is largely due to higher activity and/or accuracy of various DNA repair pathways in early-replicating, active chromosomal domains [1,2].

These segments with variable mutation rates tend to correspond to topologically associating domains (TADs) and RT domains [3–5]. Regional mutation density (RMD) of mutations in megabase-sized domains in the human genome correlates with the RT of the domain, as well as the local gene expression levels, chromatin accessibility (e.g. as measured by DNAse hypersensitive sites (DHS)), density of inactive histone marks such as H3K9me3 and, in the opposite direction, with density of active marks such as H3K4me3 [1,6–8]. The global RMD landscape in somatic cells has been shown to be to some extent tissue-specific, sufficiently so that it can be used to predict cancer type at high accuracy [9,10]. The tissue-specificity of RMD in a domain is paralleled in the tissue-specificity of (in)activity of the chromatin in the domain. For instance, the domain that switches from late-RT to early-RT, or where genes increase in expression levels, or that gets more accessible chromatin in a particular tissue, also exhibits a reduced rate of somatic mutations in that tissue [1,6]; this property can identify the cell-of-origin of some cancers [11].

Apart from variation in accessible chromatin and gene expression and RT between tissues, there is also known systematic variation in chromatin within the same tissue, but across cells or across individuals. For instance, various statistical studies of quantitative trait loci (QTL) have demonstrated genetic determinants of changes in RT or chromatin accessibility or DNA methylation [12–15]. Further, recent work suggests existence of gene expression programs that are variably active between tumors originating from the same tissue (and also between individual cells), but are recurrently seen across many different tissues [16,17]. Such recurrent programs may conceivably drive, or be driven by chromatin remodeling that activates or silences chromosomal domains. Indeed, chromatin remodeling was widely reported to occur during tumor evolution, and this can manifest as changes in RT between normal and cancerous cells, loss of DNA methylation, or changes in heterochromatin marks in some chromosomal domains upon transformation [18–22]. This large-scale chromatin remodeling occuring in various chromosomal domains of cancer cells may plausibly affect local and regional chromosomal stability, given the mechanistic links of various DNA damage and repair processes and different features of chromatin organization [1,2,20,23–25].

Here, we hypothesized that chromatin remodeling that occurs variably between tumors may generate inter-individual variation in regional mutation rates, beyond the known cell-of-origin identity effects on local mutagenesis.

We study the regional profiles of somatic mutations from ~4200 tumor whole-genome sequences, modeling them as a mixture of several underlying regional distributions, which correspond to different mutation risk mechanisms acting preferentially in some genomic domains. Some of these RMD signatures represent the expected differences between tissues/cell types stemming from chromatin organization in the cell-of-origin, or consequences of common DNA repair failures, However two commonly-occuring RMD signature patterns are novel and are associated with with large-scale chromatin remodeling. These RMD mutation signatures can be accurately predicted from local switches observed in RT programs that we characterized across ~400 tumors, and which were associated with cell cycling gene expression programs.

Such RT-switching and chromatin remodeling-associated mutation risk changes (i.e. RMD signatures) are ubiquitous amongst human cancers, and they associate with alterations in cell cycle genes *RB1* and *TP53*. The resulting wide-spread mutation rate redistribution

across chromosomal domains can increase or decrease mutation supply towards regions harboring cancer genes, potentially generating driver events and genetic interactions.

Results

Megabase-scale regional mutation density in human tumors can refine tissue-of-origin classification of tumors

We hypothesized that, in addition to the known RMD variation between cancer types, the RMD patterns may encompass variability between individual tumors observed independently of tissue-of-origin. To test this, we performed a global unsupervised analysis of diversity in one megabase (1 Mb) mutation density patterns across 4221 whole-genome sequenced tumors. To prevent confounding by the variable SNV mutational signatures across tumors [26], we controlled for trinucleotide composition across the 1 Mb windows by sampling sites and mutations (Methods). We additionally normalized the RMDs at chromosome arm-level to control for possible confounding by large-scale copy-number alterations (CNA). Finally we removed known mutation hotspots (e.g. CTCF binding sites, see Methods), and also exons of all protein-coding genes to avoid selected mutations.

To validate the obtained tumor RMD features, we applied a Principal Component (PC) analysis on the RMD profiles across all tumor samples (n=4221), expecting separation by tissue-of-origin since the RMDs mirror the local chromatin organization and the RT program in the particular tissue [1,6]. Indeed, clustering on these RMD mutational profile PCs largely reflected tissue identity (Fig S1a-e). Additionally we note cases where the similarity of RMD profiles suggested to 'merge' some plausibly similar cancer types. For example the RMD_cluster3 contains various digestive tract cancers, likely because of the broad similarities in chromatin organization of epithelial cells in different gastrointestinal organs. The RMD_cluster11 contains head-and-neck cancers, the non-melanoma skin cancers and some esophagus and lung cancers and so suggests a similar chromatin organization in the (putatively squamous) cell-of-origin for these diverse tumors (Fig S1e). Conversely, RMD profiles may subdivide some cancer types into plausible subtypes.For instance, breast cancer samples in (ovarian-like) RMD_cluster6 have visually distinct RMD mutation profiles from the (typical breast-like) RMD_cluster9 (Fig 1ab; Fig S1ef); the former tumors are strongly enriched in the triple-negative breast cancer subtype, which has similar gene expression to ovarian cancer [27]. This example suggests that RMD profiles may be more generally useful for reclassifying or subtyping various cancer types. One salient case is head-and-neck cancer, which might be split into RMD_cluster11 (squamous-like, includes non-melanoma skin cancers) and RMD_cluster13 (this group also contains some lung cancers and so might be considered lung-like) (Fig 1ac; Fig S1e).

As a control, our RMD features captured two known examples of regional redistribution of mutation rates: one affects specific genomic loci (somatic hypermutation regions in B-lymphocytes (Fig S1gh)), and the other causes a global homogenization ('flattening') of the RMDs along the genome in MMR-deficient tumors [1] (Fig S1g) and in APOBEC-mutagenized tumors [28] of various cancer types (addressed in more detail below).

Overall, while the RMD profiles expectedly reflect tissue-specific signal, there is additional systematic RMD variability in certain tumor genomes, which may be used to re-classify cancer types and subtypes based on mutation patterns. Next, we further hypothesized that there may be RMD variation that occurs independently of cancer type, which necessitated a more rigorous method for RMD analysis of regional mutation risk.

A statistical method to detect inter-individual variation in regional mutation density

Aiming to separate the tissue-specific RMD variability from the inter-tumor mutation patterns independent of tissue-of-origin, we devised a methodology analogous to that recently used for extracting trinucleotide SNV mutational signatures [26,29,30], however here applied to 2540 megabase-sized domains instead of the typical 96-channel trinucleotide SNV spectrum. In brief, non-negative matrix factorization (NMF) is repeatedly applied to bootstrap samples of mutational density per 1 Mb windows, normalized for trinucleotide composition, to find sets of NMF factors that are recovered consistently across bootstrap runs, and thus deemed robust to noise in the data. Each of these NMF factors corresponds to one "RMD signature", with a spectrum consisting of RMD window weights (for all 2540 windows), and additionally the RMD signature has data on tumor sample 'exposures' or activities (the NMF weight of each tumor genome for that RMD signature).

To test whether our NMF method is sufficiently powered to capture RMD inter-individual variability, we simulated cancer genomes containing ground-truth patterns of RMD. These were generated to affect a variable number of windows, to be present in variable number of tumor samples, and to be present at variable intensity (fold-increase over default mutation rate at each window) (Fig S2a, see detailed description in Methods). We ran our NMF methodology for these different simulated data scenarios independently. We selected the number of factors and clusters based on the silhouette index (SI), estimating reproducibility over repeated runs of NMF (Fig S2b), and then matching the known ground-truth signatures to estimate accuracy (Methods, Fig S3). We show an example of an extracted RMD signature compared to its ground-truth signature in Fig 1d.

By comparing the different scenarios (Fig S4a), encouragingly, we observed that even with a small fraction of tumor samples affected by a signature (5%), the ground-truth RMD signatures can be identified reliably, as long as the contribution of the RMD signature to the total mutation burden is reasonably high (>=20%). In addition, we observed that the NMF setup is robust to the number of windows affected, since it is usually able to recover RMD signatures that affect as little as 10% of all 1 Mb windows. Out of other characteristics that may affect power to recover RMD signatures, we identified the signature exposure strength (fold-enrichment over baseline mutation rate) as showing the highest effect, thus the signatures with subtle effects on RMD might not be recovered (Fig S4a). In summary, our simulations support that our NMF methodology can recover the genome-wide RMD signatures in a wide variety of scenarios.

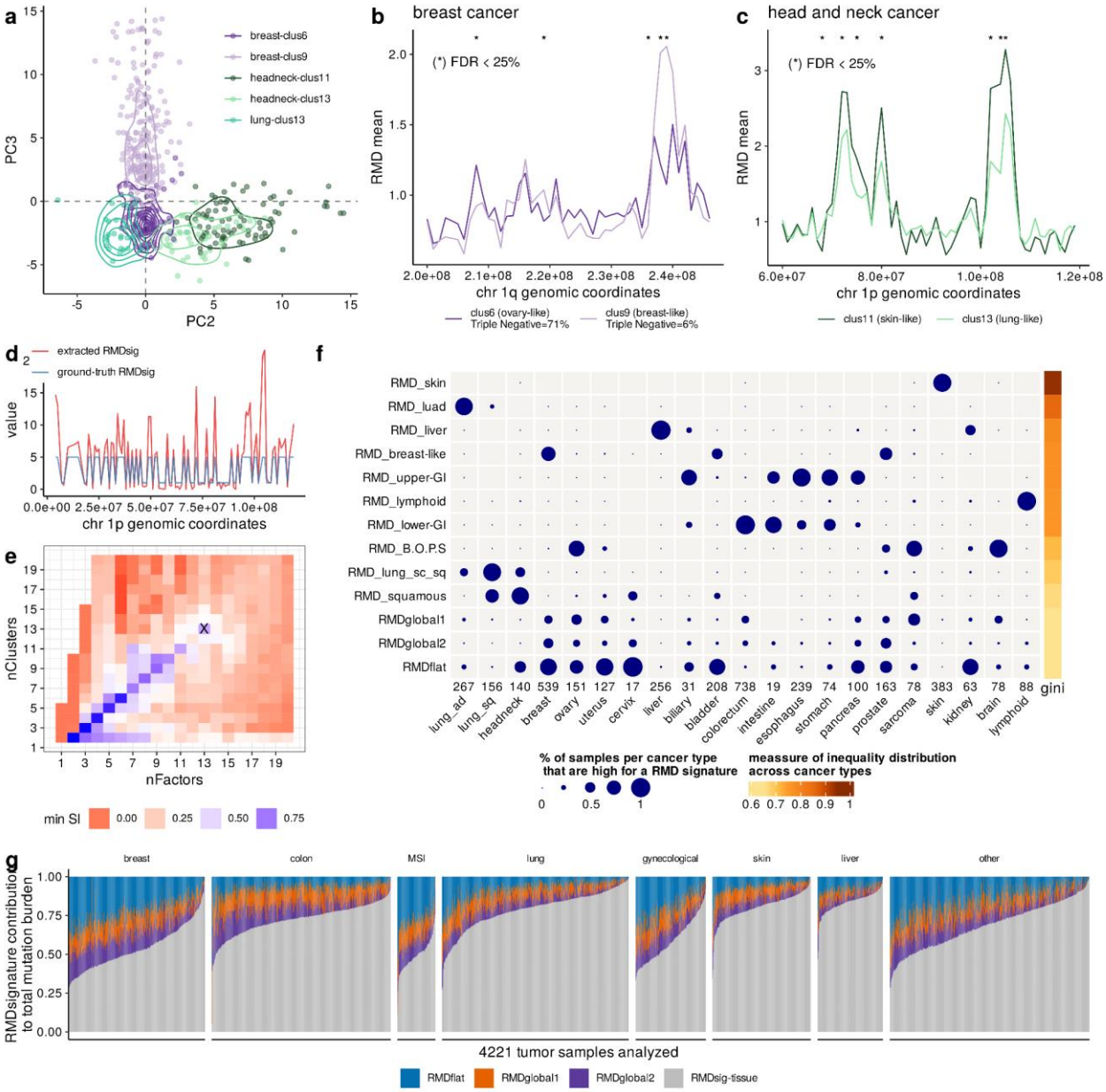


Figure 1. Identifying RMD signatures by applying an NMF-based method to megabase-scale mutation density profiles of human tumor genomes. a) A principal component (PC) analysis of the RMD profiles, containing normalized mutation frequencies in 2540 non-overlapping 1 Mb windows. The PC2 and PC3 from a PCA breast, head-and-neck and lung tumor WGS (n=1408) separate the RMD subtypes. **b)** Mean RMD profiles (shown on chr 1q) for breast cancers in the PCA cluster 6 (n = 76) and those in cluster 9 (n= 211), shows enrichment of the triple-negative subtype in the former. Stars mark windows that are significantly different at FDR<25% by wilcoxon.test (two-tailed). **c)** Mean RMD profiles (shown on chr 1p) for head-and-neck squamous cell cancers in the PCA cluster 11 (n = 81) and cluster 13 (n= 41), where the former cluster with skin cancers and the latter with lung cancers, see Fig S1e. . **d)** Example RMD signature from a simulation study (see Methods), comparing 1 Mb window weights for an extracted NMF signature and its matching simulated ground-truth signature along chr 1p. See Supplementary Figs S2-S4 for additional simulation data. **e)** Quality scores for various parameters for NMF run on 4221 tumor genomes. The minimum silhouette index (SI) across clusters (RMD signatures) for different numbers of NMF factors and clusters from k-medoids. The selected case (nFactor=13, nCluster=13) is marked with a cross. **f)** Overview of the 13 RMD signatures extracted (rows) and their distribution across

different cancer types (columns). The circle size corresponds to the fraction of tumors in a cancer type exhibiting a high activity of a specific signature (exposure >= 0.12, corresponding to the 1st percentile of the exposure of the RMDflat signature in microsatellite-instable [MSI] cancers). Total number of samples per cancer type written beneath table. Gini index quantifies the cancer type specificity, where lower Gini means shared across different cancer types (the "global" signatures). **g)** Relative contribution of each RMD signature to the total mutation burden across the 4221 tumor genomes analyzed. The The RMDsig-tissue category represents all 10 tissue-specific RMD signatures, pooled together. MSI tumors from various cancer types all shown together.

A catalog of tissue-specific mutation patterns in human cancer types

We applied the NMF methodology to the somatic RMD profiles of 4221 tumor WGS, here requiring a minimum of 3 mutations/Mb, thereby restricting to tumors with less noisy RMD profiles (as a limitation, we note that this may deplete low mutation burden cancer types). A simulation study indicated that 3 mutations/Mb are well sufficient for analysis at this resolution (Fig S4bc). In total, we extracted a total of 13 RMD signatures based on the SI, which scores the reproducibility of solutions (Fig 1ef, Fig S5).

We observed that the RMD signatures from NMF span a continuum from very tissue-specific (high Gini index, Fig 1f), to global signatures seen across many cancer types (low Gini index). We named the 10 tissue-specific RMD signatures according to the tissue or tissues they affect (e.g. RMD_upper-GI, RMD_liver), while the three global signatures were named RMDglobal1, RMDglobal2 and RMDflat (Fig 1f, Fig S5); the latter is named by the visually recognizable pattern, and also has in part known mechanisms (see below) while RMDglobal1 and RMDglobal2 are to our knowledge novel.

While some RMD signatures are extremely tissue-specific and capture the genomic regions with high mutation risk only in that particular organ , many RMD signatures are observed in several cancer types, which appear related (Fig 1f, Fig S5). For instance, RMD_upper-GI signature is present in most esophagus, stomach, pancreas and biliary tumor samples, and a few intestine tumors. The RMD_lower-GI signature, in turn, is active in mainly the colorectal and most of the intestinal tumors, broadly consistent with the subdivision by developmental origin into the foregut (RMD_upper-GI) and the midgut/hindgut (RMD_lower-GI; Fig 1f). The RMD_squamous signature is active in some squamous lung cancers, head-and-neck cancers, some bladder cancers (consistent with reports based on gene expression data [31]), also expectedly some cervical and esophageal tumors, however surprisingly includes some sarcomas and uterus cancers. These commonalities in RMD probably reflect similarity of chromatin organization in the cell-of-origin of tumor types, in accord with from anatomical site and/or cell type similarity. Our NMF-based 2540-channel RMD signatures support the proposed uses of mutational profiles for elucidating cell-of-origin and cancer development trajectories (e.g. metaplasia and/or invasion) [11] by matching to chromatin profiles.

Three prevalent patterns of megabase-scale mutation rate variation observed across most somatic tissues

In addition to tissue-specific RMD landscapes, we identified 3 global RMD signatures that occurred in a substantial subset of tumors within most cancer types (Fig 1f, Fig S5).

Firstly, the profile of the RMDflat signature captures the known reduction in variation in mutation rates between domains, previously associated with MMR and NER pathway deficiencies [1,2] and with high mutagenic activity of the APOBEC3 enzymes [28] . Based on these known associations, which were recapitulated in our data (Fig S6) we hypothesized that additional RMDflat-high tumors (only 52% were explained by known factors) may result from additional DNA repair deficiencies. Indeed we observed that deficiencies in the homologous recombination (HR) DNA repair were commonly associated with RMDflat (Fig S6b), consistent with a distribution of the HR deficiency-associated trinucleotide signature SBS3 towards early-replicating domains [32], opposite of the canonical RMD distribution; see Supplementary Text S1 for a detailed analysis and discussion. Thus various DNA repair-related defects converge onto the RMDflat mutational phenotype, with varying prevalence depending on the cancer type (Fig S6c).

Unlike the homogeneous pattern resulting when the RMDflat signature profile is superimposed onto the canonical RMD landscape, the RMDglobal1 and RMDglobal2 profiles have a complex pattern with their peaks appearing scattered throughout the chromosomes. We can rule out that RMDglobal1 and 2 resulted from random noise, because (a) their silhouette index (robustness of profile to noise introduced in NMF runs) is comparable to the other RMD signatures, and (b) their autocorrelation (similarity in weights of consecutive 1 Mb windows) is comparable to the tissue-associated RMD signatures, which have a known biological basis (Fig S7ab). The RMD signatures were determined using single-nucleotide variant (SNV) mutations; as a validation, we observed that the regional biases of the three global RMD signatures are similarly observed in indel mutation distributions (Fig S7c).

As support to the pan-cancer analysis, the tissue-specific NMF runs also robustly recovered the three RMDglobal signatures (Fig S8). We found signatures in breast, lung and esophagus at a cosine similarity > 0.84 with RMDglobal1, and in colon, uterus and breast at > 0.89 with RMDglobal2. Thus the global signatures capture inter-tumoral RMD heterogeneity in mutation risk of chromosomal domains, which is recurrently observed in various human somatic tissues.

RMDglobal1 signature increases mutation rate in regions with plastic replication timing and heterochromatin

We were interested in the mechanism underlying the widespread RMDglobal1 signature, which was significant in 25% of the tumor genomes (Fig 1f; using a conservative threshold, see Fig S5), and contributed quite variable mutation burden across individual tumors (Fig 1g). Based on tissue-specific RMD patterns reflecting tissue-specific chromatin organization [1,6,9], we hypothesized that other, tissue-independent variation in chromatin across domains may underlie the tissue-independent RMDglobal1.First, we tried to predict RMDglobal1 signature spectrum (the 1 Mb window weights) from epigenomic features previously reported to associate with megabase-scale mutation rates (reviewed in [33]): replication timing (RT), density of accessible chromatin (DNAse hypersensitive sites, DHS) and ChipSeq data for a variety of histone marks including the heterochromatin marks

H3K9me3 and H3K27me3 (Fig 2a). Attempting to predict the RMDglobal1 signature using the average of each feature (RT, DHS, histone mark) per 1 Mb window across many epigenomic datasets was not successful(Fig 2a). Predicting RMDglobal1 from each RT/DHS/histone mark dataset individually identified only moderate associations (R^2 ~= 0.2) for certain datasets with regional density of facultative heterochromatin (H3K27me3) and constitutive heterochromatin (H3K9me3 mark) (Fig 2a), suggesting a role of heterochromatin organization in determining RMDglobal1.

In stark contrast with the above, we observed that RMDglobal1 spectrum can be highly accurately predicted (R^2 up to 0.7) from either RT, or DHS, or either of the two heterochromatin marks above, however only if predicting using multiple tissue samples jointly (Fig 2a). This suggests that RMDglobal1 spectrum is explained by some pattern in the variation across the cell/tissue samples for one chromatin feature. In other words, differences between the individual heterochromatin profiles across the genome can be highly predictive of RMDglobal1 mutagenesis (individual examples shown in Fig 2b), while the average heterochromatin profile across the genome is not predictive (Fig 2a). Similar is the case for RT and DHS (Fig 2a), and as a validation we observed the same trend using regional density of chromHMM segmentation states (Fig S9).

Global analysis of heterochromatin restructuring at the domain scale across human tissues and cell types

Next, we aimed to quantify the systematic variation in heterochromatin states at the domain level, recurrently observed across diverse human tissues and cell types including tumor cells. Our analysis included intact tissues, cultured primary cells, and immortalized and cancer cell lines from ENCODE, with the goal of identifying heterochromatin variation that predicts RMDglobal1 mutation risk changes. To this end, we performed PCAs on the megabase-window signal of the H3K9me3 and H3K27me3 heterochromatin marks, RT, DHS, and Hi-C compartments [34]. While each feature was analyzed independently, the PCs that resulted -- representing chromatin restructuring across domains -- were often recurrently observed across the analyses (Fig 2c). In particular the RMDglobal1 mutagenesis pattern was in a tight cluster of chromatin restructuring PCs, most correlated with the H3K9me3_PC3 and Hi-C_PC2 and DHS_PC3 (R=0.53, -0.53 and 0.43, respectively), and additionally also with RT_PC4 and H3K27me3_PC2 (Fig 2c). These PCs, representing domain-scale chromatin reorganization in certain cell types, exhibited a regional distribution whose peaks colocated with the peaks in RMDglobal1 regional mutagenesis spectrum (example shown in Fig 2d).

Next, to understand the mechanism driving the H3K9me3_PC3 and related heterochromatin restructuring programs, we analyzed RNA-Seq gene expression levels of ENCODE samples with higher versus lower values of the heterochromatin remodeling score (H3K9me3_PC3). By interrogating the MSigDB hallmark gene sets [35], we found the H3K9me3_PC3-high samples were strongly associated with gene expression of MYC target genes, E2F target genes and G2M checkpoint genes (GSEA FDRs = 10^{-41}, 10^{-41} and 10^{-33} respectively; Fig 2e), indicating an association with increased anabolism, DNA replication, and mitotic processes, respectively. Similarly, a significant enrichment of cell cycling-associated genes was observed when using another gene set categorization, the single cell-derived RHP programs [16] (Fig S10a). These pathway enrichments were entirely consistent with those from

contrasting ENCODE samples by the chromatin accessibility shifts (DHS_PC3), or by Polycomb repressive mark shifts (H3K27me3_PC2) across domains (all FDR<1% by GSEA, Fig 2e), thus converging onto a model of genome-wide chromatin restructuring program linked with rapid cell proliferation.

These chromatin restructuring program PCs were differentially active between ENCODE intact tissues (lower scores) and cultured primary cells (higher score) ($p=1.2 \cdot 10^{-12}$ and $<1.5 \cdot 10^{-23}$ for H3K9me3 and DHS, respectively, by Wilcoxon test Fig 2f); tissues typically contain few cells that divide, while primary cell culture selects for proliferation-competent cells and is expected to usually have a higher proportion of cycling cells than tissues do. Consistently, immortal cell lines had more similar chromatin domain signatures to the primary cell cultures than to tissues (Fig 2f). Comparing the cancerous cell lines to noncancerous cell lines reveals considerable overlap between them, both in the H3K9me3 mark PC and also in DHS density PC (Fig 2f), suggesting these chromatin restructuring programs do not strongly reflect cancerous transformation per se (Fig 2f).

We further asked if these chromatin programs reflect tissue specificity. We considered the chromatin features of ENCODE cell types broadly matching the tissue-of-origin for the major cancer types in our WGS mutational data: B-lymphocytes, breast, bladder, breast, colon, esophagus/stomach, kidney, liver, lung, muscle, nervous system, ovary, pancreas, prostate, skin, and uterus; additionally we include 3 non-matched tissue types - other blood cells, heart and all other cell types combined as a reference group. The important chromatin remodeling PCs (H3K9me3, DHS PCs) did not exhibit a tissue-specificity: there was considerable overlap with the "other" mixed-tissues group in all the matched cancer types, and moreover most tissues overlapped each other in the activity of the chromatin remodeling signatures (Fig S10de). Some differences were noted in skin, pancreas, stomach and B-cells however these were not mirrored across H3K9me3 and the DHS remodelling PCs (Fig S10de). Not even the nervous system nor the muscle cells (known to have rather distinctive tissue-specific patterns of active chromatin [36]) showed a notable difference. Thus, our heterochromatin PCs of interest largely do not reflect tissue-specific chromatin organization, similarly as RMDglobal1 does not reflect tissue-specific mutation risk. As a further control, we performed a paired analysis between RMDglobal1 mutagenesis in a cancer type and the tissue-matched H3K9me3 PC signatures, finding similarly strong associations as for a mixed-tissue analysis of H3K9me3 mark (Fig S10f). Overall, tissue specificity in heterochromatin organization appears largely orthogonal to the chromatin remodelling signatures we extracted and linked with RMDglobal1 mutation risk.

Next, we considered the expression of individual genes. Differences in levels of some selected cell proliferation genes (belonging to 3 hallmarks mentioned above) were striking when comparing the ENCODE samples on the two ends of either the H3K9me3_PC3 constitutive heterochromatin reorganization, or DHS accessible chromatin domain-level reorganization (Fig 2g; also holds for H3K27me3_PC3 Polycomb repressed chromatin reorganization, Fig S10c). Taken together, this data suggests that likely the cell proliferation itself, rather than the tissue/cell type identity or the oncogenic transformation status, predicts the domain-scale chromatin reorganization that is mirrored in RMDglobal1 mutation rates.

RT switching profiles of tumors and cultured cells link chromatin changes in cancer with RMDglobal1 mutation risk signature

The above analyses of chromatin features at the domain-scale was performed over data from various ENCODE cell types, which include some cancer cell lines (e.g. 74 out of 256 H3K9me3 and 153 out of 676 DHS datasets are on cancer cell lines), but do not include tumor tissue. Therefore we turned to examine chromatin domain restructuring directly in tumor tissue, as relevant to our measured RMDglobal1 regional mutation risk signature. Here, we draw on accessible chromatin (via ATAC-Seq) measurements in 410 TCGA tumors [37]. The local distributions of accessible chromatin sites can be used to accurately infer RT programs [38]; indeed inferring RT thusly was applied to large-scale studies of RT in tissues other than tumors [39,40].

RT is a chromatin feature of interest because in our analysis of various chromatin features in the ENCODE reference data above, variation in RT matched the accuracy in predicting RMDglobal1 of the variation in the two heterochromatin histone marks (Fig 2a). Moreover the domain switches in RT correspond to switches between euchromatin and heterochromatin (see correlations between experimentally measured RT-PCs and DHS-PCs and Hi-C-PCs in Fig 2c). Thus we consider the RT programs, either measured or inferred from ATAC-Seq, to reflect domain-scale organization of euchromatin/heterochromatin in tumor tissues.

In particular, in this analysis we consider (i) a collection of RT profiles from experiments (RepliChip or RepliSeq) in multiple cell types (expRT, n = 158), (ii) predicted RT in a collection of noncancerous tissues, cultured primary cells and cell lines including cancer and stem cell lines (predRT-ENCODE, n = 597), and (iii) predicted RT in human tumors (predRT-TCGA, n = 410). For the latter two RT datasets, we inferred RT from DHS [36] or ATAC-seq data [37], respectively, using the Replicon tool [38], which was highly accurate in our tests (R between predicted RT and RepliSeq = 0.87, mean over 5 cell lines; Fig S11a) (see Methods). The modeling of RMDglobal1 mutagenesis is similarly accurate using predicted RT (from ENCODE DHS) as is with experimental RT (from various datasets) (Fig 3a), attesting to the accuracy of RT prediction by the Replicon tool. Additionally, the predicted RT profiles in TCGA tumors better model the RMDglobal1 mutagenesis than predicted RT profiles from ENCODE cell types do (Fig 3a). This suggests that our RT analysis may capture tumor-specific RT switching programs that are relevant to tumor RMDglobal1 mutagenesis.

Next, we asked what is the mechanism underlying variability in RT across individuals or tissue types that predicts RMDglobal1 mutagenesis.

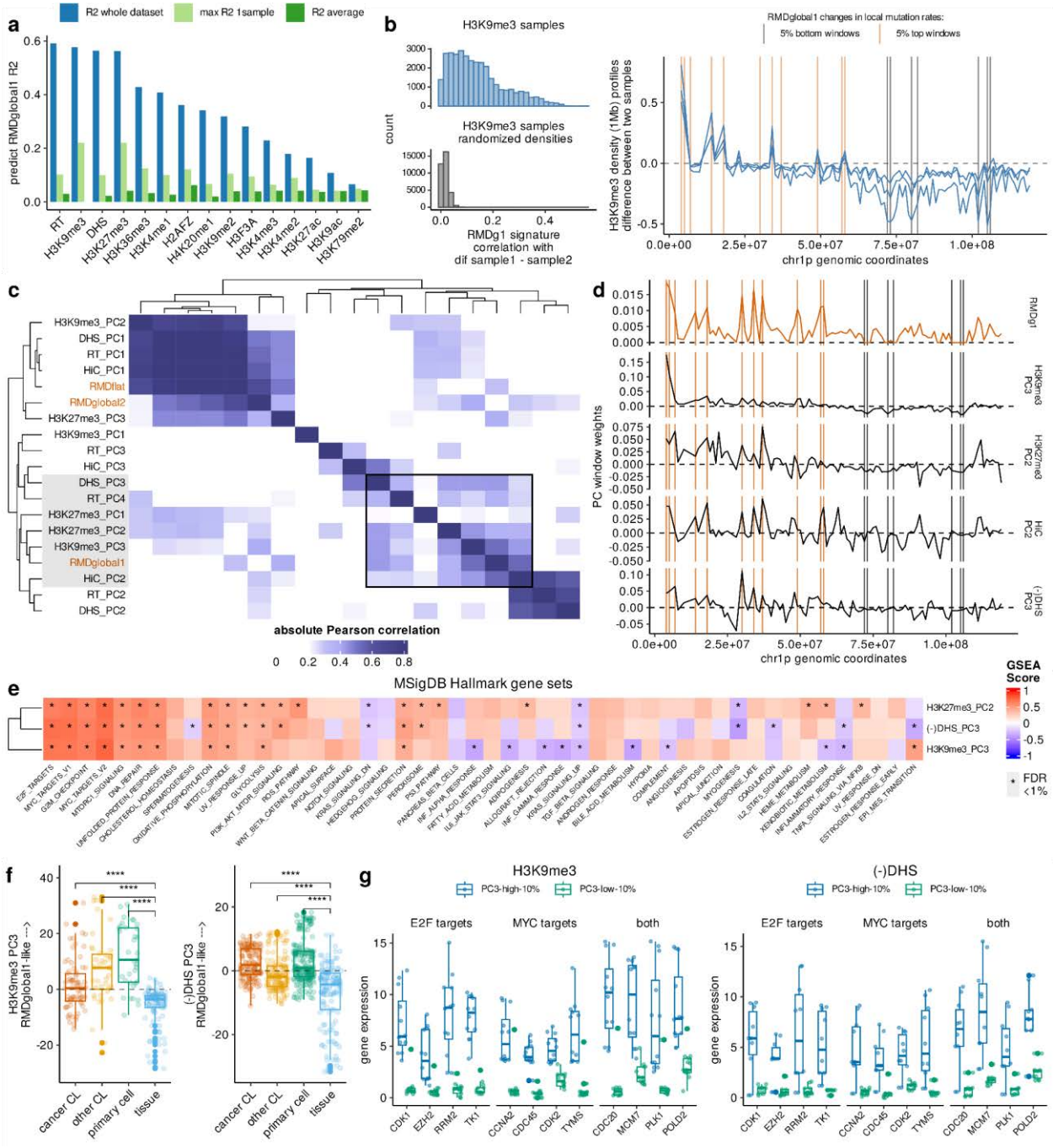


Figure 2. The RMDglobal1 mutation risk redistribution signature is linked to domain-scale variability in heterochromatin. a) Modeling the RMDglobal1 spectrum (2540 window weights) using genome-wide profiles of various epigenomic features (x axis) using either the whole dataset (multiple samples) jointly, or selecting the best-predicting individual sample (maximum R2) in the dataset, or using the average values across the samples. Dataset described in Table S2. **b)** Correlation between RMDglobal1 spectrum, and the difference between each pair of H3K9me3 profiles from ENCODE (blue) and for the ENCODE samples with randomized profiles (grey). Right panel shows the difference in H3K9me3 density for 3 example tumor pairs. The vertical lines represent the locations of top 5% and bottom 5% windows for the RMDglobal1 spectrum (i.e. where changes in cancer mutation rates are highest and lowest, respectively). **c)** Extracting domain-scale chromatin remodeling programs. One PCA was performed independently for each epigenomic feature, and their correlations

shown between the resulting chromatin remodeling PCs. . The black square denotes the cluster of RMDglobal1-associated chromatin PCs. **d)** Example of the window weights on chromosome 1p for the four RMDglobal1-associated chromatin PCs and RMDglobal1 mutagenesis itself. Vertical bars as in panel b. **e)** Gene expression associated with the 3 relevant chromatin PCs. Gene set enrichment analysis (GSEA) scores for the MSigDB hallmark gene set for an ordered list of genes, derived by associating the chromatin PC levels across ENCODE samples with the expression of each gene separately (coefficient for the gene variable from the regression is used as effect size). **f)** Differences in activities of the H3K9me3_PC3 and DHS_PC3 chromatin signatures between the relevant biosample types in ENCODE, with tissue being significantly different from the other 3 types (**** is p<=0.0001 by Wilcoxon test). **g)** Gene expression (square root TPMs) for example genes from the proliferation-associated E2F targets and MYC targets categories, comparing the ENCODE samples with high and low H3K9me3_PC3 or DHS_PC3 chromatin remodeling signatures. The DHS_PC3 was inverted, denoted by "(-)", as an interpretation aid.

To systematize the types of variability within RT profiles of human tumors (that explains RMDglobal1 signature), we applied a PCA with the predRT-TCGA dataset of 410 TCGA tumors, and independently of that also performed PCAs on experimentally measured RT data (expRT), and on predicted RT in ENCODE nontumor tissues and primary cells and varied cell lines (predRT-ENCODE). By analogy to the domain-wise PCA of various heterochromatin marks (Fig 2c), here we also find that certain RT-PCs from these varied datasets converge onto the same global pattern of alteration in the RT program: predRT-TCGA_PC5, predRT-ENCODE_PC3 and expRT_PC4 (Fig 3b); median pairwise correlations in RT profiles R=0.38). These RT PCs also correlated with the heterochromatin PCs highlighted in previous analysis (DHS_PC3, H3K27me3_PC2, H3K9me_PC3 and HiC_PC2) (Fig 3b; median pairwise correlation R = 0.43); thus these RT PCs represent the archetypes of systematic variation in RT program and heterochromatin observed across tumors, cultured cells (either cancerous or not), and healthy tissues.

We additionally considered a data set of RT measured in single cells [19], identifying the RT variation scRT-PC5 which correlates moderately (R=-0.36) with the TCGA RT program mentioned above (Fig 3b). This suggests that the non-stochastic variation of RT programs between individual human cells [40] -- presumably indicating those chromosomal domains that have intrinsically more labile RT -- may also predispose these same domains to the systemic RT switches we observe between tumors.

Next, we asked whether these shifts in RT observed across tumors can explain the local shifts in domain mutation rates observed across tumor WGS. In particular, we correlated each RT-PC with the profile of RMDglobal1 mutation rate variation across megabase windows (Fig 3b). Expectedly, the RT PCs with highest amount of variance explained either represent the average RT profile (predRT-TCGA-PC1, PC2), or the following PC3 and PC4 are cancer type-associated RT programs, separating breast from kidney and brain tumors (Fig S11b). However, the following strongest pattern of systematic RT variation, the TCGA-RT_PC5, did not exhibit a tissue signal (Fig S11c), but instead correlated strongly with RMDglobal1 mutation risk redistribution (R=-0.49) (Fig 3b). Indeed, when contrasting the RT profiles for the top-ranked RT_PC5 and bottom-ranked TCGA RT_PC5 tumors, we observed that the stronger local RT differences overlap with the RMDglobal1-relevant windows i.e. those where

mutation rate changes notably (shown for 3 example cancer types in (Fig 3c). The next best correlation of the RMDglobal1 spectrum with tumor RT-PCs was with PC6 (R=0.35), while the other RT-PCs up to PC10 did not correlate above R=0.2.

Given that RT changes were highly predictive of RMDglobal1 mutagenesis (slightly ahead of heterochromatin mark and DHS changes; Fig 2a, Fig 3a) we asked if RMDglobal1 mutation risk uniquely associates with RT, rather than with other large-scale genomic features of open chromatin. Upon performing a test using regression on TCGA-RT_PC5 and ENCODE-RT_PC3 jointly with the relevant PCs for DHS, H3K27me3, H3K9me3 and Hi-C compartments, we find that not only RT but multiple other features are independently predictive of RMDglobal1 mutagenesis (with exception of H3K27me3, which can be supplanted by other features; Fig S11d).

To "orient" these PCs from the RT and the epigenetic analyses above (Fig 2c, Fig 3b), we considered the direction of the correlation with RMDglobal1 mutagenesis. The samples with proliferative gene expression are H3K9me3-PC3 positive after orienting, and also later-replicating in the RMDglobal1 top-ranked domains and have higher H3K27me3 and lower DHS. In summary, the chromosomal domains with highest RMDglobal1 mutation rate weights become later-replicating and heterochromatinized in cells exhibiting a stronger signal of proliferation, mirrored in an increase of relative mutation rates in these domains.

Cell proliferation-associated RT shifts in tumors are mirrored in mutation risk redistribution across domains

To understand the biology underlying this RT-PC5 that is relevant to RMDglobal1 mutagenesis, we asked how gene expression changes between the TCGA tumors with high values of a RT-PC versus tumors with low values. As with heterochromatin analysis, above, considering the Hallmarks gene sets revealed that both the RT-TCGA_PC5 and the correlated RT-ENCODE_PC3 strongly associate with gene expression of E2F target, MYC target and G2M checkpoint genes (all FDR<0.1%, Fig 3d); this was replicated independently in TCGA (tumor) gene expression data and in ENCODE (majority noncancerous) gene expression. Thus, this RT switching pattern represents an RT program characteristic of tumors bearing transcriptional programs of rapid cell cycling.

Additionally, we considered the "recurrent heterogeneous programs" (RHP) gene sets, representing gene expression programs variable between individual cells, and that were recurrently observed in different cancer cell lines [16]. Again, the TCGA RT-PC5 correlates strongly with gene expression of cell cycle genes [16] (there are two such RHP gene sets, the G2/M and G1/S genes, correlated at GSEA FDR=5.2·10-21 and 1.2·10-10, respectively) (Fig S11e), while the expression of other RHP gene sets correlate less well with RT-PC5 (next strongest p=5e-5)..

To more directly support the association of the RMDglobal1 mutation redistribution with cell cycle gene expression, we next considered RNA-Seq gene expression for tumors whose WGS we analyzed for mutation patterns (subset of Hartwig Medical Foundation, where RNA-Seq was available). Expression of E2F target genes and of G2M checkpoint genes in tumors significantly associated with their RMDglobal1 status (at FDR<1%; Fig 3d), and expression of

MYC target genes had a positive trend (FDR=31%). Prominent examples of genes linked to cell division and with potential uses as cell proliferation markers -- such as TK1, TYMS and RRM2 nucleotide metabolism genes and the PLK1, CCNA2 and CDK1 regulator genes -- were associated with mutagenesis-relevant RT programs, with these associations replicating in TCGA tumors and in ENCODE tissues (Fig. 3e) and also independently validated in their association with RMDglobal1 mutagenesis pattern in our set of tumors with WGS (Fig. 3f).

Overall, this data converges onto RMDglobal1 regional mutability signature reflecting the global RT program changes, as associated with a higher expression of proliferation-characteristic genes, and thus likely altered cell cycle controls across different tumors. The chromosomal domains with higher mutation rate changes in RMDglobal1 are those domains that undergo changes in RT in tumor samples with more proliferative-like transcriptomes, compared to tumor samples with less proliferative-like transcriptomes.

The spatial chromatin compartments that are prone to remodeling associate with the mutation risk changes

We further asked what characterizes these chromosomal regions where chromatin is more malleable i.e. more easily affected by the RT/chromatin remodeling identified herein (and which mirrors the RMDglobal1 mutation rates). To this end, we analyzed data from diverse genomic assays of chromatin-related features from various prior studies (listed in (Table S3)) that reported correlations with RT. We compared the regional density of each epigenomic feature with our RMDglobal1 spectrum window weights (Table S3). Consistently with the chromatin and RT restructuring analysis above, we noted strong correlations with Hi-C subcompartments (Fig 3g), here inferred from chromatin interactions measured at fine resolution (25 kb) [43], thus allowing a more fine-grained view of the standard A and B compartments. In particular, the B1 inactive subcompartment, rather than B2 and B3 heterochromatin, was most associated with RMDglobal1. The B1 subcompartment replicates during middle S phase, and correlates positively with the Polycomb H3K27me3 mark but negatively with H3K36me3 suggesting that it represents facultative heterochromatin [43]. Next, we observed a correlation with two SPIN states (Fig 3h), the intranuclear territories derived by integrating nuclear compartment mapping assay data with chromatin interaction data [44]. RMDglobal1 signature regions are enriched in the two “Interior repressed” SPIN states [44], marking inactive regions that are (unlike other, typical heterochromatic regions) located centrally in the nucleus, rather than lamina-associated) [44]. Additionally RMDglobal1 windows are enriched in subtelomeric parts of chromosomes (Fig 3j).

In addition to RMDglobal1 mutagenesis from cancers, also the chromatin and RT restructuring programs that we identified (Fig 2, H3K9me3_PC3, DHS_PC3, H3K27me3_PC2; Fig 3, predRT-TCGA_PC5 and predRT-ENCODE_PC3) were notably enriched in the same nuclear territories: B1 subcompartment, “interior repressed” SPIN state, and subtelomeric regions (Fig S12). This suggests that these nuclear territories contain chromatin prone to restructuring that is associated with expression of cell proliferation genes (Fig 2, Fig 3).

Next, we considered associations with how spatial chromatin interactions change during evolution from normal to cancerous cells via a whole-genome duplication (WGD)

intermediate [45]. We found a correspondence between the “CORES” regions [45] i.e. domains that change chromatin conformation upon WGD, and our RMDglobal1 mutation redistribution domains (Fig 3j) and also the chromatin restructuring PC signatures (Fig S12).

In summary, the genome domains undergoing mutation rate change as per RMDglobal1 signature are enriched in the B1 facultative heterochromatin subcompartment, in the nuclear interior located repressed chromatin, and in the CORES Hi-C domains that switch chromatin state during cancerous WGD. Considered together, this analysis suggests that certain heterochromatin compartments may be intrinsically more malleable, undergoing remodeling that determines mutation rates.

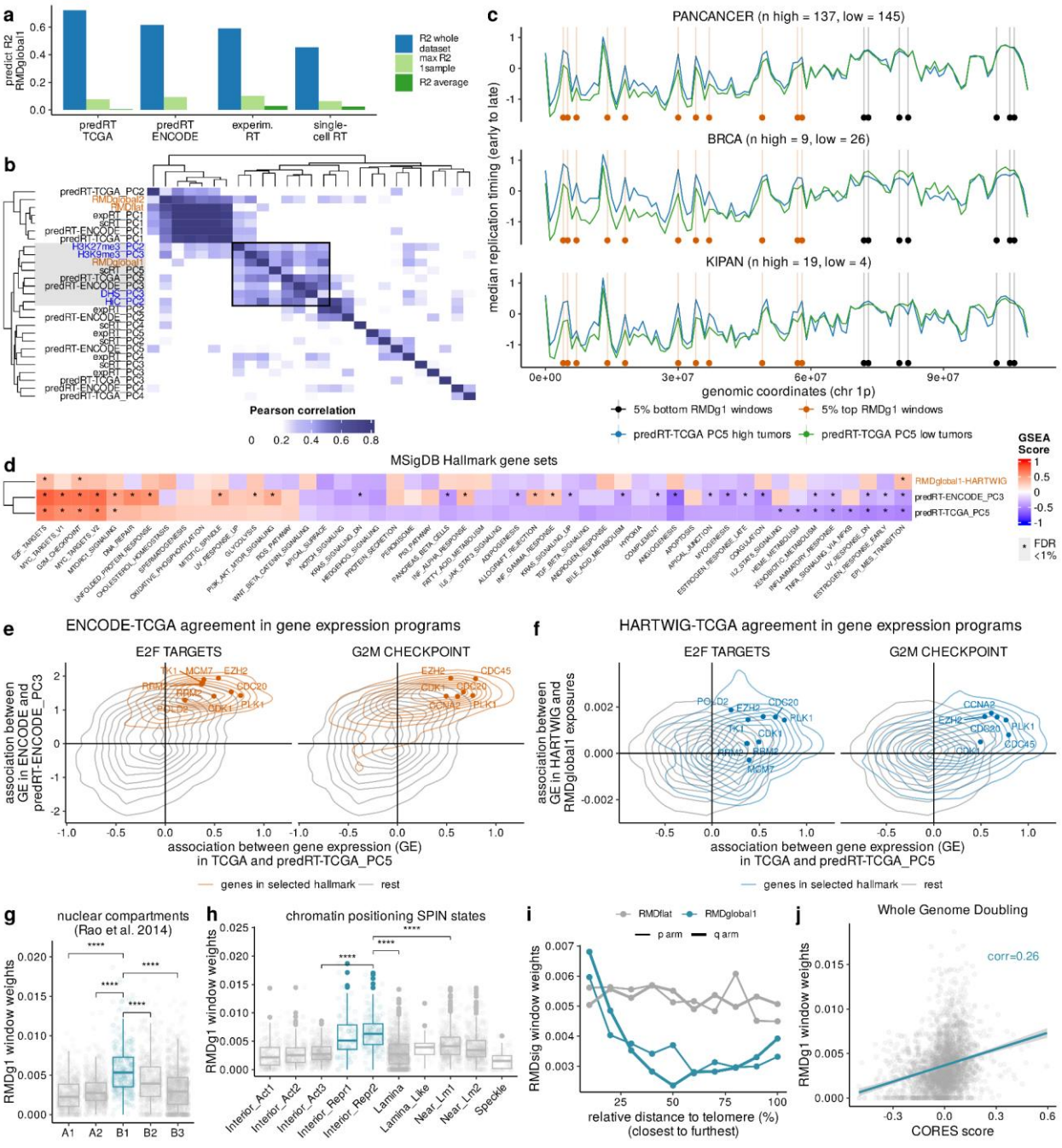


Figure 3. The RMDglobal1 mutation risk redistribution is linked with RT program remodeling in cancers. a) Modelling the RMDglobal1 spectrum (2540 window weights) using genome-wide profiles of various RT features (x axis) using either the whole dataset (multiple samples) jointly, or selecting the best-predicting inidividual sample (maximum R2) in the dataset, or using the average values across the samples. Dataset described in Table S2. **b)** Extracting RT remodelling signatures, with one PCA performed for each RT dataset. Correlations shown between the RT PCs, the 3 RMDglobal mutagenesis signatures, and the 4 relevant chromatin PCs from Fig 2 (marked in blue). The square denotes the cluster of RMDglobal1-associated RT and chromatin PCs. **c)** Median RT profiles (inferred from ATAC-Seq) across the tumor samples with high versus low predRT-TCGA_PC5 signature scores, in pan-cancer and 2 selected cancer types. Windows with top 5% and bottom 5% RMDglobal1 mutagenesis weights are marked with vertical lines. **d)** Gene expression associated with RT PCs (in TCGA and in ENCODE), as well as with RMDglobal1 (in Hartwig Medical Foundation). Gene set enrichment analysis (GSEA) scores for the MSigDB hallmark gene set for an ordered list of genes, derived by associating the 2 RT-PCs and the RMDglobal1 with expression of each gene separately (regression coefficient for the gene variable is used as effect size. **e)** Replicated associations of gene expression in ENCODE and the RT signature predRT-ENCODE_PC3 (y-axis) and gene expression in TCGA tumors and RT signature predRT-TCGA_PC5 (x-axis). The distribution for the genes in two significant proliferation-associated hallmark sets (E2F targets or G2M checkpoint genes) are shown in orange and rest-of-genes in gray; example genes are shown as individual points. **f)** Associations of gene expression with a RT remodeling signature in TCGA (x axis, same as panel e), but here replicated in association of gene expression with the RMDglobal1 mutation risk redistribution in the Hartwig Medical Foundation WGS (y-axis) **g, h)** RMDglobal1 spectrum (window weights for n=2540 windows) across Hi-C nuclear subcompartments (g) and SPIN nuclear positioning states (h); **** denotes p<=0.0001 by Wilcoxon test. **i)** RMDglobal1 and RMDflat signature window weights, stratified by distance to telomeres into 10 bins. Mean value across all chromosomes shown, separately for p and q arms. **j)** Correlation of RMDglobal1 spectrum with the CORES score describing Hi-C alterations that a domain undergoes during a whole genome doubling (WGD) phenomena [45].

RMDglobal1 mutation risk redistribution signature associates with *RB1* pathway disruption

We further hypothesized that certain genetic alterations may drive the changes in tumoral RT that we found linked with cell cycling gene expression, and with the RMDglobal1 mutation risk redistribution. We thus performed an association analysis, detecting somatic copy number alteration (CNA) events and deleterious point mutations that are significantly associated with RMDglobal1 mutation risk exposure (while adjusting for cancer type and for confounding between linked neighboring CNAs; qq plots in Fig S14a; Methods for details). Here, we considered a set of 1543 chromatin modifier genes, cell cycle genes, DNA replication and repair genes and cancer genes, compared against a background of 1000 randomly chosen control genes (Methods).

For CNA, we found a strong positive association of RMDglobal1 with deletions of the *RB1* tumor suppressor that has important roles in cell cycle control and also in chromatin organization (FDR=0.05%, and better p-value than all 1000 control genes) (Fig 4abc, S13a). Because CNA often affects large chromosomal segments, we also tested associations with *RB1* neighboring genes (Fig 4d), noting that *RB1* is at the CNA frequency peak (by the mean estimated copy-number across tumors), meaning it is the likely causal gene in the CNA segment. Strength of RMDglobal1 association with *RB1 alterations* is gene dosage dependent

(Fig S13b), solidifying the causal link of *RB1* loss with mutation rate redistribution. Further, the effect of *RB1* point mutations shows a trend in the same direction as the *RB1* deletions (*RB1* mutations are rarer, thus this trend is nonsignificant) (Fig S13c). As independent supporting evidence, we identified deletions in *CDK6*, a negative regulator upstream of *RB1*, as the CNA event negatively associated with RMDglobal1 with the strongest p-value, exceeding any of the control genes considered (Fig 4b).

The *RB1* alterations were anticipated to associate with certain gene expression patterns; we tested this in two tumor datasets: TCGA (with WES and RNA-Seq data), and Hartwig Medical Foundation (tumor WGS where we inferred RMDglobal1 signature; some have matched RNA-Seq data). In both, *RB1* deletions associated with higher expression of E2F target genes (as expected from pRb function in inhibiting the E2F transcription factors) and G2M checkpoint genes and MYC target genes (GSEA all FDR<1%; Fig 4i; additionally the mitotic spindle genes and DNA repair genes were upregulated here). These gene expression signatures of *RB1* deletion are fully consistent with gene expression signatures linked with RT and heterochromatin restructuring across tumors, tissues and cells (e.g. H3K9me3_PC3 or predRT-TCGA_PC5 signatures, Fig 2e, 3d). This solidifies the links of the widespread chromatin restructuring signatures with cell cycle-associated transcriptomic programs that resemble those resulting from *RB1* deletions.

In addition to its effects on cell cycle regulation, RB1 has additional roles in chromatin organization [23,46–48]. In specific, *RB1* deletions change heterochromatin marks H3K9me3 and H3K27me3, affecting more the regions enriched at subtelomeres and this associates with propensity to acquiring DNA damage in those regions[23]. We found these same two heterochromatin histone marks more highly correlated to RMDglobal1 than other tested marks (Fig 2a), and interestingly the RMDglobal1 spectrum window weights were also strongly enriched approximately 25% of chromosome arm length, proximal to telomere (Fig 3i).

Prompted by the above, we asked if the location of heterochromatin restructuring upon *RB1* disruption [23], matches the locations of the RMDglobal1 mutation risk changes. Indeed, the overlap of RMDglobal1-high risk domains in cancer genomes with the H3K9me3-switching regions in *RB1* wild-type *versus* isogenic *RB1* k.o. cells [23] is substantial (OR=3.25, 95% C.I [2.3-4.5], $p<10^{-13}$ for overlap in the top-10% regions), however there is no enrichment with H3K27me3-switching regions (Fig 4e). . Next, to elucidate the mechanism, we asked if the RMDglobal1 mutation risk results, at least in part, from the increase in DNA damage in these heterochromatin-switching regions upon RB1 loss (measured as CPD lesions after UV exposure [23]). The overlap between the top-10% RMDglobal1 mutation risk domains and the top-10% DNA damage-sensitized domains (upon RB1 KO) was strong (OR=5.05, $p<10^{-29}$), and similarly the overlap in bottom-10% RMDglobal1 and the top-10% damage-protected domains (OR=8.17, $p<10^{-54}$; Fig. 4f). The overlap with UV damage-sensitized domains was seen in regional mutation risk both in skin cancers, which are UV mutagenized, but similarly so in lung cancers, which are tobacco smoking chemical-mutagenized (Fig S14c), suggesting the link extends to multiple types of DNA damage. We further asked if the telomere-proximal enrichment of RMDglobal1 mutation risk (Fig 3i) associates with the RB1-associated chromatin remodeling. Indeed, only the chromosome arms exhibiting a DNA damage-increase upon RB1 KO [23], but not others, show a clear gradual increase in RMDglobal1 mutation risk towards the telomere; the affected regions span approximately 10 Mb telomere-proximal DNA on average (Fig 4g). Overall, this overlap of heterochromatin remodelling loci as well as DNA

damage sensitive loci upon *RB1* loss-of-function [23] with the RMDglobal1 mutation risk loci in tumors further implicates *RB1* activity in shaping the somatic mutation rate landscape.

In addition to the CNA analysis above (that identified RB1 and CDK6), we also tested associations between the presence of deleterious somatic point mutations in cancer genes and chromatin and DNA repair genes, with the 'exposure' to the RMDglobal1 mutagenic pattern. Here, we found the *KRAS* mutation to positively associate with RMDglobal1, at FDR=1% (Fig 4h, Fig S14d-f), and this is observed consistently across individual cancer types (Fig S14c) and significantly in colon, uterus and bladder (see Fig S14 for comment on lung adenocarcinoma about confounding by tobacco smoking signatures). Of note, the *KRAS* gene acts downstream of *RB1* loss-of-function with *RB1* in developmental and in tumor mouse phenotypes [49,50]. Consistently, *KRAS* mutation and *RB1* loss (either deletion or mutation) are mutually exclusive in our tumor dataset (chi-square $p < 2.2e-16$), supporting that the driver alterations in *RB1* and *KRAS* may converge onto the same mutation rate redistribution phenotype, the RMDglobal1. Consistently with the RMDglobal1 pattern resulting from certain driver gene alterations, we find the subclonal, later-occurring mutations are enriched in the RMDglobal1 pattern compared to the clonal mutations in various cancer types (Fig. S15).

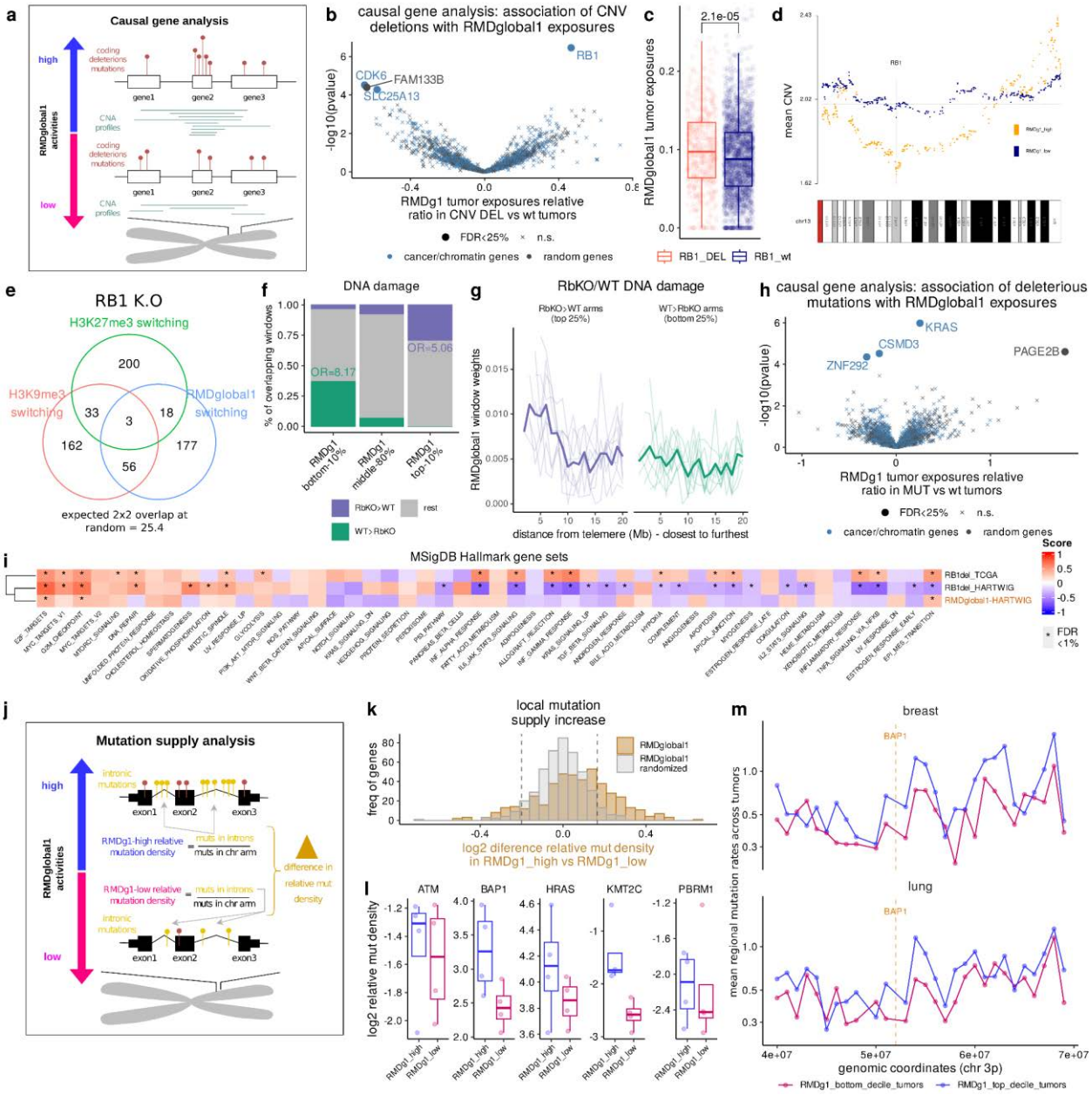


Figure 4. Genetic alterations associated with the activity of RMDglobal1 mutation redistribution signature. Analyses shown either aim to identify possible causal alterations underlying RMDglobal1 (panels a – g), or the downstream consequences of RMDglobal1 mutation risk redistribution to mutation supply towards cancer genes (panels h – k). a) A Schematic of the causal gene analysis in following panels b-d (Methods). b) Associations between somatic CNA deletions and a higher RMDglobal1 exposures in a pan-cancer analysis (n = 2875), adjusting for cancer type and for large-scale CNA patterns (Methods). n=1543 cancer genes and chromatin-related genes are shown (dots), as well as a control set of n=1000 randomly chosen genes (crosses). **c**) Differences in RMDglobal1 exposures between RB1 deletion (-1 or -2 CNA state) and the RB1 wild-type; separated by cancer type in Fig S13. n=1662 tumors. P-values from Mann-Whitney test, two-tailed. **d)** Mean local CNV profile in groups of tumors, grouped by RMDglobal1- high (n=81) and low exposure (n=93), showing of the segment of chromosome 13 containing the gene *RB1*. Each dot represents one gene. **e-f)** Overlap between the RMDglobal1 redistribution-affected domains, and the domains affected by heterochromatin remodeling (**e**) or DNA damage redistribution (**f**) in an isogenic pair of RB1 k.o. cell lines [23]. **e**) The domains where H3K9me3 is altered upon RB1

k.o., but not those where H3K27me3 is altered, overlap with the RMDglobal1 domains. **f**) The top 10% regions where DNA damage (by UV) increases upon RB1 k.o. overlap with the regions where RMDglobal1 mutagenesis increases (top 10%); additionally, the converse association also holds. **g**) RMDglobal1 window weights in the proximity of telomere separately by chromosome arms with a RB KO > WT change versus a WT > RB KO change. h) Associations between deleterious SNV and indel mutations in the same sets of genes as in panel **b**, and the RMDglobal1-high versus RMDglobal1-low activity of tumor samples (n = 2785), in a pan-cancer analysis (Methods). **i**) Gene set enrichment Analysis (GSEA) for the MSigDB hallmark gene sets, considering genes ordered by their association with RB1 deletions (in Hartwig Medical Foundation and in TCGA studies), and ordered by their association with the RMDglobal1 mutation risk redistribution itself.. The positive GSEA scores means that the genes from that gene set are overall higher expressed in the *RB1* deleted tumors compared to the *RB1*wild-type tumors. **j)** A schematic of the analysis in following panels k-m (Methods). **k)** Distribution for the difference in intronic mutation density, a measure of differential mutation supply, shown for 460 cancer genes, comparing between RMDglobal1-high and low tumors. Shown separately using the actual values of RMDglobal1 to stratify tumors, and as a baseline using the randomized RMDglobal1**.** Vertical lines show 5th and 95th percentile of the randomized distribution, used as cutoffs for significance. **l)** Mutation density for RMDglobal1-high versus low tumor samples (top tertile versus bottom tertile) for 5 example genes (common cancer drivers with the highest effect size [mutation supply increase]); dots are cancer types. **m)** RMD profile in a region on chromosome 3p, showing the mean RMD across the RMDglobal1-high versus low tumor groups (here, top and bottom decile by RMDglobal1), for two example cancer types. Vertical lines mark the position for the *BAP1* tumor suppressor gene.

Mutation supply towards cancer genes is altered by activity of global RMD risk redistribution signatures

Since RMDglobal1 captures a redistribution of mutation rates genome-wide, it follows that this will affect the local supply of mutations towards loci harboring some cancer genes. To test this, we considered 460 cancer genes and the intronic mutation density thereof (to avoid effects of selection, as we are interested in measuring the rate of mutation accrual), further normalizing to the chromosome arm-level mutation burden thus measuring mutation risk redistribution across regions (see Methods). We tested whether there is a difference between tumors with a high RMDglobal1 activity (top tertile) versus low RMDglobal1 activity (bottom tertile) (Fig 4jk). When compared to a randomized baseline (5th and 95th percentile of the random distribution used as cutoff), the mutation supply was significantly increased towards 28% of the 460 cancer genes in RMDglobal1-high tumors (Fig 4k), and in comparison mutation supply decreased towards 11% of cancer genes. The genes that increase mutation rates do so on average by 1.21-fold in the RMDglobal1-high tertile tumors, with bigger increases for some genes (shown for 5 example genes Fig 4l). For instance the median mutation supply for *BAP1* increases by 1.78-fold, for the *KMT2C* by 1.79-fold, and for *ATM* by 1.18-fold. Importantly, the mutation supply is here measured only towards introns, and thus presumably does not reflect selected mutations in *ATM*, *BAP1*, *KMT2C* etc. but instead only the relative rate of mutations appearing in the given region, which changes upon RMDglobal1 activity.

Next, we similarly considered the redistribution effects of the RMDflat signature, associated with DNA repair deficiencies (see above), and which reflects a strong increase in relative mutation rates in early replicating, euchromatic regions [28,33], and consequently

reduced correlation of domain mutation risk with RT (Fig S6d). These early RT regions also have a higher gene density. Indeed, RMDflat commonly affects the mutation supply to many cancer driver genes, with 75% of the 460 tested cancer genes [51] exhibiting an increased supply comparing RMDflat-low to RMDflat-high tumors ((Fig S6e)). The converse case was rare: a few cancer genes decreased in mutation supply in RMDflat-high tumors - 9% genes were below the 5th percentile of the random distribution. We considered 5 example common driver genes, for which mutation supply was increased 1.8-2.5 fold by RMDflat (Fig S6f). For instance the *ARID1A* tumor suppressor gene, located in a lowly-mutated region in chromosome 1p, has its mutation supply increased 1.8-fold, 2.1-fold and 2.4-fold in MSI (i.e. MMR-deficient), HR-deficient and APOBEC tumors (all cases of RMDflat-high), respectively, compared to the *ARID1A* baseline mutation supply in tumors without DNA repair deficiencies (Fig S6fg). Similarly, the *BRAF* oncogene (where driver mutations are highly enriched in MSI compared to MSS colorectal tumors [52]) has a considerably increased mutation supply in the RMDflat-high tumors (Fig S6fg).

TP53 pathway disruption reduces relative mutation supply towards late replicating regions

In addition to RMDglobal1 and RMDflat, there is a third, commonly occurring mutation rate redistribution signature observed across 21% of tumor genomes across multiple tissues (Fig 1f, Fig S5), the RMDglobal2. Its 1 Mb domain mutation rates do follow a distribution resembling the canonical RMD landscape, i.e. increasing mutation density in later RT overall, however except for late RT windows. These acquire fewer mutations than expected in the RMDglobal2 pattern (Fig 5ab). As a consequence, mutation rates increase approximately linearly with RT bins in tumors with high RMDglobal2, while in tumors with a low RMDglobal2 exposure the RT relationship to mutation rates is better described by a quadratic function (Fig 5c, Fig S16a). In other words, the RMDglobal2 redistribution "linearizes" the association of RMD to RT, by suppressing the more prominent peaks in regional mutation rates.

We aimed to identify the causal driver event behind this RMDglobal2 redistribution of mutation risk away from the latest-replicating DNA domains. As we did for RMDglobal1 above, we proceeded to test for associations of RMDglobal2-high (top tertile) versus low (bottom tertile) tumor samples, with CNAs and deleterious (coding) mutations in cancer driver, DNA repair and chromatin modifier genes. Strikingly, we found *TP53* mutation to be uniquely strongly associated with RMDglobal2 signature (FDR = $9 \cdot 10^{-10}$; next strongest positive association is PKHD1 FDR=$2.2 \cdot 10^{-6}$) (Fig 5d). As supporting evidence, we found that *TP53* deletions were also positively associated (Fig 5e) these trends were observed consistently across many cancer types (Fig S16b). Independently, the amplifications in known oncogenes that phenocopy TP53 loss (*MDM2*, *MDM4* and *PPM1D*) are all also positively associated with the activity of the RMDglobal2 mutation redistribution signature (Fig 5e, Fig S16b). This rules out that the *TP53* driver mutation occurrence is the consequence of the RMDglobal2 redistribution changing local mutation supply, but rather provides evidence for a causal effect of *TP53* pathway inactivation in this RMDglobal2 mutation risk redistribution.

Since *TP53* mutations were reported to be associated with increased burdens of CNA events [53], we tested whether RMDglobal2 RMD signature could be due to confounding from a multiplicity of focal CNA events, which might modify apparent local mutation rates. However,

there is only a weak correlation between the CNA burden and RMDglobal2 signature levels upon stratifying for *TP53* status (R<=0.11), suggesting that RMDglobal2 largely does not reflect changes in local DNA copy number but rather results from TP53 loss-of-function (Fig S17). As with RMDglobal1, also RMDglobal2 is enriched in subclonal, late-occurring mutations (Fig S15), consistent with it being triggered by *TP53* alterations, which are unlikely to be present in noncancerous cells while they accumulate mutations. The activity of the RMDglobal2 signature, inferred from SNV mutations, is also mirrored in the regional pattern of indel mutations and also of rearrangements (Fig S7cf).

Overall, the above convergent genetic associations strongly implicated the deficiencies in TP53 pathway in the RMDglobal2 mutation redistribution, similarly as the deficiencies in RB1 pathway -- another cancer gene controlling the cell cycle -- were implicated in the RMDglobal1 mutation redistribution (Fig 4). Interestingly, the RMDglobal2 mutation risk redistribution was strongly negatively correlated with the activity of the clock-like trinucleotide mutational signature SBS1 (C>T changes at CpG dinucleotides), consistently across cancer types (Fig S18ab), and there was also a positive association with the SBS93 trinucleotide signature (Fig S18ac) of unknown aetiology. We did not identify associations of similar magnitude and consistency across cancer types between the RMDglobal1 redistribution and trinucleotide SBS signatures (Fig S18a; some tentative associations are in Fig S18de).

Changes to local mutation supply because of redistribution can result in artefactual epistasis-like phenomena

RMDglobal2 signature describes mutation risk variation in certain genome regions, which may affect mutation supply to genes residing therein. We tested whether there is a difference in mutation rate in the cancer genes for RMDglobal2-high (top tertile) versus low (bottom tertile) tumor samples (Fig 5f).; again we considered only intronic regions, to measure (neutral) mutation supply and not selection. When compared to randomized data (5th percentile), 26% of cancer genes exhibited a decreased relative mutation supply in high-RMDglobal2 tumors (which are often TP53 mutant). In contrast, only 6% genes exhibit an increased mutation supply with high RMDglobal2 (Fig 5f). As an example, we show the coding mutation density in *ARID1A* and *GATA3*, two genes which decreased in local mutation supply due to activity of RMDglobal2 redistribution (as above, measured using intronic rates; the decrease implies the genes are below the 5th percentile of the randomized distribution) (Fig 5g).

We hypothesized that apparent genetic interactions involving ARID1A, GATA3 and similarly other RMDglobal2-affected genes -- for example mutual exclusivity with *TP53 mutations that are drivers of RMDglobal2* -- might arise due to redistribution of local mutation supply to genes. We thus considered 13 genes bearing coding mutations mutually exclusive with *TP53* mutations [54] , hypothesizing that this might in fact be due to altered local mutation supply via RMDglobal2 in *TP53*-mutant cancers, rather than a selected genetic interaction. Indeed we found that nearly half (6/13) of these genes were below the 5th percentile of the random distribution of local mutation rates. This supports the hypothesis that what appears to be epistatic interaction is in fact commonly just a diversion of the mutational supply away from the 6 genes by a TP53-dependant mutation risk redistribution (Fig 5fh). Upon inspection of the raw RMD profiles for RMDglobal2-high tumors, for several cancer types, we noted a difference in the mutation supply towards the locus where *ARID1A* resides (Fig 5i). Overall, this illustrates

how a global redistribution of mutation risk, here mediated by *TP53* loss, can create apparent genetic interactions that may not indicate selection on functional effects. Thus, regional mutation risk redistribution signatures, which can vary considerably between tumors, should be explicitly controlled for in statistical studies of driver gene selection and of epistasis in cancer genomes.

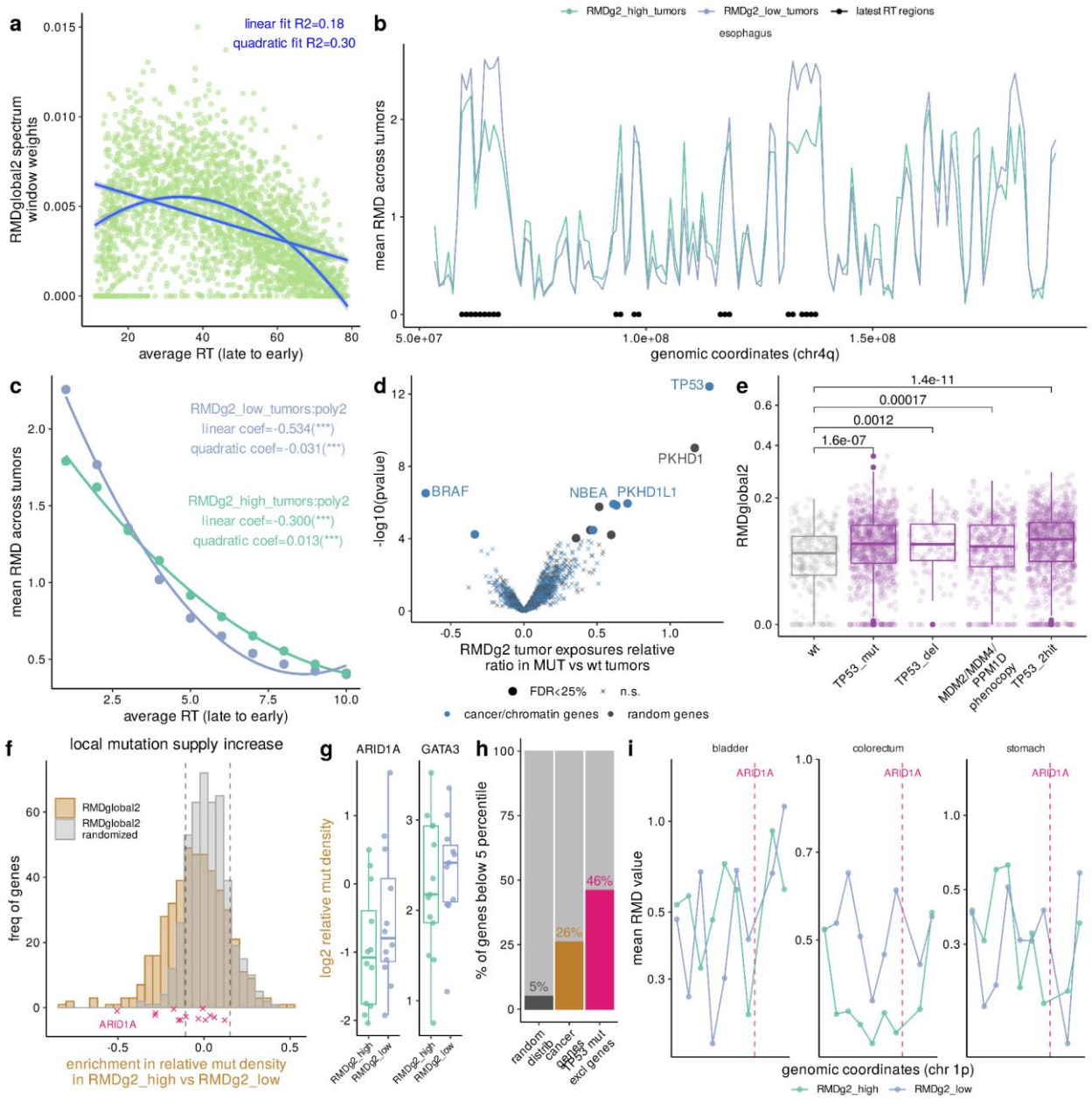


Figure 5. A TP53 loss-associated mechanism underlies the RMDglobal2 mutation risk redistribution signature. a) A quadratic association of RMDglobal2 spectrum (window weights, n=2540) with RT. **b)** Mean mutation density profiles on chromosome 4q for the RMDglobal2-high (n=7) versus low (n=129) tumors in esophagus cancer. Latest RT windows are marked with black dots. **c)** A linearization of the mapping between RT and domain

mutation rates by the RMDglobal2 signature. The relative RMD profile across 10 RT bins, showing the mean for tumors that are RMDglobal2-high (exposure > 0.17, n = 30 tumors) versus RMDglobal2-low (exposures < 0.01, n=78 tumors). Cancer types considered were the ones with many tumors with high RMDglobal2: breast, lower-GI, upper-GI and prostate. **d)** Associations between RMDglobal2 exposures in 2785 tumors, and the deleterious SNV mutations in cancer genes and chromatin genes (dots) and a control set of 1000 randomly chosen genes (gray dots); p-values from Z-test on regression coefficient. **e)** RMDglobal2 signature exposures of 2785 tumors stratified by: wild-type for *TP53* (wt), TP53 with 1 mutation (TP53_mut), TP53 with 1 deletion (TP53_del), TP53-loss phenocopy via a CNA gain in *MDM2*, *MDM4* or *PPM1D* oncogenes (TP53_pheno), or *TP53* with any two hits of these alterations (TP53_2hit). **f)** RMDglobal2 alters the mutation supply away from the loci harboring various cancer genes. Distribution of the enrichment in the relative intronic mutation density for 460 cancer genes, comparing between RMDglobal2 high tumors and RMDglobal2 low tumors. Histograms are shown using the actual values of mutation supply difference ("RMDglobal2" histogram), and using randomized values as a baseline ("RMDglobal2 randomized" histogram). Position of the genes known to have mutually exclusive mutations with TP53 mutations are marked with crosses (Methods) **g)** Percentage of genes with significantly lowered mutation supply (below 5th percentile of a random distribution) upon RMDglobal2 redistribution, shown for the random genes, cancer genes, and the known TP53 mutually exclusively mutated genes. **h)** Log2 relative intronic mutation density (normalized to flanking DNA in same chromosome arm, see Fig 4j), estimating mutation supply, for RMDglobal2-high versus RMDglobal2-low tumors for two example genes known to be TP53 mutually exclusive. Each dot is a cancer type. **i)** Mean RMD profile across the RMDglobal2-high versus low tumor groups in a region of chromosome 1p harboring the *ARID1A* gene.

Discussion and concluding remarks

Mutation rates are lower in early-replicating, euchromatic DNA compared to late-replicating heterochromatic DNA [8,55–58]. If either RT or heterochromatin (or both) are causal to mutation rates, which is likely the case and is often mediated by differential DNA repair [1,2,6,59,60] and/or differential DNA damage [25,61] then local changes in RT or in heterochromatin status would change local mutation risk. Our study provides robust evidence this is commonly the case: the regions in the human genome with switching RT and switching heterochromatin status across many biological samples overlap the regions where somatic mutation rates vary across tumors (as summarized in our RMDglobal1 mutation risk redistribution signature). These regions often correspond to Polycomb-marked facultative heterochromatin residing in the B1 nuclear subcompartment [43] , which is inactive but centrally located in the nucleus, and predisposed to be more malleable across cell types [62].

This locally variable RT that we inferred across ~400 tumors was associated with gene expression of cell proliferation genes esp. E2F target genes, and locally variable heterochromatin was associated with *RB1* disruption in an experimental model [23] , plausibly reflecting various molecular consequences of accelerated and/or dysregulated cell cycles on RT and heterochromatin organization. Consistently, the changes in regional mutation rates – observed at the chromosomal domains that are prone to switch RT/heterochromatin status – are strongly associated with somatic alterations of *RB1* and *CDK6* genes in tumors. Additionally to RB1, we implicate disruptions in another tumor suppressor pathway -- including TP53 and associated genes -- in altering local somatic mutation rates in tumors, in this case most prominently in the latest-replicating heterochromatic domains (as per RMDGlobal2 signature). Our data converges onto a mechanism where altered cell cycling, commonly

occurring via somatic alterations of tumor suppressor genes RB1 and TP53 and functionally related driver genes, can trigger RT changes and large-scale heterochromatin restructuring, which in turn alters the local distribution of mutation rates. This affects differential mutation supply to disease genes, altering the likelihood of acquiring pathogenic mutations and steering the course of somatic evolution towards particular outcomes.

Methods

WGS mutation data collection and processing

We collected whole genome sequencing (WGS) somatic mutations from 6 different cohorts (Table S1). First, we downloaded 1950 WGS somatic single-nucleotide variants (SNVs) from the Pan-cancer Analysis of Whole Genomes (PCAWG) study at the International Cancer Genome Consortium [63] Data portal (https://dcc.icgc.org/pcawg). Second, we obtained 4823 WGS somatic SNVs from the Hartwig Medical Foundation (HMF) project [64] (https://www.hartwigmedicalfoundation.nl/en/). Third, we downloaded 570 WGS somatic SNVs from the Personal Oncogenomics (POG) project [65] from BC Cancer (https://www.bcgsc.ca/downloads/POG570/). Fourth, we obtained 724 WGS somatic SNVs from The Cancer Genome Atlas (TCGA) study as in [9]; we used QSS_NT>=12 mutation calling threshold in this study.

Finally, we downloaded bam files for 781 WGS samples from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) project [66,67] and bam files for 758 tumor samples from the MMRF COMMPASS project [68] from the GDC data portal (https://portal.gdc.cancer.gov/). Somatic variants were called using Illumina's Strelka2 caller [69], using the variant calling threshold SomaticEVS >=6. Additionally, for these samples we performed a liftOver from GRCh38 to the hg19 reference genome.

We collected the samples' metadata (MSI status, purity, ploidy, smoking history, gender) from data portals and/or from the supplementary data of the corresponding publications. Additionally, we harmonized the cancer type labels across cohorts. Here, since lung tumors in HMF data are not divided into lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) types, we used a CNA-based classifier to tentatively annotate them in the HMF data. We downloaded copy number alteration data from HMF and TCGA for lung tumor samples and adjusted for batch effects between cohorts using ComBat as described in our previous work [70]. We trained a Ridge regression model with TCGA data to discriminate between LUSC and LUAD and applied the model to predict LUSC/LUAD in the HMF lung samples. We did not assign a label to samples with an ambiguous prediction score between 0.4 and 0.6.

Similarly, since POG breast cancer (BRCA) samples are not divided into subtypes (luminal A, luminal B, HER2+ and triple-negative) we used a gene expression classifier to annotate them. We downloaded gene expression data for TCGA and POG breast tumors and adjusted the data for batch effect using ComBat as previously described [70]. We trained a Ridge regression model with TCGA data to discriminate between the breast cancer subtypes (one-versus-rest) and applied the model to the POG breast samples to assign them to a

subtype. We did not assign 23 samples that are predicted as two subtypes and 8 that are not predicted as any subtype.

Defining windows and filtered regions

We divided the hg19 assembly of the human genome into 1 Mb-sized windows. These divisions are performed on each chromosome arm separately. To minimize errors due to misalignment of short reads, we masked out all regions in the genome defined in the 'CRG Alignability 75' track [71] with alignability <1.0. In addition, we removed the regions that are unstable when converting between GRCh37 and GRCh38 [72] and the ENCODE blacklist of problematic regions of the genome [73].

Additionally, to minimize the effect of known sources of mutation rates variability at the sub-gene scale we removed CTCF binding site regions (downloaded from the Table Browser), ETS binding regions (downloaded from http://funseq2.gersteinlab.org/data/2.1.0) and APOBEC mutagenized hairpins downloaded from [74]. Finally, we removed all coding exon regions (+-2nts, downloaded from the Table Browser) to minimize the effect of selection on mutation rates.

Matching trinucleotide composition across megabase windows

To minimize the variability in mutational spectra confounding the analyses, we accounted for the trinucleotide composition of each window. For this, we removed trinucleotide positions from the genome in an iterative manner to reduce the difference in trinucleotide composition across windows. We selected 800,000 iterations that reach a tolerance <0.0005 (difference in relative frequency of trinucleotides between the windows). After the matching, we removed all windows that end up with less than 500,000 usable bps. The final number of analyzed windows is 2,540.

Calculating the Regional Mutation Density (RMD) of each window

For our WGS tumor sample set (n=9,606 WGS) we counted the number of mutations in the above-defined windows. We required a minimum number of mutations per sample of 5,876, which corresponds to 3 muts/Mb (total genome = 1,958,707,652 bp). In total, 4221 tumor samples remain, which we use for the downstream analyses.

To calculate the RMD, we normalized the counts of each window by: (i) the nt-at-risk available for analysis in each window and (ii) the sum of mutation densities in each chromosome arm. To control for whole arm copy number alterations.

To calculate the RMD applied to NMF analysis, we first subsample mutations from the few hyper-mutator tumors, to prevent undue influence on overall analysis. We allow a maximum of 20 muts/Mb that is 39,174 muts. If the tumor mutation burden is higher we subsample the mutations to reduce it to that maximum value. Then, as above, we normalized the RMD by: (i) the nt-at-risk in each window [RMD = counts * average_nt_risk / nt_at_risk] and (ii) the sum of mutation density in each chromosome arm [RMD * row_mean_WG / rowMeans by chr arm]. We multiply by the average nucleotides at risk and the mean whole genome to maintain the values range of each sample for the bootstrapping.

Applying NMF to extract RMD signatures

We applied bootstrap resampling (R function UPmultinomial from package sampling) to the RMD scores that we calculated for NMF as above. The result for each tumor sample is a vector of counts with a tumor mutation burden close to the original one but normalized by the nucleotides at risk by window and for the possible chromosome arm copy number alterations (CNA). Then, we applied NMF (R function: nmf) to the bootstrapped RMD matrices, testing different values of the rank parameter (1 to 20), herein referred to as nFact.

We repeated the bootstrapping and NMF 100 times for each nFact. We pooled all the results by nFact and performed a k-medoids clustering (R function pam), with different number-of-clusters k values (1 to 20). We calculated the silhouette index value, a clustering quality score (which here measures, effectively, how reproducible are the NMF solutions across runs), for each clustering to select the best nFact and k values.

Additionally, we also applied the same NMF methodology to each cancer type separately (n = 12 cancer types that had >100 samples available).

Simulated data with ground-truth RMD signatures

For each cancer type, we calculated a vector of RMD values (i.e. regional mutation density mean of all samples from that cancer type) based on observed data, and super-imposed simulated ground-truth signatures onto these cancer type-derived canonical RMD patterns. We generated 9 simulated ground-truth RMD signatures with different characteristics, varying the number of windows affected by the signature (10, 20 or 50% of 2540 windows total) and the fold-enrichment of mutations in those windows (x2, x3 or x5) over the RMD window value in the canonical RMD pattern for that tissue.

In particular, we tested 9 different scenarios, varying the signature contribution to the total mutation burden (10, 20 or 40%) and the number of tumor samples affected by the signature (5, 10 or 20%). We randomly assigned the ground-truth signatures to be super-imposed onto each tumor sample (e.g. sample A will be affected by RMD signature 1 and 3 while sample B will be affected by signature 4). In total, we have simulated genomes for 9 different scenarios (different RMD signature contributions and number of tumor samples affected), each of them containing the 9 simulated ground-truth RMD signatures.

We applied the NMF methodology for the 9 different scenarios independently and obtained NMF signatures. For each case, we selected an NMF nFact and k-medoids clustering k, based on the minimum cluster silhouette index (SI) quality score. To assess the method, we compared the extracted NMF signatures with the ground-truth simulated signatures. In particular, we considered that an extracted NMF signature matches the ground-truth simulated signatures when the cosine similarity is >=0.75 only for that ground-truth simulated signature and < 0.75 for the rest.

Analysis of differential mutation supply towards cancer genes

For 460 cancer genes from the MutPanning list [51] (http://www.cancer-genes.org/), we tested if they are enriched in intronic mutations in tumor samples with high RMDflat, RMDglobal1 or RMDglobal2. An enrichment will mean that there is a higher supply of mutations in the intron

regions of those genes when the RMDsignature is high. For this, we considered the counts of mutations in the intronic regions of the gene, normalized to the number of mutations in the whole chromosome arm, comparing pooled tumor samples with RMD signatures high or low, by tissue. Note that the possibly different number of eligible nucleotides-at-risk in the central window, nor the length of the flanking chromosome arm are relevant in this analysis, because they cancel out when comparing one group of tumor samples (split by the RMD signature) to another group of tumor samples. We binarized the tumor samples by RMDflat, RMDglobal1 and RMDglobal2 by dividing each of them into tertiles, and keeping 1st tertile *versus* 3rd tertile for further analysis. We applied a Poisson regression with the following formula:

Count_gene_intron ~ offset(count_chr_arm) + RMDflat + RMDglobal1 + RMDglobal2 + tissue

where "count" refers to mutation counts. By including the tissue as a variable in the regression, we controlled for possible confounding by cancer type. The log fold-difference in mutation supply between RMD signature high versus low tumor samples is estimated by the regression coefficients for RMDflat, RMDglobal1 and RMDglobal2 variables. As a control, we repeated the exact same analysis but randomizing the tertile assignment for the three RMD signatures prior to the regression.

Association analysis of gene mutations with RMD global signatures

We created a subset of 1543 relevant genes: cancer genes from the MutPanning list [51] and Cancer Gene Census list [75], and furthermore we included genes associated with chromatin and DNA damage [76]. As control, we used a subset of 1000 random genes selected as in [76].

We applied the analysis for two different features: copy number alterations (CNA) and deleterious point mutations. For CNA, we use the CN values by gene, using a score of -2, -1, 0, 1 or 2 for each gene. We considered a gene to be amplified if CNA value was +1 or +2 and deleted if the CNA value was -1 or -2. For deleterious mutations, we selected mutations predicted as moderate or high impact in the Hartwig (HMF) variant calls, (https://github.com/hartwigmedical/hmftools). We binarized the feature into 1 if the sample has the feature (CNA, or deleterious mutations present) or 0 if it has not. We considered CNA deletions and amplifications as two independent features. We binarized RMDflat, RMDglobal1 and RMDglobal2 by dividing each of them in tertiles and comparing tumor samples in 1st tertile *versus* 3rd tertile, by tissue.

We fit a linear model to test whether the binary genetic feature (amplification CNA, deletion CNA or deleterious mutation in a particular gene) can be explained by the RMD signatures activity being high *versus* low (i.e. upper tertile versus lower tertile). We controlled for tissue by including it as covariate. The regression formula was:

genetic_feature ~ RMDflat + RMDglobal1 + RMDglobal2 + tissue

We used the regression coefficients, and p-values (according to the R function "summary") from the variables RMDflat, RMDglobal1 and RMDglobal2 to identify genetic events associated with high levels of each RMD global signatures, suggesting possible RMD signature generating events. In the case of CNAs, to adjust for the linkage between CNA resulting in confounding, we added to the regression the PCs from a PCA on the CNA landscape across all genes. We calculated the lambda (inflation factor) for the p-value

distribution of associations, while including PCs from 1 to 100 to decide the best number of PCs to include so as to minimize lambda. We included the first 55 PCs for the deletion CNA and the first 63 PCs for the amplification CNA association study.

Epigenomic and related data sources

ENCODE data. We downloaded from ENCODE (https://www.encodeproject.org/) all data available for *Homo sapiens* in the genome assembly hg19 for DHS, H3F3A, H3K27me3, H3K4me1, H3K4me3, H3K9ac, H3K9me3, HiC, DNA methylation (WGBS), H2AFZ, H3K27ac, H3K36me3, H3K4me2, H3K79me2, H3K9me2 and H4K20me1 marks. Data is described in Table S2. For each of these features, we downloaded the narrow peaks, calculated their weighted density for each 1Mb window as the width of the peak multiplied by the peak value.

HiC data. We downloaded chromatin domain hierarchies and compartment scores generated by the Calder method from Hi-C data from 114 cell lines [34].

ChromHMM chromatin states. We downloaded the 25 ChromHMM states segmented files ("imputed12marks_segments") for the 129 cell types available from Roadmap epigenomics [77](http://compbio.mit.edu/ChromHMM/). We calculated the density of each state for each 1Mb window as the fraction of the window covered by the chromatin state.

Other epigenomic data. We downloaded RT variability genomic data describing RT heterogeneity [78], Constitutive and Developmental RT domains [79], RT changes upon overexpression of the oncogene *KDM4A* [80], RT signatures of replication stress [81], RT signatures of tissues [42], RT states [82], changes in RT upon *RIF1* knock-out [83] and RT changes due to RT QTLs [84]. In addition, we downloaded data for variability in DNA methylation [19,85], HMD and PMD regions [20], CpG density, gene density, lamina associated domains (LADs), asynchronous replication domains [86], early replicating fragile sites [87], SPIN states [44], A/B subcompartments [43], DHS signatures [88] and H3K27me3 and H3K9me profiles for *RB1* wild-type and knock-out [23]. Data described in Table S3. We calculated the density for each feature for each 1 Mb window, and correlated this with the RMDglobal1 signature windows weights.

Replication timing data sources and generation

We downloaded experimental RT data, from RepliChip or RepliSeq assays, from the Replication Domain database (https://www2.replicationdomain.com/index.php) [79] in multiple human cell types (n = 158 samples). In addition, we predicted RT using the Replicon software [38] from two type datasets: (i) in noncancerous tissues, cultured primary cells and cell lines including cancer and stem cells (n = 597 samples) using the DHS chromatin accessibility data downloaded from ENCODE; and (ii) in human tumors (n = 410 samples, most of them with technical replicates) using ATAC-seq data of TCGA tumors downloaded from [37]. We used the Replicon tool with the default settings.

Gene expression analyses

Gene expression data. For the genomes from the HMF data set, we downloaded gene expression data (as adjusted TPM values) from Hartwig [64], available for a subset of samples

for which we derived the RMD signatures. In total, we had gene expression data for 1534 samples and 18889 protein coding genes. For the genomes from the TCGA data set, we downloaded gene expression data (as TPM values) from the Genomic Data Commons data portal (https://dcc.icgc.org/pcawg) for the same TCGA samples for which we predicted RT. In total, we have gene expression data for 399 overlapping samples and 20092 genes.

For the samples from the ENCODE data set, we downloaded RNAseq expression data (as TPM values) from ENCODE (https://www.encodeproject.org/). We linked the RNAseq experiments with the DHS and chromatin marks by the donor id + the tissue of origin. There are several cases for which we have more than one experiment per donor id + tissue, in those cases we matched at random the replicates from RNAseq with the replicates from the DHS or chromatin marks with the same donor id (without repeating any experiment id).

Gene expression association with the RMDglobal1 and chromatin signatures. For each signature (RMDglobal1_exposures, H3K9me3_PC3, etc) we predicted the signature from the gene expression of 1 gene: signature_x ~ gene1 (TPM). The coefficient from the regression indicates the gene effect (upregulated or downregulated) with respect to the signature and the p-value. We performed this analysis for all genes.

Gene Set Enrichment Analysis (GSEA). We used the regression coefficients for the association with a particular signature to order the genes and applied a GSEA analysis. We consider two gene sets: MSigDB Hallmarks gene set [35] and the Recurrent heterogeneity pathways (RHP) from a single cell study [16].

Clustering of RMD profiles

For RMD profiles we applied a PCA to the centered data, where rows were tumor samples and the columns were megabase windows. Next, we applied a clustering on the PC1 to PC21 using the R function tclust for robust clustering. We tested different numbers of clusters and alpha value (number of outliers removed). In addition, we tested the clustering using all PCs (PC1 to PC21) and without PC1 (PC2 to PC21), selecting the clustering for k=18 and alpha = 0.02 without PC1 based on the log likelihood measurement.

Indel and structural variant mutation risk profiles

For Hartwig Medical Foundation tumor WGS, we calculated the indels and structural variants (SV) regional mutation risk profiles by pooling samples with similar activity levels of the RMD signature of interest. For SVs, we separated duplications (DUP), inversions (INV) and deletions (DEL). Firstly, we binned the tumor WGS samples according to their RMD signature "exposure" levels. We created 20 bins for the SV analysis, and ~60 bins for indel analysis (20 tumors per bin). This binning is to increase statistical power in detecting regional mutation rate enrichment, as the numbers of indels and SVs are lower than SNVs. Then, we calculated the indels/SVs profile (regional indel/SV event density) across the 1Mb windows, for the indels/SVs pooled across the tumor WGS in that bin. For SVs, we count each SV only once by taking into account only the reported start position by the SV caller software. Finally, we correlated these indel-profiles / SV-profiles (separately for the different bins, stratified by the

SNV RMD signature), with the 1 Mb window weights profile of the RMD signature of interest, next comparing between the matching and the mismatching RMD signatures.

RMDsignature exposures for clonal vs subclonal mutations

We separated clonal/subclonal mutations using: mutation VAF<0.4*purity for subclonal mutations for Hartwig, and a generic threshold of VAF<0.3 for CPTAC-3 (purity data not available). Per tumor genome, we randomly sample the mutations to have the same number in the clonal and subclonal category, to equalize noise from low mutation counts. Next we calculated the RMD mutation risk profiles (number of mutations per each 1 Mb window) for the subclonal and the clonal mutations separately.

Each RMD profile is a mixture of mutations arising from different processes (our RMD signatures). We used a regression to model their relative activity ("exposure") to the observed RMD profile of each tumor. We compared the exposures for RMD signatures, thus inferred, from clonal versus subclonal mutations in each tumor sample.

Data availability

In this study published datasets were reanalyzed. We downloaded 1950 WGS somatic single-nucleotide variants (SNVs) from the Pan-cancer Analysis of Whole Genomes (PCAWG) study at the International Cancer Genome Consortium [63] Data portal (https://dcc.icgc.org/pcawg). We obtained 4823 WGS somatic SNVs from the Hartwig Medical Foundation (HMF) project [64] (https://www.hartwigmedicalfoundation.nl/en/). We downloaded 570 WGS somatic SNVs from the Personal Oncogenomics (POG) project [65] from BC Cancer (https://www.bcgsc.ca/downloads/POG570/). We obtained 724 WGS somatic SNVs from The Cancer Genome Atlas (TCGA) study as in [9]; we used QSS_NT>=12 mutation calling threshold in this study. We downloaded bam files for 781 WGS samples from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) project [66,67] and bam files for 758 tumor samples from the MMRF COMMPASS project [68] from the GDC data portal (https://portal.gdc.cancer.gov/).

Code availability

Custom code available in a github repository https://github.com/marina-salvadores/RMDsig

Acknowledgements

M.S. was funded by a FPU fellowship of the Spanish government, Ministry of Universities. Work in the lab of F.S. is supported by an ERC StG "HYPER-INSIGHT" (757700), Horizon2020 project "DECIDER" (965193), Spanish government project "REPAIRSCAPE", CaixaResearch project "POTENT-IMMUNO" (HR22-00402), an ICREA professorship to F.S.,

the SGR funding of the Catalan government, and the Severo Ochoa centers of excellence award of the Spanish government to the hosting institution.
This publication and the underlying research are partly facilitated by Hartwig Medical Foundation and the Center for Personalized Cancer Treatment (CPCT) which have generated, analysed and made available data for this research. In addition, data used in this publication were generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). We acknowledge that the results published here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

References

1. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
2. Zheng, C. L. *et al.* Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. *Cell Rep.* **9**, 1228–1234 (2014).
3. Pope, B. D. *et al.* Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405 (2014).
4. Akdemir, K. C. *et al.* Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nat. Genet.* **52**, 1178–1188 (2020).
5. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci.* **107**, 139–144 (2010).
6. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
7. Makova, K. D. & Hardison, R. C. The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **16**, 213–223 (2015).
8. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
9. Salvadores, M., Mas-Ponte, D. & Supek, F. Passenger mutations accurately classify human tumors. *PLOS Comput. Biol.* **15**, e1006953 (2019).
10. Jiao, W. *et al.* A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* **11**, 1–12 (2020).
11. Kübler, K. *et al.* Tumor mutational landscape is a record of the pre-malignant state.

517565 Preprint at https://doi.org/10.1101/517565 (2019).

12. Koren, A. *et al.* Genetic Variation in Human DNA Replication Timing. *Cell* **159**, 1015–1026 (2014).
13. McRae, A. F. *et al.* Identification of 55,000 Replicated DNA Methylation QTL. *Sci. Rep.* **8**, 17605 (2018).
14. Oliva, M. *et al.* DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat. Genet.* **55**, 112–122 (2023).
15. Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).
16. Kinker, G. S. *et al.* Pan-cancer single cell RNA-seq uncovers recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
17. Barkley, D. *et al.* Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.* **54**, 1192–1201 (2022).
18. Du, Q. *et al.* Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat. Commun.* **10**, 416 (2019).
19. Du, Q. *et al.* DNA methylation is required to maintain both DNA replication timing precision and 3D genome organization integrity. *Cell Rep.* **36**, 109722 (2021).
20. Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).
21. Brinkman, A. B. *et al.* Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation. *Nat. Commun.* **10**, 1749 (2019).
22. Gurrion, C., Uriostegui, M. & Zurita, M. Heterochromatin Reduction Correlates with the Increase of the KDM4B and KDM6A Demethylases and the Expression of Pericentromeric DNA during the Acquisition of a Transformed Phenotype. *J. Cancer* **8**, 2866–2875 (2017).
23. Wong, K. M., King, D. A., Schwartz, E. K., Herrera, R. E. & Morrison, A. J. Retinoblastoma protein regulates carcinogen susceptibility at heterochromatic cancer

driver loci. *Life Sci. Alliance* **5**, e202101134 (2022).

24. Huang, Y., Gu, L. & Li, G.-M. H3K36me3-mediated mismatch repair preferentially protects actively transcribed genes from mutation. *J. Biol. Chem.* **293**, 7811–7823 (2018).
25. Poetsch, A. R., Boulton, S. J. & Luscombe, N. M. Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. *Genome Biol.* **19**, 215 (2018).
26. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
27. Jönsson, J.-M. *et al.* Molecular Subtyping of Serous Ovarian Tumors Reveals Multiple Connections to Intrinsic Breast Cancer Subtypes. *PLOS ONE* **9**, e107643 (2014).
28. Mas-Ponte, D. & Supek, F. DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. *Nat. Genet.* 1–11 (2020) doi:10.1038/s41588-020-0674-6.
29. Supek, F. & Lehner, B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell* **170**, 534-547.e23 (2017).
30. Degasperi, A. *et al.* Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. *Science* **376**, abl9283 (2022).
31. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
32. Yaacov, A. *et al.* Cancer Mutational Processes Vary in Their Association with Replication Timing and Chromatin Accessibility. *Cancer Res.* **81**, 6106–6116 (2021).
33. Supek, F. & Lehner, B. Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair* **81**, 102647 (2019).
34. Liu, Y. *et al.* Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. *Nat. Commun.* **12**, 2439 (2021).
35. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set

collection. *Cell Syst.* **1**, 417 (2015).

36. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
37. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
38. Gindin, Y., Meltzer, P. S. & Bilke, S. Replicon: a software to accurately predict DNA replication timing in metazoan cells. *Front. Genet.* **5**, (2014).
39. Pratto, F. *et al.* Meiotic recombination mirrors patterns of germline replication in mice and humans. *Cell* **184**, 4251-4267.e20 (2021).
40. Gnan, S. *et al.* Kronos scRT: a uniform framework for single-cell replication timing analysis. *Nat. Commun.* **13**, 2329 (2022).
41. Ryba, T. *et al.* Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. *Genome Res.* **22**, 1833–1844 (2012).
42. Rivera-Mulia, J. C. *et al.* Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. *Genome Res.* **25**, 1091–1103 (2015).
43. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
44. SPIN reveals genome-wide landscape of nuclear compartmentalization | Genome Biology | Full Text. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02253-3.
45. Lambuta, R. A. *et al.* Whole-genome doubling drives oncogenic loss of chromatin segregation. *Nature* **615**, 925–933 (2023).
46. Gonzalo, S. *et al.* Role of the RB1 family in stabilizing histone methylation at constitutive heterochromatin. *Nat. Cell Biol.* **7**, 420–428 (2005).
47. Krishnan, B. *et al.* Active RB causes visible changes in nuclear organization. *J. Cell Biol.* **221**, e202102144 (2022).
48. Dick, F. A., Goodrich, D. W., Sage, J. & Dyson, N. J. Non-canonical functions of the RB

protein in cancer. *Nat. Rev. Cancer* **18**, 442–451 (2018).

49. Takahashi, C., Contreras, B., Bronson, R. T., Loda, M. & Ewen, M. E. Genetic Interaction between Rb and K-ras in the Control of Differentiation and Tumor Suppression. *Mol. Cell. Biol.* **24**, 10406–10415 (2004).
50. Lee, K. Y., Ladha, M. H., McMahon, C. & Ewen, M. E. The Retinoblastoma Protein Is Linked to the Activation of Ras. *Mol. Cell. Biol.* **19**, 7724–7732 (1999).
51. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).
52. Yaeger, R. *et al.* Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer. *Cancer Cell* **33**, 125-136.e3 (2018).
53. Cramer, D., Serrano, L. & Schaefer, M. H. A network of epigenetic modifiers and DNA repair genes controls tissue-specific copy number alteration preference. *eLife* **5**, e16519 (2016).
54. Donehower, L. A. *et al.* Integrated Analysis of TP53 Gene and Pathway Alterations in The Cancer Genome Atlas. *Cell Rep.* **28**, 1370-1384.e5 (2019).
55. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
56. Hodgkinson, A., Chen, Y. & Eyre-Walker, A. The large-scale distribution of somatic mutations in cancer genomes. *Hum. Mutat.* **33**, 136–143 (2012).
57. Woo, Y. H. & Li, W.-H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat. Commun.* **3**, 1004 (2012).
58. Liu, L., De, S. & Michor, F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.* **4**, 1502 (2013).
59. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32**, 71–75 (2014).
60. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).

61. García-Nieto, P. E. *et al.* Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. *EMBO J.* **36**, 2829–2843 (2017).
62. Xiong, K. & Ma, J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat. Commun.* **10**, 5069 (2019).
63. Hudson (Chairperson), T. J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
64. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
65. Pleasance, E. *et al.* Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* **1**, 452–468 (2020).
66. Ellis, M. J. *et al.* Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov.* **3**, 1108–1112 (2013).
67. Edwards, N. J. *et al.* The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J. Proteome Res.* **14**, 2707–2713 (2015).
68. Walker, B. A. *et al.* A high-risk, Double-Hit, group of newly diagnosed myeloma identified by genomic analysis. *Leukemia* **33**, 159–170 (2019).
69. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
70. Salvadores, M., Fuster-Tormo, F. & Supek, F. Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Sci. Adv.* **6**, eaba1862 (2020).
71. Derrien, T. *et al.* Fast Computation and Applications of Genome Mappability. *PLOS ONE* **7**, e30377 (2012).
72. Ormond, C., Ryan, N. M., Corvin, A. & Heron, E. A. Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Brief. Bioinform.* **22**, bbab069 (2021).
73. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of

Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).

74. Buisson, R. *et al.* Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, (2019).
75. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
76. Vali-Pour, M., Lehner, B. & Supek, F. The impact of rare germline variants on human somatic mutation processes. *Nat. Commun.* **13**, 3724 (2022).
77. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
78. Zhao, P. A., Sasaki, T. & Gilbert, D. M. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biol.* **21**, 76 (2020).
79. Sima, J. *et al.* Identifying cis Elements for Spatiotemporal Control of Mammalian DNA Replication. *Cell* **176**, 816-830.e18 (2019).
80. Van Rechem, C. *et al.* Collective regulation of chromatin modifications predicts replication timing during cell cycle. *Cell Rep.* **37**, 109799 (2021).
81. Sarni, D. *et al.* Replication Timing and Transcription Identifies a Novel Fragility Signature Under Replication Stress. 716951 Preprint at https://doi.org/10.1101/716951 (2019).
82. Poulet, A. *et al.* RT States: systematic annotation of the human genome using cell type-specific replication timing programs. *Bioinformatics* **35**, 2167–2176 (2019).
83. Klein, K. N. *et al.* Replication timing maintains the global epigenetic state in human cells. *Science* **372**, 371–378 (2021).
84. Ding, Q. *et al.* The genetic architecture of DNA replication timing in human pluripotent stem cells. *Nat. Commun.* **12**, 6746 (2021).
85. Gunasekara, C. J. *et al.* A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol.* **20**, 105–105 (2019).
86. Mukhopadhyay, R. *et al.* Allele-Specific Genome-wide Profiling in Human Primary Erythroblasts Reveal Replication Program Organization. *PLoS Genet.* **10**, e1004319

(2014).

87. Barlow, J. H. *et al.* Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. *Cell* **152**, 620–632 (2013).

88. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).

Supplementary Figures

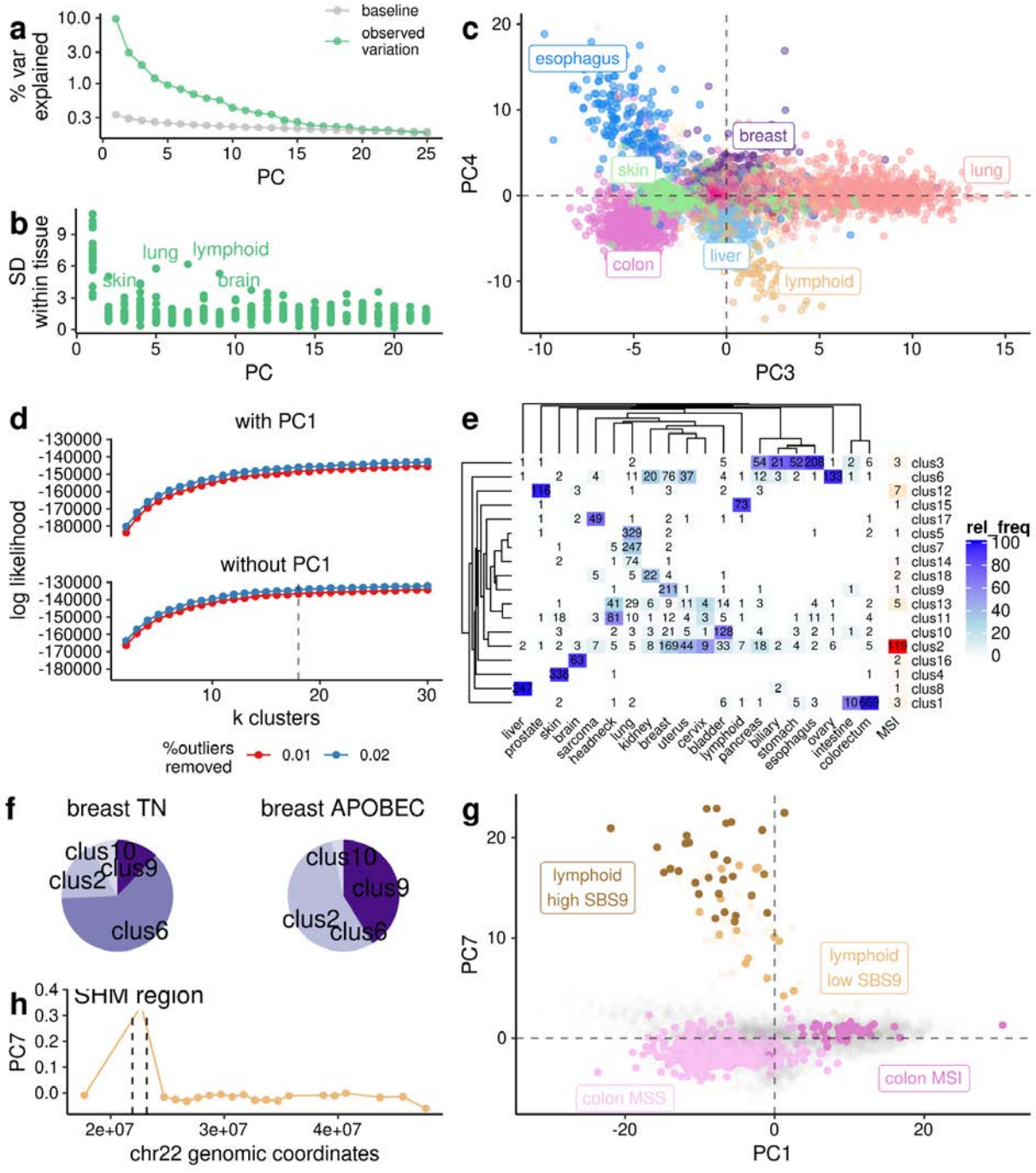


Fig S1. RMD variability across tissues and individuals. a) Variance explained for the first 25 PCs of a PCA on the RMD matrix (4221 samples x 2540 one-megabase windows), and a baseline (by the broken stick heuristic). **b)** Inter-individual variability (SD, standard deviation) within each tissue for the first 22 PCs. **c)** PC3 and 4 separate various cancer types. **d)** Log likelihood of the clustering of the first 22 PCs using different numbers of k clusters for two cases (including and excluding PC1). **e)** Clustering into 18 clusters, showing the number of tumor samples from each cancer type that are assigned to each RMD-based cluster (Methods). **f)** Cluster assignment for breast cancer samples of triple negative (TN) subtype and samples with high APOBEC (>25% of mutations are in APOBEC contexts). **g)** As controls, the PC1 separates MSI versus MSS tumors, and PC7 separates lymphoid tumors

according to their level of the SHM-associated mutational signature SBS9. **h)** PC7 window weights for chromosome 22 agree with the known SHM region.

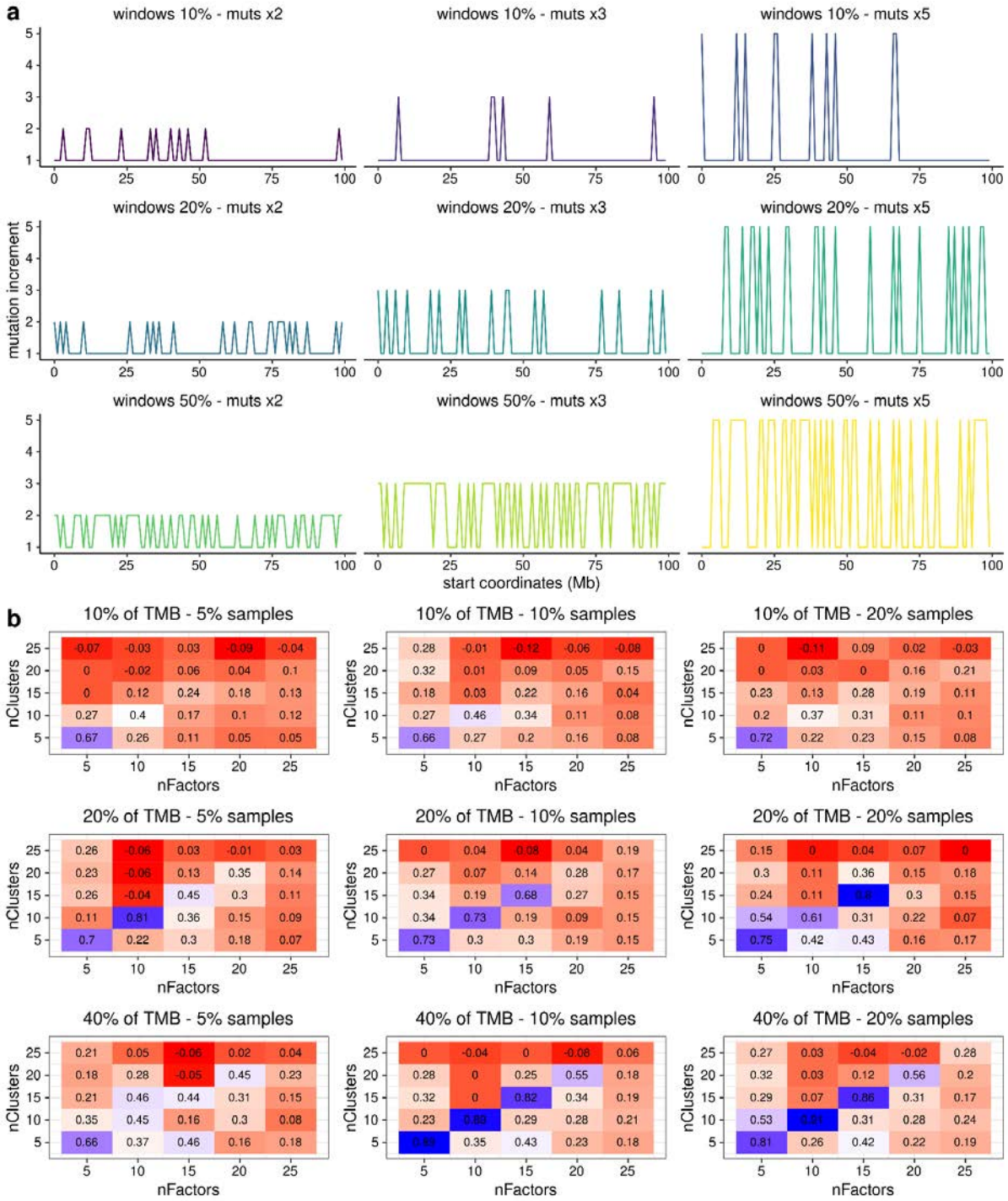


Fig S2. Simulated ground-truth RMD signatures to benchmark the NMF methodology. a) Profiles for the 9 simulated signatures generated (first 100 windows of chr1 shown). The value is the increase of mutations it will generate when multiplied by the cancer type vector (x2, x3 or x5). In various scenarios considered, the number of

windows affected by the increase can be 10%, 20% or 50%, as stated in the panel title: sig_[windows affected %]_[increase X]. **b)** Robustness of extracted RMD signatures, estimated as the silhouette index (SI) for the clustering of the extracted RMD signatures at different k (number k-medoids clusters) and nFact (number NMF factors). Each heatmap shows the minimum clustering SI for a particular condition (varying the signature contribution to the TMB and the number of samples affected, as specified in the title).

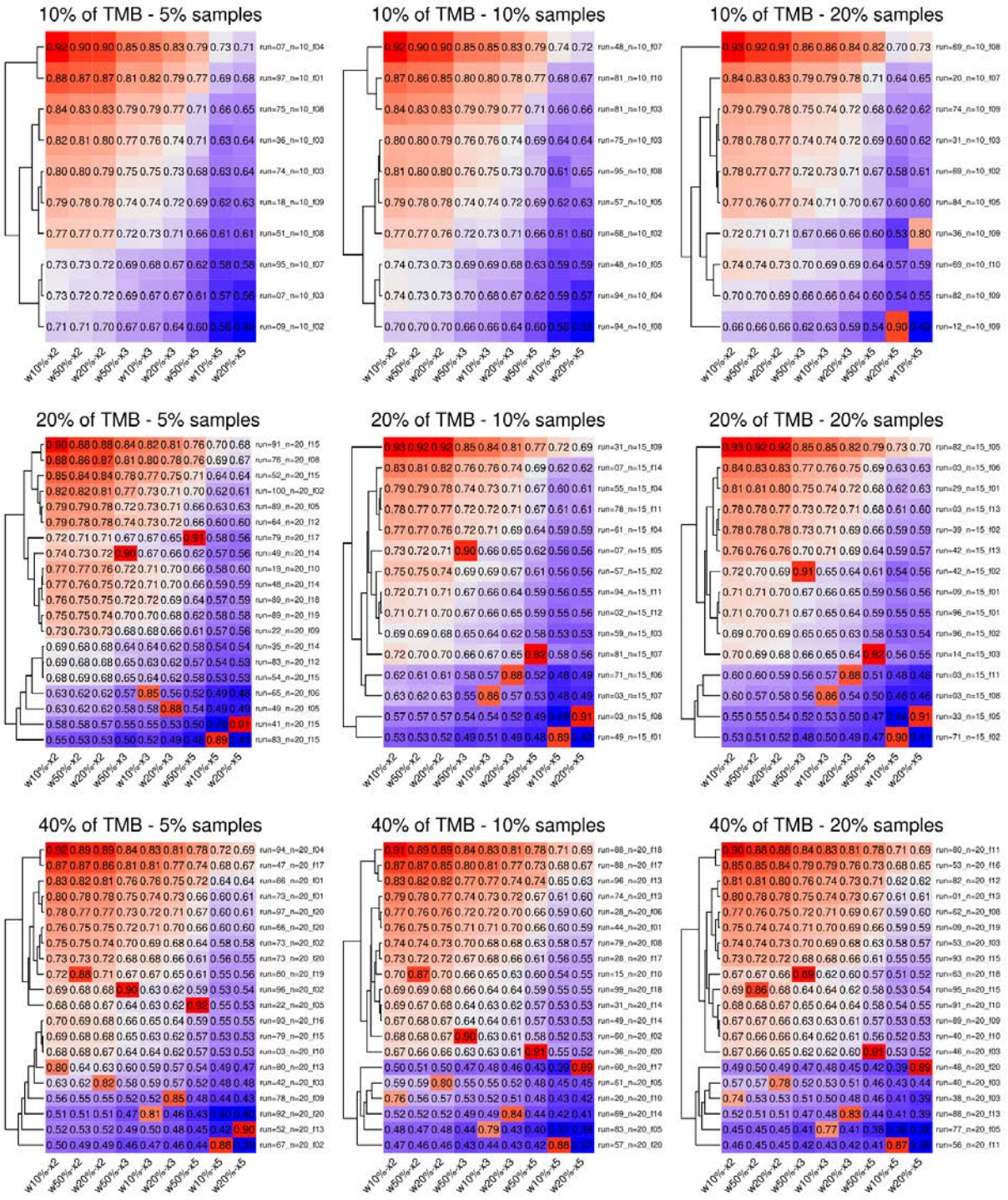


Fig S3. Testing ability to Identify NMF signatures matching the ground-truth simulated signatures. Cosine similarity values between the ground-truth RMD signatures (columns) and the extracted RMD signatures using NMF

(rows). We consider a signature to be recovered correctly when it matches exactly one ground-truth RMD signature with cosine similarity >= 0.75.

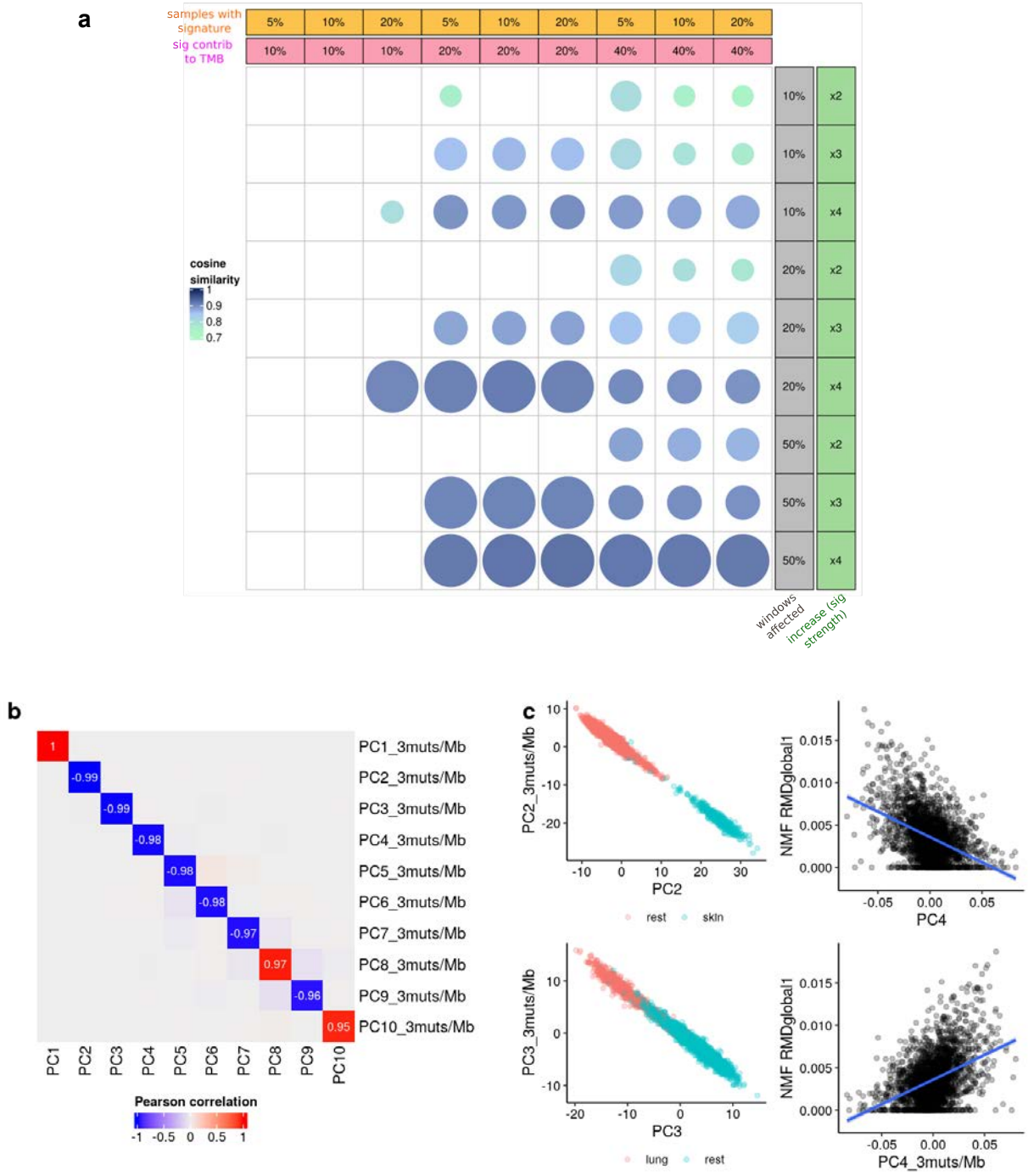


Fig S4. Summary of the benchmarking of the NMF methodology using simulated cancer genomes. a) Comparison of the number of signatures recovered by NMF depending on the 9 simulation scenarios (columns) and the characteristics of the 9 simulated signatures (rows). We can see which simulated signatures are more often recovered (with a circle) and how strong (color and size of the circle) and under which conditions. **b)** Testing whether 3 mutations / Mb is sufficient for extracting signatures. A PCA in the original dataset is compared to a PCA from a subsampled dataset (mutations removed from tumor genomes until all genomes have 3 mutations / Mb). The

spectra of the first 10 PCs correlate very highly between the original and reduced-mutation dataset, suggesting robust NMD signatures. **c)** PC2 and PC3 tumor sample exposures correlation between the original, and the reduced-mutation datasets. In addition, in both datasets PC4 window weights correlates similarly strongly with the RMDglobal1 signature.

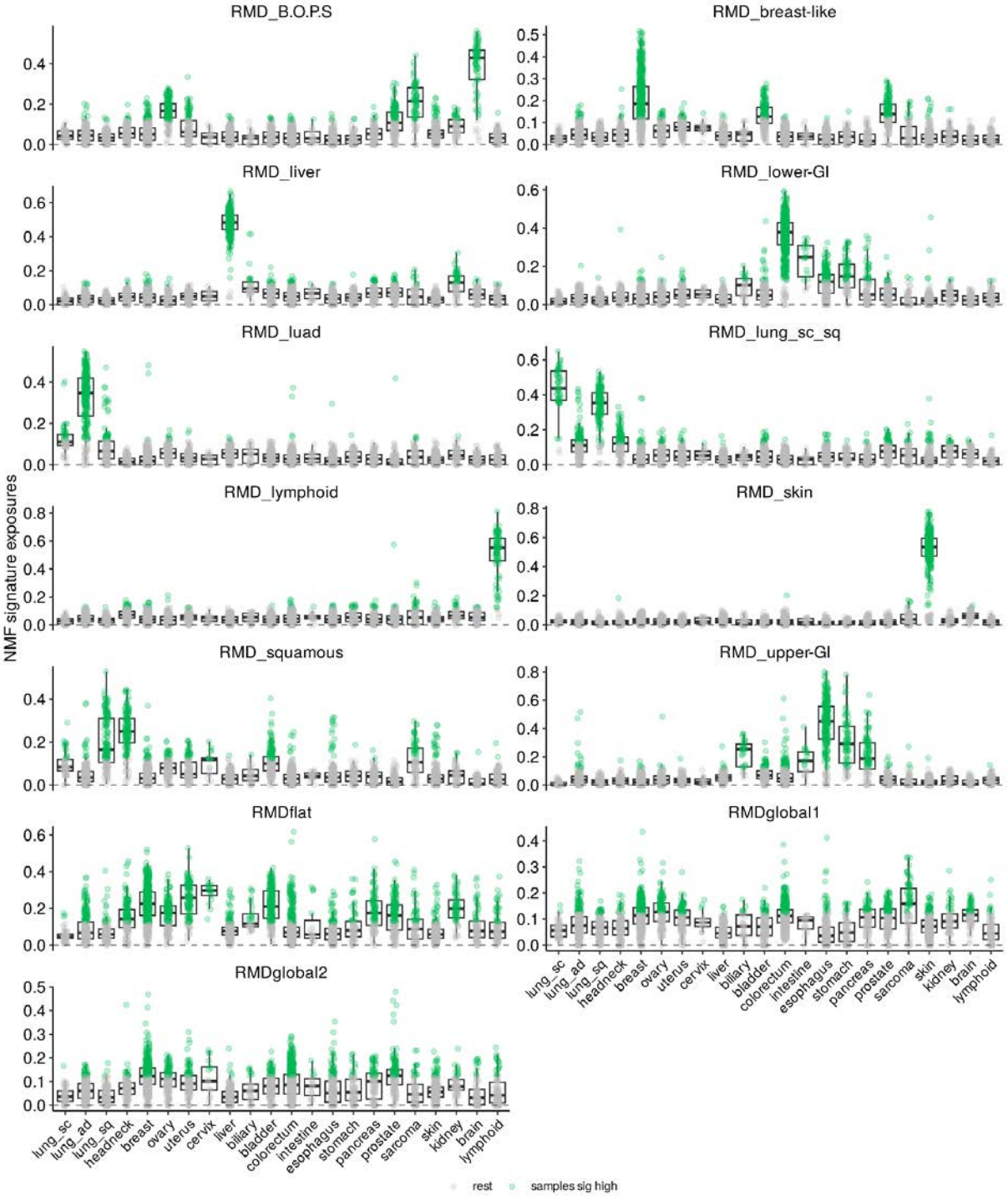


Fig S5. The activity levels of the NMF-derived RMD signatures across various human tumor types. Tumor sample “exposures" distributed across cancer types for the 13 RMD signatures extracted. Marked in green are the

tumor samples considered to have a high contribution from a particular signature. The threshold corresponds to the tumor sample exposure value for which 99% of the MSI samples are recovered in the RMDflat signature (a positive control for association of RMD signatures with a biological feature).

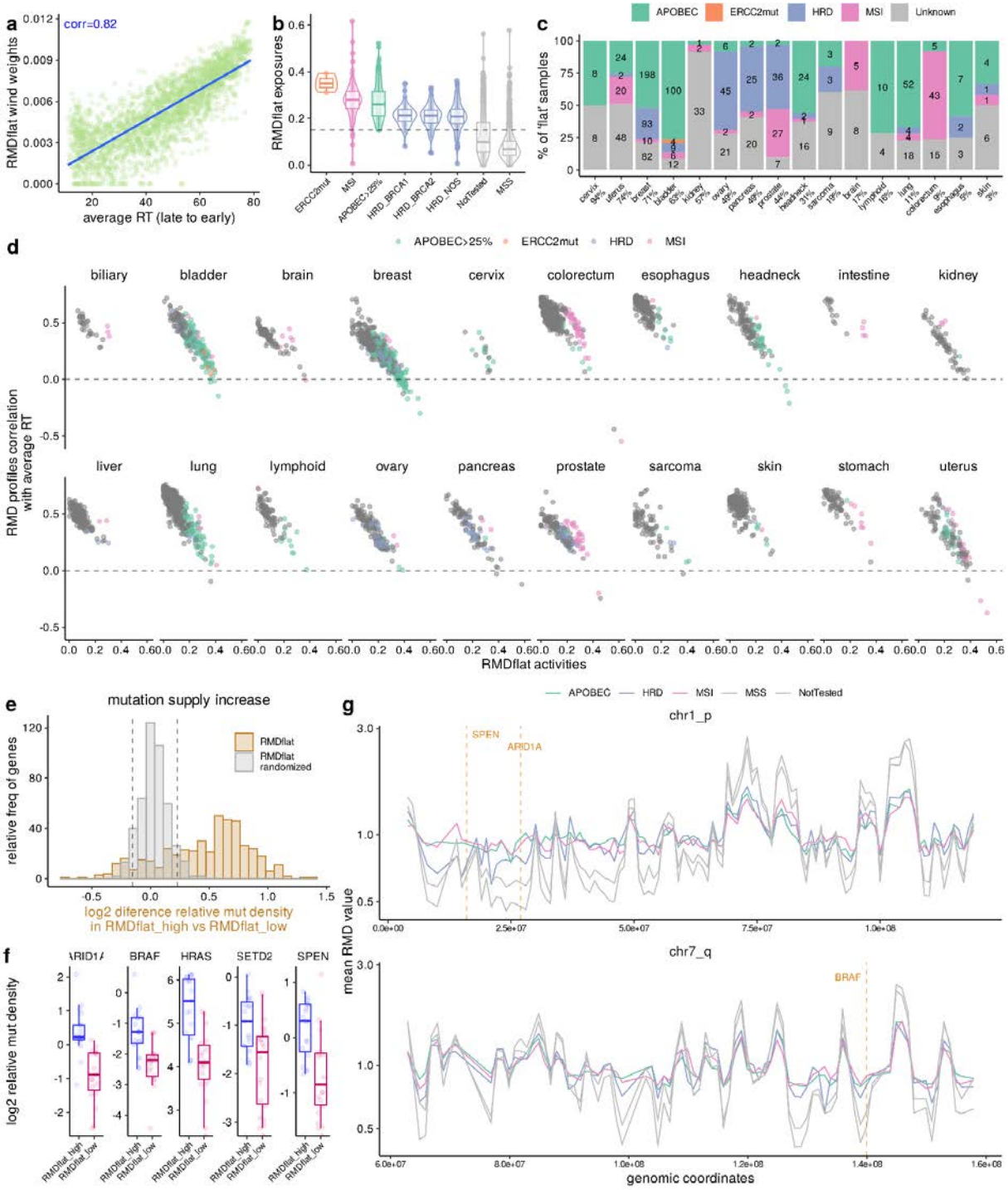


Fig S6. Characterization of the RMDflat signature, which represents a loss of megabase-scale mutation rate heterogeneity. a) Correlation between RMDflat signature NMF window weights and the DNA replication timing (RT) (here, average Repli-Seq signal across 10 cell lines). **b)** RMDflat signature exposures (i.e. activities) for groups of tumor samples with various DNA repair failures: (i) MSI, microsatellite instable, indicating a MMR failure; (ii) HRD, homologous recombination deficiency, split by the BRCA1-type or BRCA2-type or not otherwise specified, or (iii) high levels of APOBEC mutation signatures, (iv) *ERCC2* mutant, and (v) the remainder of the tumors are in the "not tested" or "MSS" groups (latter means that MSI was tested but was negative). **c)** Percentage of tumor samples with flat mutation rate landscapes (RMDflat exposure>0.177, a threshold that recovers 95% of known MSI samples) belonging to each of the 5 listed DNA repair categories, broken down by cancer type. The percentage of RMDflat-high samples in each cancer type is indicated in the x-axis labels. **d)** Association between the RMDflat activities and the tumor correlation with the average replication timing. **e)** Distribution of the log2 difference in the intronic mutation rates (a measure of mutation supply towards a gene locus) for 460 cancer gene loci, comparing between RMDflat-high tumors and RMDflat-low tumors, using the actual values ("RMDflat" histogram) and randomized values ("RMDflat randomized" histogram). **f)** Log2 relative intronic mutation density (normalized to flanking DNA in same chromosome arm, see panel **d**) for the RMDflat-high *versus* the RMDflat-low, for 5 example common driver genes with highest effect sizes in this test. Each point is one cancer type. **g)** Mean RMD profile across the DNA repair groups, shown examples for chr 1p and chr 7q. Vertical lines mark the position for three of the example genes from panel **f**.

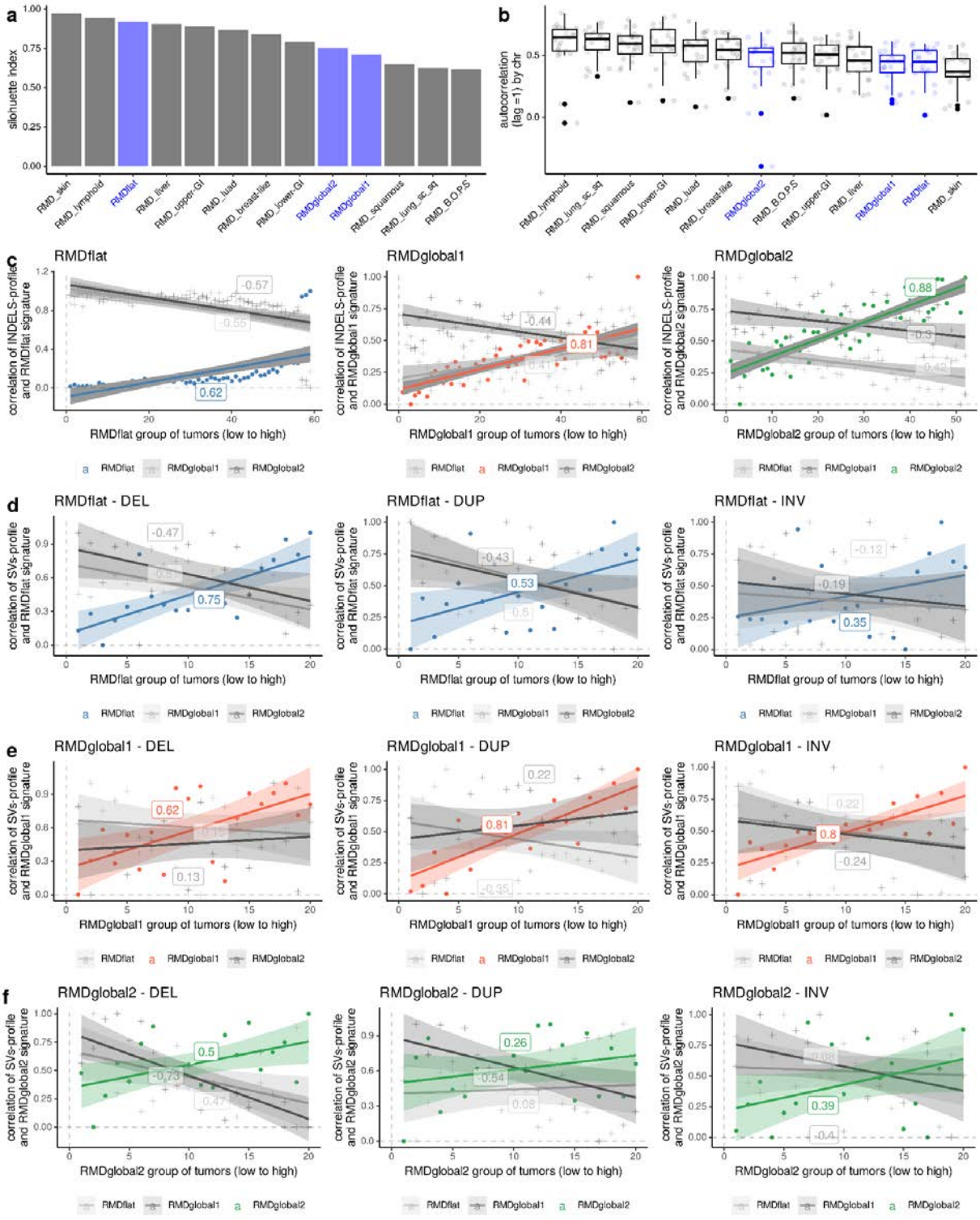


Fig S7. **Quality control for the mutation risk redistribution signatures RMDglobal1 and RMDglobal2. a)** Silhouette index (clustering quality score) of each RMD signature. **b)** Autocorrelation values with lag = 1 (calculated for each chromosome separately) for each RMD signature. **c)** Correlation of RMD-indel profiles with the 3 global RMD signatures, stratifying the tumor samples into bins (20 tumors per bin) by their RMD signature exposure (x axis; derived from SNV mutations, not indels). The slope from the regression is shown next to each regression line. In

specific, we binned the tumor samples according to their RMD signature exposure levels, and calculated the indel mutation density profile across the 1Mb windows for the pooled indels across tumor samples in each bin (pooling is to increase statistical power, as the numbers of indels are lower than of SNV mutations). Then, we correlated these RMD-indel-profiles with the profile of 1 Mb window weights of the matching RMD signature. **d-f)** Correlation of RMD-SVs profiles (inversions, deletions and duplications) with the 3 global RMD signatures, stratifying the tumor samples into 20 bins by their RMD signature exposure (x axis; derived from SNV mutations). The slope from the regression is shown next to each regression line. In specific, we binned the tumor samples according to their RMD signature exposure levels, and calculated the DNA rearrangement mutation density profile across the 1Mb windows for the pooled DNA rearrangements across tumor samples in each bin (pooling is to increase statistical power, as the numbers of DNA rearrangements are lower than of SNV mutations). We count each DNA rearrangement as 1 count (1 event) by taking into account only the start position. Then, we correlated these RMD-SVs-profiles with the profile of 1 Mb window weights of the matching RMD signature.

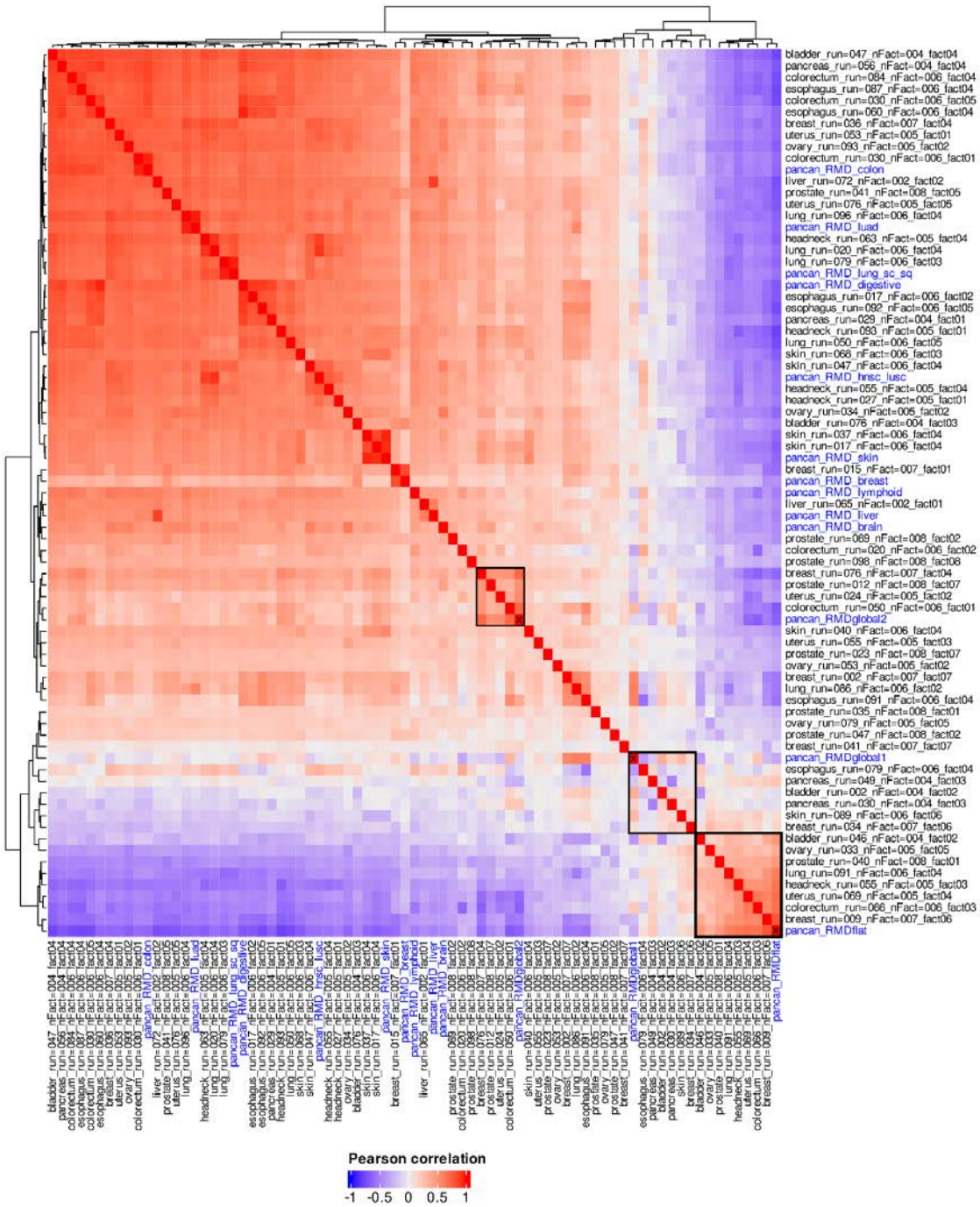


Fig S8. Replication of the pancancer RMD signatures in individual cancer types. Agreement between pan-cancer and individual cancer type extracted RMD signatures. Correlation between the RMDsignatures extracted in the pan-cancer NMF analysis (main results) and in NMF runs by cancer type (additional analysis) suggests that the 3 global RMD signatures (RMDflat, RMDglobal1 and RMDglobal2) can be recovered also from many individual cancer types (examples in black squares in the heatmap) .

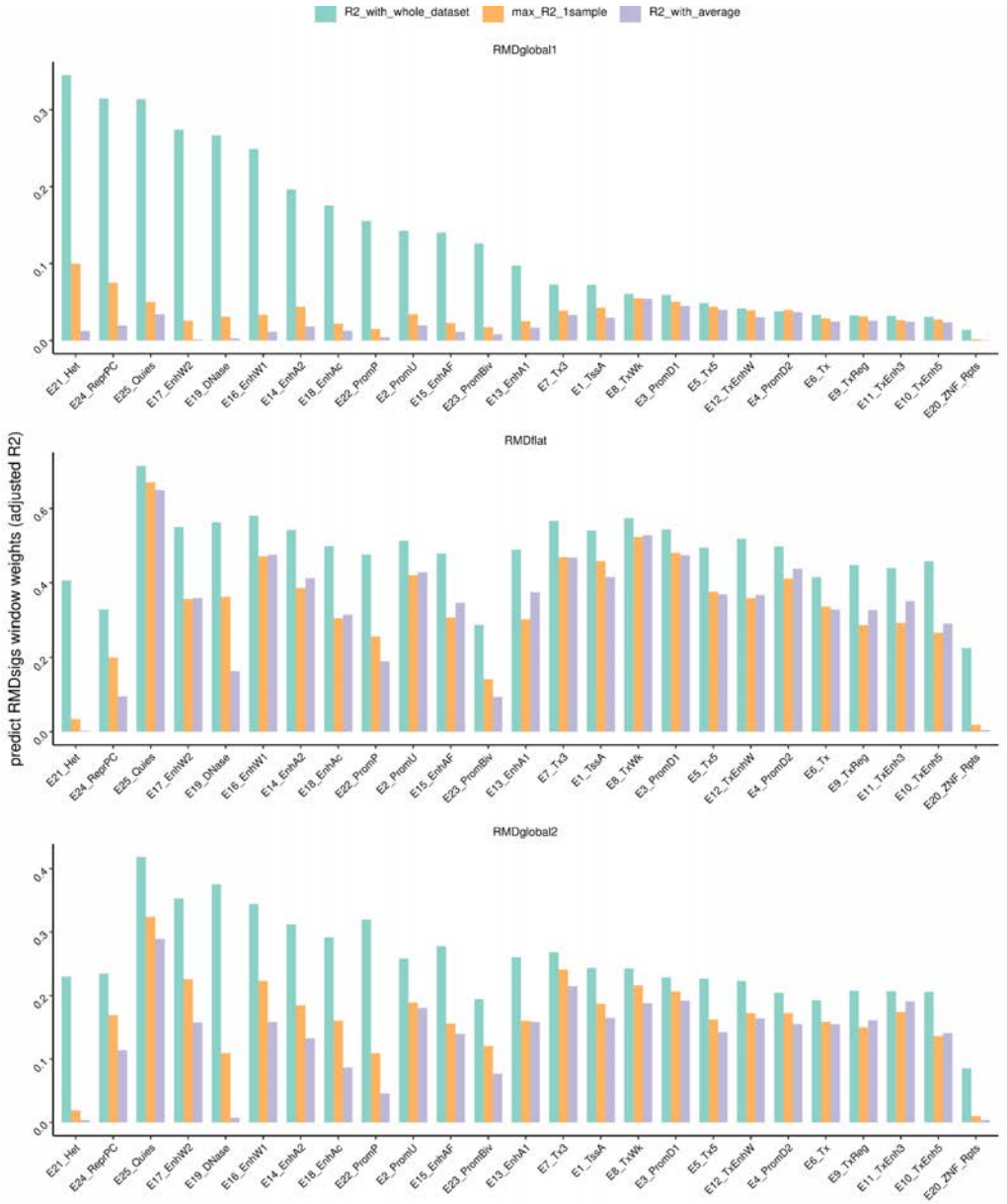


Fig S9. Prediction of RMD signatures from the density of ChromHMM states in chromosomal domains. Adjusted R2 of a regression predicting RMD signatures window weights from the chromHMM states density (% of the 1 Mb window covered by a particular state) (x axis). Predictions were made using either the whole dataset of the densities (many samples used as input for regression), the mean density of the feature (average over all samples used as input for regression) or selecting the maximum adjusted R2 of calculating R2 in regression for each sample individually.

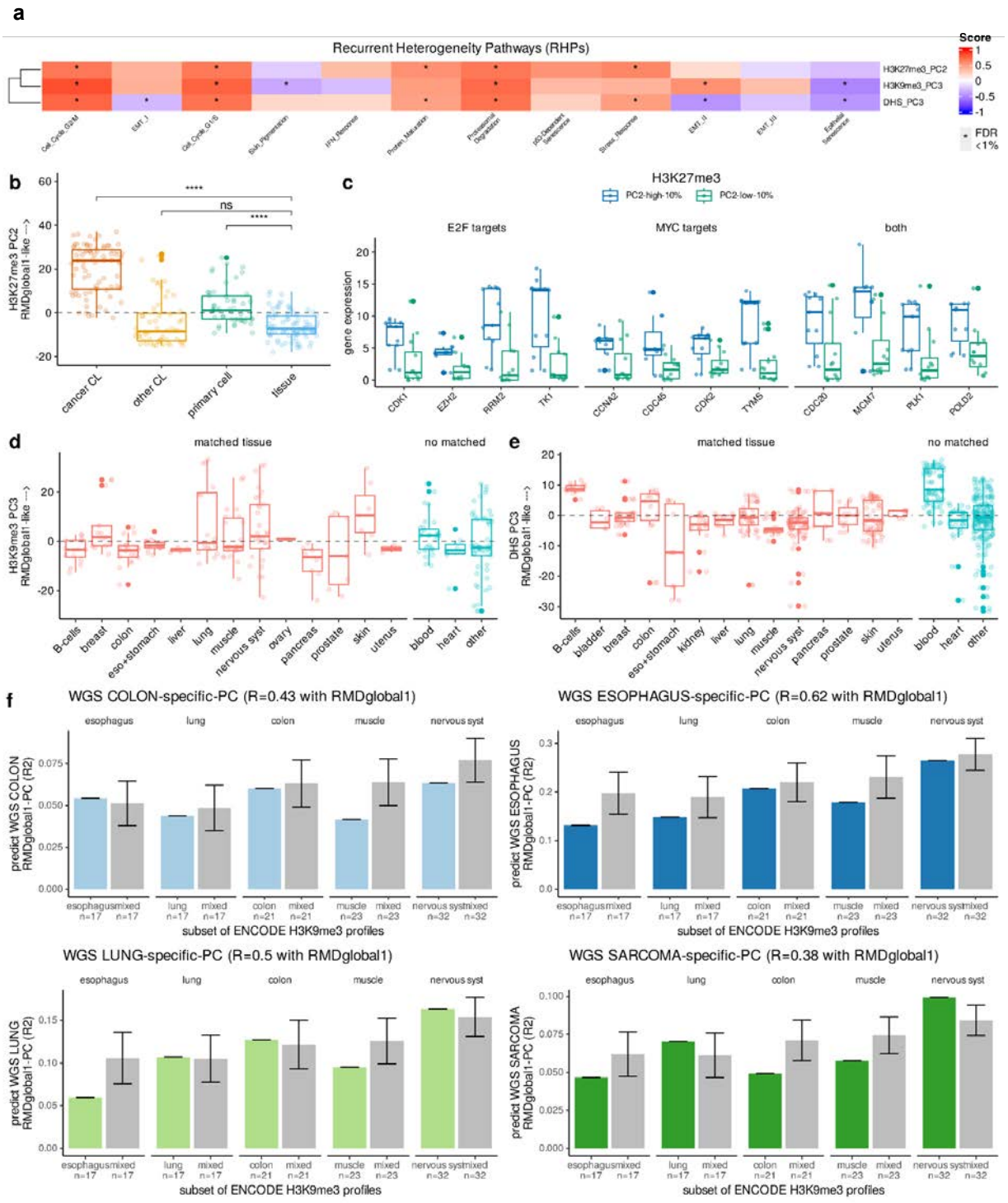


Fig S10. Variability in ENCODE gene expression and histone marks correlates with tumor RMDglobal1. a) Gene set enrichment Analysis (GSEA) scores for the Recurrent Heterogeneity Pathways (RHPs) (Kinker et al. 2020), for an ordered list of genes, based on the correlation of each gene's expression to the exposures of the 3 epigenomic PCs across ENCODE samples. For each epigenomic PC, we predicted the PCx sample exposures from the

expression of each gene separately (e.g. regression formula: H3K9me3_PC3 ~ geneA). We used the coefficient for the gene variable from the regression (i.e. the effect size) to order the genes, and applied GSEA on the gene list ordered by effect size. The GSEA positive values means that the genes from that gene set are overall higher expressed in the PCx-high samples compared to the PCx-low samples. **b)** H3K27me3_PC3 sample exposures stratified by the most relevant biosample types in ENCODE. **c)** Example of gene expression (sqrt TPMs) for a subset of genes from E2F targets and MYC targets categories of proliferation-associated genes, comparing ENCODE samples with high and low H3K27me3_PC3 exposures. **d-e)** Epigenetic-PCs ENCODE sample exposures stratified by the tissue types for H3K9me3_PC3 (d) and (-)DHS_PC3 (e). **f)** Prediction score (R-squared) of tissue-specific RMDglobal1 PCs in colon, esophagus, lung and sarcoma using different subsets of H3K9me3 profiles, grouped by tissue or mixed or tissues with the sample sample size.

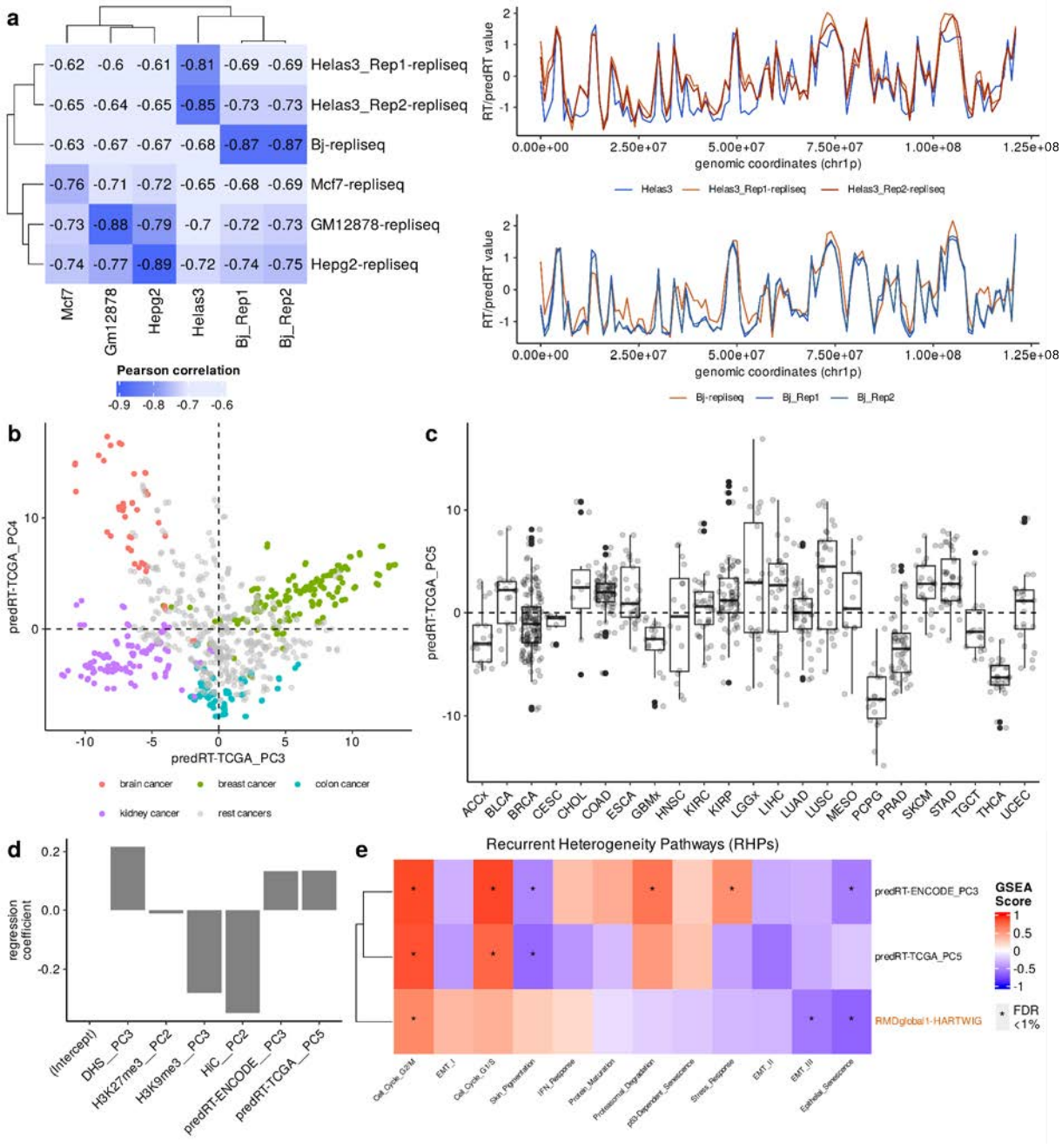


Fig S11. Accuracy of RT prediction and the predicted RT profile correlation with RMDglobal1 mutagenesis. a) High correlation between predicted RT from ENCODE DHS (x axis) and experimental RT data from Repli-seq in the same cell lines (y axis). Example RT profiles for DHS-predicted RT and RepliSeq RT shown for two matched cell lines in the right panel. **b)** PredRT-TCGA_PC3 and PredRT-TCGA_PC4 are tissue-specific, separating: breast versus kidney cancer and brain versus colon cancer, respectively. **c)** The tumor sample exposures of the PredRT-TCGA_PC5, a RT-PC that correlates with RMDglobal1 mutagenesis, broken down by cancer type, showing that cancers from almost all tissue types overlap by the range of RMDglobal1 scores (with a possible exception of the very rare cancer "PCPG" pheochromocytoma/paraganglioma). **d)** Regression coefficients from a regression predicting the RMDglobal1 mutagenesis profile across 1 Mb windows jointly from multiple relevant epigenomic PCs and RT PCs. **e)** Gene set enrichment Analysis (GSEA) scores for the Recurrent Heterogeneity Pathways (RHPs) (Kinker et al. 2020), for an ordered list of genes based on the correlation of gene expression with RT-PCs, and separately of that, gene expression with RMDglobal1 mutation risk. For each PC, we predicted the PCx sample exposures from the gene expression of each gene separately (e.g. regression formula: predRT-TCGA_PC5 ~ geneA). We used the coefficient for the gene variable from the regression (i.e. the effect size) to order the genes, and applied GSEA on the gene list ordered by effect size.

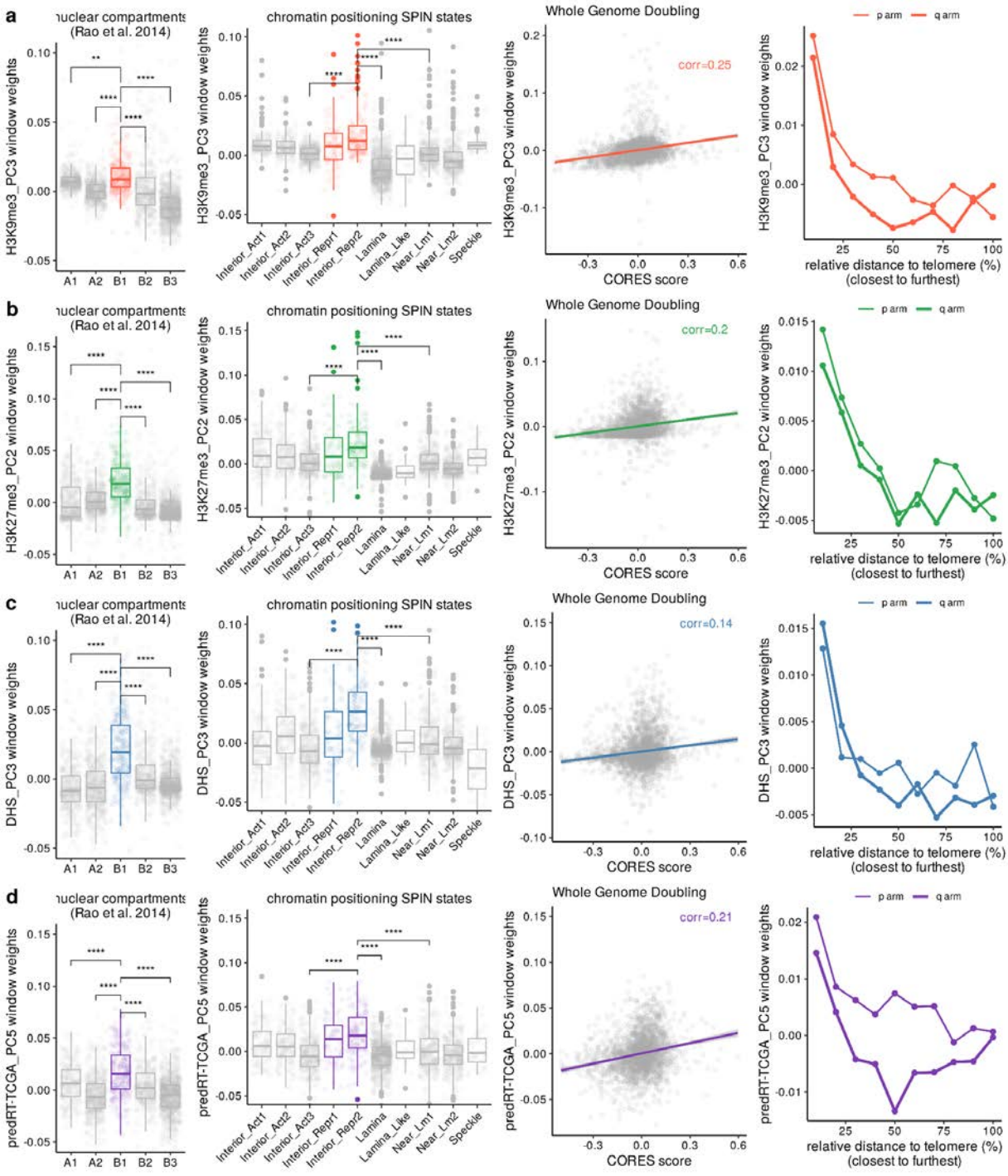


Fig S12. The RMDglobal1-associated chromatin restructuring PCs are linked with chromatin features similar to RMDglobal1 itself. The panels show the chromatin PC window weights across, from left to right, (i) different Hi-C nuclear subcompartments; (ii) different SPIN nuclear spatial positioning states; (iii) correlation with the CORES score for Hi-C changes during a whole-genome doubling; and (iv) compared to distance to telomeres, shown separately for **(a)** H3K9me3_PC3, **(b)** H3K27me3_PC2, **(c)** DHS_PC3 and **(d)** predRT-TCGA_PC5.

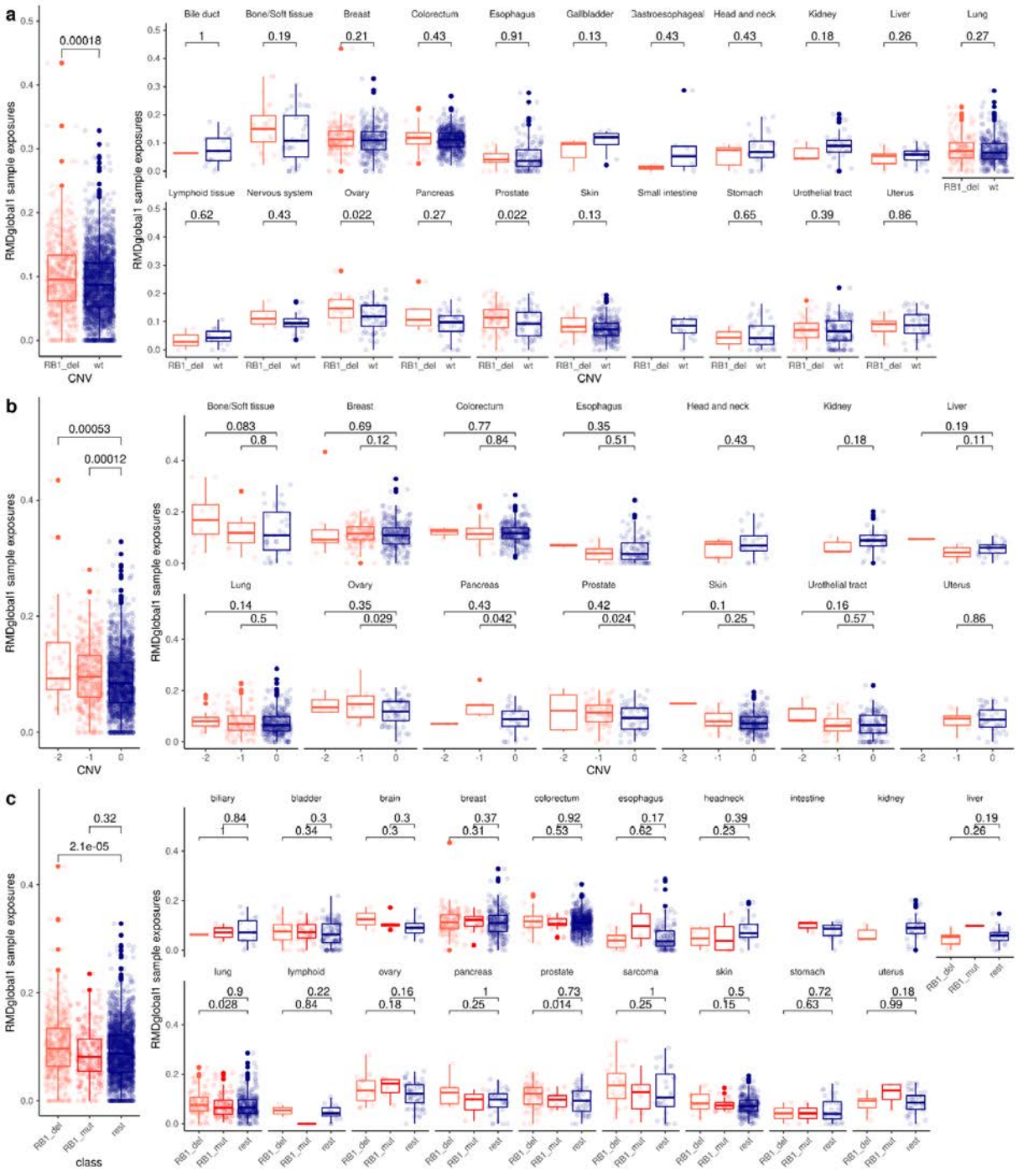


Fig S13. Associations between *RB1* deletion or *RB1* coding mutation ("del" and "mut") with RMDglobal1 mutation risk redistribution exposures. a) Differences in RMDglobal1 exposures between *RB1* deletion (-1 or -2 state) and wild-type in a pan-cancer analysis (left) or by cancer type (right). **b)** Differences in RMDglobal1 exposures between a monoallelic (-1) and biallelic (-2) RB1 deletion and the wild-type (0). **c)** Differences in RMDglobal1 exposures between RB1 deletion, RB1 mutations and wild-type.

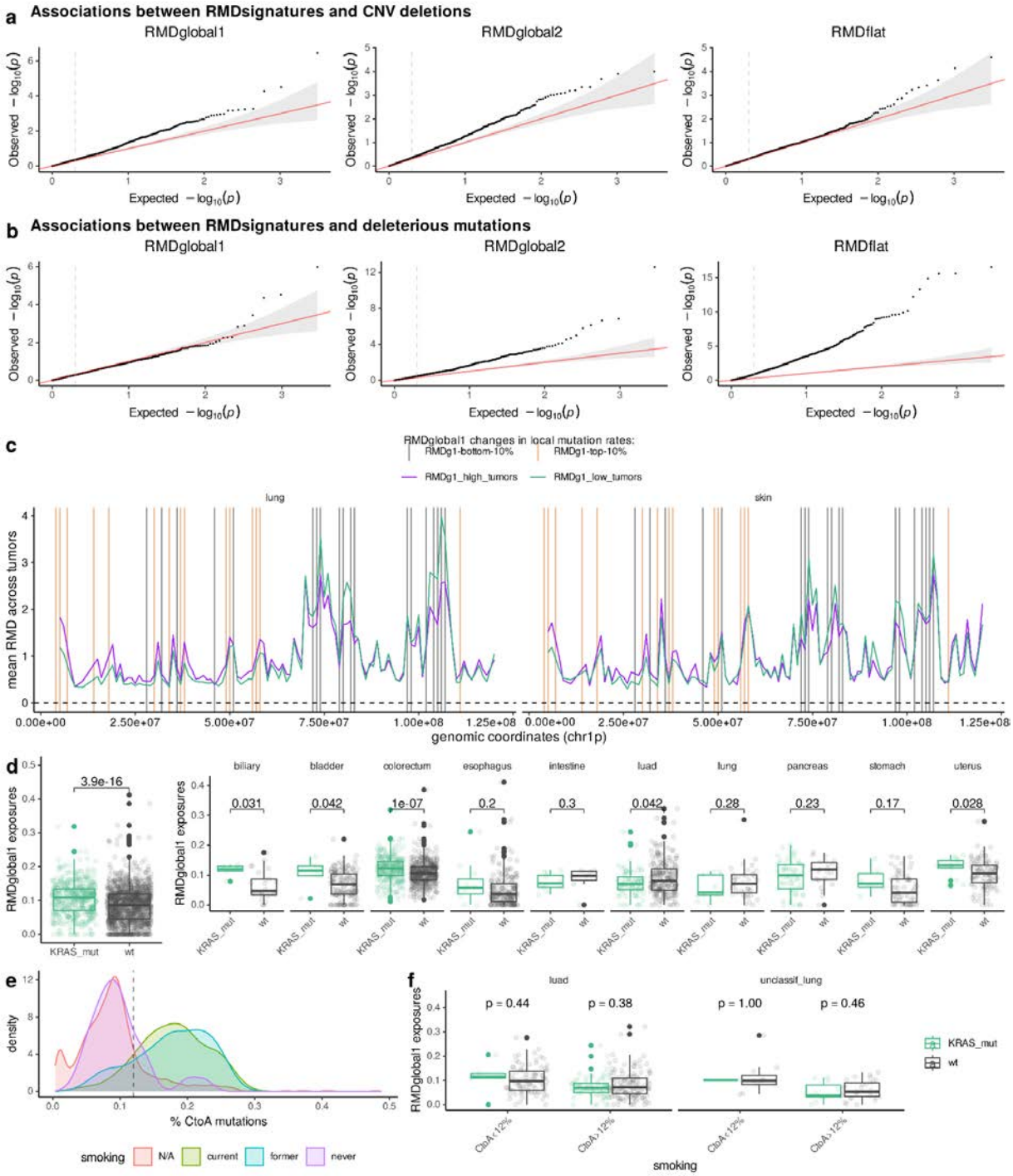


Fig S14. Association analyses between RMD mutation risk signatures, and genetic alterations in tumors. a) The qq plots for p-values in association testing of three RMD global signatures with CNA in cancer genes and chromatin modifying genes. **b)** The qq plots for p-values in association testing of three RMD global signatures with deleterious mutations in cancer genes and chromatin modifying genes. **c)** RMD profiles for skin and lung tumors grouping samples by RMDglobal1 high or low activities. Windows with lowest and highest RMDglobal1 changes marked with vertical lines. **d)** Differences in RMDglobal1 exposures between *KRAS* mutated and wild-type tumors in a pan-cancer analysis (left) or by cancer type (right); p-value by Mann-Whitney test. **e)** Distribution of fraction C>A

mutations (proxy for tobacco smoking exposure) along the patients separated by their self-reported smoking status. We classified the lung cancer samples with missing data ("N/A" in the plot) according to the C>A fraction threshold of 0.12 (vertical line). **f)** Differences in RMDglobal1 exposures between *KRAS* mutated and *KRAS* wild-type lung cancers, after stratifying by putative smoking (C>A > 12%) and non-smoking (C>A < 12%) status, shows that *KRAS* mutation is not associated with RMDglobal1 in smokers, but might be in smokers (non-significant trend, few non-smoker cancers are *KRAS* mutant).

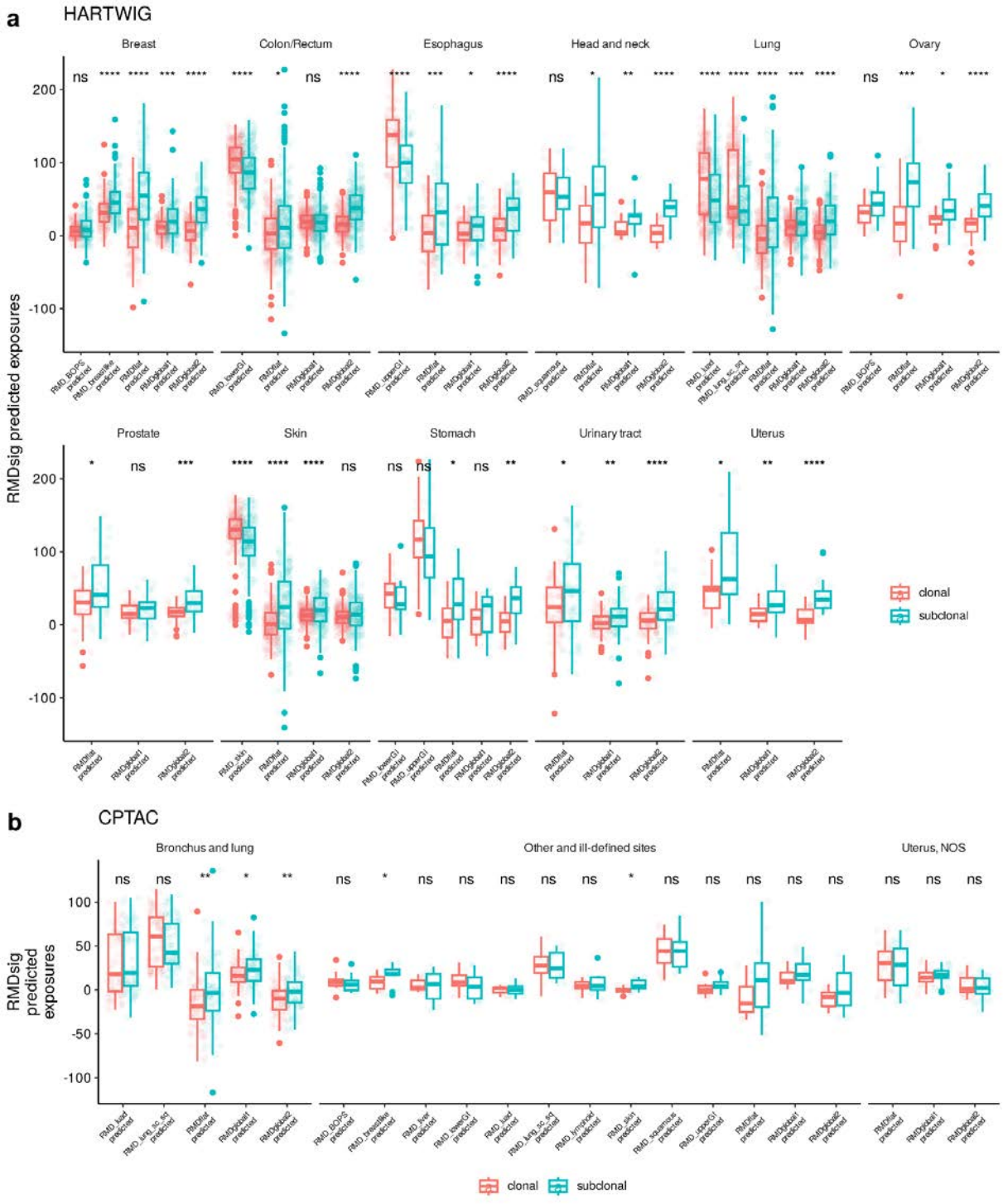


Fig S15. **Contribution of clonal versus subclonal mutations to RMDsignatures. a)** RMD signature predicted exposures for different cancer types comparing the exposure for clonal versus subclonal profiles in Hartwig. **b)** RMD signature predicted exposures for different cancer types comparing the exposure for clonal *versus* subclonal mutation profiles in CPTAC.

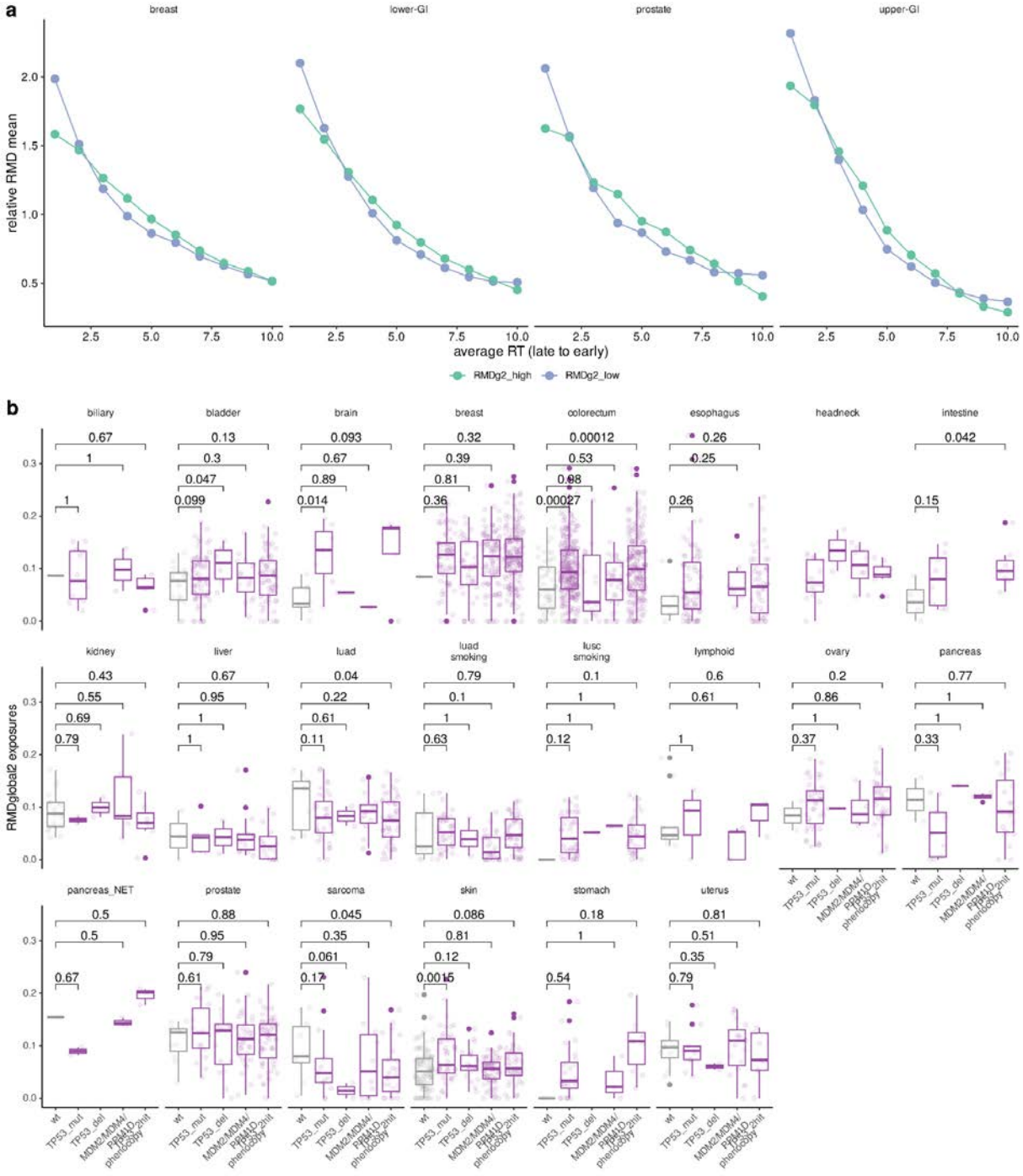


Fig S16. Association of RMDglobal2 activities and TP53 deficiency and TP53 deficiency phenocopies. a) the association of regional mutation rates with RT is altered by the activity of RMDglobal2 signature. Mean RMD values across 10 RT bins (late to early) for samples with RMDglobal2_high versus RMDglobal2_low, shown across tissues. **b)** Association of the activity of the RMDglobal2 mutation risk signature with various mechanisms of TP53 inactivation in different cancer types. RMDglobal2 exposures grouped by: wild-type (wt) TP53 pathway, *TP53* with 1 mutation (TP53_mut), *TP53* with 1 deletion (TP53_del), TP53 loss phenocopied via an amplification in the *MDM2*,

MDM4 or *PPM1D* genes (TP53_pheno), or TP53 inactivation with two hits of either of the previously mentioned mechanisms (TP53_2hit).

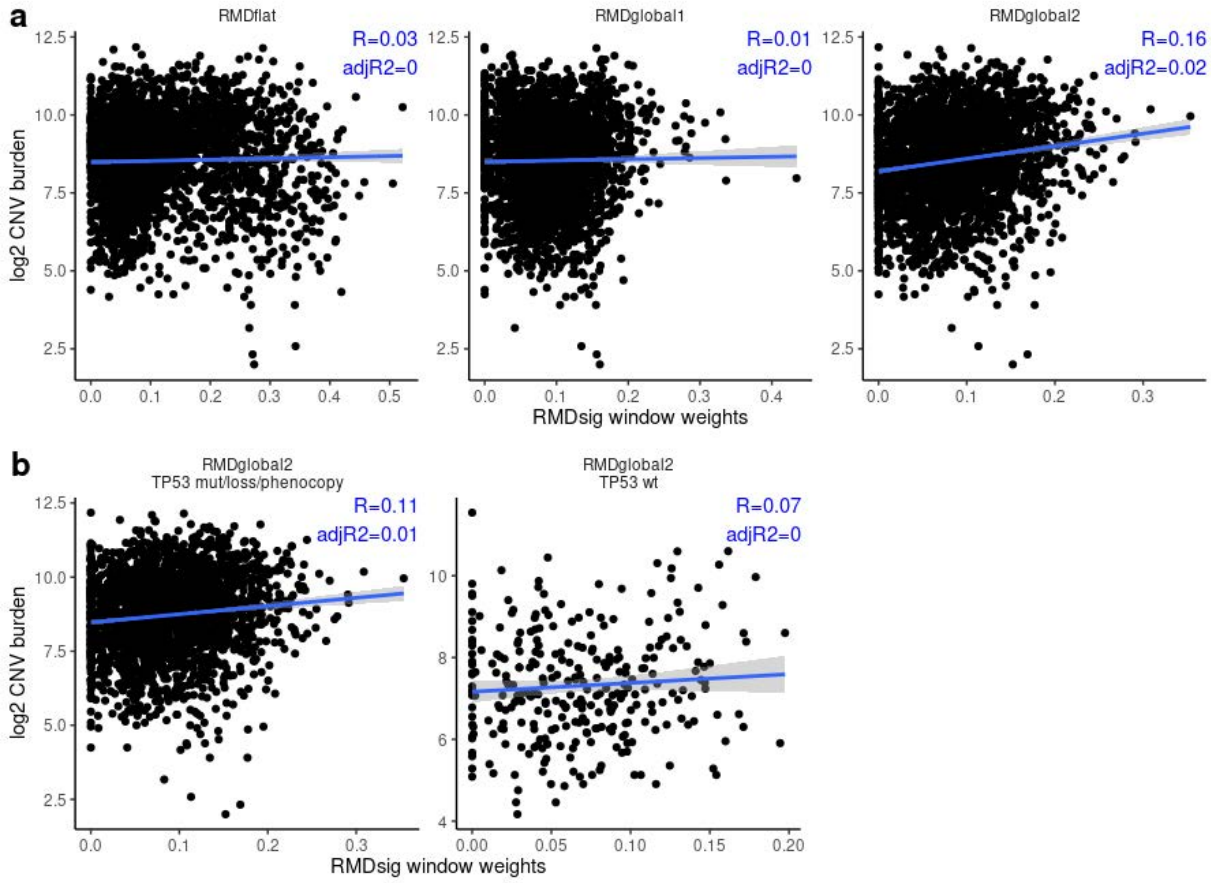


Fig S17. Burden of copy number variants does not explain global RMD signatures. a) Correlation between global RMD signatures activity and the CNV burden. **b)** Correlation between global RMDsignature2 and the CNV burden, upon stratifying by TP53 functional status.

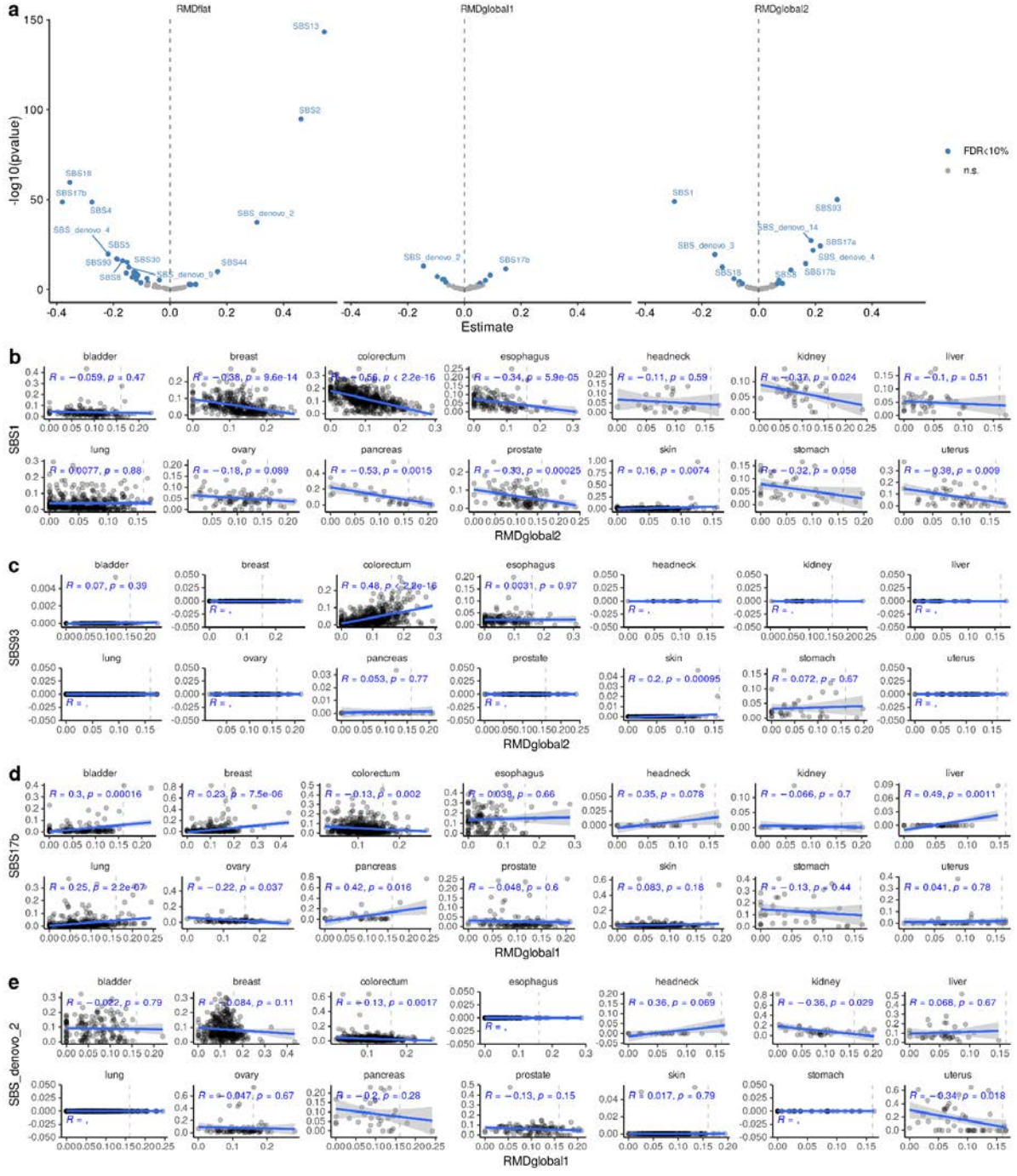


Fig S18. The association of the activity of the global RMD mutation risk redistribution signatures with the COSMIC trinucleotide mutational signatures. a) Associations between the 3 global RMD signature exposures and the trinucleotide (COSMIC) mutational signatures exposures. We included all 3 RMD signatures in a single regression when testing for associations, such as to have the associations of one RMD signature be adjusted for the other two RMD signatures. **b)** Correlations, broken down by cancer type, between RMDglobal2 and the SBS1 exposure (top positive hit). **c)** Correlations between RMDglobal2 and SBS93 (top negative hit). **d)** Correlations between RMDglobal1 and SBS17b (top negative hit). **e)** Correlations between RMDglobal1 and SBS_denovo_2 (top positive hit).

Suplementary Text S1. Homologous recombination-deficient tumors show lower regional mutation rate variability

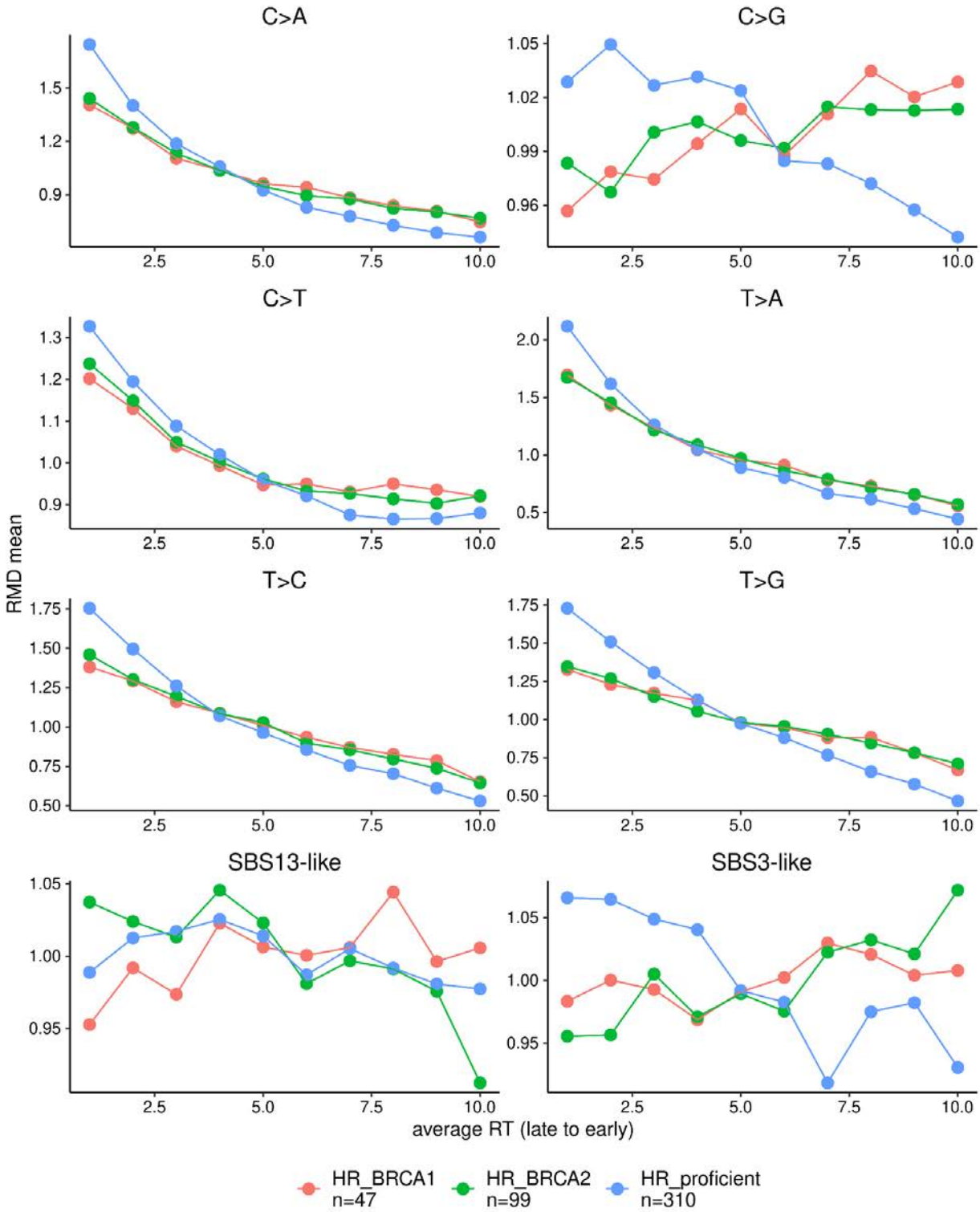
The RMDflat global signature we extracted from the NMF analysis captures the 'flat' distribution of mutations that was reported for MSI tumors, which are deficient in MMR (doi: 10.1038/nature14173). The window weights of RMDflat correlated with the average RT (Fig S6a), opposite of the canonical RMD landscape (which has few mutations in early-replicating DNA), thus the additive combination of the two results in a flat, low-variation landscape.

As expected, MSI samples showed high exposures to this RMD signature (Fig S6b), as well as bladder samples with mutations in the *ERCC2* gene, participating in the NER pathway (Fig S6b), consistent with previous reports (doi: 10.1038/nature14173 and 10.1038/ng.3557). In addition, we observed that tumor samples with high APOBEC signature mutagenesis also showed high exposures to the RMDflat signature (Fig S6b), solidifying prior reports of APOBEC mechanisms being enriched in early-replicating DNA, possibly via their association with DNA repair activity providing ssDNA substrate for APOBECs (doi: 10.1038/s41588-020-0674-6, 10.1101/gr.197046.115 and 10.7554/eLife.02001).

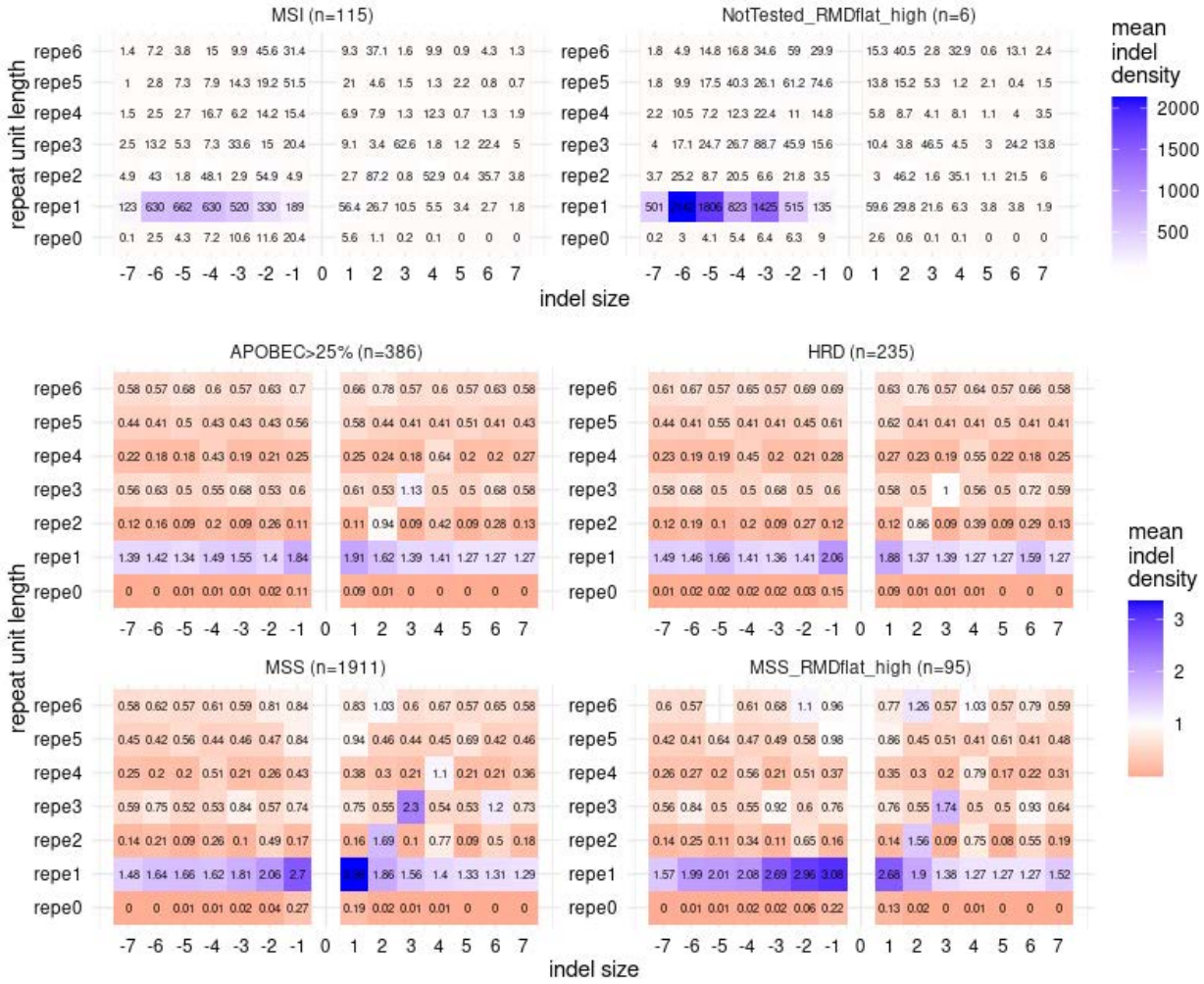
Based on these known associations involving MMR, NER, and APOBEC activity, we hypothesized that some of the remaining unexplained cases of RMDflat-high tumors (total 52% were explained) may be associated with deficiencies in another DNA repair pathway. In particular, we considered homologous recombination (HR) repair deficient samples, as ascertained by the CHORD method based on SNV and CNA (but not RMD) mutational signatures (doi: 10.1038/s41467-020-19406-4). The HR deficient samples also presented higher RMDflat exposures, both for BRCA1 and BRCA2 subtypes (Fig S6b). When HR is deficient, there is an increase in the spectrum of the trinucleotide mutational signature SBS3 (doi: 10.1038/s41467-020-19406-4 and 10.1038/nature12477) may result from activity of error-prone DNA polymerases (doi: 10.1038/s41467-021-27872-7). We observed that in HR-deficient tumor samples, the SBS3-like mutational spectrum [mutation types with high weights in SBS3, such as C>G mutations] accumulate more in early replicating DNA (i.e. opposite to canonical RMD pattern) (Supp Text Figure 1), thus contributing to the "flatness" of the RMD landscape.

Thus various DNA repair related mechanisms converge onto the RMDflat phenotype, with considerable variation in prevalence depending on the tissue: in colorectal tumors the main mechanism is the MMR deficiency, while in ovary and pancreas it is the HR deficiency, and APOBEC mutagenesis is the main mechanism in bladder and lung (Fig S6c).

For the final 28% of RMDflat tumor samples that are unexplained, we suggest this is unlikely to be due to false negatives in the MMR or HR deficiency tests, since these tumors had, on average, an indel spectrum (in microsatellite loci and elsewhere) not obviously different from the general indel spectrum of same cancer types (Supp Text Figure 2). Thus we suggest there are other mechanism(s) involved, for instance in kidney cancer there were many unexplained RMDflat signature samples, however this cancer type very rarely has known MMR, HR deficiencies or APOBEC mutagenesis; a possible explanation is a particular mutational process in kidney (doi: 10.1186/s13059-019-1892-z) that may evades DNA repair mechanisms operative in early-RT domains.



Supplementary Text S1 Figure 1. RMD association with RT for HR deficient and HR proficient tumors. Mutations are stratified by mutation types. SBS13-like are T(C>G)N mutations (corresponding to APOBEC mutagenesis), while SBS3_proxy are [CAG](C>T)[CTA] mutations (corresponding to the previously HR deficiency-associated mutations).



Supplementary Text S1 Figure 2. The indel mutational spectra for various groupings of tumor by the RMDflat global signature. As controls, the MSI tumors (upper left panel) show very high numbers of deletions of sizes -2 to -6, in the "repe1" category (repeats with repeat unit 1 nucleotide long i.e. homopolymer microsatellites), while the known MSS tumors (lower left panel) have considerably lower numbers thereof and have more bias towards -1 deletions and insertions. The MSS_RMDflat-high category of tumors, which are high in the RMDflat signature but not known to be MSI nor APOBEC-high nor HR-deficient, have an indel spectrum resembling that of the MSS tumors, suggesting that it is not likely that the RMDflat-high category often results from undetected MMR failures similar to those in MSI cancers. For the NotTested_RMDflat_high we have 6 samples that showed an indel spectrum similar to MSI suggesting they are MSI samples.

Chapter 4

Passenger mutations accurately classify human tumors

The following chapter has been selected from the publication:

M Salvadores, D Mas-Ponte and F Supek (2019). Passenger mutations accurately classify human tumors. PloS Comp Biol 15 (4), e1006953

Access: https://doi.org/10.1371/journal.pcbi.1006953

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Passenger mutations accurately classify human tumors

Marina Salvadores[1], **David Mas-Ponte**[1], **Fran Supek**[1,2]*

1 Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, Barcelona, Spain, **2** Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

* fran.supek@irbbarcelona.org

OPEN ACCESS

Citation: Salvadores M, Mas-Ponte D, Supek F (2019) Passenger mutations accurately classify human tumors. PLoS Comput Biol 15(4): e1006953. https://doi.org/10.1371/journal.pcbi.1006953

Editor: Maricel G. Kann, University of Maryland Baltimore County, UNITED STATES

Received: October 16, 2018

Accepted: March 15, 2019

Published: April 15, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: FS was funded by the ERC StG 757700 HYPER-INSIGHT (https://erc.europa.eu/) and by the MINECO grant BFU2017-89833-P (http://www.ciencia.gob.es/portal/site/MICINN/). We acknowledge funding from the Severo Ochoa Center of Excellence award (http://www.ciencia.gob.es/portal/site/MICINN/excellentinstitutions) to the IRB Barcelona. The funders had no role in

Abstract

Determining the cancer type and molecular subtype has important clinical implications. The primary site is however unknown for some malignancies discovered in the metastatic stage. Moreover liquid biopsies may be used to screen for tumoral DNA, which upon detection needs to be assigned to a site-of-origin. Classifiers based on genomic features are a promising approach to prioritize the tumor anatomical site, type and subtype. We examined the predictive ability of causal (driver) somatic mutations in this task, comparing it against global patterns of non-selected (passenger) mutations, including features based on regional mutation density (RMD). In the task of distinguishing 18 cancer types, the driver mutations–mutated oncogenes or tumor suppressors, pathways and hotspots–classified 36% of the patients to the correct cancer type. In contrast, the features based on passenger mutations did so at 92% accuracy, with similar contribution from the RMD and the trinucleotide mutation spectra. The RMD and the spectra covered distinct sets of patients with predictions. In particular, introducing the RMD features into a combined classification model increased the fraction of diagnosed patients by 50 percentage points (at 20% FDR). Furthermore, RMD was able to discriminate molecular subtypes and/or anatomical site of six major cancers. The advantage of passenger mutations was upheld under high rates of false negative mutation calls and with exome sequencing, even though overall accuracy decreased. We suggest whole genome sequencing is valuable for classifying tumors because it captures global patterns emanating from mutational processes, which are informative of the underlying tumor biology.

Author summary

Mutations accumulate throughout the lifetime of human somatic cells. While some may affect oncogenes or tumor suppressor genes and cause tumors–the 'driver' mutations–most are thought to be of no consequence. The density of such 'passenger' mutations across the human chromosomes is very uneven and is correlated with replication time and gene expression in the cell type the tumor had originated from. This property can be used to classify a tumor, assigning it to a tissue of origin and also the molecular subtype.

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

This is useful in cases of those metastatic cancers where the location of the primary tumor is unknown and is also of interest for the upcoming 'liquid biopsy' diagnostic approaches, where DNA is directly sequenced from bodily fluids to detect the presence of a cancer. The ability to type and subtype tumors is important to guide more detailed diagnostics and therapy, because the organ and the cell type which generated the tumor determines response to a variety of therapies, including targeted drugs.

Introduction

The rapid development of genomic techniques has brought considerable advances to the diagnosis and treatment of cancer. Most prominently, genetic variants in cancer genes can serve as markers for targeted therapeutics, in certain cases resulting in an impressive clinical response. Nevertheless, such cases are still not prevalent [1,2], and the cancer tissue-of-origin is also a major factor in deciding on therapeutic approaches [3]. For example, the common V600E mutation in the BRAF oncogene is a marker of the response to the BRAF V600E-targeting drug vemurafenib in melanoma. However, colorectal cancer bearing the same mutation does not respond to the drug [4], illustrating how both the particular oncogenic mutation and the tissue of origin are important for predicting the response to therapy. Consistently, large-scale drug screens on cancer cell lines suggest that drug response is often determined by gene expression patterns that stem from the cancer type [5], independently of the mutated oncogenes. The cancer type classification is being continually refined and extended [6]: driven by large-scale transcriptome, methylome and proteome analyses, new molecular subtypes are being proposed for various cancers [7–9]. Because these subtypes may have important clinical implications such as survival differences, it is advantageous that the cancer type and subtype are accurately determined for each patient. Genome sequencing provides an opportunity for development of approaches to meet this need, yielding cancer classifiers which may be useful to guide diagnostic work and decisions on treatment in certain scenarios. First, in approximately 3% of metastatic cancers the standard diagnostic procedure cannot identify the site of origin [10]. While there exist algorithms to classify such 'cancers of unknown primary' (CUP) based on gene expression [11,12] or DNA methylation [13], genome sequencing provides an opportunity to obtain independent predictions. Second, genomic classifiers of cancer type are relevant for liquid biopsies, where cell-free DNA or circulating tumor cells are retrieved from blood and DNA sequencing is performed. Recently, this approach was shown to hold much potential for screening the general population, or persons at risk for cancer [14,15]. After having determined that a tumor may be present, the genomic data from a liquid biopsy may be used to prioritize suspected primary sites, minimizing invasive diagnostic tests and reducing the time to therapy. Third, genomic classifiers may be valuable in the common case where the primary site of the tumor is known, because mutational patterns provide an additional means to determine the molecular subtype of the tumor. Standard subtyping is based on histopathology and immunochemistry tests and on gene expression panels. However, it is increasingly appreciated that integrating diverse omics data types leads to more robust subtyping [6,16], motivating research into how somatic mutation data may complement transcriptome, DNA methylation or proteome data for assigning a clinically relevant subtype to each tumor [17].

Therefore, a class of emerging approaches aims to classify cancers based on somatic mutations observed by comparing genomes of tumor and healthy tissue from the same patient. This genomic data is complex and features useful for classification need to be extracted from it. One important set of features describe the presence of specific driver mutations in a particular

tumor. These mutations are positively selected during tumor evolution because they promote growth of the mutated clones, meaning they are causal to carcinogenesis. Cancer driver genes affected by mutations are known to differ between tissues, where for instance the KRAS oncogene is often mutated in pancreatic, lung and colorectal cancer, but rarely in brain, breast and skin cancer. This provides a rationale for use of driver mutations in existing tumor classifiers [15,18].

In addition to a small number of driver mutations, each cancer contains orders of magnitude more passenger mutations, which are not selected during tumor evolution [19]. Because they are not shaped by selection, passenger mutation patterns provide a track record of the mutagenic processes that a tumor has undergone. These differ between tissue of origin, for instance melanoma has a very high proportion of C>T changes resulting from UV radiation [20]. More generally, it has been shown that considering the differential mutability of trinucleotides can provide a 'mutation signature' which results from exogenous and endogenous mutagenic exposures and, consistently, varies between cell types [21]. Indeed, trinucleotide and also pentanucleotide mutation frequencies were previously applied as predictive features to classify a tumor to a tissue-of-origin [22,23].

In this work, we examined the predictive utility of these existing features and evaluated a novel set of genomic features, based on the global patterns of passenger mutations: the regional mutation density (RMD). In particular, somatic mutation rates are known to exhibit striking variability at the megabase scale, wherein late-replicating, heterochromatic regions mutate faster due to reduced DNA repair [24–26]. The pattern of RMD, measured in megabase-scale chromosomal domains, is sufficiently variable across tumor types to allow their discrimination. In particular, the domains which change towards earlier replication timing and higher average gene expression in a tissue also mutate less in that tissue [26], and similarly so if their chromatin is more accessible [27]. While these associations do not reveal the causal factors underlying mutability, they nevertheless suggest that global RMD patterns emanating from thousands of passenger mutations are a useful marker for tissue of origin.

Therefore, we have systematically evaluated the ability of the passenger mutation-derived RMD features to classify 18 tumor types and subtypes thereof. Next, we asked if the predictions are complementary to those obtained by established passenger mutation-derived features, the trinucleotide mutation spectra (henceforth: MS96). Finally, we compared both types of passenger features–the RMD and the MS96 –with features describing the occurrence of specific driver mutations in a tumor (oncogenic mutations, OGM). Overall, passenger mutations are substantially more predictive than drivers in the task of classifying cancer type and subtype, and the RMD are an important component of a combined tumor type classifier based on global patterns of passenger mutations.

Results

Classification of tumors using global RMD features

We systematically evaluated whether cancer types can be classified using features describing regional mutation density (RMD), which are simply the normalized mutation counts across 2655 megabase-sized chromosomal domains. For this analysis we used a dataset which contains 18,863,479 mutations (SNVs and short indels) from whole-genome sequences (WGS) of 2267 tumor samples from 18 cancer types (S1A Fig and S1 Table). We calculated the RMD features and supplied them to an SVM classifier to generate 18 models that differentiate between each cancer type and the rest (One-vs-Rest scheme). To assess the accuracy of the models, we calculated the Receiver Operating Characteristic (ROC) curve and the area under the ROC curve (AUC) for each cancer type using five-fold crossvalidation (Fig 1A). Classification

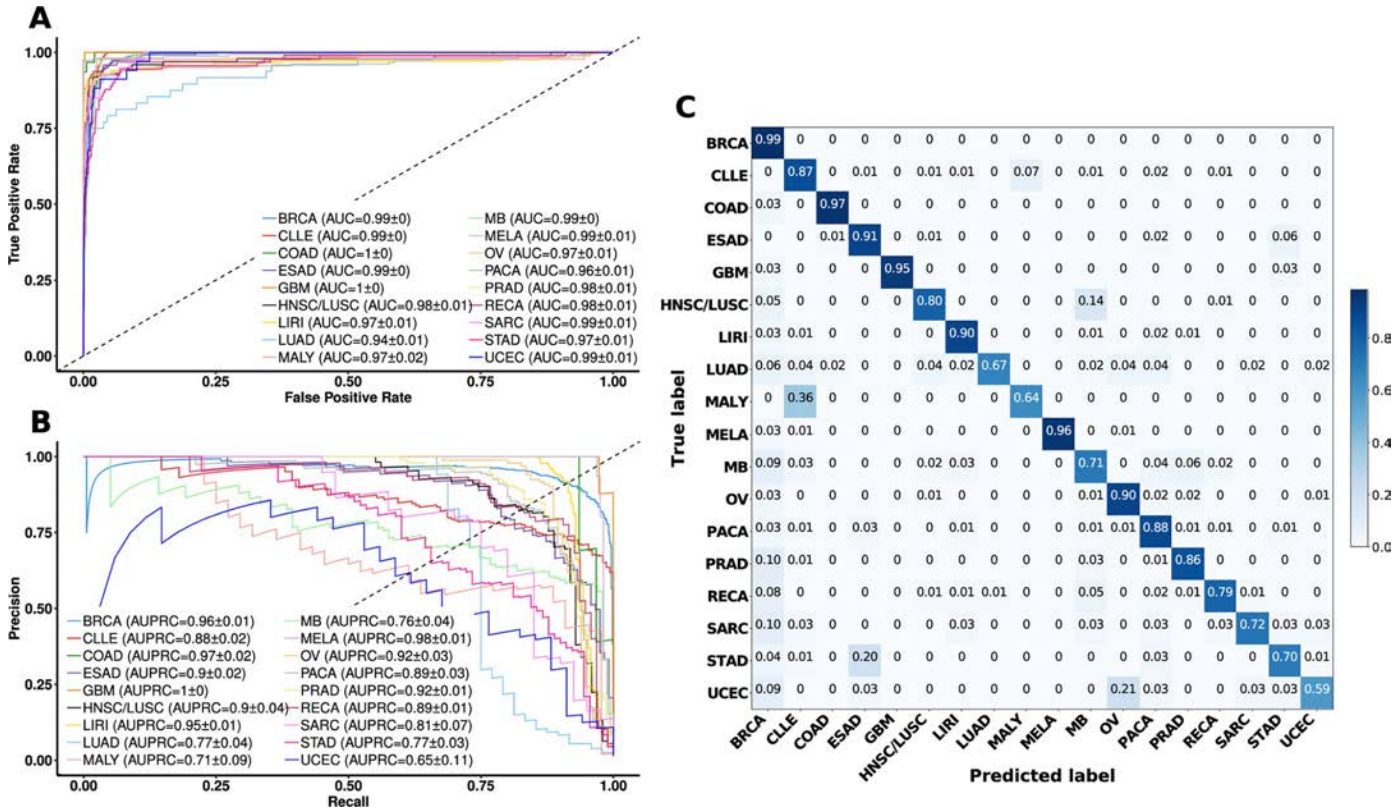


Fig 1. Accuracy of classifiers based on regional mutation density (RMD) in predicting cancer type. (A) Receiver Operating Characteristic (ROC) curve for each cancer type versus the rest. Area under the ROC curve mean and standard error of the mean across crossvalidation folds for each cancer type are reported in the legend. (B) Precision-recall (PR) curve for each cancer type versus the rest. Area under the PR curve mean and standard error of the mean across crossvalidation folds for each cancer type is reported in the legend. (C) Normalized confusion matrix, showing the fraction of misclassified tumors.

https://doi.org/10.1371/journal.pcbi.1006953.g001

models for 16 out of 18 cancer types showed a crossvalidation AUC≥0.97, and the two remaining cancer types ≥0.93, indicating RMD are indeed highly informative of cancer type; see S1 Table for a list of cancer type acronyms. While the AUC is a useful measure for comparing the accuracy between data sets with different class sizes (here, ranging from 37 glioblastoma to 560 breast cancer), in case an interpretable accuracy measure for imbalanced datasets is desired, the Area Under the Precision-Recall Curve (AUPRC) is preferable [28,29]. Our models yielded crossvalidation AUPRC scores ranging from 0.65 for UCEC to 1.00 for GBM cancer types (Fig 1B). In addition to AUC and AUPRC measures that integrate over a range of stringency thresholds, we also provide point estimates of precision, recall and F-score for each cancer type (S1B Fig). F-score, the harmonic mean of precision and recall, is very high in certain cancer types (≥0.90 in MELA, GBM, COAD, LIRI and BRCA), while it is lower for MALY (B-cell lymphoma) and UCEC (endometrioid uterine carcinoma), being 0.68 and 0.63, respectively. To investigate the cause for these differences, we examined the confusion matrix (Fig 1C) to ascertain which types of errors are being made. MALY and CLLE (chronic lymphocytic leukemia) were confounded by the models, consistent with both being blood cancers derived from B-cells. Next, STAD and ESAD are also confounded, which is consistent with the observation that the esophageal tumors close to the gastro-esophageal junction have molecular characteristics similar to gastric adenocarcinoma [30]. UCEC is the cancer type confounded with another gynecological cancer, OV (ovarian serous adenocarcinoma), suggesting our classifiers might

capture gender-specific features. Overall, many of the apparent errors of the RMD classification models reflect genuine tumor biology and not noise and/or biases which might stem from how the genomic data was obtained or processed.

To further investigate the known sources of bias, we evaluated if the models are influenced by batch effects (sequencing center, age and ancestry) by examining the first five principal components (PCs) of a PC analysis of the RMD features (S2–S4 Figs). The tumor samples tend to cluster by cancer type, but the samples from different sequencing centers, of different age and ancestry are intermixed (S2–S4 Figs) and do not present an obvious pattern. For these groups, we obtained similar accuracies when training on one group and testing on the others, *versus* training and testing in a mixture of the groups (S5 Fig), suggesting the grouping does not confound overall accuracy estimates. Additionally, we validated the models on external data sets, in order to investigate if overfitting is evident in the RMD-based predictive models. For this, we considered all cancer types represented by at least two datasets originating from different sources, distributing them in a secondary training set of 1749 tumor samples and an external validation set of 640 samples, covering 11 cancer types (S2 Table and S6A Fig). We then generated SVM models using the secondary training set and tested it (i) using crossvalidation and (ii) on the external validation dataset (S6B and S6C Fig and S5 Table). Overall, the crossvalidation AUC of 0.99 ± 0.02 matches the external validation set AUC of 0.99 ± 0.04 (median ± IQR across cancer types) implying no detectable overfitting (p-value = 0.22 for AUC difference, Wilcoxon test). Finally, we randomized the labels of the RMD classifiers and obtained AUC scores of approximately 0.5 (S7 Fig), further substantiating that the models do not fit to noise. In summary, the global patterns of passenger mutations captured by the RMD features provide a robust cancer type classifier.

Classification of cancer molecular subtypes using RMD

In addition to classifying cancer type, being able to identify the molecular subtype is important to guide further diagnostics, monitoring or treatment. Therefore, we asked whether the global patterns in RMD could be applied to tumor subtyping. To this end, we trained predictive models to distinguish subtypes for six major cancer types and evaluated them using crossvalidation in the same manner as the between-cancer type classifiers. For breast cancer, the classifier is able to separate the four molecular subtypes with an AUC>0.8 for every subtype and in particular the triple-negative subtype can be recognized more accurately with AUC = 0.92 (Fig 2A). For liver cancer, the model distinguishes biliary and hepatocellular tumors with AUC = 0.77 (S8D Fig). For stomach cancer, apart from its molecular subtypes (diffuse and intestinal) we also included the MSI status as a subtype because it is an indicator of response to cytotoxic chemotherapy such as 5-FU as well as immunotherapy [31]. All three subtypes are classified with AUC≥0.80, particularly the MSI subtype with AUC≈1 (S8F Fig), which is expected given the broad disruption of RMD previously observed in MSI cancers [26]. We note that the MSI subtype can be more directly observed by quantifying microsatellite indels in genomic data without necessity for RMD analysis, but it nonetheless provides a demonstration how RMD may be a useful tool for subtyping of tumors. Similarly to stomach cancer, we divided colorectal cancer by anatomical location (left *versus* right colon) and also based on the presence of MSI and POLE hypermutation, which have therapeutic relevance, yielding AUCs of 0.61 to 0.99 (Fig 2B). The modest AUC of 0.61 of left *versus* right colon suggests it might be possible to use RMD to classify intestinal tumors based on anatomical location, but given the small sample size this is not a significant result (p = 0.14, Mann-Whitney test on AUC>0.5). In case of head-and-neck squamous cell carcinoma and of melanoma, RMD clearly separated the cancer subtypes by anatomical location: oral *versus* non-oral groups (incl. alveolar ridge, larynx,

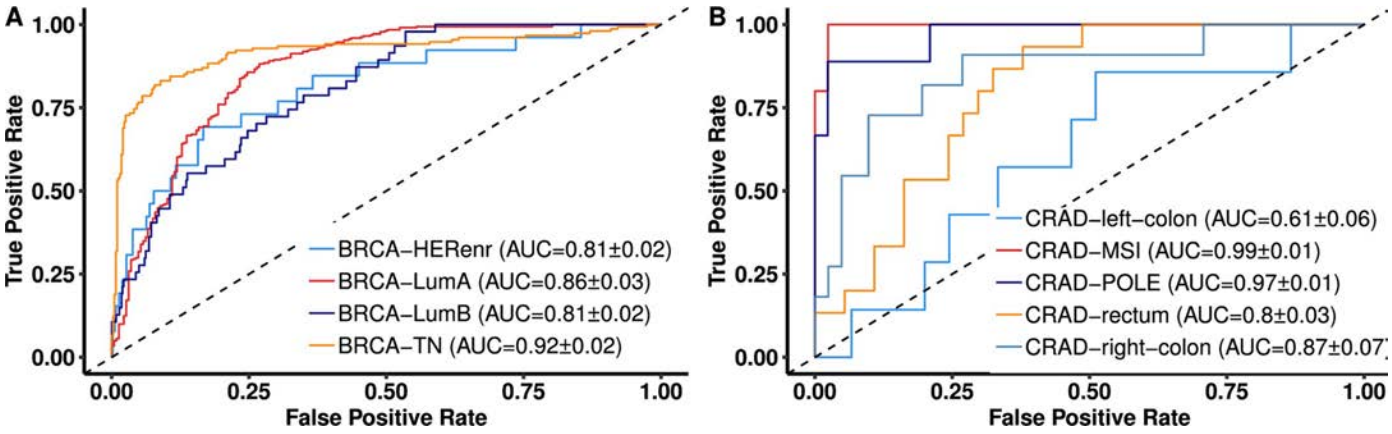


Fig 2. Accuracy of classifiers based on regional mutation density (RMD) in predicting cancer subtypes. Receiver Operating Characteristic (ROC) curves for discriminating each tumor subtype and/or anatomical location, versus the rest within that cancer type, shown for breast cancer (A) and colorectal cancer (B). See S8 Fig for additional cancer types. TPR and FPR calculated in five-fold cross validation.

https://doi.org/10.1371/journal.pcbi.1006953.g002

oropharynx and tonsil) at AUC = 0.89 and the three anatomical sites of melanoma at AUC $\geq$ 0.83 for each (S8E and S8F Fig).

In summary, the global mutational patterns across the genome are systematically different for the subtypes and anatomical locations within the same cancer type in a manner which can be recognized by RMD classifier.

Predictive power of driver *versus* passenger mutations

As described above, other genomic features have been reported to be able to classify cancer types [18,22,23]. These consist of features that are–similarly to RMD–derived from global patterns of passenger mutations, in particular the relative frequency of mutation types sorted by trinucleotide mutation spectra, herein referred to as MS96 (mutation spectrum with 96 components). In addition, the previously utilized features also include those derived from occurrence of cancer driver mutations (herein referred to as OGM, for oncogenic mutations). We turn to compare the accuracy of models based on novel RMD features with the models based on MS96 and based on OGM in their ability to distinguish cancer type, subtype and anatomical location.

First, we evaluated how each set of features performed individually in terms of the crossvalidation AUPRC (Fig 3A). Strikingly, the genomic features derived from global, genome-wide patterns of passenger mutations (RMD and MS96) performed notably better than the presence/absence of known driver mutations (OGM). In particular, passenger RMD yielded an AUPRC of 0.90 ± 0.13 (median ± IQR across 18 cancer types), passenger MS96 yielded 0.66 ± 0.40, while driver OGM only 0.21 ± 0.16, when using a Support Vector Machine classifier.

To ascertain this result is not specific to the SVM classifier, we repeated the above analyses by applying a Random Forest algorithm, which yielded crossvalidation AUPRC of 0.70 for RMD (median across cancer types), 0.77 for MS96 and 0.23 for driver OGM features. This supports the notion that RMD features are informative using various algorithms, but also highlights the utility of the SVM in producing most accurate models possible using this data. Moreover, the RMD features are more numerous than others, and we thus investigated if this affects the results. Upon removing correlated RMD features (see Methods) to reduce their number to 500 –thus matching the OGM–the predictive accuracy of the RMD remained high

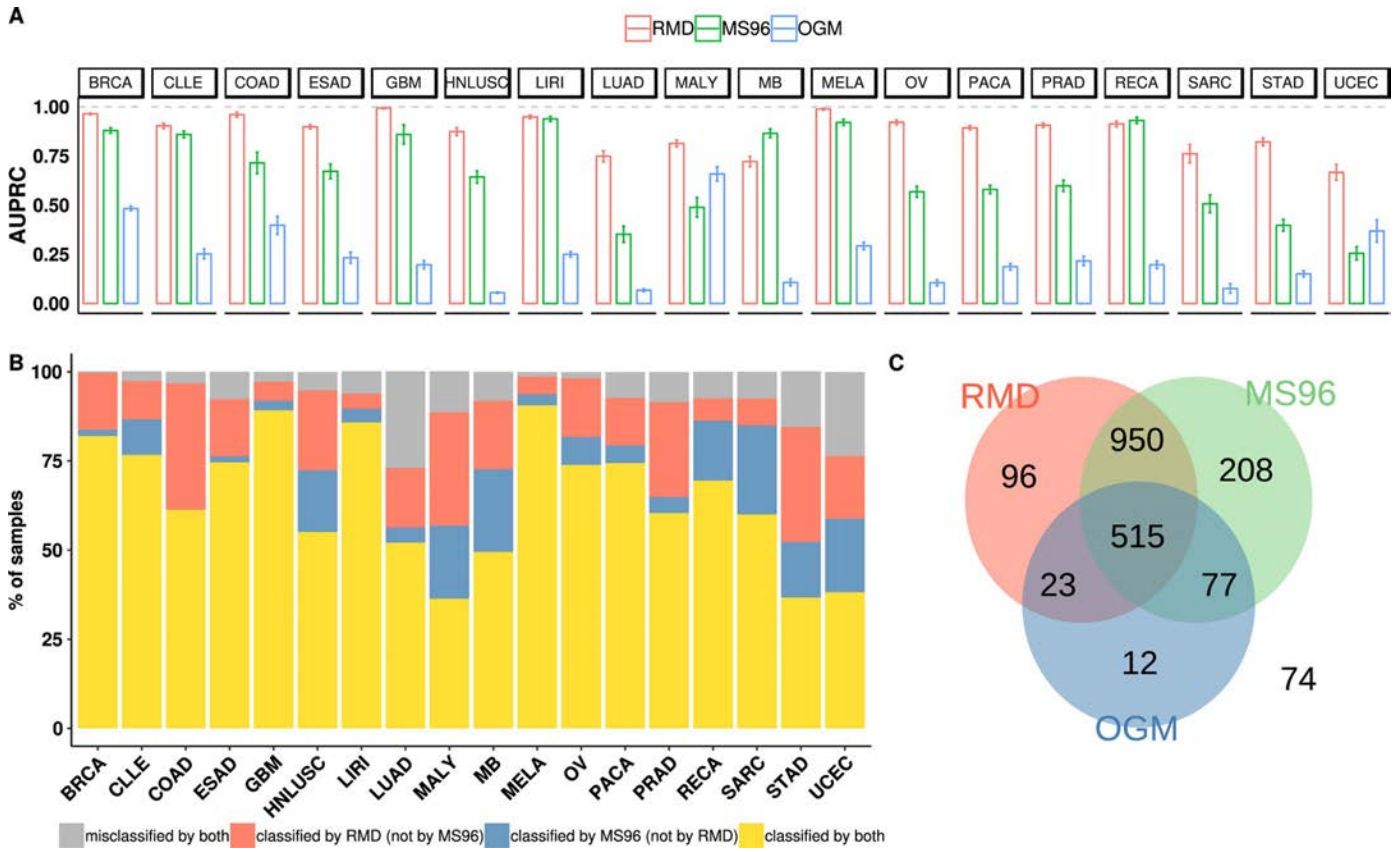


Fig 3. Predictive accuracy and complementarity of driver and passenger mutation features. (A) Mean Area Under the Precision-Recall Curve (AUPRC) for each cancer type, shown for classifiers derived from: regional mutation density (RMD, in red), trinucleotide mutation spectra (MS96, in green) and presence/absence of driver oncogenic mutations (OGM, in blue). Error bars are S.E.M. (B) Proportion of tumor samples that are: correctly classified by both RMD and MS96 (yellow), misclassified by both methods (gray), correctly classified by the MS96 but not by the RMD (blue) and vice versa (red). (C) Overlap between tumor samples correctly recognized by classifiers based on RMD, MS96 or OGM features independently.

https://doi.org/10.1371/journal.pcbi.1006953.g003

(median AUPRC = 0.80; S9 Fig), still exceeding MS96 and OGM. We supply the most highly ranking RMD, MS96 and OGM features for every cancer type in S7 Table. The use of feature selection prior to training the SVM classifiers did not improve the accuracy of the models (S10 Fig).

The fact that oncogenic driver mutations (OGM) were, in relative terms, poorly discriminative across cancer types suggests that no combination of oncogenes or tumor suppressor genes is uniquely informative of each individual cancer type, even though many cancer genes are well known to be enriched in certain groups of cancer types. We note that the classification algorithm we use (SVM using a Gaussian kernel) is capable of modeling statistical interactions between features, meaning it can in principle draw on epistatic effects between driver mutation occurrence that are cancer type-specific [32]. We recognize however that many cancer genes are mutated infrequently [33] or may be specific to some subtypes of a cancer type [7,8] and therefore the number of instances in the training dataset might be limiting for recovering complex patterns linking driver mutations and cancer site. We therefore examined the performance of the OGM features on a larger set of 6403 whole-exome sequences from the 15 cancer types matched to the original dataset (1971 WGS in those types). Indeed this did result in an improvement (AUPRC 0.54 ± 0.47 versus 0.24 ± 0.15, respectively, median ± IQR), but the accuracy nevertheless remained substantially lower than RMD (0.93) and MS96 (0.75) on

WGS in these 15 cancer types (S11 Fig). Finally, in the task of tumor subtyping the driver mutations were again less informative than passengers (S12A Fig), even though the difference is less striking than for cancer type classification. We have additionally tested classifiers that draw on refined OGM features that account for mutation impact, by: (i) thresholding or (ii) by weighting variants by CADD score [34], (iii) by using weighted pathway scores from the SAMBAR tool [35] and (iv) by restricting to the high-confidence set of driver mutations identified by Bailey *et al.* [33] (see Methods). Although this yielded slight gains in accuracy for some prioritization schemes, there were no significant overall improvements (best p = 0.11 for thresholding by CADD, S13 and S14 Figs). The OGM features had only modest accuracy in classifying tumors even after accounting for predicted mutation impact.

Complementarity between passenger mutations features

Given that the two types of passenger mutation features–RMD and MS96 –appear quite informative in discriminating cancer types, we wondered if the information in them is redundant, or instead if each of them is uniquely proficient in classifying a particular set of tumor samples. The correlation between AUPRC scores of RMD and MS96 across the 18 tissues is modest (R^2 = 0.24) suggesting that they might indeed be independently valuable in classifying tumors. For example, the RMD is more accurate in classifying colon, prostate and pancreatic cancer, while the MS96 is more proficient for kidney and medulloblastoma (S15 Fig).

Motivated by the above differences, we systematically evaluated whether the MS96 and the RMD are complementary in terms of individual tumor samples for which correct predictions can be obtained. To this end, we performed a classification with the RMD and the MS96 independently. Afterwards, we calculated how many of the misclassified tumor samples by the MS96 had been correctly classified by RMD and vice versa. Using the MS96 model, 468 samples out of 2264 were misclassified. Out of these 468, the majority (345) were correctly re-classified using RMD features (Fig 3B and 3C). On the other hand, starting with the RMD model, 287 samples out of 2264 were misclassified, and 164 out of the 287 samples could be correctly classified by applying the MS96 classifier (Fig 3B and 3C). We note that the two classifiers provided many complementary predictions even for tissues when overall accuracy was similar: for example, in uterine cancer RMD uniquely covered 17.6% of the predicted instances and MS96 20.9% of the instances (Fig 3B). This trend extends to cancer subtyping (S12B and S12C Fig).

This complementarity suggests that a combination of both types of passenger mutation features would be beneficial for making a joint classification model. Thus, we compared the accuracy of a baseline OGM driver model to models drawing on both the drivers and the MS96 and RMD passengers, considered alone and in combination (Fig 4). As expected, the baseline and the combination models presented very different AUPRC scores of 0.21 ± 0.16 and 0.96 ± 0.11, respectively (median ± IQR across 18 cancer types, $p<10^{-15}$, Wilcoxon test). To provide insight into practical implications of this observed increase in the AUPRC in the joint model, we compared the precision-recall curves for each cancer type (Fig 4A and S16 Fig) to estimate the number of cancer patients who would receive a confident diagnosis by the model after having introduced additional features. In particular, we estimated the *recall* score of the model–the proportion of all tumors of the relevant cancer type which was diagnosed–at the fixed *precision* of 0.8 (equivalent to false discovery rate, FDR = 20%) for that cancer type (grey line in Fig 4A and S16 Fig). Indeed for most cancer types a large increase in discovered cancer cases was enabled by use of the passenger features–first by introducing the MS96 and then by introducing both MS96 and RMD into a joint model. Notably, for ovarian cancer we observed an 8 percentage points (pp) increase in the number of patients correctly classified at FDR = 20% upon the addition of the MS96 (passengers) to the baseline OGM (drivers) and then a

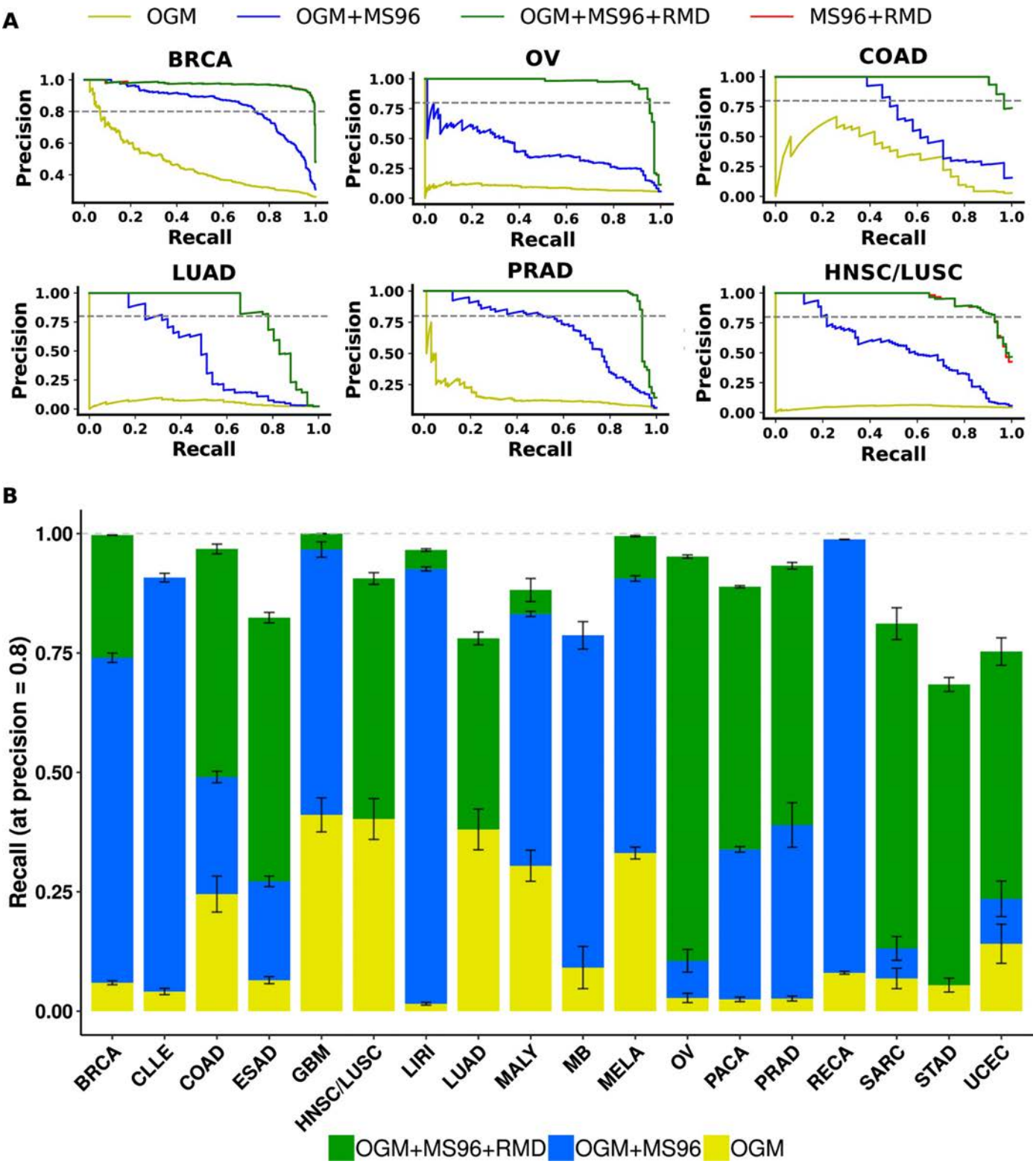


Fig 4. Combined classifiers provide substantially increased coverage with confident predictions. (A) Precision-Recall Curve for six example cancers (see S16 Fig for others) for the OGM driver mutations (yellow), for the combination of OGM and MS96 passenger mutations (blue), for the combination of MS96 and RMD (red) and for the combination of the OGM, MS96 and the RMD passenger mutations (green). Grey line indicates the threshold where precision = 0.8, equivalent to FDR = 20%.

(B) Recall at FDR = 20% for the classification models trained on: OGM (yellow), OGM+MS96 (blue) and OGM+MS96+RMD features (green); height of the stacked bars indicates the excess proportion of patients receiving correct predictions upon introducing additional features to the classification model. See S17 and S18 Figs for the same analysis using only features from passenger mutations.

https://doi.org/10.1371/journal.pcbi.1006953.g004

striking 85 pp increase upon further introducing the RMD passengers into the joint MS96 +OGM model. Lung and esophagus adenocarcinoma, squamous cell carcinoma group (LUSC/HNSC), sarcoma, pancreatic, prostate, stomach and uterine cancer showed remarkable increases of ≥50 pp tumors discovered at FDR = 20% by including the RMD in addition to other features, while colorectal adenocarcinoma and breast cancer show an increase of approximately 40 and 10 pp, respectively, from use of RMD. Even cancers with good accuracy for the OGM+MS96 classifier, such as brain, lymphocytic leukemia, liver and melanoma, showed improvements by including the RMD (Fig 4B, S16 Fig and S5 Table), resulting in an overall gain of 50 pp (median over 18 cancer types) at FDR = 20% by use of RMD.

In summary, the passenger mutation features including the trinucleotide mutation spectra (MS96) and the domain-scale mutation rates (RMD) provide unique, non-redundant information for classifying individual tumors and should therefore be used in combination to provide coverage for a maximum number of patients.

Mutation dropout and use of exome sequencing data

Whole-exome sequencing (WES) is becoming common among the diagnostic tests performed on cancer patients, offering cost savings over WGS while capturing most mutations known to be clinically relevant. Therefore, being able to predict cancer type of a metastatic tumor from this type of genomic data would be useful. Despite the fact that the exome only encompasses ~2% of the whole genome, the global passenger mutation features studied here (MS96 and RMD) are aggregate statistics across many genomic regions and thus have the potential to maintain the informative signal even with sparse data. Hence, we evaluated to which extent cancer type can be predicted from MS96 and the RMD features using exomes, by generating simulated WES data from the main WGS dataset and re-training the predictive models on such data. When using WES data, overall, a substantial loss of accuracy was observed for the passenger mutation features: only lymphoma, melanoma, colorectal and esophagus cancer show crossvalidation AUC≥0.9, and 10 of 18 cancer types dropped below AUC of 0.80 (Fig 5A).

Of note, the accuracy of the OGM driver features we used is generally not affected by moving from whole-genome to exome (there are minor differences due to low sequencing coverage in a certain number of exons; see Methods). This is because we consider only exonic coding mutations here, and additionally TERT promoter mutations, in accord with the latest estimates that non-coding cancer driver mutations appear to be rare [36]. Importantly, even this reduced accuracy of the passenger features (RMD and MS96) in WES is still higher than the accuracy of the driver mutations: 0.24 ± 0.35 (AUPRC median ± IQR across 18 cancer types) for RMD+MS96 passengers *versus* 0.17 ± 0.13 for OGM drivers in the simulated exomes. Therefore, for reduced representations of the genome, passenger mutation patterns still outperform drivers in classifying tumors. Finally, we evaluated the complementarity of the driver and the passenger features for the WES data and found that different classifiers are again able to classify the samples misclassified by the other method. In addition, greater proportions of tumors are uniquely correctly classified by passengers but not drivers, than by the drivers but not passengers (S19A Fig). Overall, we suggest that also for WES there is a tangible accuracy benefit from using combined classification models. In addition, we performed an external validation using the datasets from actual WES data from the TCGA consortium, while matching

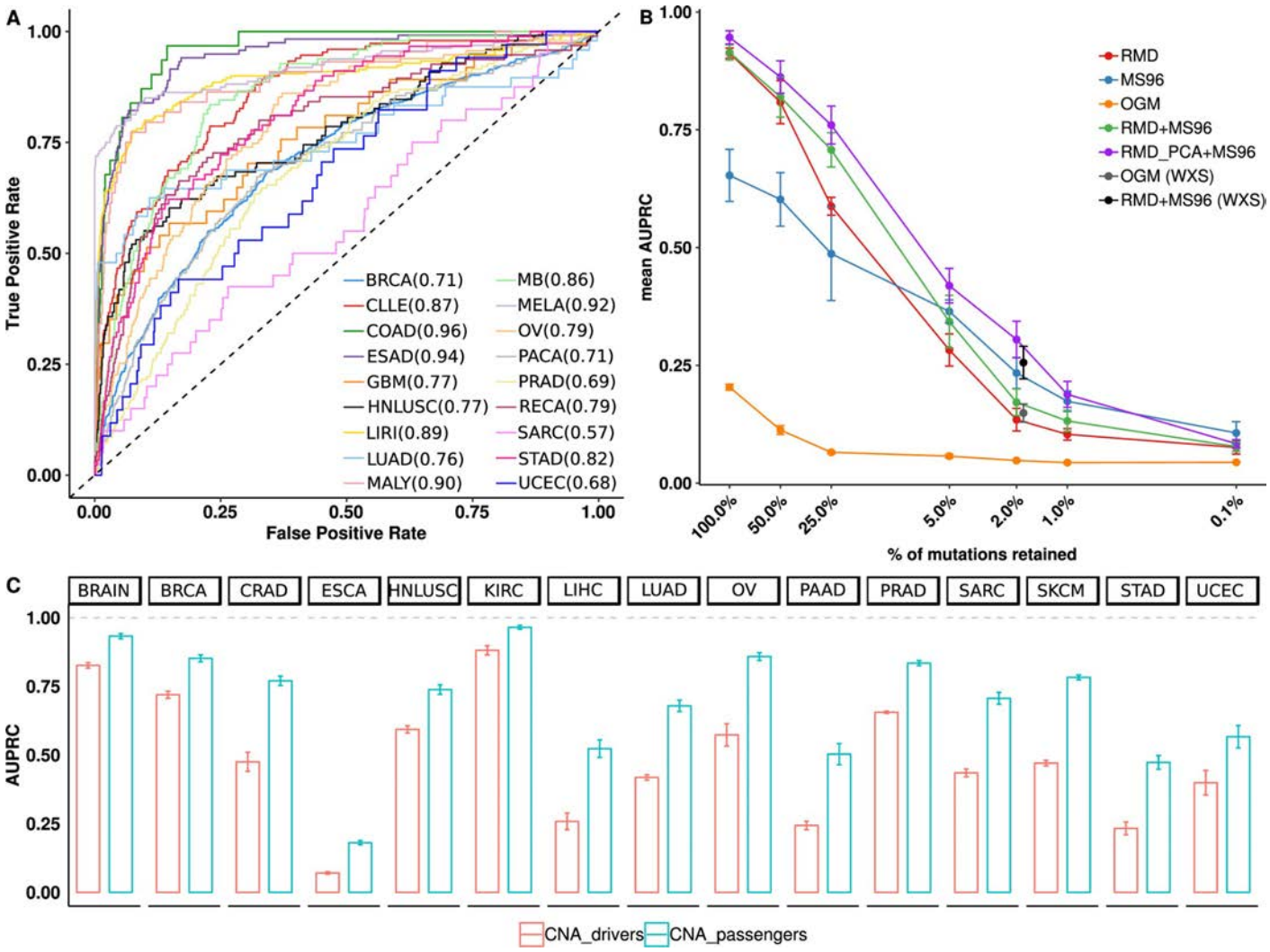


Fig 5. Predictive accuracy using data types other than whole-genome sequencing. (A) ROC curves for simulated whole-exome sequencing data, for one cancer type versus all others. Area under the ROC curve (mean across crossvalidation folds) for each cancer type is reported in the legend. (B) Crossvalidation AUPRC score for classifiers trained on various feature types under conditions of false negative mutation calls. (C) Crossvalidation AUPRC score for copy number states in 1Mb segments, estimated from SNP array data and tentatively divided into driver and passenger events. In (B) and (C), mean ± standard deviations are shown over multiple crossvalidation runs; (B) depicts the median score across all cancer types, and mean and s.d. thereof across crossvalidation runs are shown.

https://doi.org/10.1371/journal.pcbi.1006953.g005

the cancer types to those we have in the main training dataset. For both datasets, the AUPRC scores were similar (S19B and S19C Fig) and showed no significant difference in a Wilcoxon test for the driver features ($p = 0.27$) nor the passenger features ($p = 0.42$).

An important application of genomic classifiers of cancer type and subtype would be to liquid biopsies, wherein tumoral DNA can be extracted at low purity (in many cases at $<10\%$ [14]). Moreover, in practice WGS is often applied at low coverage, for reasons of cost efficiency. Both of these factors result in the decreased ability to call somatic mutations, which adds noise and would affect the performance of genomic classifiers such as those we employ. Motivated by this, we tested the performance under conditions of false negative mutation calls (dropout), generating simulated genomes with 50% to 99.9% of all mutations removed at random. Expectedly, mutation dropout lowers performance of all classifiers, yet the passenger-

mutation classifiers (RMD and MS96) (Fig 5B) are more robust. In particular, with 75% dropout, the RMD decreases 1.49-fold in its predictive performance (ratio of crossvalidation AUPRC), the MS96 decreases 1.24-fold, while driver mutations decrease 3-fold. This means that at this level the passenger mutations retain utility for classification (median AUPRC = 0.72 for RMD+MS96) while the drivers do not (median AUPRC = 0.07). At higher dropout (95%, 98%, 99% and 99.9% tested), the performance of all classifiers drops severely, but the relative advantage of passenger mutations is upheld (Fig 5B).

Copy number alteration-based tumor type classifiers

In addition to analyzing the distribution of mutations, here implying single-nucleotide variants (SNVs) and small indels, we briefly examined the genomic distribution of copy-number alterations (CNA), motivated by the previous use of CNA data for tumor type classification [18,37]. Upon tentatively dividing CNA into driver and passenger events based on whether they affect a dosage-sensitive cancer gene (Supplementary Methods in S1 Text), we observed that in 15 of 15 tested cancer types, the global pattern of passenger CNA is more predictive than the 68 individual driver CNA (Fig 5C). Of note, passenger mutation patterns (RMD +MS96) are more predictive than either type of CNA feature. We acknowledge that it is difficult to disentangle driver from passenger CNA and that our data likely reflects a mix of the two types; see Supplementary Results in S1 Text.

Discussion

There is a need for methods to detect tumor primary site, type and subtype in order to guide diagnostics and therapy. This is necessary for metastatic cancers of unknown primary (CUP) but also for screening the population-at-risk using emerging liquid biopsy techniques, wherein upon detection of tumoral DNA in bodily fluids, anatomical sites need to be prioritized. While there are established methods for tissue classification based on gene expression or DNA methylation in CUPs [11–13], genomic classifiers might provide an attractive alternative. This is because unlike the transcriptome and the epigenome, which diverge as the cells become malignant or when they metastasize to a new environment [38], the accumulated somatic mutations rarely revert. Moreover the tally of mutations arriving after tumorigenesis appears to be less numerous, compared to the mutations that had accumulated in the healthy cell before it turned cancerous [39]. In light of this, it is perhaps not surprising that the global patterns of passenger mutations, many arriving prior to cancer onset, accurately reflect the identity of the cell-of-origin. This is in contrast to the smaller number of driver mutations, which do affect cancer types differently to some extent but do not appear to possess sufficient information to distinguish the full diversity of cancers. In other words, mutational processes are more diverse between somatic cell types than are the selective pressures on oncogenic events, which tend to be shared across cancers.

A further attractive property of genomic classifiers is that mutations could plausibly be reliably detected in very impure samples–such as those originating from liquid biopsies–which might be challenging for extracting epigenomic or transcriptomic features. Given that approx. 30% of cancer patients were estimated to have enough circulating tumor DNA to make genome sequencing feasible [14], this opens the opportunity to use this genomic data for prioritizing cancer types, minimizing costly and invasive diagnostic tests and ensuring speedy access to therapy. Here, two important considerations are sequencing depth (coverage) and breadth (whole genomes *versus* exomes *versus* gene panels). Our simulation studies suggest that the passenger mutation patterns are to a certain extent robust to false negatives that might result from e.g. low sample purity or low-coverage sequencing, even though it remains to be

established how the number and distribution of false negatives scales with reduced coverage and purity in realistic settings. The robustness stems from the fact that RMD and MS96 features do not rely strongly on detecting individual mutations, which could be missed in a particular patient, but instead draw on aggregate statistics that describe global distributions of mutation types. Therefore, methods based on genome-wide mutation patterns have the potential to generate models robust to noise in mutation calls, with applications to low-purity or low-coverage DNA sequencing.

Similarly, we find that also in case of exome sequences, the global mutation patterns derived from passengers outperform the driver mutations. Nevertheless accuracy is overall compromised, suggesting that WGS is superior to WES for tumor classification because it better captures the global patterns in passenger changes. By extrapolation, we expect what is relevant for WES to be even more so for gene panel sequencing, which has been gaining traction as a diagnostic tool to prioritize patients for therapies. The panels cover coding regions of $\leq$500 genes [40] and, given high sequencing coverage, have excellent ability to detect driver mutations. However, it is very difficult for a gene panel to capture a pattern of mutation rate variability across chromosomal domains (here, RMD), because many mutations in the panel will be subject to strong positive selection. Since panel sequencing provides only the driver features, it would therefore have a limited ability to classify cancer types (approximated by our OGM results), because it is blind to global mutation patterns emanating from passenger mutations. In addition, while it may be possible to infer driver CNA from panel sequencing data [41] thus boosting predictions of cancer type, the accuracy of such approaches remains to be established.

Our work highlights the surprising accuracy of the novel RMD features for classifying tumors. The tissue-specific signal in the global distribution of mutations across genomic domains was suggested to stem from differential replication timing programs in the cell-of-origin [26] or from differential chromatin accessibility [27]. Given how closely correlated these two variables are, causality still remains to be established, but the resulting mutation pattern nevertheless constitutes a very useful marker of the cell-of-origin. We currently use a simple representation for RMD: normalized mutation counts for each 1 Mb-sized genomic window. Given the reported correlation between neighboring and also distant 1Mb windows [26], we speculate that a simpler representation of RMD, more robust to noise, might be feasible. Indeed we tested a PC analysis on the RMD, as introduced previously [26], in the tissue classification task and found that it modestly increases accuracy (Fig 5B, label "RMD_PCA+MS96"), particularly for more noisy data resulting from false-negative mutation calls.

While the predictive power of RMD for cancer type, subtype and anatomical location is substantial, we suggest that RMD are best used in tandem with another type of global mutation pattern–the trinucleotide mutation spectrum (MS96)–since each provides predictions for distinct sets of tumor samples. This fits well with the current understanding of mechanisms that underlie these patterns: MS96 are thought to reflect the mutagenic exposures–either exogeneous like UV light [20], or endogeneous like defective DNA repair [42], which vary between cell types [21]. The RMD on the other hand describes the epigenome organization in the cell-of-origin [26,27]. This is reflected in the mutations that accumulate with cell divisions, even without necessitating a particular mutagenic exposure. The RMD thus provide a track record of the normal functioning of the cell, and MS96 reflect the extraordinary circumstances it has encountered on its road to cancer, describing different aspects of the natural history of each specific tumor.

Our machine learning classifiers using RMD, interestingly, made ocasional mis-classifications which appeared to be informative of the underlying cancer biology. This suggests the utility of RMD for defining novel cancer subtypes or refining existing ones, similary to how

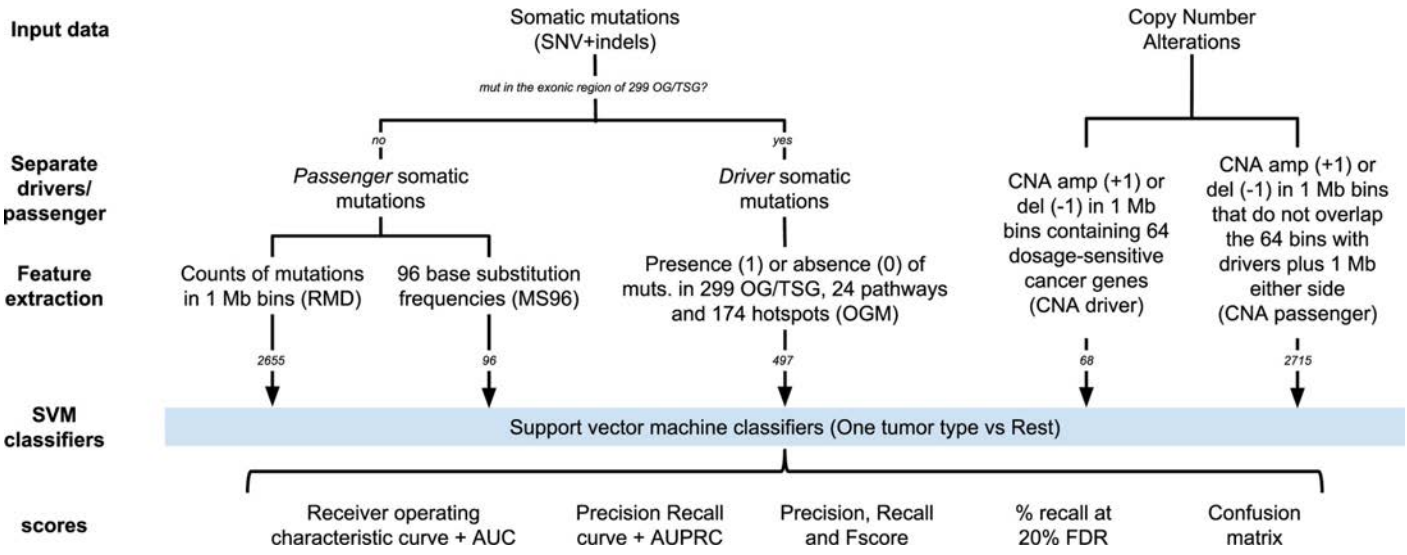


Fig 6. Schematic overview of the analyses.

https://doi.org/10.1371/journal.pcbi.1006953.g006

diverse omics data types reinforce and correct each other to yield robust molecular subtypes [6]. In this analysis, we have mainly focused in the application of RMD for identifying the tissue-of-origin of tumors, and known subtypes thereof. However, other clinical variables that are known to be associated with molecular subtypes–prognosis, propensity to metastasize and response to therapy–might be also possible to infer directly from RMD features, which remains as a direction for future work.

Materials and methods

Data collection and preparation

We collated whole-genome sequencing (WGS) datasets of tumor somatic mutations from diverse sources to create a main training dataset containing cancer types with at least 20 tumor samples (Supplementary Methods in S1 Text and S1 Table). These were further subdivided into secondary training and validation sets with matching tumor types selected from different data sources or sequencing centers (S2 Table) and into subtypes based on molecular features and/or anatomical location (S3 Table). We masked out all regions in the genome defined in the 'CRG Alignability 36' track [43] as having imperfect mappability (<1.0), retaining 1.9 Gb of the *hg19* genome assembly. Additionally, we simulated whole-exome sequencing (WES) data from the WGS datasets by observing the actual sequencing coverage for TCGA exomes; details in Supplementary Methods in S1 Text. A schematic overview of the methodology is given in Fig 6.

Genomic features calculation

To calculate regional mutation density (RMD) of passenger mutations, we filtered out all mutations in the coding regions of commonly mutated oncogenes (OG) and tumor suppressor genes (TSG) [33]. For each tumor, the RMD profile was calculated dividing each chromosome into 1 Mb windows and counting the number of mutations (SNVs and indels) in each window. Upon masking out regions of the genome with alignability <1.0, we normalized each count by dividing by the size of alignable regions (in Mb) in that window for WGS data, and by size of

the capture regions (in Mb) for WES data (see below). We discarded 242 windows with less than 100 kb with alignability = 1 for WGS data, and 1472 windows with less than 10 kb covered by the capture regions for WES data. Chromosomes X and Y were excluded. The final RMD vector consisting of 2655 features (WGS) or 1597 (WES) was normalized by the average number of mutations per window in that tumor. We also considered reduced-redundancy versions of the WGS based features, by performing a k-medoids clustering (*pam* function in R) on the genomic windows and selecting 500 medoids, and moreover we performed a PCA (*prcomp* function in R) and selected the first 100 principal components.

We obtained the mutation spectrum for each tumor by calculating the relative frequencies of six mutation types (considered DNA strand-symmetrically: C>G, C>T, C>A, A>T, A>G, and A>C) in every trinucleotide context, yielding a set of 96 mutational spectrum (MS96) features. The mutations in 299 cancer driver genes were excluded from calculation.

To create the matrix of driver mutations we first examined the list of 299 significantly mutated OG and TSG [33], which we matched to a set of 1744 UCSC transcripts (using the UCSC Table browser) that were further examined for occurrence of mutations. For each sample, we generated a binary vector reflecting whether the exonic regions ±5 flanking nucleotides of any transcript of a given gene contains a somatic mutation (SNV or indel) or not. Secondly, we checked the presence of mutations in 24 cancer-related pathways [33]. For each sample, we generated a binary vector describing whether the exonic regions ±5nt of any of the genes in a particular pathway harbor a somatic mutation (SNV or indel). Thirdly, we checked the presence of mutations in individual hotspots. We downloaded 'All Mutations in Census Genes' dataset from COSMIC [44] and selected as hotspot positions the coordinates where mutations were observed at least 100 times ±5nt. For each sample, we generated a binary vector describing whether each hotspot position was mutated, yielding 174 features. Finally, we concatenated the three binary matrices.

In addition, we generated six additional sets of refined OGM features to look into whether prioritizing the mutations according to functional impact improves ability to classify cancer type. First, to adjust for the influence of the overall tumor mutation burden and of gene length on the occurrence of mutations in cancer genes, we substituted the original pathway features for those obtained from the SAMBAR tool [35]. Secondly, to account for the functional impact of mutations, we stratified the mutations into putatively high and low impact using CADD scores [34], and created the OGM features only using putative high-impact mutations while ignoring the low-impact mutations, using two different stringency thresholds (>10 and >20; a mutation with a CADD score of 10 means that it is among the top 10% most impactful variants, and score of 20 top 1%). We additionally tested using CADD score directly as a weight, replacing the 0/1 indicator in the input data matrix. Thirdly, we generate the OGM features using only the ~3400 driver mutations reported by Bailey *et al.* [33] and additionally a set of features that represents each of these driver mutation loci individually as a feature (1 if mutation present, 0 if absent).

Training predictive models

For cancer type classification, we generated one model per cancer type to discriminate it from the rest of tumors of other cancer types pooled together (one-vs-rest). For training the model and assessing its performance, we used the models and functions in the *sklearn* library [45].

Machine learning algorithms. We applied the Support Vector Machine (SVM) algorithm for classification [46]. Briefly, SVM is a supervised machine learning method that searches for the hyperplane that maximizes the width of a margin between the instances of opposite classes. We used the SVC function of *sklearn.svm*, combined with OneVsRestClassifier function from

sklearn.multiclass module. To account for the class imbalance, we introduced class weights, calculated as the total number of samples divided by the number of classes multiplied by the number of samples of a class. In addition, the metaparameters of the SVM radial basis function kernel, *C* and *gamma*, were optimized by a grid search to maximize crossvalidation accuracy as recommended in the LibSVM best practices [47], for each training set separately. In addition, we applied the RandomForestClassifier function as implemented in *sklearn.ensemble*, where no hyperparameters were optimized and a forest size of 500 trees was trained.

Evaluating model accuracy. The performance of the model was assessed for each method by calculating the ROC curve and the Area Under the ROC Curve (AUC) for each cancer type using five-fold crossvalidation. Each AUC score was obtained in 10 runs of crossvalidation, and the mean, median, standard deviation (sd) and interquartile range (IQR) thereof were calculated. Similarly, we calculated the precision-recall (P-R) curve and the Area Under the P-R Curve (AUPRC) for each cancer type using five-fold crossvalidation. In addition, the precision, recall and F-score values were obtained 10 times and the previous statistics were calculated. Additionally, to quantify the improvement of the classification accuracy by introducing additional sets of features into a joint model, we calculated the precision-recall curve with five-fold cross validation, repeated 10 times. We determined the recall score at the precision = 0.8 (equivalent to 20% FDR) for the classifiers derived from OGM, from OGM+MS96, from OGM+MS96+RMD and from RMD+MS96 features. Precision and recall are TP/(TP+FP) and TP/(TP+FN), respectively, where TP is the number of true positives, FP false positives and FN false negatives. The mean and standard deviation of recall scores across crossvalidation runs were determined. The difference in recall can be interpreted as the percentage point (pp) increase of correctly classified patients due to the addition of a new set of features to the baseline set of features (OGM). In addition to crossvalidation, we also provide results on independent genomic data sets, which originate from a different study and/or sequencing center (see S2 and S4 Tables).

Feature importance calculation. We generated a list of most the informative features obtained by four different supervised feature selection methods. These are (i) Elastic Net regularized regression, (ii) Random Forest feature importance, (iii) the Relief-F method and (iv) Information Gain, which we applied to RMD classifiers and to OGM+RMD+MS96 classifiers for each cancer type. For the RMD classifiers, we reported the RMD features that were found to be in the top 25 consistently by two or more different methods. Additionally, for the OGM +RMD+MS96 classifiers we reported the top 25 features for each set of features in each cancer type, along with their rank within the combined classifier. Furthermore, we applied feature selection to the classifier during crossvalidation to examine effects of reduced feature sets on accuracy of models. For this, in each fold of the crossvalidation we fit an Elastic Net model in the training set and select the features whose coefficients in the EN were different from zero.

Supporting information

S1 Fig. Summary statistics of the main training dataset. (A) Distribution of the total number of mutations per sample for each cancer type. (B) Precision, Recall and F-score values for each cancer type. Each bar represents the mean of the corresponding score obtained from five independent runs (5-fold cross validation in each run) for each cancer type. Error bars represent the standard error of the mean for each cancer type.
(PDF)

S2 Fig. Analysis of possible batch effects: Sequencing center. Biplots for combinations of the 5 first principal components (PCs) for the RMD features, in a subset of cancer types with at least 5 samples per annotation group. Scree plot shows variance explained by each PC. The

above-diagonal part of the scatterplot is colored by cancer type. The below-diagonal part of the scatterplot is colored by sequencing center.
(PDF)

S3 Fig. Analysis of possible batch effects: Age. Biplots for combinations of the 5 first principal components (PCs) for the RMD features, in a subset of cancer types with at least 5 samples per annotation group. Scree plot shows the variance explained by each PC. The above-diagonal part of the scatterplot is colored by cancer type. The below-diagonal part of the scatterplot is colored by age (patients stratified in old and young groups according to whether their age is greater or lower than the median for each cancer type respectively).
(PDF)

S4 Fig. Analysis of possible batch effects: Ancestry. Biplots for combinations of the first 5 principal components (PCs) for the RMD features, in a subset of cancer types with at least 5 samples per annotation group. Scree plot shows the variance explained by each PC. The above-diagonal part of the scatterplot is colored by cancer type. The below-diagonal part of the scatterplot is colored by the ancestry group of the patient.
(PDF)

S5 Fig. Analysis of possible batch effects: Ancestry and age. (A) Area Under the Precision Recall curve (AUPRC) for RMD features in a subset of three cancer types training in samples from one group (european) and testing on another group (pooled asian, african and Hawaiian) in red; and training and testing in a mixture of groups maintaining the proportions between training and testing sets in blue. (B) AUPRC for RMD features in a subset of 3 cancer types training in samples from one group (old—age above the median of the cancer type) and testing on another group (young—age below the median of the cancer type) in red; and training and testing in a mixture of groups maintaining the proportions between training and testing sets in blue.
(PDF)

S6 Fig. Description of the secondary training and external validation datasets. (A) Distribution of the total number of mutations per sample for each cancer type of the secondary training dataset (red) and the external validation dataset (blue). (B) For RMD features AUC mean of 5 independent runs obtained by training in the secondary training dataset and testing in the external validation dataset (blue) and by crossvalidation in the secondary training dataset (red) for each cancer type. Error bars represents the standard error of the mean of each cancer type. (C) For RMD features AUPRC mean of 5 independent runs obtained by training in the secondary training dataset and testing in the external validation dataset (blue) and by crossvalidation in the secondary training dataset (red) for each cancer type. Error bars represents the standard error of the mean of each cancer type.
(PDF)

S7 Fig. Evaluation of RMD with randomized labels. Area Under the ROC curve (AUC) for each cancer type calculated with RMD classifiers with the class labels (cancer type) randomized. Each dot corresponds to one independent run (10 runs in total).
(PDF)

S8 Fig. Evaluation of subtype classifiers performance. Receiver Operating Characteristic (ROC) curve for each subtype versus the rest of samples, in breast cancer (A), colorectal cancer (B), head and neck squamous cell adenocarcinoma (C), liver cancer (D), melanoma (E) and stomach cancer (F) datasets. Area Under the ROC curve (AUC) reported (mean and standard

error of the mean across 5-fold cross validation rounds) in the legend of each panel.
(PDF)

S9 Fig. Evaluation of predictive accuracy of RMD features after dimensionality reduction. (A) Mean Area Under the ROC Curve (AUC) for each cancer type for the full set regional mutation density (RMD) features in yellow and the reduced set of 500 RMD features (by k-medoid clustering) in blue. (B) Mean Area Under the Precision Recall Curve (AUPRC) for each cancer type for the full set of RMD features in yellow and a reduced set of 500 RMD features (by k-medoid clustering) in blue. Error bars are the standard error of the mean.
(PDF)

S10 Fig. Evaluation of RMD with and without feature selection. Area Under the Precision Recall Curve (AUPRC) for the main training dataset for each cancer type with all RMD features (in green); and applying feature selection to each training subset using Elastic Net and testing with those selected features within each fold of the crossvalidation (in orange). Each dot represents one cancer types (median AUPRC across the 3 folds of the crossvalidation, distinguishing that cancer type versus the rest of cancer types).
(PDF)

S11 Fig. Evaluation of model accuracy using passenger *versus* driver mutation features. Mean Area Under the Precision Recall Curve (AUPRC) scores for each cancer type in the main training dataset, using regional mutation density (RMD) features in red, and 96 mutation spectra (MS96) features in green. The presence/absence of mutations in cancer genes (OGM features) was determined for a dataset of real WES with equivalent cancer types (shown in blue).
(PDF)

S12 Fig. Comparison and complementarity analysis for subtypes using RMD features. (A) Mean Area Under the Precision Recall Curve (AUPRC) scores for each cancer type for the subtypes of 6 major cancer types, using different sets of features: regional mutation density (RMD) in red, 96 mutation spectra (MS96) in green and presence/absence of oncogenic mutations (OGM) in blue. (B) For the subtypes datasets of six cancer types % of samples that are: (i) correctly classified by both RMD and MS96 (yellow), (ii) misclassified by both methods (gray), (iii) correctly classified by the MS96 but not by the RMD (blue) and (iv) correctly classified by the RMD but not by the MS96 (red). (C) Venn diagram of samples correctly classified by MS96, RMD or OGM features and their intersections for the subtypes classification.
(PDF)

S13 Fig. Evaluation of features that account for mutation impact. (A) Area Under the Precision Recall curve (AUPRC) for five sets of oncogenic mutation (OGM) features (see *Methods*) on WGS data. P-value reported for each set of features compares with the default OGM features as a baseline, using one-tailed Wilcoxon signed rank test (with alternative set to "less" in the R function *wilcox.test*). (B) Area Under the Precision Recall curve (AUPRC) for three different sets of OGM features (see *Methods*) in a subset of 560 patients (only considering genomes with at least one mutation from the Bailey *et al.* list). P-values for each set of features compare with the default OGM features as a baseline one, using a statistical test as in (A).
(PDF)

S14 Fig. Evaluation of features that account for mutation impact. (A) Area Under the Precision Recall curve (AUPRC) for five sets of oncogenic mutation (OGM) features (see *Methods*) on WES data. P-value reported for each set of features compares against the baseline OGM features using a Wilcoxon signed rank test, one-tailed (alternative set to "less" in the R function

wilcox.test).
(PDF)

S15 Fig. Comparison of accuracy of RMD and MS96 features per cancer type. Mean Area Under the Precision Recall curve (AUPRC) score of the RMD features (x axis) versus MS96 features (y axis) of the main training dataset, in crossvalidation. Error bars are the standard deviation of AUPRC for RMD features (x axis) and MS96 features (y axis).
(PDF)

S16 Fig. Gains in classifier accuracy by introducing additional sets of features. Precision-Recall Curve for each cancer type for the OGM (yellow), for the combination of OGM and MS96 (blue) and for the combination of the OGM, MS96 and the RMD (green), and MS96 and RMD without OGM (red) features for the main training dataset. In most cancer types the red curve overlaps the green curve perfectly and is thus hidden on the plots. Grey line indicates the threshold where precision = 0.8.
(PDF)

S17 Fig. Improvement in accuracy of RMD-based classifiers by the addition of MS96 features. Precision-Recall Curve for each cancer type for the RMD (blue) and for the combination of RMD and MS96 (purple) for the main training dataset. Grey line indicates the threshold where precision is equal to 0.8
(PDF)

S18 Fig. Improvement of RMD classifiers by the addition of MS96 features. Recall at FDR = 20% for the classification models trained on RMD (blue) and RMD+MS96 features (purple); height of the stacked bars indicates the excess proportion of patients receiving correct predictions upon introducing the additional features to the classification model. Bars show the mean of five cross-validation runs, and error bars are standard deviations.
(PDF)

S19 Fig. Complementarity and external validation of features with simulated WES datasets. (A) For the main training dataset of simulated WES, fraction of samples that are: (1) correctly classified by both the passengers (RMD+MS96) and the drivers (OGM) (yellow), (2) misclassified by both methods (gray), (3) correctly classified by the passengers but not by the drivers (orange) and (4) correctly classified by the drivers but not by the passengers (green). (B) For RMD+MS96 features dataset, mean AUPRC of five classification runs obtained by training on the simulated WES dataset and testing on the real WES dataset as an external validation (blue) and by crossvalidation in the secondary training dataset (red) for each cancer type. Error bars represents the standard error of the mean of each cancer type. (C) As above, but for OGM features. Error bars represents the standard error of the mean of each cancer type.
(PDF)

S1 Table. Description of the datasets that form the main training dataset.
(XLSX)

S2 Table. Paired datasets from different sources or sequencing centers for external validation. In column 1 the cancer type, in column 2 and 3, the dataset name, the number of samples in brackets and the source.
(XLSX)

S3 Table. Description of the subtypes of the six major cancer types and their source.
(XLSX)

S4 Table. Description of the data used for simulated WES and real WES from TCGA.
(XLSX)

S5 Table. Comparison of AUC score for secondary training dataset in crossvalidation (AUC crossvalidation) and AUC score training in secondary training dataset and testing in External validation dataset (AUC external validation). Refer to S2 Table for a detail description of each dataset.
(XLSX)

S6 Table. Mean Recall at precision 0.8 for different combination of features for the main training dataset.
(XLSX)

S7 Table. Most informative features for tissue classification. (A) Top 25 features ranked by 4 different suppervised feature selection methods for RMD classifiers for each cancer type. (B) Top 25 features ranked by 4 different suppervised feature selection methods for RMD classifiers for each subtype (separate classifiers for each cancer type). (C) Top 25 features ranked by 4 different suppervised feature selection methods for RMD+MS96+OGM classifiers for each cancer type. In each tab the top 25 features for each set of features are reported, along with their rank in the combined classifier.
(XLSX)

S1 Text. Methods: Collection and preparation of genomic data, obtaining copy number alteration data.
(PDF)

Acknowledgments

The results published here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/ and other members of the ICGC initiative. We thank Eduard Porta for sharing the list of cancer driver mutations, and the members of the Genome Data Science group at the IRB Barcelona for discussions and support.

Author Contributions

Conceptualization: Marina Salvadores, David Mas-Ponte, Fran Supek.

Data curation: Marina Salvadores, David Mas-Ponte.

Formal analysis: Marina Salvadores.

Funding acquisition: Fran Supek.

Investigation: Marina Salvadores.

Methodology: Marina Salvadores, David Mas-Ponte, Fran Supek.

Project administration: Fran Supek.

Supervision: Fran Supek.

Validation: Marina Salvadores.

Visualization: Marina Salvadores.

Writing – original draft: Marina Salvadores, Fran Supek.

Writing – review & editing: Marina Salvadores, David Mas-Ponte, Fran Supek.

References

1. Prasad V. Perspective: The precision-oncology illusion. Nature. 2016; 537: S63–S63. https://doi.org/10.1038/537S63a PMID: 27602743
2. Le Tourneau C, Delord J-P, Gonçalves A, Gavoille C, Dubot C, Isambert N, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. Lancet Oncol. Elsevier; 2015; 16: 1324–34. https://doi.org/10.1016/S1470-2045(15)00188-6 PMID: 26342236
3. Greco FA. Molecular Diagnosis of the Tissue of Origin in Cancer of Unknown Primary Site: Useful in Patient Management. Curr Treat Options Oncol. Springer US; 2013; 14: 634–642. https://doi.org/10.1007/s11864-013-0257-1 PMID: 23990214
4. Kopetz S, Desai J, Chan E, Hecht JR, O'Dwyer PJ, Maru D, et al. Phase II Pilot Study of Vemurafenib in Patients With Metastatic BRAF-Mutated Colorectal Cancer. J Clin Oncol. American Society of Clinical Oncology; 2015; 33: 4032–8. https://doi.org/10.1200/JCO.2015.63.2497 PMID: 26460303
5. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell. Elsevier; 2016; 166: 740–754. https://doi.org/10.1016/j.cell.2016.06.017 PMID: 27397505
6. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. Elsevier; 2014; 158: 929–944. https://doi.org/10.1016/j.cell.2014.06.049 PMID: 25109877
7. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. Nature Publishing Group; 2012; 487: 330–337. https://doi.org/10.1038/nature11252 PMID: 22810696
8. The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. Nature Publishing Group; 2015; 517: 576–582. https://doi.org/10.1038/nature14129 PMID: 25631445
9. The Cancer Genome Atlas Network. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. Cell Press; 2017; 169: 1327–1341.e23. https://doi.org/10.1016/j.cell.2017.05.046 PMID: 28622513
10. Fizazi K, Greco FA, Pavlidis N, Daugaard G, Oien K, Pentheroudakis G. Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2015; 26: 133–138. https://doi.org/10.1093/annonc/mdv305 PMID: 26314775
11. Weiss LM, Chu P, Schroeder BE, Singh V, Zhang Y, Erlander MG, et al. Blinded comparator study of immunohistochemical analysis versus a 92-gene cancer classifier in the diagnosis of the primary site in metastatic tumors. J Mol Diagn. Elsevier; 2013; 15: 263–9. https://doi.org/10.1016/j.jmoldx.2012.10.001 PMID: 23287002
12. Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, et al. MicroRNAs accurately identify cancer tissue origin. Nat Biotechnol. 2008; 26: 462–469. https://doi.org/10.1038/nbt1392 PMID: 18362881
13. Moran S, Martínez-Cardús A, Sayols S, Musulén E, Balañá C, Estival-Gonzalez A, et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. Lancet Oncol. 2016; 17: 1386–1395. https://doi.org/10.1016/S1470-2045(16)30297-2 PMID: 27575023
14. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. Nature Publishing Group; 2017; 8: 1324. https://doi.org/10.1038/s41467-017-00965-y PMID: 29109393
15. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science (80-). 2018; 359: 926–930. https://doi.org/10.1126/science.aar3247 PMID: 29348365
16. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. Elsevier; 2018; 173: 291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022 PMID: 29625048
17. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. Nature Publishing Group; 2013; 10: 1108–1115. https://doi.org/10.1038/nmeth.2651 PMID: 24037242
18. Marquard AM, Birkbak NJ, Thomas CE, Favero F, Krzystanek M, Lefebvre C, et al. TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. BMC Med Genomics. BioMed Central; 2015; 8: 58. https://doi.org/10.1186/s12920-015-0130-0 PMID: 26429708

19. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. Elsevier; 2017; 171: 1029–1041.e21. https://doi.org/10.1016/j.cell.2017.09.042 PMID: 29056346

20. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat J-P, et al. A Landscape of Driver Mutations in Melanoma. Cell. Cell Press; 2012; 150: 251–263. https://doi.org/10.1016/j.cell.2012.06.024 PMID: 22817889

21. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., et al. Signatures of mutational processes in human cancer. Nature. Nature Publishing Group; 2013; 500: 415–421. https://doi.org/10.1038/nature12477 PMID: 23945592

22. Temiz NA, Donohue DE, Bacolla A, Vasquez KM, Cooper DN, Mudunuri U, et al. The somatic autosomal mutation matrix in cancer genomes. Hum Genet. Springer; 2015; 134: 851–64. https://doi.org/10.1007/s00439-015-1566-1 PMID: 26001532

23. Cario CL, Witte JS, Hancock J. Orchid: a novel management, annotation and machine learning framework for analyzing cancer mutations. Hancock J, editor. Bioinformatics. Oxford University Press; 2018; 34: 936–942. https://doi.org/10.1093/bioinformatics/btx709 PMID: 29106441

24. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012; 488: 504–507. https://doi.org/10.1038/nature11273 PMID: 22820252

25. Zheng CL, Wang NJ, Chung J, Moslehi H, Sanborn JZ, Hur JS, et al. Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. Cell Rep. NIH Public Access; 2014; 9: 1228–34. https://doi.org/10.1016/j.celrep.2014.10.031 PMID: 25456125

26. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. Nature Publishing Group; 2015; 521: 81–84. https://doi.org/10.1038/nature14173 PMID: 25707793

27. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. Nature Publishing Group; 2015; 518: 360–364. https://doi.org/10.1038/nature14221 PMID: 25693567

28. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. Public Library of Science; 2015; 10: e0118432. https://doi.org/10.1371/journal.pone.0118432 PMID: 25738806

29. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves [Internet]. Available: https://www.biostat.wisc.edu/~page/rocpr.pdf

30. The Cancer Genome Atlas Network. Integrated genomic characterization of oesophageal carcinoma. Nature. Nature Publishing Group; 2017; 541: 169–175. https://doi.org/10.1038/nature20805 PMID: 28052061

31. Jo W-S, Carethers JM. Chemotherapeutic implications in microsatellite unstable colorectal cancer. Cancer Biomark. NIH Public Access; 2006; 2: 51–60. Available: http://www.ncbi.nlm.nih.gov/pubmed/17192059 PMID: 17192059

32. Park S, Lehner B. Cancer type-dependent genetic interactions between cancer driver alterations indicate plasticity of epistasis across cell types. Mol Syst Biol. European Molecular Biology Organization; 2015; 11: 824. https://doi.org/10.15252/msb.20156102 PMID: 26227665

33. Bailey HM, Tokheim C, Porta-Pardo E, Mills GB, Karchin R, Ding L, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018; 173: 371–385. https://doi.org/10.1016/j.cell.2018.02.060 PMID: 29625053

34. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. Oxford University Press; 2019; 47: D886–D894. https://doi.org/10.1093/nar/gky1016 PMID: 30371827

35. Kuijjer ML, Paulson JN, Salzman P, Ding W, Quackenbush J. Cancer subtype identification using somatic mutation data. Br J Cancer. Nature Publishing Group; 2018; 118: 1492–1501. https://doi.org/10.1038/s41416-018-0109-7 PMID: 29765148

36. Rheinbay E, Nielsen MM, Abascal F, Tiao G, Hornshøj H, Hess JM, et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. bioRxiv. Cold Spring Harbor Laboratory; 2017; 237313. https://doi.org/10.1101/237313

37. Molparia B, Nichani E, Torkamani A. Assessment of circulating copy number variant detection for cancer screening. Galli A, editor. PLoS One. 2017; 12: e0180647. https://doi.org/10.1371/journal.pone.0180647 PMID: 28686671

38. Hartung F, Wang Y, Aronow B, Weber GF. A core program of gene expression characterizes cancer metastases. Oncotarget. Impact Journals, LLC; 2017; 8: 102161–102175. https://doi.org/10.18632/oncotarget.22240 PMID: 29254233

39. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci. 2013; https://doi.org/10.1073/pnas.1221068110 PMID: 23345422

40. AACR Project GENIE Consortium TAPG. AACR Project GENIE: Powering Precision Medicine through an International Consortium. Cancer Discov. American Association for Cancer Research; 2017; 7: 818–831. https://doi.org/10.1158/2159-8290.CD-17-0151 PMID: 28572459

41. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016; 44: e131–e131. https://doi.org/10.1093/nar/gkw520 PMID: 27270079

42. Kim T-M, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell. Elsevier; 2013; 155: 858–68. https://doi.org/10.1016/j.cell.2013.10.015 PMID: 24209623

43. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, et al. Fast Computation and Applications of Genome Mappability. Ouzounis CA, editor. PLoS One. Public Library of Science; 2012; 7: e30377. https://doi.org/10.1371/journal.pone.0030377 PMID: 22276185

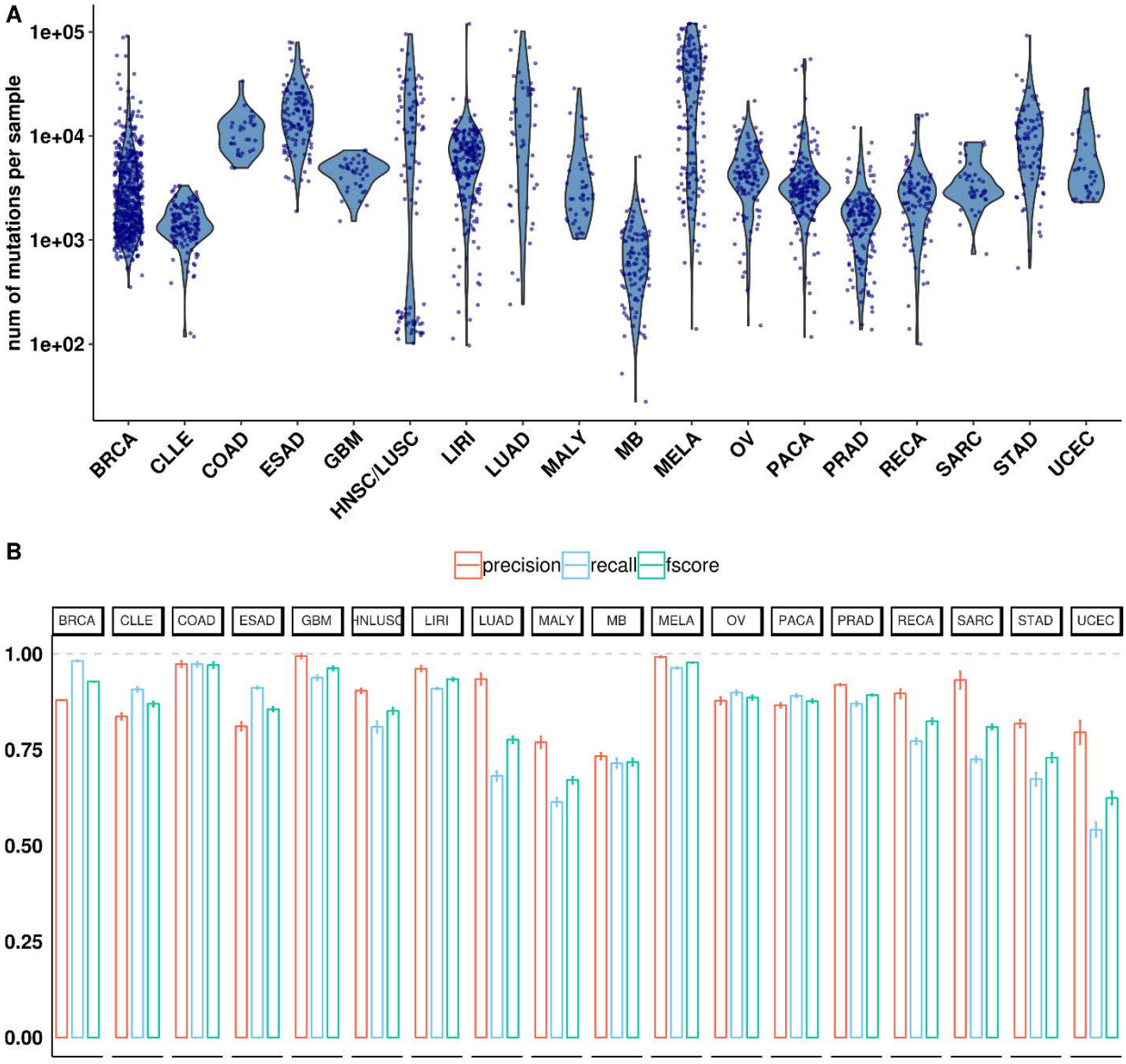
44. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. Oxford University Press; 2017; 45: D777–D783. https://doi.org/10.1093/nar/gkw1121 PMID: 27899578

45. Pedregosa F. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011; 12: 2825–2830. Available: http://delivery.acm.org/10.1145/2080000/2078195/p2825-pedregosa.pdf?ip=161.116.222.16&id=2078195&acc=OPEN&key=4D4702B0C3E38B35.4D4702B0C3E38B35.4D4702B0C3E38B35.6D218144511F3437&__acm__=1525430851_d83b9595f2c38819779f3ec95216a7f9

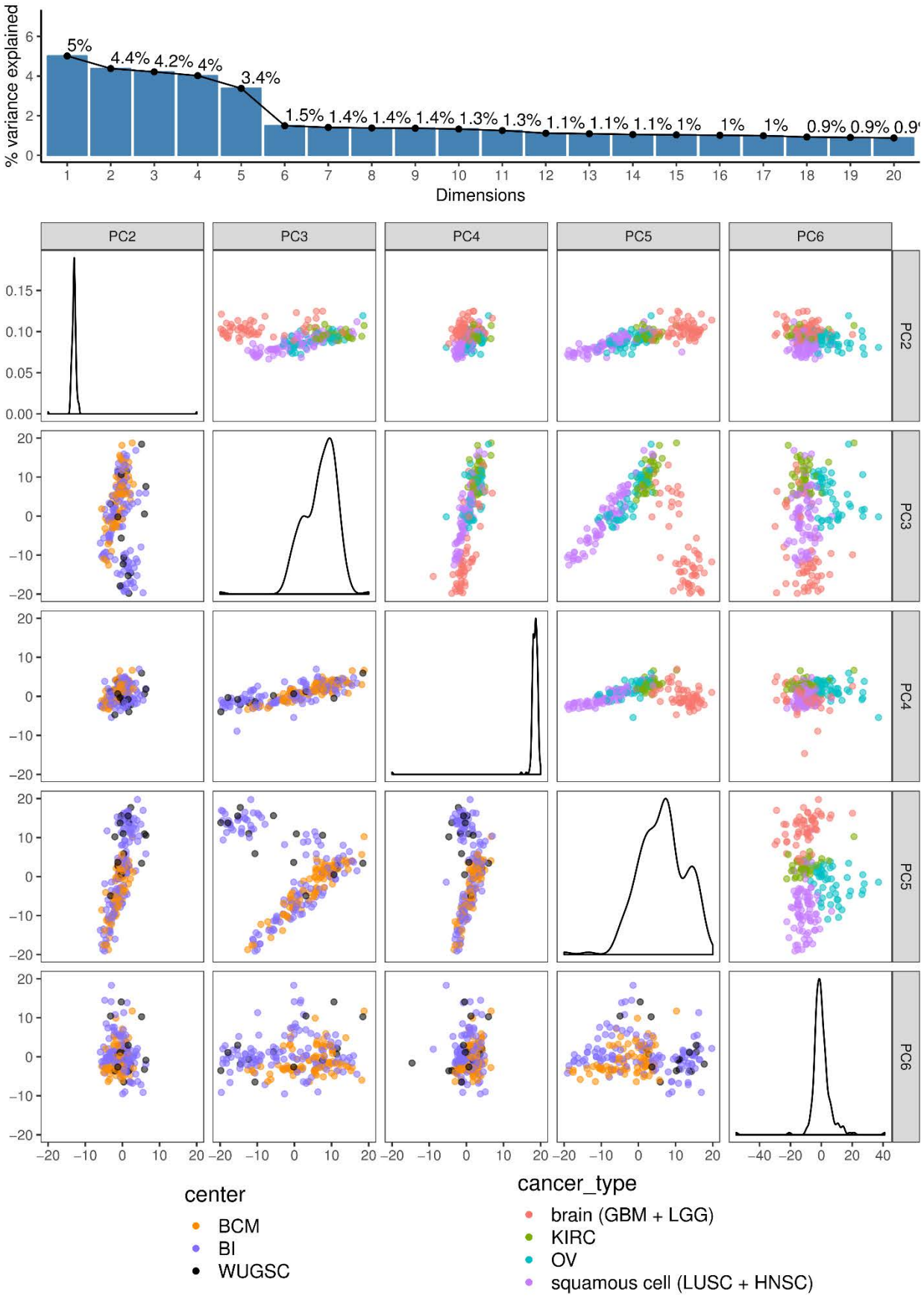
46. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. Kluwer Academic Publishers-Plenum Publishers; 1995; 20: 273–297. https://doi.org/10.1023/A:1022627411411

47. C. Hsu, Chang C, Lin C. No Title. In: A Practical Guide to Support Vector Classification [Internet]. 2003. Available: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Supplementary Figures

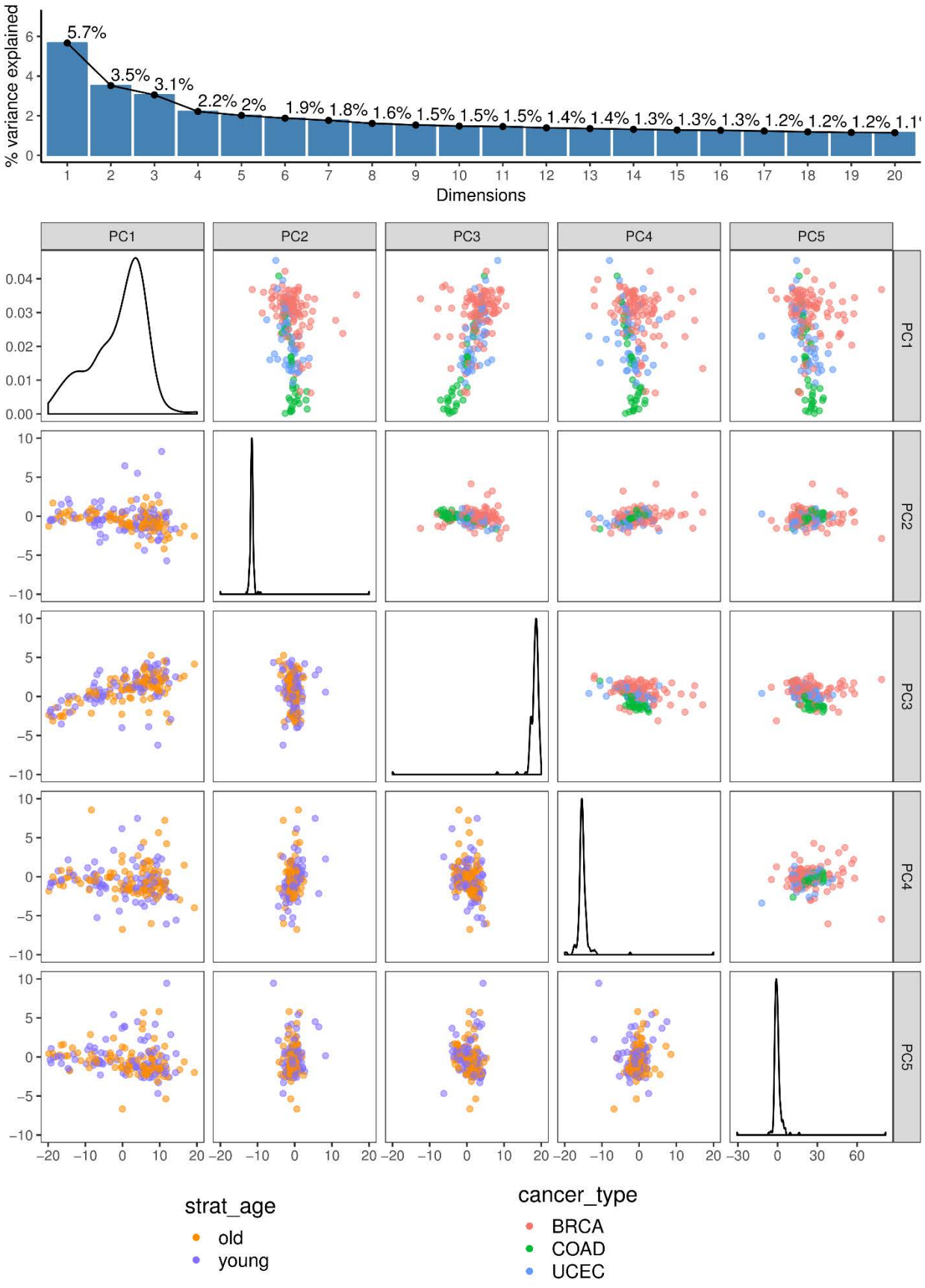


S1 Fig. **Summary statistics of the main training dataset.** (A) Distribution of the total number of mutations per sample for each cancer type. (B) Precision, Recall and F-score values for each cancer type. Each bar represents the mean of the corresponding score obtained from five independent runs (5-fold cross validation in each run) for each cancer type. Error bars represent the standard error of the mean for each cancer type.



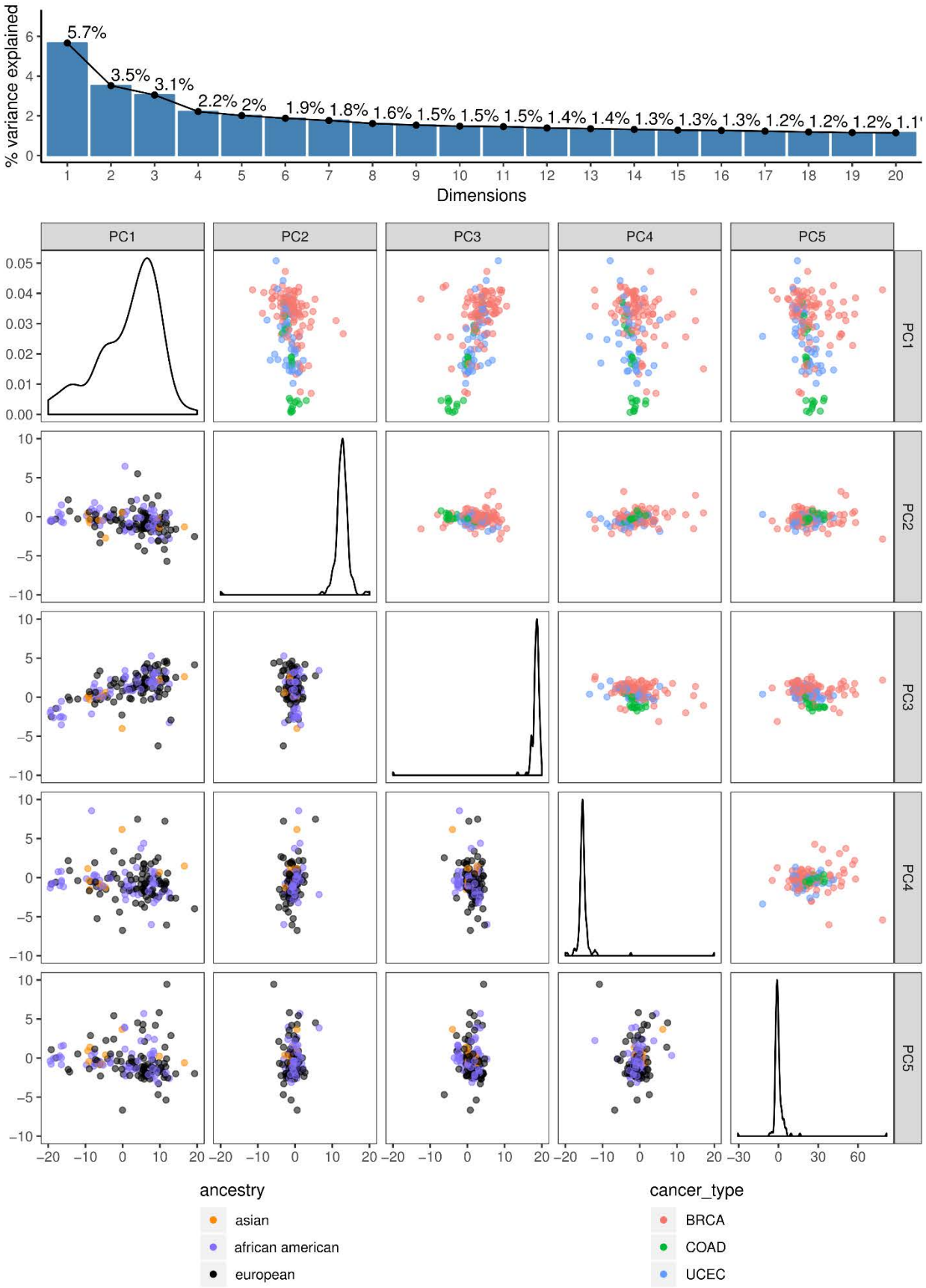
S2 Fig. **Analysis of possible batch effects: Sequencing center.** Biplots for combinations of the 5 first principal components (PCs) for the RMD features, in a subset of cancer types with at least 5 samples per annotation group. Scree plot shows variance

explained by each PC. The above-diagonal part of the scatterplot is colored by cancer type. The below-diagonal part of the scatterplot is colored by sequencing center.

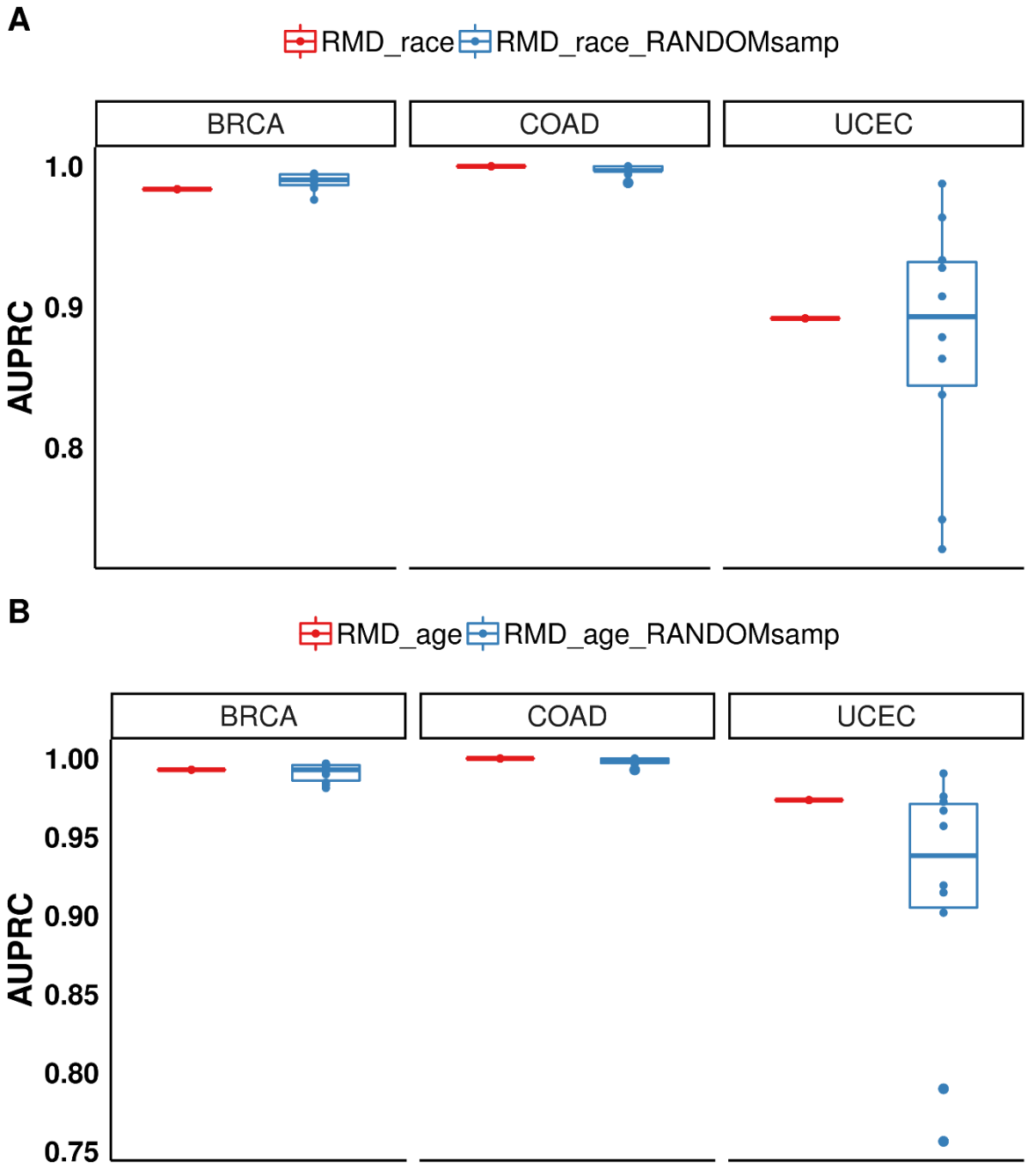


S3 Fig. Analysis of possible batch effects: Age.

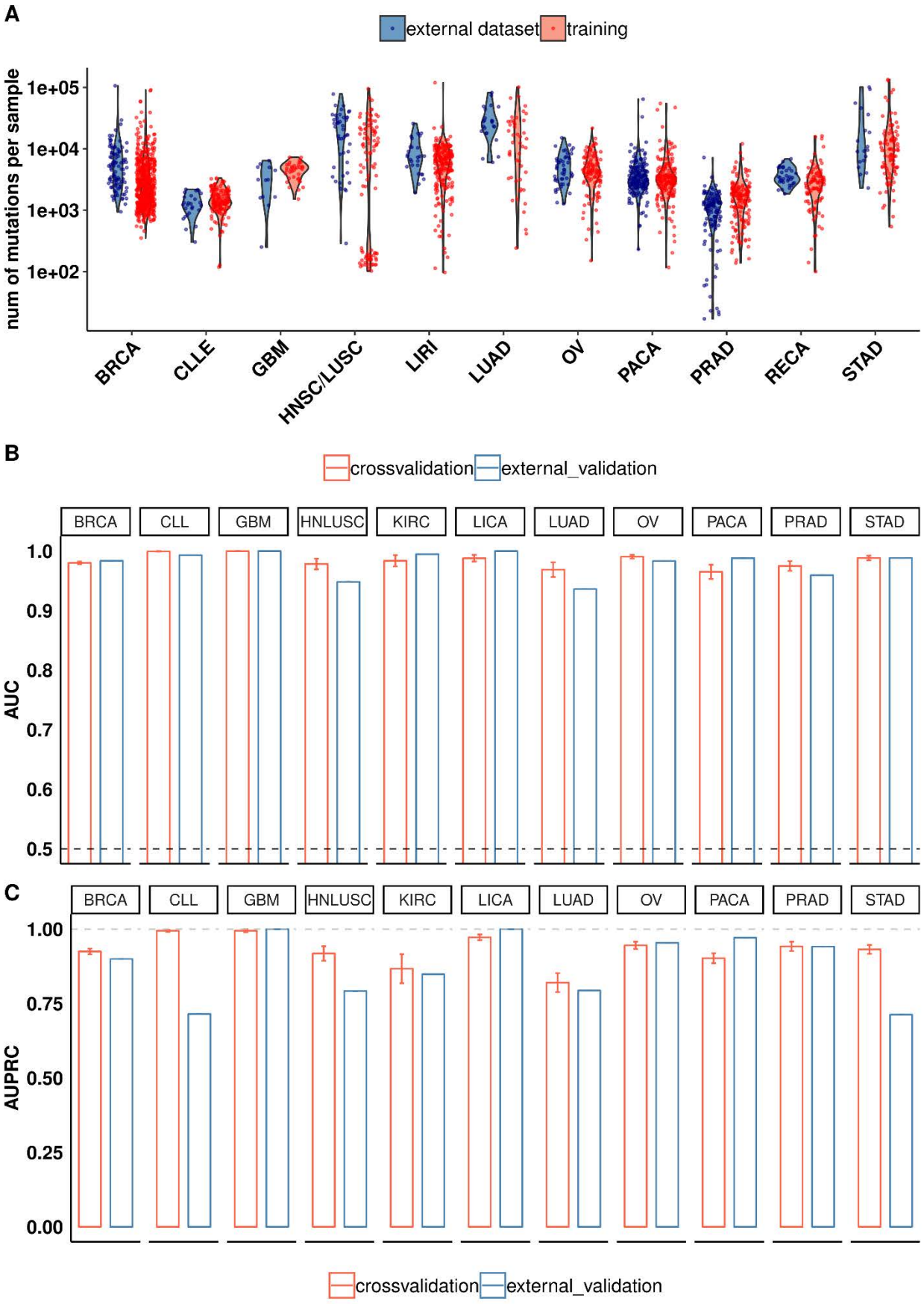
Biplots for combinations of the 5 first principal components (PCs) for the RMD features, in a subset of cancer types with at least 5 samples per annotation group. Scree plot shows the variance explained by each PC. The above-diagonal part of the scatterplot is colored by cancer type. The below-diagonal part of the scatterplot is colored by age (patients stratified in old and young groups according to whether their age is greater or lower than the median for each cancer type respectively).



S4 Fig. **Analysis of possible batch effects: Ancestry.** Biplots for combinations of the first 5 principal components (PCs) for the RMD features, in a subset of cancer types with at least 5 samples per annotation group. Scree plot shows the variance explained by each PC. The above-diagonal part of the scatterplot is colored by cancer type. The below-diagonal part of the scatterplot is colored by the ancestry group of the patient.

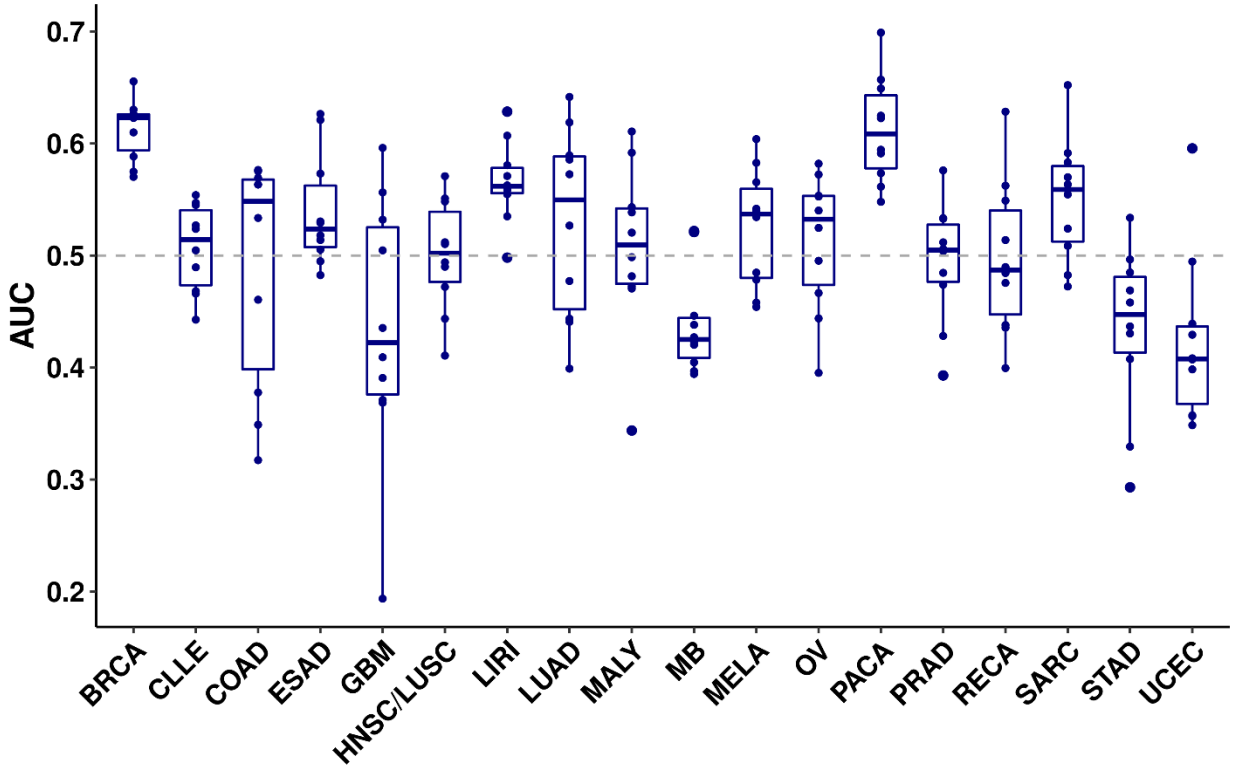


S5 Fig. **Analysis of possible batch effects: Ancestry and age.** (A) Area Under the Precision Recall curve (AUPRC) for RMD features in a subset of three cancer types training in samples from one group (european) and testing on another group (pooled asian, african and Hawaiian) in red; and training and testing in a mixture of groups maintaining the proportions between training and testing sets in blue. (B) AUPRC for RMD features in a subset of 3 cancer types training in samples from one group (old—age above the median of the cancer type) and testing on another group (young—age below the median of the cancer type) in red; and training and testing in a mixture of groups maintaining the proportions between training and testing sets in blue.

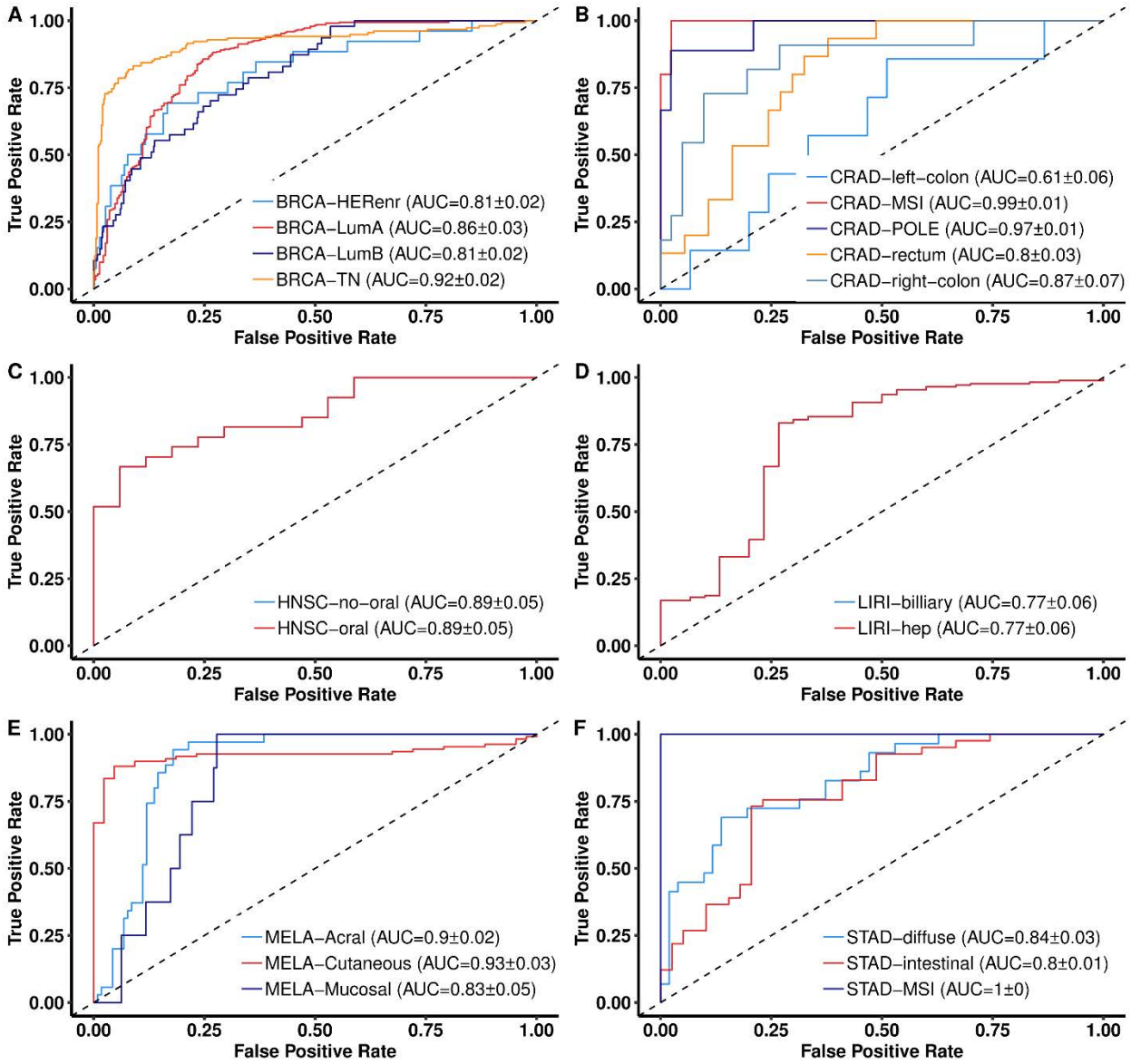


S6 Fig. Description of the secondary training and external validation datasets. (A) Distribution of the total number of mutations per sample for each cancer type of the secondary training dataset (red) and the external validation dataset (blue). (B) For RMD features AUC mean of 5 independent runs obtained by training in the secondary training dataset and testing in the external validation dataset (blue) and by crossvalidation in the secondary training dataset (red) for each cancer type. Error bars

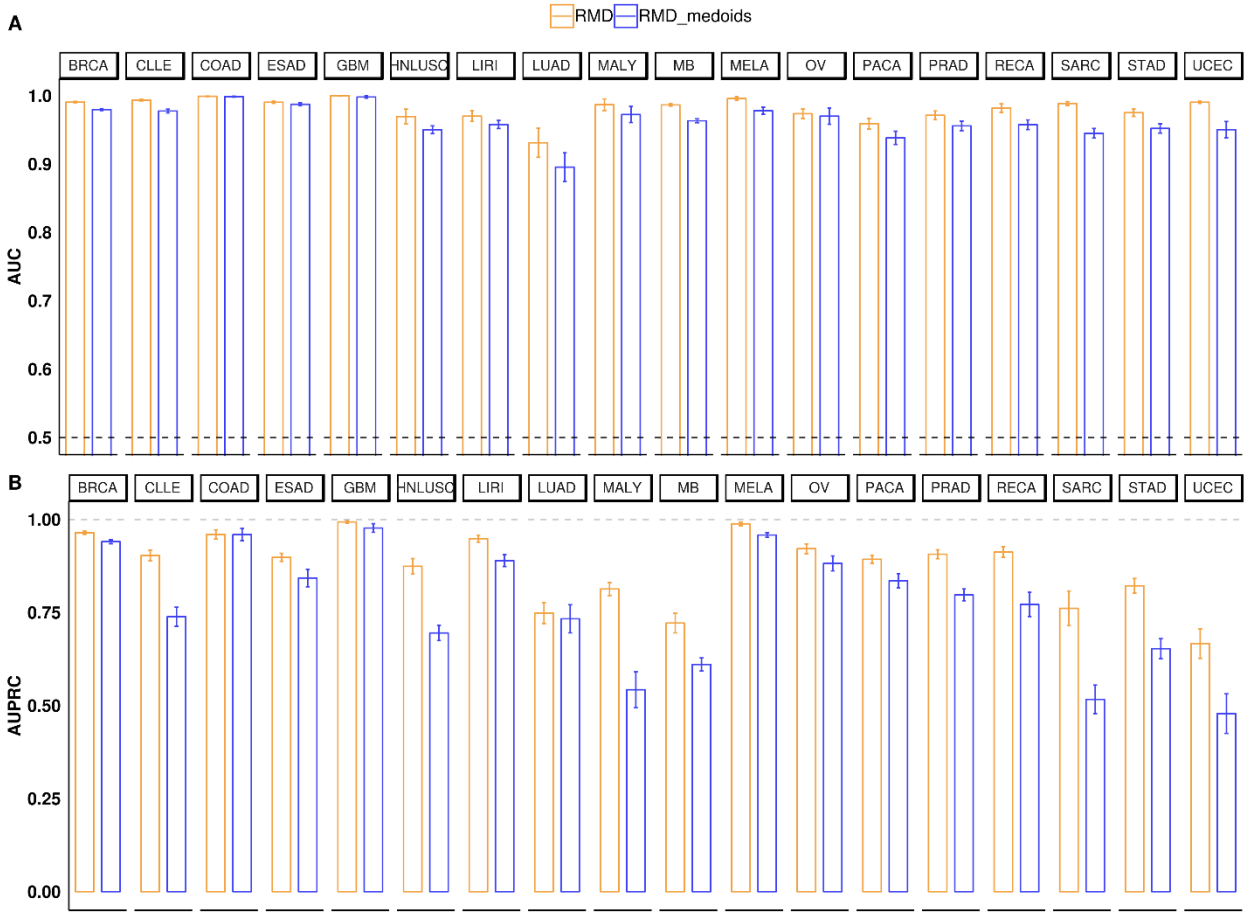
represents the standard error of the mean of each cancer type. (C) For RMD features AUPRC mean of 5 independent runs obtained by training in the secondary training dataset and testing in the external validation dataset (blue) and by crossvalidation in the secondary training dataset (red) for each cancer type. Error bars represents the standard error of the mean of each cancer type.



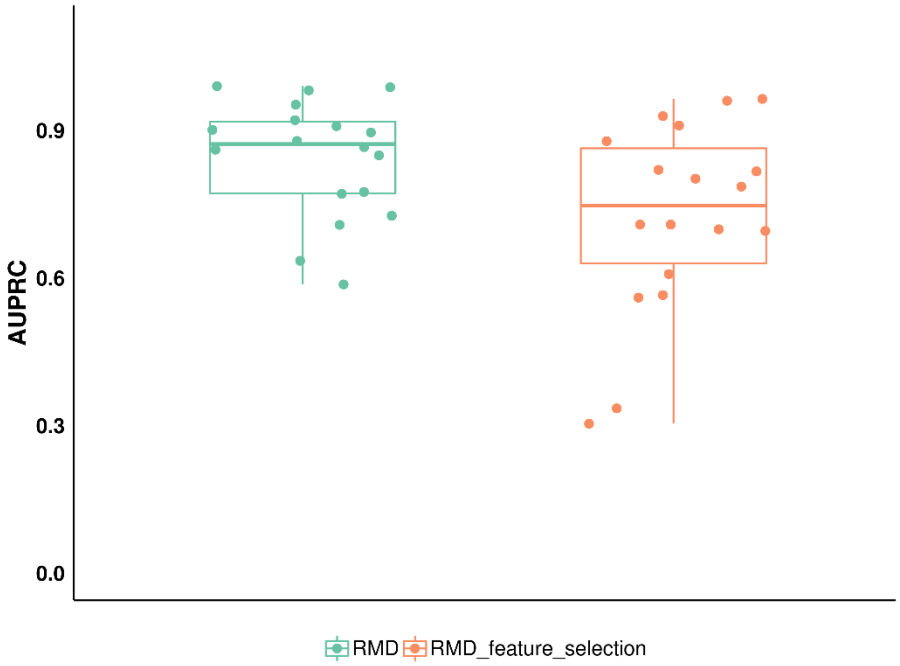
S7 Fig. **Evaluation of RMD with randomized labels.** Area Under the ROC curve (AUC) for each cancer type calculated with RMD classifiers with the class labels (cancer type) randomized. Each dot corresponds to one independent run (10 runs in total).



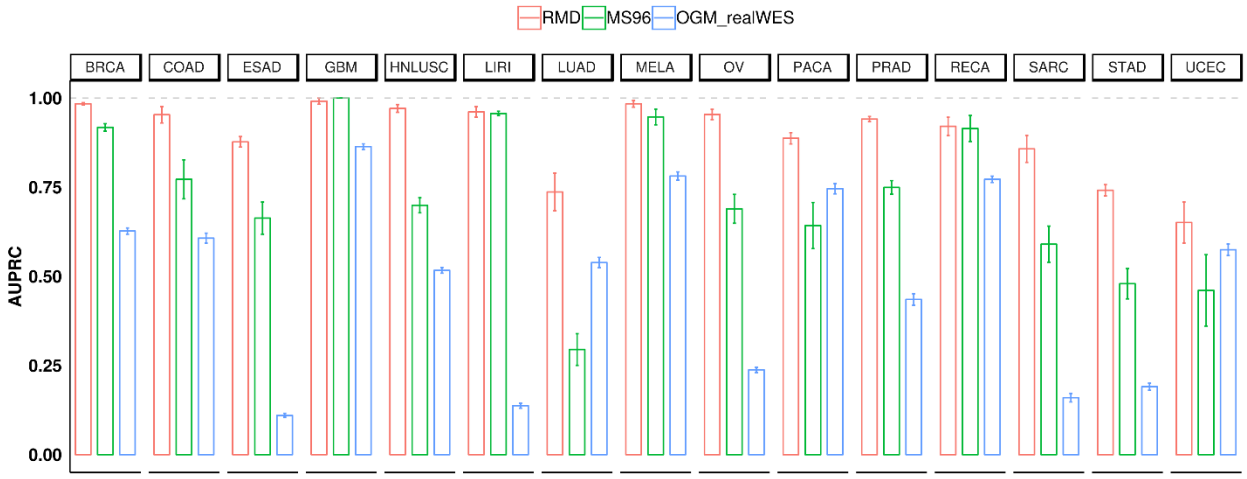
S8 Fig. Evaluation of subtype classifiers performance. Receiver Operating Characteristic (ROC) curve for each subtype versus the rest of samples, in breast cancer (A), colorectal cancer (B), head and neck squamous cell adenocarcinoma (C), liver cancer (D), melanoma (E) and stomach cancer (F) datasets. Area Under the ROC curve (AUC) reported (mean and standard error of the mean across 5-fold cross validation rounds) in the legend of each panel.



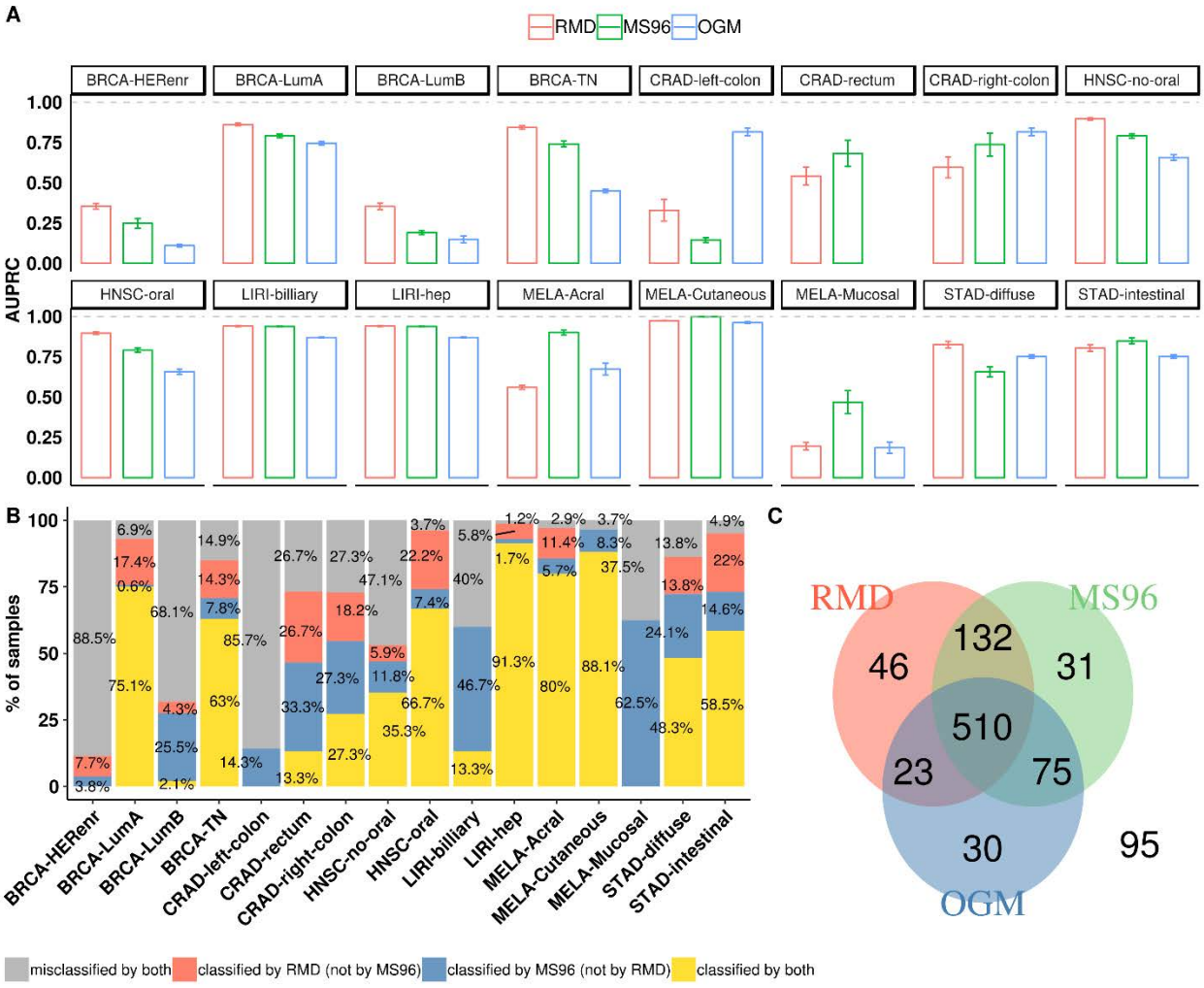
S9 Fig. Evaluation of predictive accuracy of RMD features after dimensionality reduction. (A) Mean Area Under the ROC Curve (AUC) for each cancer type for the full set regional mutation density (RMD) features in yellow and the reduced set of 500 RMD features (by k-medoid clustering) in blue. (B) Mean Area Under the Precision Recall Curve (AUPRC) for each cancer type for the full set of RMD features in yellow and a reduced set of 500 RMD features (by k-medoid clustering) in blue. Error bars are the standard error of the mean.



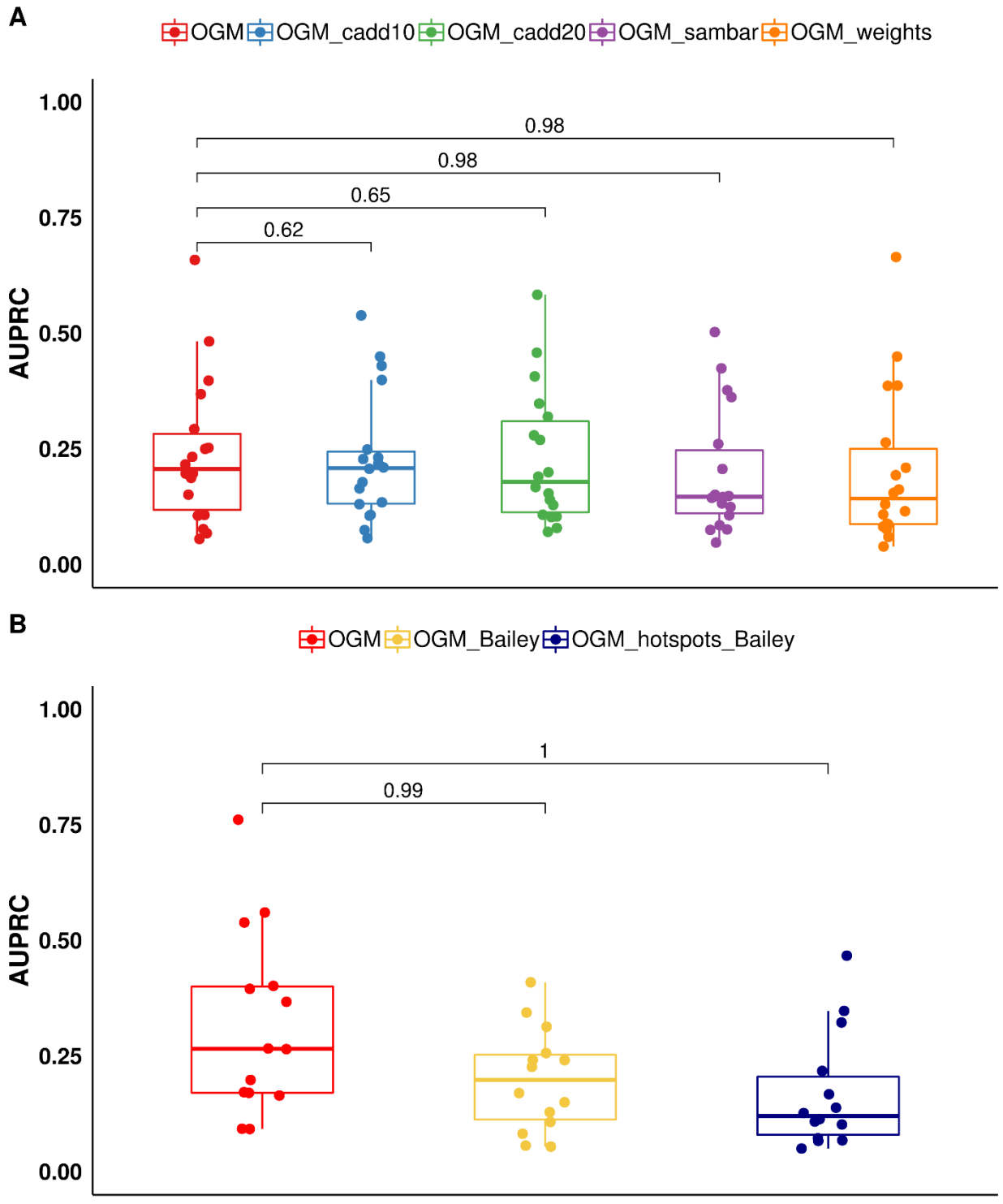
S10 Fig. Evaluation of RMD with and without feature selection. Area Under the Precision Recall Curve (AUPRC) for the main training dataset for each cancer type with all RMD features (in green); and applying feature selection to each training subset using Elastic Net and testing with those selected features within each fold of the crossvalidation (in orange). Each dot represents one cancer types (median AUPRC across the 3 folds of the crossvalidation, distinguishing that cancer type versus the rest of cancer types).



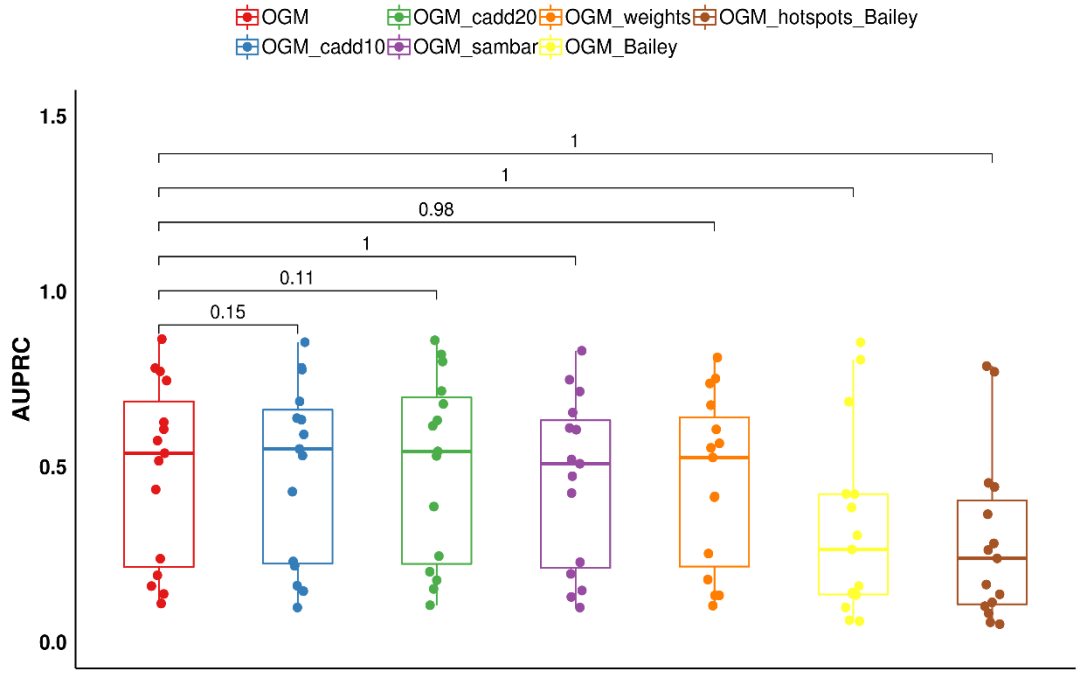
S11 Fig. Evaluation of model accuracy using passenger *versus* driver mutation features. Mean Area Under the Precision Recall Curve (AUPRC) scores for each cancer type in the main training dataset, using regional mutation density (RMD) features in red, and 96 mutation spectra (MS96) features in green. The presence/absence of mutations in cancer genes (OGM features) was determined for a dataset of real WES with equivalent cancer types (shown in blue).



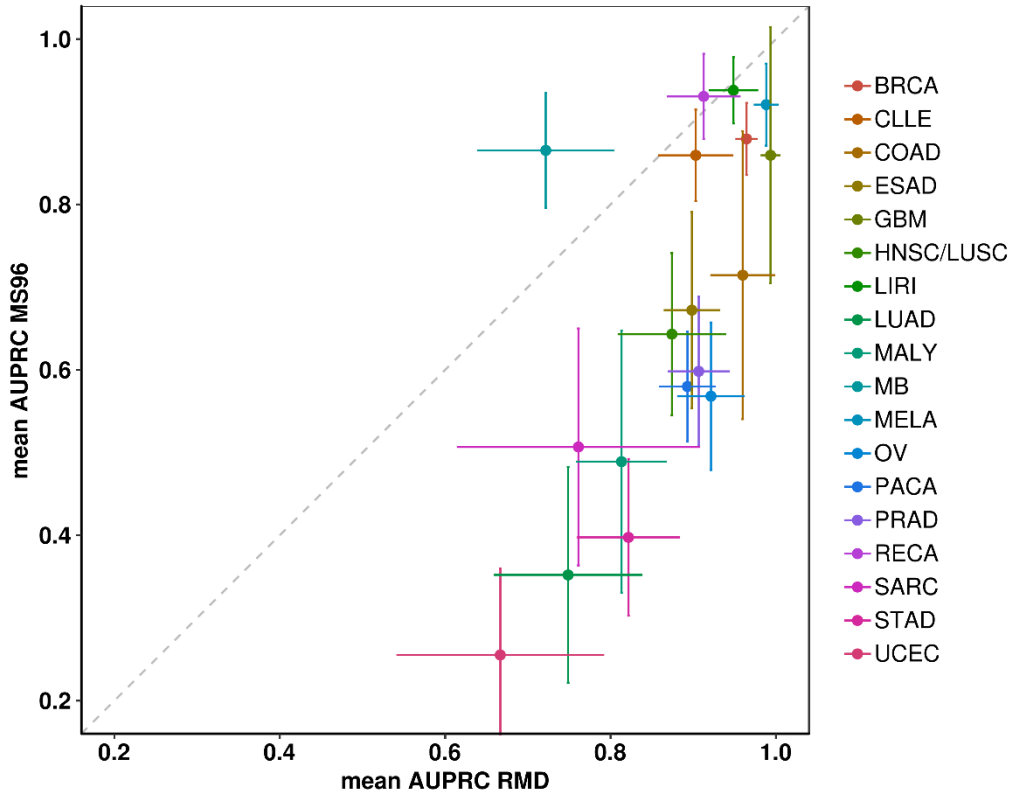
S12 Fig. Comparison and complementarity analysis for subtypes using RMD features. (A) Mean Area Under the Precision Recall Curve (AUPRC) scores for each cancer type for the subtypes of 6 major cancer types, using different sets of features: regional mutation density (RMD) in red, 96 mutation spectra (MS96) in green and presence/absence of oncogenic mutations (OGM) in blue. (B) For the subtypes datasets of six cancer types % of samples that are: (i) correctly classified by both RMD and MS96 (yellow), (ii) misclassified by both methods (gray), (iii) correctly classified by the MS96 but not by the RMD (blue) and (iv) correctly classified by the RMD but not by the MS96 (red). (C) Venn diagram of samples correctly classified by MS96, RMD or OGM features and their intersections for the subtypes classification.



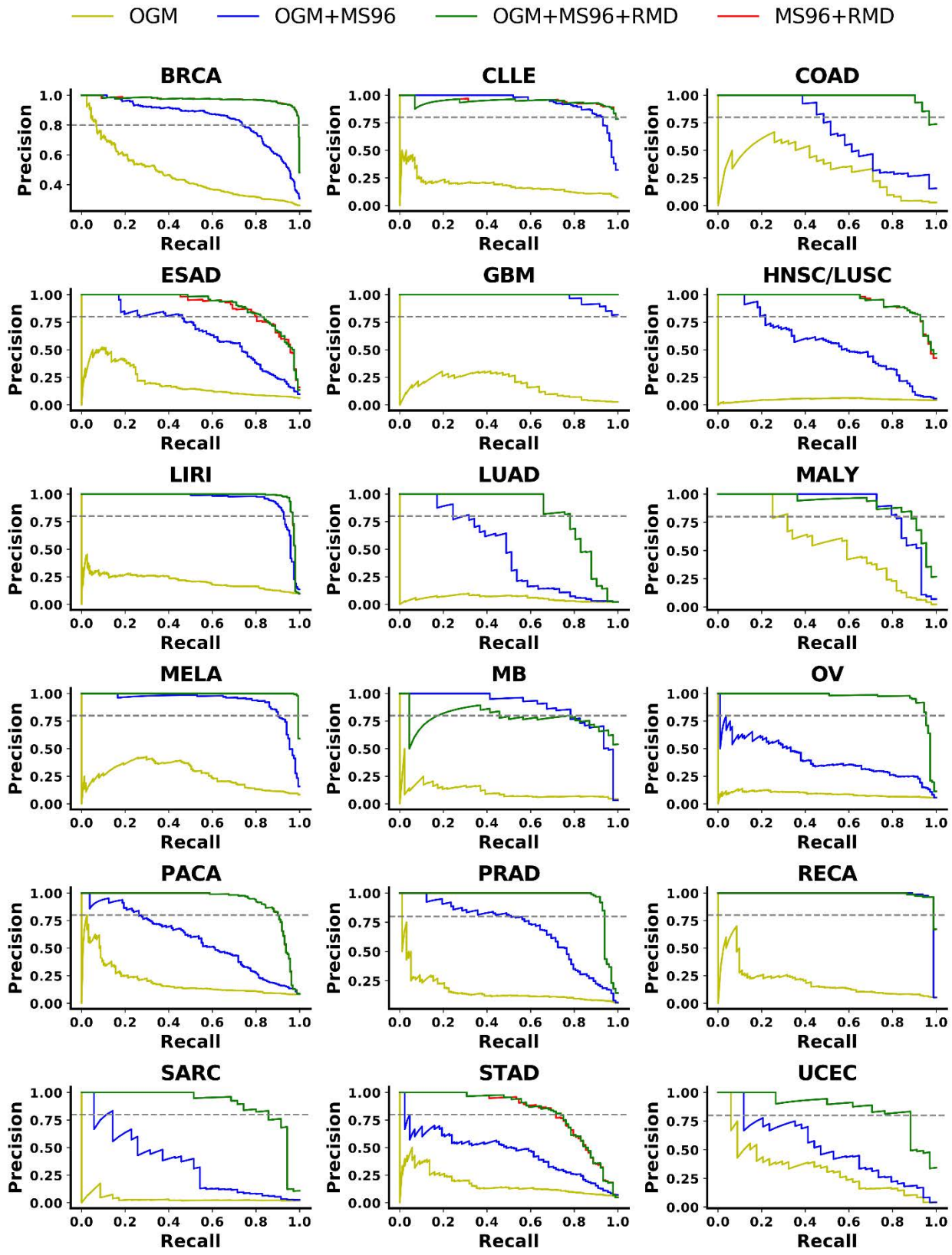
S13 Fig. **Evaluation of features that account for mutation impact.** (A) Area Under the Precision Recall curve (AUPRC) for five sets of oncogenic mutation (OGM) features (see *Methods*) on WGS data. P-value reported for each set of features compares with the default OGM features as a baseline, using one-tailed Wilcoxon signed rank test (with alternative set to "less" in the R function *wilcox.test*). (B) Area Under the Precision Recall curve (AUPRC) for three different sets of OGM features (see *Methods*) in a subset of 560 patients (only considering genomes with at least one mutation from the Bailey *et al*. list). P-values for each set of features compare with the default OGM features as a baseline one, using a statistical test as in (A).



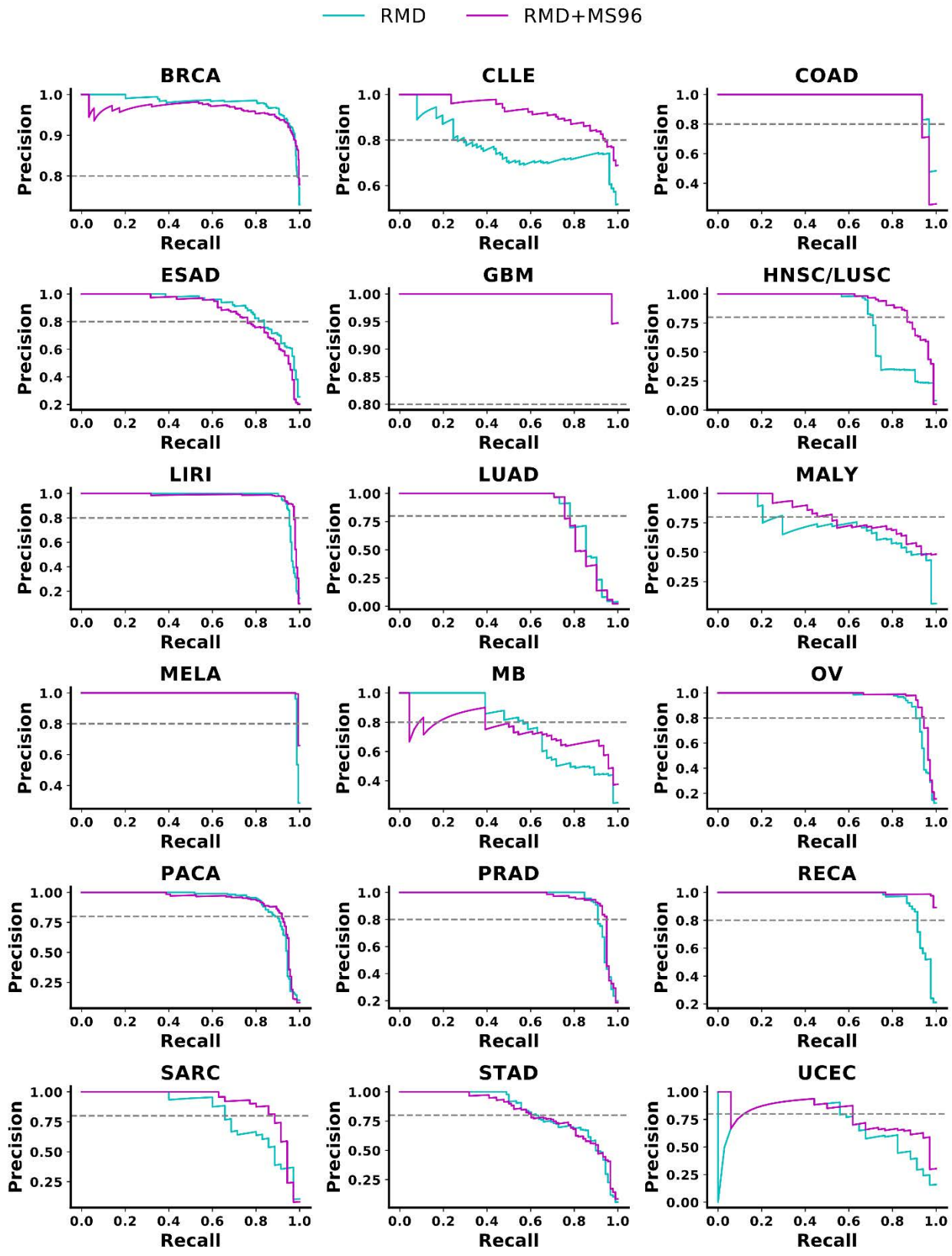
S14 Fig. Evaluation of features that account for mutation impact. (A) Area Under the Precision Recall curve (AUPRC) for five sets of oncogenic mutation (OGM) features (see *Methods*) on WES data. P-value reported for each set of features compares against the baseline OGM features using a Wilcoxon signed rank test, one-tailed (alternative set to "less" in the R function *wilcox.test*).



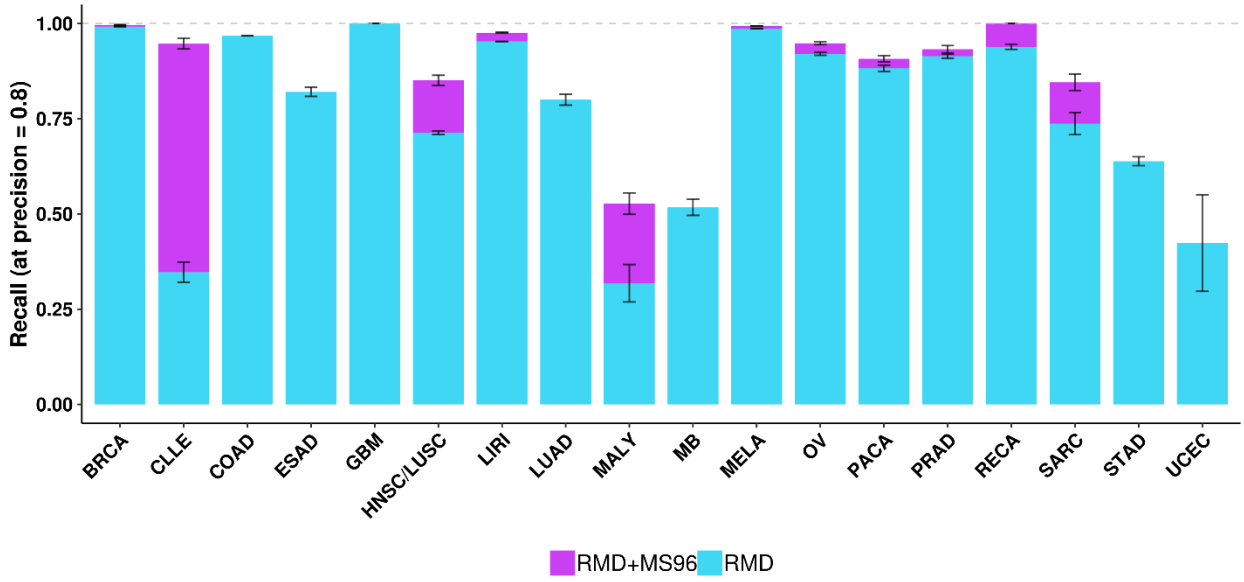
S15 Fig. **Comparison of accuracy of RMD and MS96 features per cancer type.** Mean Area Under the Precision Recall curve (AUPRC) score of the RMD features (x axis) versus MS96 features (y axis) of the main training dataset, in crossvalidation. Error bars are the standard deviation of AUPRC for RMD features (x axis) and MS96 features (y axis).



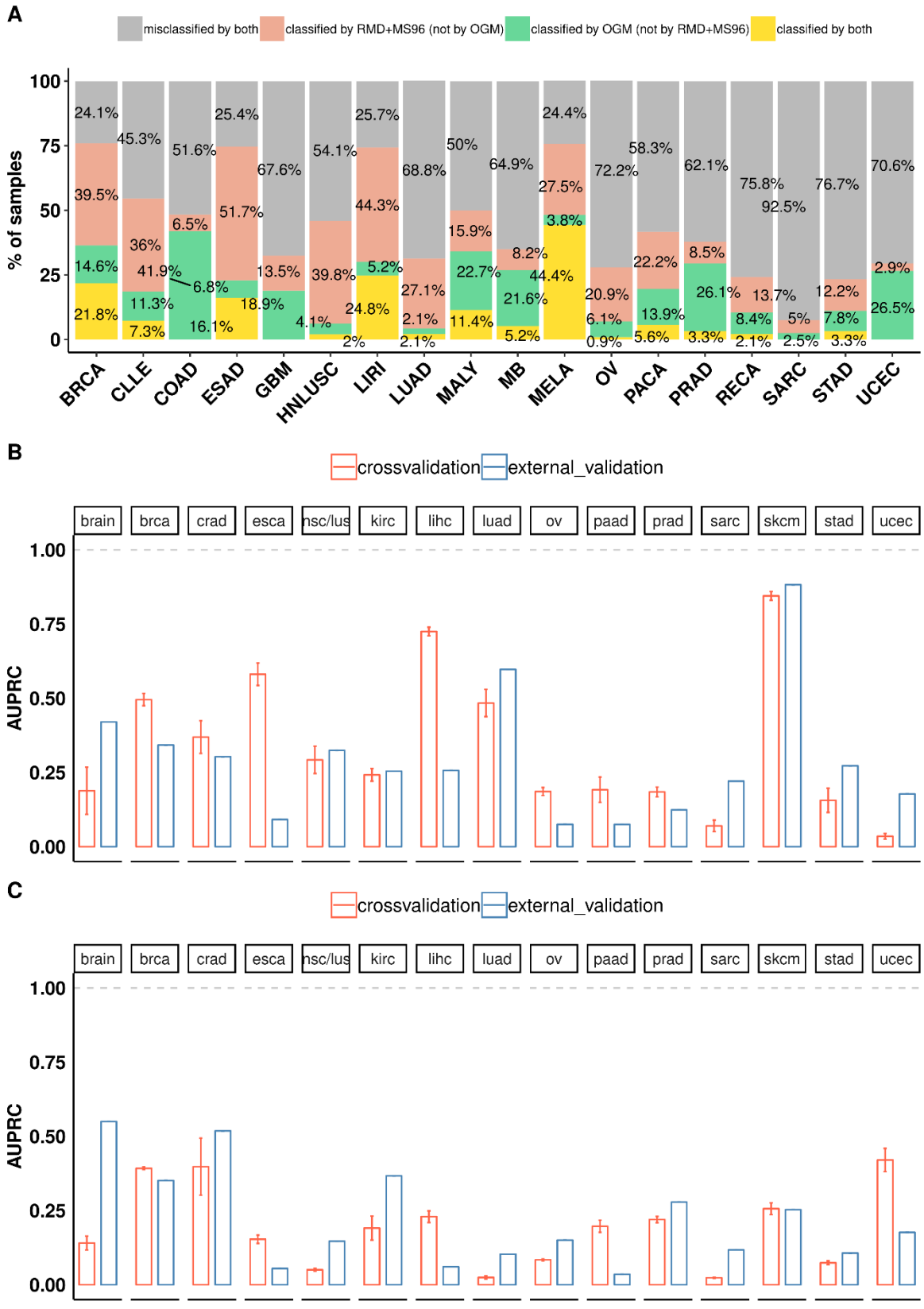
S16 Fig. **Gains in classifier accuracy by introducing additional sets of features.** Precision-Recall Curve for each cancer type for the OGM (yellow), for the combination of OGM and MS96 (blue) and for the combination of the OGM, MS96 and the RMD (green), and MS96 and RMD without OGM (red) features for the main training dataset. In most cancer types the red curve overlaps the green curve perfectly and is thus hidden on the plots. Grey line indicates the threshold where precision = 0.8.



S17 Fig. Improvement in accuracy of RMD-based classifiers by the addition of MS96 features. Precision-Recall Curve for each cancer type for the RMD (blue) and for the combination of RMD and MS96 (purple) for the main training dataset. Grey line indicates the threshold where precision is equal to 0.8



S18 Fig. Improvement of RMD classifiers by the addition of MS96 features. Recall at FDR = 20% for the classification models trained on RMD (blue) and RMD+MS96 features (purple); height of the stacked bars indicates the excess proportion of patients receiving correct predictions upon introducing the additional features to the classification model. Bars show the mean of five cross-validation runs, and error bars are standard deviations.



[S19 Fig.](#) **Complementarity and external validation of features with simulated WES datasets.** (A) For the main training dataset of simulated WES, fraction of samples that are: (1) correctly classified by both the passengers (RMD+MS96) and the drivers (OGM) (yellow), (2) misclassified by both methods (gray), (3) correctly classified by the passengers but not by the drivers (orange) and (4) correctly classified by the drivers but not by the passengers (green). (B) For RMD+MS96 features dataset, mean AUPRC of five classification runs obtained by training on the simulated WES dataset and testing on the real WES dataset as an external validation (blue) and by crossvalidation in the secondary training dataset (red) for each cancer type. Error bars represents the standard error of the mean of each cancer type. (C) As above, but for OGM features. Error bars represents the standard error of the mean of each cancer type.

Supplementary Methods

Collection and preparation of genomic data

We combined datasets from four different sources to maximize the coverage of cancer types and subtypes, encompassing 24 datasets with somatic mutations obtained by whole-genome sequencing (WGS), including: data from the ICGC Data portal [1], the WGS somatic single-nucleotide variants (SNVs) from Alexandrov *et al.* [2] and the somatic mutations of 100 WGS samples of stomach cancer from Wang *et al*. [3]. Additionally, data was obtained as aligned short reads (to the hg19/GRCh37 assembly) for the WGS of tumours and the matched normal tissue from the The Cancer Genome Atlas (TCGA) repository, formerly at CGHub (now decommissioned), and somatic mutations were called as described previously [4]. Finally, the files containing somatic mutations for whole-exome sequences (WXS) were downloaded from the Genomic Data Commons (TCGA). To minimize errors due to misalignment of short reads, we masked out all regions in the genome defined in the 'CRG Alignability 36' track [5] as having alignability <1.0, for all datasets.

For the WGS analysis, we created a main training dataset including all available cancer types with more than 20 tumor samples (S1 Table). LUSC and HNSC were merged into a single category called HNSC/LUSC [6]. For validation, we examined all cancer types from the main training dataset that had an equivalent dataset from a different data source or sequencing center, which therefore constitute the external validation dataset (S2 Table). Additionally, we separated six cancer types into subtypes based on molecular subtyping, their microsatellite instability (MSI) status and presence of cancer-associated viruses HPV for HNSC or EBV for STAD (S3 Table). For the analysis of robustness towards noise, we simulated genomes with false negative mutation calls (dropout) by randomly sampling 100, 50, 25, 5, 2, 1 and 0.1% of the mutations in each tumor.

Additionally, we simulated WXS data from the WGS datasets by retaining only mutations inside the regions that reflect the actual WXS sequencing coverage for the three TCGA sequencing centers: BI, BCM and WUGSC, requiring at least 8 reads in at least 90% samples from each sequencing center, as previously [7]. Of note, we used the intersection of these three regions to reduce biases when comparing cancer types from different sequencing centers. This filtering is more stringent (~1.8% mutations retained) than using only the WXS regions from one sequencing center (~2.2% mutations retained). Regarding validation of the WXS analyses, we used the same main training dataset for generating simulated WXS data to train the classification model, which were then validated by comparing with tissue-matched real WXS from TCGA, which thus constituted an external validation dataset (Supplementary Table 4).

Obtaining Copy Number Alteration (CNA) data

We downloaded CNAs data (SNP6) from FireBrowse [8]. For each CNA segment we use a threshold of segment_mean>0.1 and <-0.1 to classify the segment as +1 or -1 respectively (the rest were set to 0). For each sample, the CNAs matrix was calculated dividing each chromosome into 1 Mb windows, counting the number of CNA events per window and labelling them as +1 if amplification counts was higher or -1 if deletion counts was higher in the window. Afterwards, we divided the CNA matrix into CNA drivers and CNA passengers matrices, where drivers affect the 1 Mb windows containing the 64 known dosage-sensitive cancer genes from the Cancer Gene Census (68 features). The passengers are the remaining CNA that do not overlap the windows containing those genes nor are in their immediate vicinity (excluding 1 Mb either side) (2715 features).

Supplementary Results

Copy number alteration-based tumor type classifiers

Our analysis up to this point adressed the distribution of mutations encompassing single-nucleotide variants (SNVs) and small indels. This is because previous work suggested that SNVs and indels might convey tissue-specific signal in form of trinucleotide spectra [2] and in form of domain-scale mutation rates [4,9]. Another very common type of genetic change in tumors are somatic copy-number alterations (CNA), constituting amplifications or deletions affecting cancer driver genes which are dosage-sensitive [10] and additionally many other alterations which have a lower or no selective advantage. It is indeed known that some CNA recur more often in certain cancer types, providing a rationale for previous use of CNA features for tumor type classification [11,12]. We briefly investigated their predictive power, comparing it to the mutational features derived from driver mutations and from global patterns of passenger mutations.

We tentatively divided the CNA into drivers and passengers, where drivers affect the 64 known dosage-sensitive cancer genes from the Cancer Gene Census. The passengers are the remaining CNA that do not overlap those genes nor are in their immediate vicinity (1 Mb either side), and are represented as copy-numbers in megabase-sized windows across the genome, by analogy with RMD features. An important consideration is that it is very difficult to resolve driver from passenger CNA events because some CNA may span very large swaths of a chromosome, affecting many genes – drivers and passengers alike – and therefore the signal we observe will necessarily be mixed. Furthermore it has been argued that the copy number changes that affect many genes that are not conventional drivers may still have selective advantages to the tumor [13]. Another consideration is that the CNA estimates we analyze are obtained from SNP arrays (from the TCGA project) while in realistic applications, such as liquid biopsies, the CNAs would often be estimated from DNA sequencing [14]. This may not yield similar quality CNA estimates, particulary in case of exome or panel sequencing.

By analogy with examining mutational features, we find that driver CNA are less predictive of cancer type than the putative passenger CNA profiles, with AUPRC 0.46 and 0.75 (median across 15 cancer types) for drivers versus passengers (Fig S15). On a matched set of whole genome sequences from 15 cancer types, we observed median AUPRC 0.22 for SNV/indel drivers (OGM) versus 0.92 for SNV/indel passengers (RMD+MS96). Overall, two trends emerge. First, both for CNA and for SNV/indels, the genome-wide pattern emanating from non-selected alterations provides a better tumor classifier than the known oncogenic CNAs, SNVs or indels. Second, that CNA in fact have a high potential for classifying cancer type, at least when measured on high-purity tumor samples using SNP arrays. Further analyses will elucidate determine how the CNA classifiers fare on lower quality data; simulation studies may prove useful in this task [11,12].

Supplementary Text References

1. The International Cancer Genome Consortium. International network of cancer genome projects. Nature. Nature Publishing Group; 2010;464: 993–998. doi:10.1038/nature08987

2. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., et al. Signatures of mutational processes in human cancer. Nature. Nature Publishing Group; 2013;500: 415–421. doi:10.1038/nature12477

3. Wang K, Yuen ST, Xu J, Lee SP, Yan HHN, Shi ST, et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nat Genet. Nature Publishing Group; 2014;46: 573–582. doi:10.1038/ng.2983

4. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. Nature Publishing Group; 2015;521: 81–84. doi:10.1038/nature14173

5. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, et al. Fast Computation and Applications of Genome Mappability. Ouzounis CA, editor. PLoS One. Public Library of Science; 2012;7: e30377. doi:10.1371/journal.pone.0030377

6. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. NIH Public Access; 2014;158: 929–944. doi:10.1016/j.cell.2014.06.049

7. Park S, Supek F, Lehner B. Systematic discovery of germline cancer predisposition genes through the identification of somatic second hits. Nat

Commun. Nature Publishing Group; 2018;9: 2601. doi:10.1038/s41467-018-04900-7

8. Broad Institute of MIT and Harvard. Broad Institute TCGA Genome Data Analysis Center. In: Firehose stddata__2016_01_28 run [Internet]. 2016 [cited 24 Aug 2018]. doi:doi:10.7908/C11G0KM9

9. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. Nature Publishing Group; 2015;518: 360–364. doi:10.1038/nature14221

10. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. Nature Publishing Group; 2013;45: 1134–1140. doi:10.1038/ng.2760

11. Marquard AM, Birkbak NJ, Thomas CE, Favero F, Krzystanek M, Lefebvre C, et al. TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. BMC Med Genomics. BioMed Central; 2015;8: 58. doi:10.1186/s12920-015-0130-0

12. Molparia B, Nichani E, Torkamani A. Assessment of circulating copy number variant detection for cancer screening. Galli A, editor. PLoS One. 2017;12: e0180647. doi:10.1371/journal.pone.0180647

13. Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. Elsevier; 2013;155: 948–62. doi:10.1016/j.cell.2013.10.011

14. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. Nature Publishing Group; 2017;8: 1324. doi:10.1038/s41467-017-00965-y

Chapter 5

Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns

---

The following chapter has been selected from the publication:

M Salvadores, F Fuster-Tormo and F Supek (2020). Matching cell lines with cancer type and subtype of origin via mutational, epigenomic and transcriptomic patterns. Sci Adv. 6 (27), eaba1862

Access: https://doi.org/10.1126/sciadv.aba1862

CANCER

Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns

Marina Salvadores[1], Francisco Fuster-Tormo[1,2], Fran Supek[1,3]*

Cell lines are commonly used as cancer models. The tissue of origin provides context for understanding biological mechanisms and predicting therapy response. We therefore systematically examined whether cancer cell lines exhibit features matching the presumed cancer type of origin. Gene expression and DNA methylation classifiers trained on ~9000 tumors identified 35 (of 614 examined) cell lines that better matched a different tissue or cell type than the one originally assigned. Mutational patterns further supported most reassignments. For instance, cell lines identified as originating from the skin often exhibited a UV mutational signature. We cataloged 366 "golden set" cell lines in which transcriptomic and epigenomic profiles strongly resemble the cancer type of origin, further proposing their assignments to subtypes. Accounting for the uncertain tissue of origin in cell line panels can change the interpretation of drug screening and genetic screening data, revealing previously unknown genomic determinants of sensitivity or resistance.

INTRODUCTION

Cell lines are an important research tool, often used in place of primary cells and intact organisms to study biological processes. Cell lines are used for various applications such as testing drug metabolism and cytotoxicity, study of gene function, generation of artificial tissues, and synthesis of biological compounds (*1*). In cancer research, cell lines derived from tumors are commonly used as models, because they are presumed to carry the genomic and epigenomic alterations that arise in tumors (*2*). To understand the response of tumors to therapy, many studies have linked genetic and/or epigenetic alterations with drug response across cell line panels, generating datasets such as the Genomics of Drug Sensitivity in Cancer (GDSC) (*3*) and the Cancer Cell Line Encyclopedia (CCLE) (*4*). These efforts have advanced our understanding of tumor biology by generating a massive resource of genomic, transcriptomic, epigenomic, and drug response data for hundreds of cell lines (*2*).

As a model for cancer, cell lines are cost-effective, convenient, and amenable to high-throughput screening (*1*, *2*). However, a major question associated with the use of cell lines is whether they are representative of the cancer they are meant to model, which may be complicated by issues of misidentification (*1*, *2*, *5*).

Misidentified cell lines may lead to inconsistent conclusions across studies using the affected cell lines. For instance, the cell lines referred to as HEp-2 and INT 407 in the literature are commonly cross-contaminated with HeLa (cervical cancer) cells, rather than being laryngeal cancer and normal intestinal epithelium cells, respectively (*6*, *7*). Because of the potential for contamination, demonstrating cell line identity via genetic markers is now a routine quality-control step. Current resources based on large-scale cancer cell panels are therefore largely unaffected by this issue (*4*).

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [2]MDS Research Group, Institut de Recerca Contra la Leucèmia Josep Carreras, Institut Català d'Oncologia-Hospital Germans Trias i Pujol, Universitat Autònoma de Barcelona, Badalona, Spain. [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.
*Corresponding author. Email: fran.supek@irbbarcelona.org

However, even if the genetic identity of the cell line is correct, its properties may not match the cancer type it is meant to model. In particular, the tissue of origin might be incorrect. One way in which this error could arise is that tumors thought to originate in a certain tissue might be metastatic lesions originating from a distal site (*8*). Cell lines derived from these tumors would have a different tissue/cell type identity than that assigned at isolation, constituting a case of mislabeling. It is conceivable that, also in the case of primary tumors, ambiguous histological or anatomical features may cause the tumor type or subtype to be misdiagnosed and therefore also for a cell line derived from that tumor. Furthermore, the process of establishing the culture might select for a rare cell type that is not representative of the tumor isolate on the whole, meaning that the derived cell line would again effectively be mislabeled with a different cell type (*9*). In addition to the initial changes upon adaptation to culture, cell lines evolve over time due to selection and due to genetic drift, potentially diverging from the characteristics of the originating tissue (*1*).

Tissue and/or cell type is a key determinant of response of cultured cells to a variety of experimental conditions, including drug exposure and genetic perturbation (*10*, *11*). Therefore, having accurate information on the tissue and cell type identity of a tumoral cell line is important for interpreting the experimental results obtained using these cell lines.

Recent work has examined cell line panels of certain cancer types, showing discrepancies between the features of cell lines and corresponding tumor types or subtypes. For example, a gene expression analysis of lung tumors and lung cell lines (*9*) suggested that some lung adenocarcinoma cell lines did not resemble adenocarcinoma tumors but instead clustered with other lung tumor subtypes (small cell and squamous cell tumors). A study of high-grade serous ovarian cancer cell lines that used gene expression, driver gene mutations, and copy number alteration (CNA) data reported that two frequently used cell lines showed poor genetic similarity to molecular profiles of this ovarian cancer subtype (*12*). A study of a panel of renal cancer cell lines compared their CNA to kidney tumors, finding that some cell lines used as models of the clear cell

carcinoma more closely resemble papillary renal cell cancer (*13*). These examples highlight the need to systematically identify the cell lines whose genotype and/or molecular phenotypes do not resemble the characteristics of the matched human tumor type. A major challenge in the use of human tumor molecular data to classify cell lines are the widespread global changes in gene regulation between cell lines and tumors that arise in cell culture conditions.

We performed a systematic analysis that aligned mRNA expression and DNA methylation data between ~600 cancer cell lines and ~9000 tumors from 22 different cancer types, adjusting for global differences in transcriptomes and epigenomes. Classifiers trained on human tumor mRNA and DNA methylation profiles were used to identify those cell lines whose genomic and epigenomic profiles are highly consistent with human tumors of their declared cancer type of origin. Conversely, we used the same classifiers to identify those cell lines that might be mislabeled with an incorrect cancer type or that might have diverged from their original tissue and/or cell type identity. Our data suggest that tens of cell lines might be epigenetically and/or genetically not consistent with their stated tissue or cell type of origin, which is an important consideration for experiments that use these cell lines. We demonstrate this by reanalyzing associations between drug sensitivity and genetic variation in a large panel of cell lines. After explicitly accounting for putative cases of cell lines with mislabeled tissue identity, many previously unknown associations of genes with drug sensitivity or resistance were revealed.

RESULTS

Identification of tissue/cell type of origin for cell lines by a joint analysis with tumors

The tissue that originated a tumor is well known to be a major determinant of drug responses, including drugs targeted to specific genetic mutations, both in vitro (*10*, *11*) and in vivo (*14*, *15*). Tissue of origin is an important factor in shaping the networks of genetic interactions in cancer (*16*), and it also determines the phenotypes resulting from genetic perturbation (*11*). Therefore, ascertaining the tissue/cell type identity of cell lines is relevant for interpreting results of various experiments. For this reason, here, we have systematically examined the global features of the transcriptome and epigenome to identify the tissue of origin of tumoral cell lines.

During the process of adaptation to cell culture, the cells undergo changes in gene regulation that affect many genes (*17*, *18*). The global alterations in gene expression and DNA methylation mean that it is not straightforward to directly compare cell line transcriptomes and epigenomes with data obtained from tumors. To adjust for these cell culture–associated shifts, we introduce a computational methodology—HyperTracker—which can unify transcriptome, epigenome, and mutational data across tumors and cell lines and provide robust predictions of tissue, cell type, and subtype identity.

In particular, we collected gene expression [RNA sequencing (RNA-Seq)] and DNA methylation data (microarrays) for 9681 and 9039 human tumors, respectively (TCGA), and additionally for 614 cell lines (CL) of various solid cancer types and 69 CL of blood tumors. For gene expression data (GE), we examined transcript-per-million (TPM) normalized counts for the 12,419 genes, where RNA-Seq data could be linked between cell lines and tumors. For DNA methylation data (MET), we examined beta values for 10,141 probes from methylation arrays after selecting a single probe per gene promoter with the highest variance across the dataset. To align human tumor and cell line data, we quantile-normalized the data and applied ComBat, a batch effect correction method (*19*), which is highly performant compared to other related methods (*20*). In brief, ComBat estimates parameters for location and scale adjustment of each batch (TCGA and CL in our case) for each gene. Then, it removes the variability that is particular to the CL but not present in TCGA, while retaining the intra-dataset variability of the tumors, which should presumably be evident in both the tumor and in the cell line datasets.

A principal components analysis (PCA) in the data (before and after adjustment) suggests that there were strong global differences between TCGA and CL, which are largely removed by our approach (Fig. 1A and fig. S1A). To quantify this, we calculated the effect size (Cohen's *d* statistic) for each gene/probe between TCGA and CL before and after adjustment. It can be observed that these differences are reduced to the minimum after adjustment (fig. S1C). In addition, we trained a classification model that predicts the CL versus TCGA origin of the data points based on GE and MET (Fig. 1B and fig. S1B). The model is able to distinguish CL versus TCGA perfectly when using the preadjustment datasets [area under the receiver operating characteristic (ROC) curve (AUC) = 1/area under the precision-recall (PR) curve (AUPRC) = 1], while the post-adjustment datasets do not perform better than random (~0.5 for AUC and ~0 for AUPRC), suggesting that the cell type–specific signal has been largely removed. Last, we tested the optimal number of features (genes/probes) using tumor classifiers and calculating the accuracy in the cell line data (table S1A); we selected 5000 features with the highest SD for subsequent analyses.

Once the data were aligned, we set out to determine which cell lines have tissue identity not matching the declared tissue of origin (henceforth, "suspect set") and, conversely, which cell lines have largely retained their tissue identity (henceforth, "golden set"), by comparing against a large set of tumors from 17 tissues in the TCGA (Fig. 1C). Using TCGA data, we derived one-versus-rest classification models (using ridge regression), separately for the GE and the MET data. These two data types were used because they yielded higher accuracies for the one-versus-rest setting than the four different mutation-based classifier types we tested (fig. S1G); the mutation-based classifiers were nonetheless useful for subsequent validation analyses (see below). Some pairs of cancer types were considered jointly based on their overall similarity, for example, stomach adenocarcinoma (TCGA code: STAD) and esophageal adenocarcinoma (subset of samples from TCGA code: ESAD); see Materials and Methods for a full list. Our study examines solid cancer types and blood cancers in separate analyses.

Differential tumor purity across the TCGA does not have a notable bearing on our tissue classification: Accuracies of models from higher-purity tumor samples were similar to models from other tumor samples (fig. S1D). Moreover, introducing GE data of healthy tissues [from Genotype-Tissue Expression (GTEx)] into a joint analysis with the TCGA tumor data did not further improve the GE classifier (fig. S1F). GE data from healthy tissues, considered by themselves, were less informative for assigning cancer cell lines to tissues of origin than GE data from tumors (fig. S1E), consistent with a tumoral origin of the cell lines.

Next, we obtained predictions of cancer type identity for each cell line. For every cancer type, we split TCGA data randomly into training and testing sets, and we used the calculated PR curve of the TCGA testing dataset to obtain the precision score for every cell line

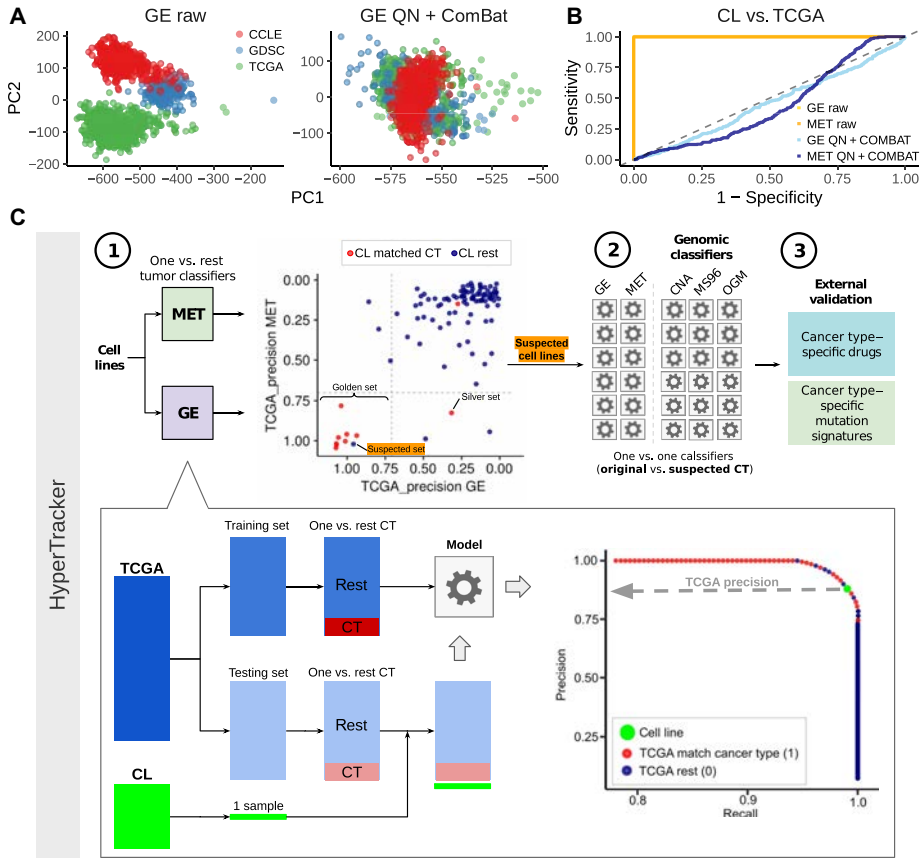


Fig. 1. Methodology for data alignment and cancer type classification. (**A**) Principal component (PC) 1 and PC2 of a PCA, in the gene expression (GE) data, before adjustment for batch effects (raw data) and after adjustment [quantile normalization (QN) + ComBat] [see fig. S1 for PCA of DNA methylation data (MET)]. Colors represent the dataset sources (GDSC and CCLE are two sources for the cell line data, and TCGA is the source for the tumor data). (**B**) ROC curves for classifying tumors versus cell lines in the data before adjustment (orange) and after adjustment (blue) for GE and MET. (**C**) Schematic overview of the HyperTracker methodology. First, we systematically identified possible mislabeled cell lines using GE and MET data, independently. Second, we used various types of mutation-based data to corroborate the predictions. Third, we further validated the cell lines (CL) suspected to originate from skin using independent data, such as drug sensitivity. CT, cancer type.

(details in Materials and Methods; all TCGA-derived precision score values are listed in table S1, B and C). The higher the precision, the more likely the cell line is to belong to that particular cancer type. As expected, most of the cancer type labels of the cell lines match the declared tissue of origin of that cell line—they tend to cluster at high precision values for the cognate cancer type (red dots in Fig. 2A and fig. S2). However, among these many correctly classified cell lines (red dots), there are some with similarly high precision scores, but which were originally annotated as belonging to another cancer type (Fig. 2A, blue dots with labels shown). A clustering analysis of the GE and MET values for the genes with the highest weights in the classification models (Fig. 2B and fig. S3) showed that the samples generally cluster by cancer type, but not by CL versus TCGA label. Moreover, we observed that the suspect cell lines (i.e., cell lines with highly confident precision scores to a different cancer type) tend to cluster with the newly assigned cancer type, rather than with the original one (Fig. 2B).

In further analyses, we designated as the golden set those cell lines that have precision ≥0.7 (see fig. S4 and Materials and Methods for threshold selection) for both GE and, independently, MET in their originally declared cancer type (n = 366 of 614 examined cell lines, 60%). For these cell lines, two independent types of evidence—transcriptomes and epigenomes—support that they match their expected cancer type well, suggesting that these cell lines would be preferred experimental models. Further, we designated as the "silver set" those cell lines that have precision ≥0.7 for one classifier (either GE or MET but not both) (n = 131 of 614, 21%). From the remaining 117 cell lines, we selected as suspect set those CL that exhibit a precision of ≥0.7 for both GE and for MET, but for a different cancer type than declared for that cell line (n = 43 of 614, 7% of analyzed cell lines) (Fig. 2C and fig. S4C). This set of cell lines may consist either of mislabeled cell lines, where the cancer type of origin is different than initially thought, or of heavily diverged cell lines, where the genomic and/or epigenomic alterations accumulating during culture have overridden the original cancer type identity. Notably, cell line cross-contamination issues (*21*) cannot commonly underlie the trends we observe, because the repositories that provided GE and MET data have used genetic markers to ascertain the identity of the cell lines (*4*). The fact that two classifiers based on independent data types—one transcriptomic and one epigenomic—reached the same predictions adds confidence that these are bona fide cases of mistaken tissue/cell type identity. In case of blood cancer classifiers,

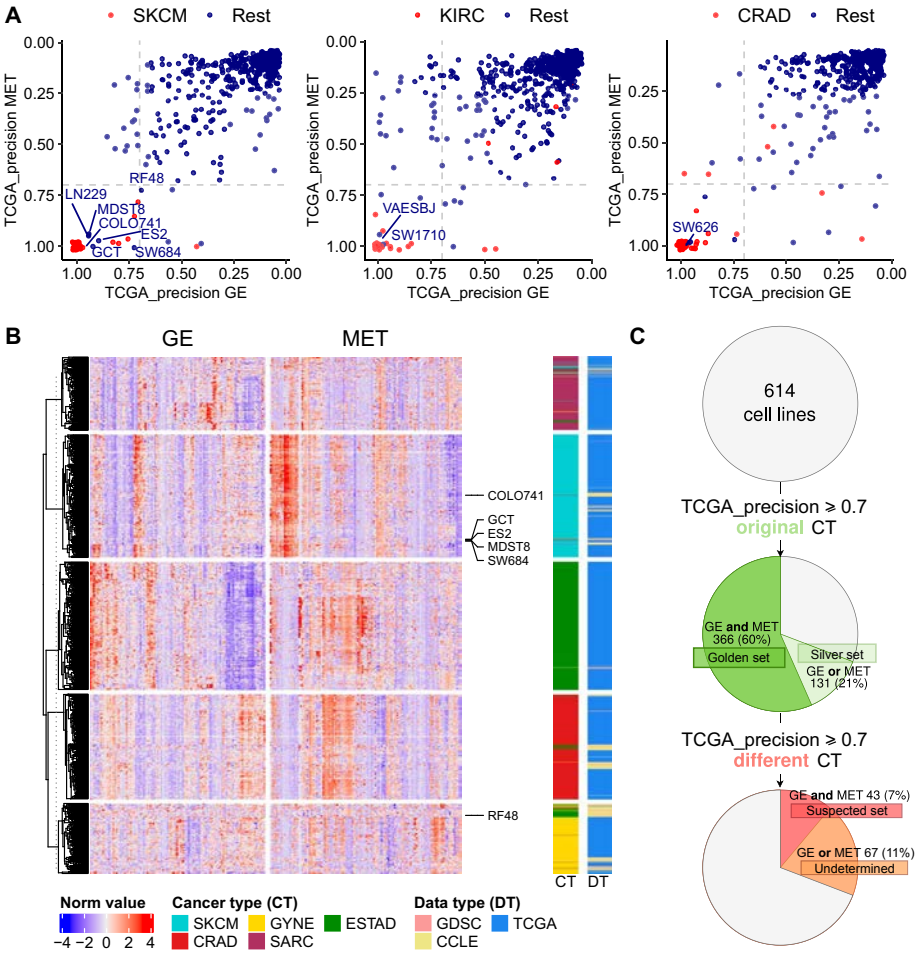


Fig. 2. Detection of cell lines suspected to be mislabeled with a different cancer type. (**A**) TCGA-based precision scores for 614 cell lines were calculated in the MET and GE cancer type classifiers (one-versus-rest) and for 69 blood cancer cell lines. The higher the precision, the higher the confidence that the sample belongs to that particular cancer type (here, showing cases of SKCM, KIRC, and CRAD from left to right; see fig. S2 for the other cancer types). The cell lines that were originally annotated as the cancer type that is being tested are shown in red, and the rest in blue. (**B**) Heat map showing the 25 genes (GE) and CpG probes (MET) with the highest absolute values of ridge regression coefficients for each of the cancer types in the plot in one-versus-rest classifiers. The suspected skin cell lines are labeled. The cancer types shown are the suspected cancer type [melanoma (SKCM) in this case] and, additionally, the originally declared cancer types of the suspected cell lines [here, esophagus and stomach cancer (ESTAD), sarcoma (SARC), colorectal cancer (CRAD), and ovarian and uterus cancer (GYNE)]. See fig. S3 for the heat maps for the rest of the suspected cell lines. (**C**) Overview of the results from the systematic mislabeling testing of all cell lines. Cell lines with a TCGA_precision ≥ 0.7 to its original cancer type in (i) both in GE and in MET are assigned to the golden set group and (ii) either in GE or in MET are assigned to the silver set. If, however, the TCGA-based precision ≥ 0.7 to a different cancer type in GE and in MET, the cell line is assigned to the suspect set.

which were highly accurate in distinguishing myeloid and lymphoid lineages, there were no cell lines suspected to be mislabeled between the lineages (fig. S2B).

Validation of individual examples of suspected mislabeled cell lines using genomic classifiers

We detected 43 cell lines that bear both transcriptomic and epigenomic features of a different cancer type than the one they were originally annotated with. We next turned to support individual examples of reassigned tissue identity by analyzing independent data. In particular, we used genomic sequence–based classifiers, which are able to predict the tissue of origin based on somatic mutation patterns (*22*, *23*), in particular, the trinucleotide mutation spectra and the presence of oncogenic mutations and CNA profiles (*22*, *23*). In this validation setting, we applied these genomic classifiers to a problem of one-versus-one classification, where we contrasted the originally assigned cancer type versus the newly proposed cancer type for each reassignment. We found that these one-versus-one classifiers based on genomic data had satisfactory accuracy with whole-exome sequencing (WES) datasets that we used [fig. S5; notably, our recent work (*22*) suggests that whole-genome sequences, when available, would provide more powerful classifiers that draw on regional mutation density (RMD) patterns].

For the 43 examples of suspect cell line tissue identity, we first derived one-versus-one classification models separately for GE and MET and prioritized reassignments that were consistently observed across multiple runs of the classification algorithm. We randomized the labels to obtain a background model of expected values of the

consistency score for the reclassification (Fig. 3B and fig. S6A). From the 43 suspect cell lines, 35 are consistently reassigned to the other tissue (score > 10) (Fig. 3A and fig. S6B). Next, we calculated the same score for the genomic classifiers (based on mutations and CNA, as described above) on these 35 suspect cell lines (Fig. 3A).

Of these, approximately two-thirds (n = 22 cell lines) received high support for the new tissue label by one or more genomic classifiers [Fig. 3A; consistency score ≥ 15, corresponding to randomization-based false discovery rates (FDRs) of 0, 0, and 18% for the CNA, OGM, and MS96 classifiers, respectively; Fig. 3B]. These data suggest that 22 cell lines are candidates for assignment to another cancer type, based on converging evidence from the levels of the genome, epigenome, and transcriptome, which provides higher confidence. Reassuringly, this list contains two cell lines that have been previously shown to be misclassified: SW626, which was initially annotated as ovarian cancer but later discovered to be derived from colon cancer (*24*), and COLO741, which was originally thought to be a colon adenocarcinoma cell line but later shown to originate from a melanoma (*25*). The fact that these two known examples were detected and reassigned to the correct cancer type provides evidence that our HyperTracker method is overall reliable.

The two plausible reasons why a cell line thought to originate from one cell type would be reassigned to a different cell type are (i) that, at the time of isolation, the cell line was not of the type that it was thought to be (mislabeling) and (ii) that, during prolonged cell culture, the cell line diverged greatly and now resembles another

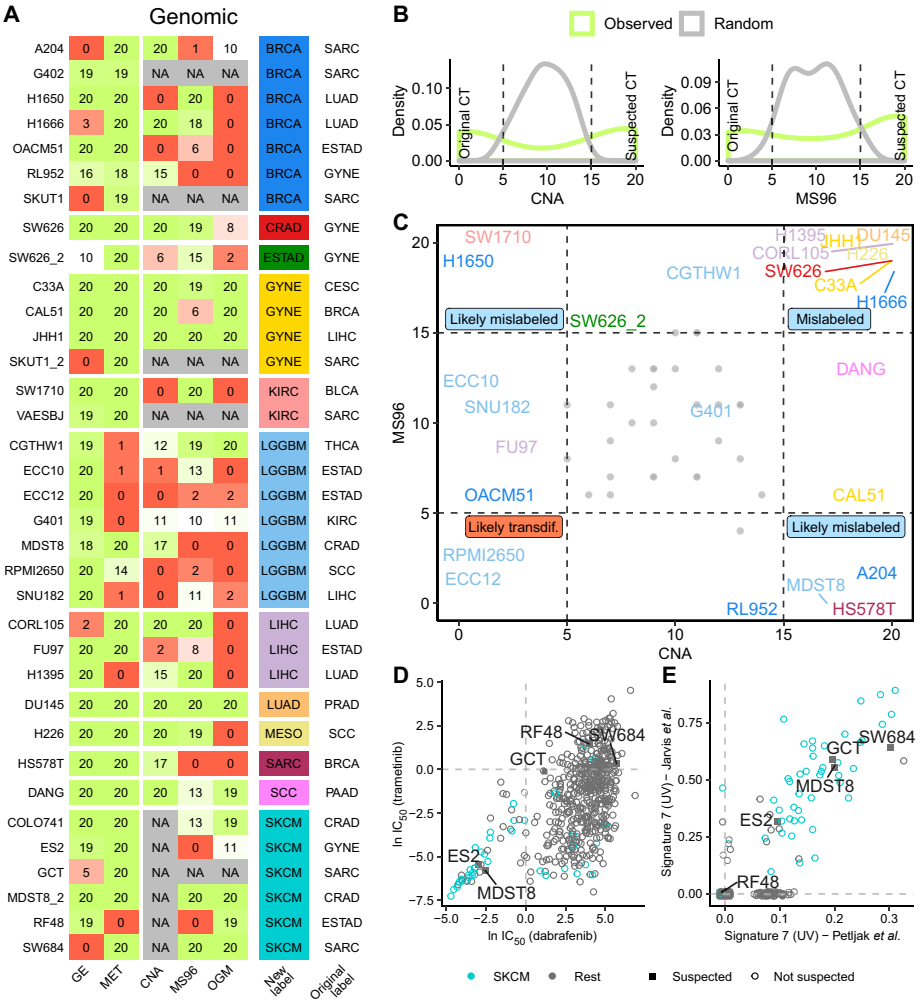


Fig. 3. Additional evidence supporting tissue identity of the suspected mislabeled cell lines. (**A**) Prediction consistency score (0 to 20) for each suspected cell line for 20 runs of one-versus-one classifiers that predicted suspected versus original cancer type in GE, MET, CNA, trinucleotide mutation spectrum (MS96), and oncogenic mutations (OGM). A value of 20 means that the cell line is predicted as suspected consistently in the 20 runs of the classification algorithm, and a value of 0 means that it is predicted as original cancer type. (**B**) Histograms of the consistency scores for CNA and MS96 classifiers for the models based on actual data and a baseline expectation on randomized data. (**C**) Prediction scores for MS96 and CNA for the suspected cell lines. Colors represent the suspected cancer type [see column "new" in (A)]. Gray dots represent the random values. (**D**) Drug sensitivity (IC_{50}) for mutant BRAF-targeting drugs dabrafenib and trametinib for 614 cell lines. Cell lines originally labeled as skin cancer are shown in turquoise, and skin-suspected cell lines are marked with a square and their name. (**E**) Burden of UV-associated mutation signature 7 (estimated from two different sources) in 614 cell lines. Cell lines originally labeled as skin cancer are shown in turquoise, and skin-suspected cell lines are marked with a square and the name label.

cell type (transdifferentiation). Our data allow us to examine how prevalent each case is: Mislabeling is expected to be reflected equally in both the epigenome and the genome, while transdifferentiation is expected to be reflected more strongly in the (presumably more malleable) epigenome, and less so in the genome, which retains the mutations from the original tumor. We suggest that mislabeling at isolation is a much more common scenario (Fig. 3C, many reassigned cell lines are in the upper right corner). However, it is possible that there exist individual examples of cell lines that have effectively transdifferentiated during culture, because their genomic features are consistent with the original tissue identity, while the epigenomic features are consistent with another tissue (Fig. 3C, lower left corner; e.g., the RPMI2650 and OACM51 cell lines are possible candidates).

Validation of cell lines suspected to originate from the skin

From the previous analysis, we identified six cell lines that are reassigned from various cancer types to skin cancer. We note that, of skin cancers, the TCGA study contains only melanoma but not the nonmelanoma skin cancers, so we were not able to distinguish between cell type identities of different skin cancers.

To further support that these cell lines are skin cells, we performed an independent analysis based on mutational signatures. Large-scale analyses of trinucleotide mutation spectra across human tumors have revealed at least 30 different types of mutational signatures (*26*). Of these, signature 7 (C>T changes in CC and TC contexts) was associated with exposure to ultraviolet (UV) light and is highly abundant in sun-exposed melanoma tumors (*27*). The same signatures were recently estimated in cancer cell lines (*28*, *29*), which enabled us to use the existence of UV-linked signature 7 to examine whether these cell lines originated from the skin. On the basis of mutational burden of signature 7, the known melanoma cell lines are clearly separated from the rest (Fig. 3E), meaning the approach can distinguish skin-derived cells. Among the melanoma cell lines with high mutational burden of signature 7, we found four of five of the suspected cell lines (Fig. 3E), in particular, GCT, SW684, ES2, and MDST8 are very likely skin cell lines, and not sarcoma, sarcoma, ovarian cancer, or colorectal cancer, respectively, as originally thought. For the sixth suspected cell line COLO741, the mutational signature data are not available; however, COLO741 has previously been reported to express skin-specific genes (*25*).

The RF48 cell line (originally considered stomach, here putatively reassigned to skin) does not exhibit the UV signature or the DNA methylation patterns of skin; therefore, a highly confident call cannot be made. Nonetheless, a pattern of cancer driver mutations in RF48 suggests that it is not a stomach cell line (Fig. 3A). Previous work based on gene expression proposed a lymphoid origin for RF48 (*30*).

Next, we sought to substantiate these findings using drug sensitivity data. In particular, two drugs (dabrafenib and trametinib) that target mutant BRAF are approved for treating melanoma in the clinic. These drugs are known to have poor efficacy in other cancer types bearing *BRAF* mutations, such as in colon cancers (*31*). Therefore, sensitivity to these drugs adds confidence that we are looking at a melanoma cell line (note that the converse does not necessarily hold here: resistance does not imply that the cell line is not a melanoma). Therefore, we compared the median inhibitory concentration (IC_{50}) of these two drugs for all cell lines (Fig. 3D). As expected, many melanoma cell lines cluster at low values of IC_{50} for the two drugs, meaning that these cells are sensitive to the drug. This includes two of five of our suspected cell lines (ES2 and MDST8), providing further supporting evidence that these are of skin, likely melanoma, origin.

In conclusion, out of six cell lines suspected to originate from skin, four were confirmed by the UV mutational signatures and two were additionally confirmed by the drug sensitivity to BRAF/MEK (mitogen-activated protein kinase kinase) inhibitors. This notable example demonstrates how the transcriptome- and epigenome-based tissue/cell type classifiers are able to link cultured human cell lines with their correct cancer type of origin.

In addition to these examples of skin cell lines, we were able to support several other cancer type reassignments using drug sensitivity (results are summarized in table S1D) (*32*). The DANG cell line is consistent with squamous cell carcinoma of the lung or of the head and neck (SCC), rather than with its original assignment of pancreatic adenocarcinoma (notably, this reassignment is also observed with multiple genomic classifiers; Fig. 3A). Similarly, SW1710 may be a kidney, rather than a bladder, cell line, based on the original reassignment via transcriptome and epigenome, supported by the mutation patterns (Fig. 3A) and additionally supported in the global analysis of drug responses (table S1D). We note that such analyses of drug screening data can be applied to distinguish only certain pairs of tissues and not all reassignments can be reliably validated in this test (see AUC scores in table S1D).

Identification of subtypes for cell lines using multi-omics analyses

Tumors are heterogeneous, and major differences exist between tumor samples of the same cancer type. To manage this variability, cancer types are subdivided on the basis of their molecular characteristics, including global patterns in gene expression and DNA methylation (*33*–*35*). However, with the exception of a few tumor types, prominently breast cancer, molecular subtypes are still being established or refined, often with the goal of better predicting disease progression in response to particular treatments. Because drug screens and genetic screens performed on cell line panels have the goal of serving as models for actual tumors, it is useful to be able to transfer the subtype assignments from tumors to cell lines, thereby establishing which cell lines are the most appropriate model for each cancer subtype.

Previously, molecular subtypes from tumors have been inferred in cell lines using different strategies, often based on gene expression, for example, in breast (*36*), colorectal (*37*), and renal cancer (*13*). In a recent pan-cancer study, subtypes have been assigned to a set of 600 cell lines (*38*).

Our approach to assign subtypes to cell lines is to apply the same strategies that yielded accurate cancer type classifiers: first, the integration of transcriptomic and epigenomic data to boost confidence in the predictions, and second, careful adjustment of the datasets to make them comparable between TCGA tumors and cell lines (fig. S1). An important consideration here is the lack of known labels needed to assess accuracy; thus, assignments should be treated as tentative. However, for breast cancer cell lines, the subtype labels are available and can be used as a benchmark (*36*).

We examined subtypes proposed for 15 cancer types in the TCGA and generated subtype classifiers (see Materials and Methods) for each cancer type. The combination of both data types (GE and MET) achieved a higher cross-validation accuracy in the TCGA (median AUPRC across cancer types: 0.81) than GE (0.76) or MET (0.72) individually. Therefore, we used the combined datasets to

generate subtype classifiers and propose assignments of the cell lines to cancer subtypes. Most were uniquely assigned to a single subtype (fig. S7A); we used only those in further analysis. As a benchmark, we calculated the accuracy for the breast cancer cell lines with subtypes available (fig. S7B): The median AUPRC across subtypes for CL is 0.83. This suggests acceptable performance in obtaining tentative subtype assignments for cell lines in all 15 cancer types, which we provide in table S1E. This resource is complementary to a recent set of subtype predictions for nine cancer types based on transcriptomes (*38*).

Next, we examined whether the relative prevalence of subtypes is similar between tumors and cell line panels of the same cancer type. Cell line panels of some cancer types have good representation of subtypes, for instance, lung squamous cell cancer, head and neck squamous cell cancer, lung adenocarcinoma, and gastric/esophageal cancers (fig. S7C). However, the converse is the case for liver, skin, and thyroid cancer cell lines, where a single subtype predominates in cell line panels, unlike in tumors (statistics are listed in table S1F). In addition, we observe suboptimal representation (half of the tumor subtypes are not represented) in the kidney, bladder, and brain cancer cell line panels, when considering the 463 cell lines we analyzed. This suggests that, in some cancer types more than others, the commonly used cell line panels do not represent the diversity of molecular subtypes in tumors well. One possible reason for this is the relative ease of culturing certain subtypes compared to others (*2*).

Accounting for mislabeled cell lines reveals associations in drug screening data

We detected 35 cell lines that may have a tissue or cell type identity different than the one originally assigned to them. Because the cell type is an important determinant of drug response in cancer cell lines and in tumors (*10*), we hypothesized that the inclusion of this new tissue information when searching for genetic determinants of drug sensitivity may change the results. In a comprehensive study, Iorio *et al.* (*10*) searched for associations between drug response and cancer functional events (CFEs): recurrent mutations, CNA, and hypermethylation events present in human tumors. Here, we used GDSCTools (*39*) to reproduce the results of that study, however, after filtering the cell lines to those that better represent the cancer type in question. In particular, we repeated the same analysis using for each tissue (i) all the cell lines, (ii) only the cell lines in the golden set (G), and (iii) as a less stringent filtering criterion, only the cell lines in the golden set and silver set (G&S). In addition, as controls, we included a random subset of cell lines that matches (iv) the number of cell lines in golden set (r_G) and (v) the number in "golden and silver set" combined (r_G&S).

For most of the cancer types, we observed that one of the filtered subsets recovered a higher number of significant [at FDR ≤ 25%, as applied in the original study (*10*)] associations of CFE with drug sensitivity or resistance than were recovered using all cell lines (fig. S8A and table S1G). For instance, for glioblastoma, using the golden set cell lines, we found 23 associations, which were not recovered from the entire cell line panel or from the random subset controls (Fig. 4B). For example, this recovers the positive association of *CDKN2A* loss with camptothecin sensitivity (Fig. 4C), which was previously reported in an independent analysis of the NCI-60 cell line panel screening data (*40*). Similarly, for pancreatic adenocarcinoma, benefits were observed by focusing on cell lines that resemble the corresponding cancer type better: Using only the golden set plus silver set cell lines, 10 significant, previously unknown associations were found (Fig. 4B). For instance, we detected that *SMAD4*-mutant cell lines are more resistant to piperlongumine, a natural product that exerts antitumor properties via multiple pathways (*41*). Mutations in the tumor suppressor gene *EP300* were associated with higher sensitivity to three drugs in pancreatic cancer cell lines (Fig. 4D). The observation that more associations were found despite using a somewhat lower number of cell lines (thus less statistical power) emphasizes the importance of focusing on the cell lines that more closely model the tissue and/or cell type of origin of the cognate tumor.

Colorectal cancer provides an illustrative example of the importance of removing nonrepresentative cell lines from drug screening efforts. In the original study (*10*), 50 colorectal adenocarcinoma (CRAD) cell lines were tested. Of those, we strongly suspected that MDST8 derives from skin. To test the influence of this individual mislabeled cell line, we have performed association testing with all colorectal cell lines and again after excluding MDST8. For the association between dabrafenib response and *BRAF* mutation status, we observed that CRAD cell lines in general (i.e., irrespective of *BRAF* mutation) are not sensitive to dabrafenib, except for MDST8, which is strongly sensitive to dabrafenib (Fig. 4A). The FDR of the analysis of variance (ANOVA) analysis when using all (allegedly) colorectal cell lines is significant at 6%, while when removing MDST8 the FDR for *BRAF*-dabrafenib association in colon becomes nonsignificant (45%). Therefore, in this case, the presence of a single mislabeled cell line is sufficient to cause the appearance of an association between a drug and a feature, which is likely not relevant for this cancer type. This is evidenced in the clinical responses of patients: *BRAF*-mutant melanoma patients respond well to dabrafenib, while colorectal tumors with the same *BRAF* V600 mutation are not sensitive to BRAF or MEK inhibitor monotherapy (*31*).

Accounting for mislabeled cell lines in genetic screening data analyses

Motivated by the associations revealed by reanalyzing the drug screening data, we asked whether the same extends to genetic screening data in cancer cell lines, because results in genetic screens also depend on cell lineage (*11*). To further investigate, we analyzed CRISPR screening data from Project Score and Project Achilles (see Materials and Methods), from which 347 cell lines overlap our tested cell lines. Then, we applied the same association testing method, which was, however, underpowered, because the number of available overlapping cell lines was smaller. Nonetheless, in colorectal and ovarian cancer (which have the largest number of cell lines in this dataset), we observed that by focusing only on the golden set and/or silver set, the number of recovered associations increased (as a control, there were no increases in randomly chosen cell line subsets of the same size; fig. S8B and table S1H).

To illustrate the importance of removing suspect cell lines in gene dependency screenings, we provide two examples of associations that were originally not significant because of the presence of a mislabeled cell line. For ovarian cancer, the presence of SW626 [mislabeled cell line confirmed by the literature (*24*)] prevents finding the association between *MED8* dependency and a copy number gain in the region containing *ASXL1* ["cnaOV72" in (*10*)] as significant (fig. S9A). Similarly, for colorectal cancer, the presence of MDST8 (mislabeled cell line confirmed by the UV mutational signature) prevents detection of the association between *TUBB4B* dependency and a copy number gain in the region containing *STK4*

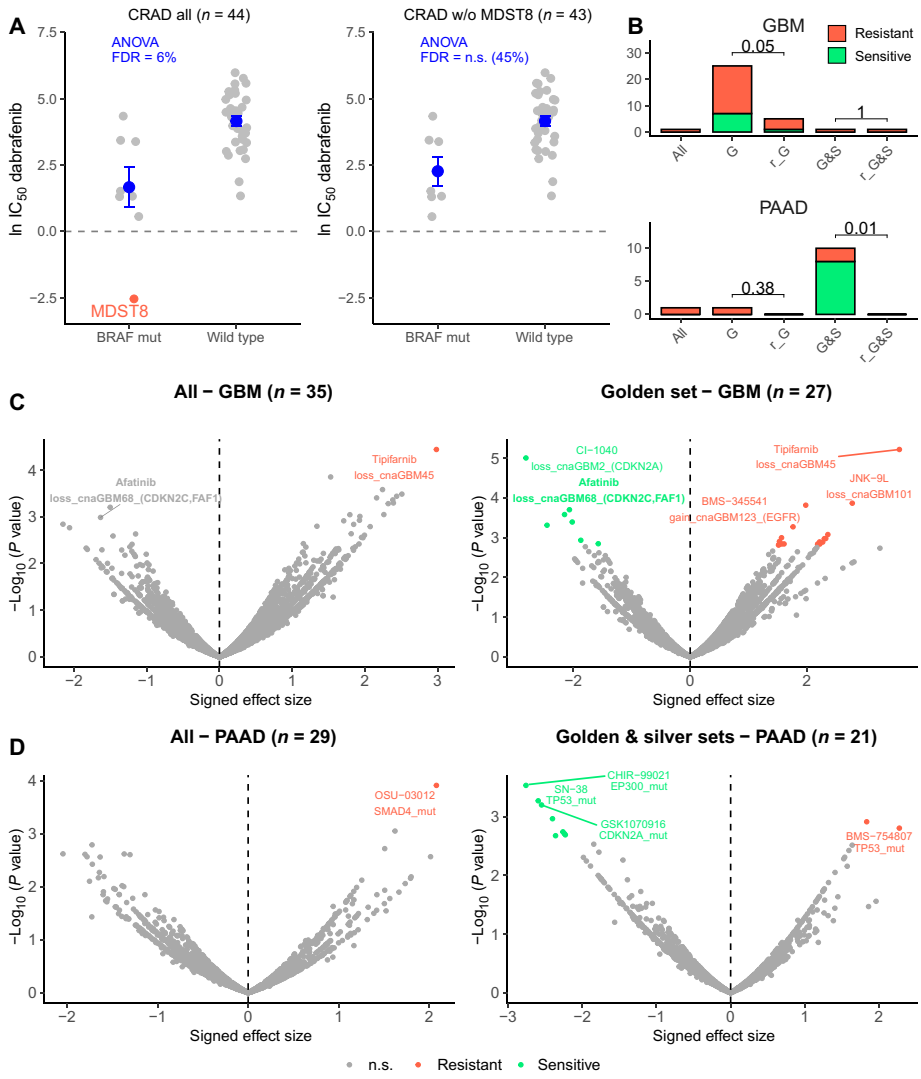


Fig. 4. Drug sensitivity association testing using high-confidence sets of cell lines. (**A**) Drug sensitivity (IC_{50}) to dabrafenib in all colorectal (CRAD) cell lines (left) and all CRAD cell lines except MDST8 (right), which is suspected of being skin cancer. Cell lines with a *BRAF* mutation and without (wild type) are compared. ANOVA FDR for this association (dabrafenib sensitivity with *BRAF* mutation status) is shown in blue for both datasets. Horizontal line is shown at 0, because score < 0 implies sensitivity to the drug. Dots and error bars represent the mean and SEM. (**B**) Number of significant associations between "CFEs" (includes mutations and CNAs in cancer genes) and drug associations detected (at FDR 25%) in the ANOVA test for all cell lines ("all"), cell lines in the golden set ("G"), cell lines in the golden plus silver sets ("G&S"), random subset of cell lines that match the number in the golden set ("r_G"), and random subset of cell lines that match the number in the golden plus silver sets ("r_G&S"). For the random subsets, the number of significant associations is calculated from 10 random selections and median shown. *P* values for a sign test (one-tailed) between the number of associations in the G/G&S and in r_G/r_G&S are shown. See fig. S7 for the remaining cancer types. (**C**) Differential sensitivity of drugs was analyzed by ANOVA for all brain cancer cell lines (left) and the brain cancer cell lines in the golden set only (right). Each point is an association between the sensitivity of a drug and a genetic feature (CFE). (**D**) Differential sensitivity of drugs was analyzed by ANOVA for all pancreatic (PAAD) cell lines (left) and PAAD cell lines in the golden and silver set only (right). Each point is an association between the sensitivity of a drug and a genetic feature (CFE). n.s., not significant.

("cnaCOREAD32") (fig. S9B). Next, a significant association between *WRN* dependency and *MLL2* (also known as *KMT2D*) gene mutation is recovered only with the filtered cell lines in ovarian cancer (fig. S9C). This *WRN-MLL2* association has been recently reported using a different set of cell lines (from Project Score) (*42*) that partially overlap our set.

Last, our reanalyses of drug screening and genetic screening data revealed an interesting association independently supported in both drug and genetic data. The drug afatinib inhibits the epidermal growth factor receptor (EGFR) protein and is clinically indicated for *EGFR*-mutated lung cancer; however, in *EGFR*-altered glioblastoma, afatinib is generally not considered to elicit a response (*43*). Consistently, afatinib sensitivity was associated with *EGFR* alterations in lung cancer previously (*10*), as well as in our reanalysis (FDR lung G&S sets = 0.6%), but not in the brain cell line panel (all cell lines, FDR ≥ 25%). However, using the focused (golden set) of brain cancer cell lines revealed a significant association (ANOVA FDR = 15%; fig. S9D) between afatinib sensitivity and a different genetic alteration: copy

number loss in a region at 1p32.3 containing the *CDKN2C* and *FAF1* genes ["cnaGBM68" in (*10*)]. The same loss at 1p32.3 is associated with sensitivity to genetic knockout of *EGFR* in brain cell line panels in two independent large-scale genetic screens (Project Scores and Project Achilles; fig. S9, F and G) and to pharmacological inhibition in another drug screen (PRISM; fig. S9E). The meta-analysis of the two drug screens and two genetic screens suggests high strength of combined evidence ($P = 0.00094$, Fisher's method of combining P values) linking the CNA loss at 1p32.3 (chr1: 51169045-51472178) with sensitivity to pharmacological or genetic EGFR inhibition in brain cells, suggesting a candidate marker for follow-up work.

In summary, the presence of cell lines with dubious or incorrect labels of tissue identity may strongly affect association studies of drug or CRISPR screening data in two different ways. First, the presence of mislabeled cell lines can cause the appearance of spurious associations that do not reflect the biology of the cancer type of interest. Second, the presence of mislabeled or divergent cell lines can prevent the recovery of true associations.

DISCUSSION

Cell lines are commonly used as models for tumors; however, it is an open question how to best apply the available cell line panels to learn about cancer biology. The availability of genomic data from large tumor cohorts and from cell line panels has spurred efforts to find which cell lines are closer to tumors by their transcriptomic (*9*, *37*, *38*) and/or genomic features (*12*, *13*), presumably making better models, and which are more distant from examples of actual tumors, thus making less good models.

Our work addresses a different question: We attempt to detect the cancer type (i.e., tissue and/or cell type) that originated the cell line to ascertain whether this matches the declared origin of the cell line. A mismatch may conceivably stem from the sampling step, for instance, a metastasis might have a different tissue of origin than thought at the time of surgical collection. The work-up after collecting the tumor sample may have inaccurately assigned the cancer type, based on unclear histological or anatomical features. Another possibility is that the mismatch might stem from the step of adaptation to cell culture, where a minority cell type that is not representative of the tumor prevails over other tumoral cells. We consider these to be cases of cell line mislabeling during isolation. In addition, we would also detect cases where the cell line might have acquired some properties of a different tissue/cell type during extended periods in culture; however, our analyses (Fig. 3C) suggest that this is a less common occurrence, although individual examples cannot be ruled out.

This phenomenon of tissue/cell type mislabeling is distinct from well-known and widespread cell line misidentification issues (*21*), where one cell line (often HeLa) was mistakenly used in place of another cell line originating from a different individual, commonly due to cross-contamination. The cell line panels that provided data used in our analyses (GDSC and CCLE) have authenticated their cell lines (*4*, *42*); thus, misidentification/cross-contamination cannot underlie our observations of mislabeling of the cancer type of origin. [We note that there were rare cases of misidentified cells reported in these panels (*42*); however, these do not overlap our results.]

Methodologically improving over previous work, we introduce the HyperTracker framework that performs global analyses, which independently examine transcriptomic, epigenomic, and mutational features. In addition, we carefully adjust for the known bulk differences between cell lines and tumors, which might have resulted, e.g., from impurities in tumors or from altered expression of cell cycle–related genes in cell lines (*38*). Parallel analyses of different omics data provide increased confidence in our inferences, which suggested, remarkably, that 5.7% (35 of total 614 considered cell lines) exhibit significant transcriptomic and epigenomic features of a different tissue/cell type than the declared cell type of origin. For 3.6% (22 cell lines), these reassignments to a different cancer type were additionally supported in at least one type of genomic evidence. This increased confidence that these were examples of cell lines with mislabeled (or, less likely, diverged) tissue/cell type identity. Notable examples are cell lines GCT, SW684, ES2, and MDST8 that we predict to originate from the skin, based on the presence of the UV mutational signature, in addition to strong evidence in transcriptome/epigenome data. These cases are reminiscent of the recent reports of UV mutational signatures found in some cases of presumed lung cancers, suggesting that they may instead be metastases originating from sun-exposed skin (*44*).

In interpreting our data, an important consideration is that the cancer sample types in TCGA may not necessarily reflect the full diversity of rarer subtypes within a cancer type, which may cause some ambiguous predictions. For instance, the ECC10 and ECC12 cell lines are assigned to STAD cancer type (stomach adenocarcinoma) when matched with TCGA tumors. These cell lines originate from gastric small cell neuroendocrine carcinomas. This might explain why, in our analysis, gene expression patterns point toward brain tissues, while mutational features suggest stomach cancer. In such cases of disagreement between different types of features, a future use of a more exhaustive set of reference tumor data, which includes rarer cancer types, may resolve the ambiguity and improve confidence in predictions.

The genomic classifiers we used here were based on whole-exome sequences and were overall less powerful than the transcriptome/DNA methylation classifiers in our data (figs. S1G and S5). Recent work by us and others (*22*, *23*) suggest that analyzing whole-genome sequences of these cell lines would permit use of additional, highly predictive features based on RMD of chromosomal domains. This may provide further genomic evidence for the identity of the cell of origin for the 35 suspected cell lines. Furthermore, targeted experimental follow-up work on these cell lines will provide further evidence to support or refute our predictions, which are based on global, multi-omics analyses.

Knowing the correct tissue-of-origin label for a cell line is important, because this has a strong bearing on the response of the cell line to drug treatment and to genetic perturbation. We demonstrate the implications of this general principle to analyses of drug and genetic screening data: By accounting for suspect cell lines, the power to discover determinants of sensitivity to pharmacological and to genetic perturbation may increase substantially for some cancer types, such as brain, lung, and pancreatic cancers. Therefore, when designing future screening efforts, it is important not only to increase the number of cell lines to gain more power but also to focus on the cell lines that best reflect the tissue and/or cell type of interest.

MATERIALS AND METHODS

Omics data collection and preparation

DNA methylation data

We downloaded DNA methylation data as beta values (platform: Illumina HumanMethylation450) from the GDC Data Portal (*45*)

for TCGA samples and from GDSC (*3*) for CL samples. We filtered out all probes outside promoter regions and probes with not available (NA) values in more than 100 samples. For the probes in promoter regions, we selected only one probe per gene that had the highest SD across samples. We transformed the beta values to m-values ($\log_2$ ratio of the intensities of methylated probe versus unmethylated probe). In total, this yielded 10,141 features for 942 CL samples and 8453 TCGA samples.

Gene expression data

We downloaded gene expression data as TPM from GDC Data Portal (*45*) for TCGA samples, from GDSC (*3*) and the CCLE (*4*) for CL samples, and from GTEx Portal for healthy data. We filtered out genes with NA values in more than 100 samples and selected the overlapping genes between the four sources. We removed low expressed genes (TPM < 1 in 90% of the samples) and applied a square-root transformation to the TPM data. In total, we have 12,419 features for 942 CL samples and 9149 TCGA samples.

For both DNA methylation (MET) and gene expression (GE), we created datasets of different sizes: 1000, 2000, 3000, 5000, and 8000 features by selecting the genes/probes with the highest SD across TCGA samples. In addition, we downloaded tumor purity data from Aran *et al.* (*46*) and used consensus measurement of purity estimations (CPE) for creating the groups of high- versus low-purity samples. High-purity samples were defined as CPE > 0.75, and for the lower-purity group, we took a subset of samples (same number of samples that are in high-purity subset) with the lowest purity.

CNA data

We downloaded CNA data (computed by gene) from GDC Data Portal (*45*) for TCGA samples and from DepMap (*47*) for CL samples. In total, we have 20,491 features for 942 CL samples and 9188 TCGA samples. To reduce the dataset, we selected CNA events in 299 known cancer driver genes (*48*).

Mutation data

For human tumors, we downloaded mutation data as WES MC3 dataset (*49*) from the GDC Data Portal for TCGA samples. For cell lines, aligned short reads (bam files) were obtained from the European Genome-phenome Archive (ID number: EGAD00001001039). Variant calling was performed using Strelka (version 2.8.4) with default parameters. Variant annotation was performed using ANNOVAR (version 2017-07-16). In samples where Strelka was unable to run, a realignment was performed using Picard tools (version 2.18.7) to convert the bams to FASTQ, and following that, the alignment was performed by executing bwa sampe (version 0.7.16a) with default parameters. The resulting bam files were sorted and indexed using Picard tools. To account for germline variants, we removed all mutations that were present in the gnomAD v2.1.1 database (*50*) at an allele frequency of $\geq$0.001 in any of the populations. Last, using the filtered somatic mutations, we calculated three sets of mutational features: RMD, mutation spectra (MS96), and oncogenic mutations (OGM) as described by Salvadores *et al.* (*22*). RMD features did not exhibit high accuracy when applied to exome-sequencing data and so were not considered further in this analysis.

For the cell line samples, we matched their cancer types to the TCGA cancer types using the cell line metadata from GDSC (*3*) and manually annotated those that did not have a TCGA cancer type label using Cellosaurus (*51*). Next, we selected the cell lines from solid tumors that had a matching cancer type in TCGA, yielding a total of 614 cell lines from 22 cancer types. Blood cancers (LAML and DLBC) are tested separately, because the cell lines there commonly grow in suspension, making them less likely to be confounded with solid tumors. For further analysis, we merged the cancer types that were overall similar: HNSC with LUSC and ESCC (jointly known as SCC), GBM with LGG (LGGBM), STAD with ESAD (ESTAD), and OV with UCEC (GYNE).

The identification of the cell line samples was performed by the laboratories generating the cell line databases, using short tandem repeat analysis (*4*, *42*). They reported a few commonly misidentified cell lines: Ca9-22, RIKEN, MKN28, KP-1N, OVMIU, and SK-MG-1 (*42*). These cell lines do not overlap with our suspected samples, and additionally, the misidentification does not affect tissue or cancer type of origin.

Data alignment between tumors and cell lines

For the alignment of TCGA and CL data, we first applied quantile normalization (R package preprocessCore 1.46.0) and next applied ComBat (R package sva 3.32.1), a batch effect correction method. We used ComBat as if our dataset was the TCGA and CL data combined, and the batch effects labels were whether a sample belongs to TCGA or CL (for MET) or a sample belongs to TCGA, GDSC, or CLLE (for GE). We applied this method for GE, MET, CNA, MS96, and RMD. For validation, we calculated a PCA, subsampling TCGA data to match the number of CL samples (stratified by cancer types). In addition, we calculated Elastic Net classifiers to predict (in the processed dataset) TCGA versus CL and calculated the AUC and AUPRC to check whether the process of alignment is being successful or not. In addition to the chosen adjustment method, we tested other approaches based on canonical correlation analysis, partial least squares, and PCA, which did not exceed accuracy of ComBat and therefore were not examined further.

Cancer type classifiers

For the TCGA dataset, we generated ridge regression model for predicting the cancer type in a one-versus-rest manner (using cv.glmnet function with alpha = 0 and family = binomial, R package glmnet 2.0.18). To calculate the accuracy, we trained classifiers in the TCGA dataset and tested in the CL dataset. In particular, we calculated the AUC and the AUPRC for each cancer type versus the rest (all the rest of cancer types combined).

TCGA_precision score

For each cell line, we calculated a TCGA_precision score of belonging to a particular cancer type. For this, we divided the TCGA data into two datasets (training and testing) of the same size, keeping the cancer type proportions. For each cancer type, we trained classifiers in the TCGA training dataset, and we introduced the cell lines one by one with the testing data and calculated the PR curve (TCGA testing + 1CL). We set the cell line precision score for that specific cancer type as the precision at the threshold where the cell line is situated in the PR curve. Overall, for every cell line, we obtained 17 precision scores, 1 for each possible cancer type. We repeated this procedure five times and calculated the median precision for every case to get more robust values. In addition, when training for one cancer type (label = 1) versus the rest of cancer types combined (label = 0), we made some exceptions and removed those cancer types that are similar, and therefore, the classifier is not good at separating them (e.g., when we calculated precision for ESTAD, we removed from the rest CRAD and PAAD, all hidden cases in table S1I). This is conservative with respect to reassigning cell lines to another cancer type; however, some resolution is traded off, because the more

closely related cancer types are, by design, not distinguished. We have further attempted to reclassify cell lines within these hidden tissues and the combined ones. However, when using one-versus-one classifiers, the accuracy is not good enough for distinguishing the two cancer types in the cell lines. We selected TCGA-based precision score ≥ 0.7 as a threshold to separate the different sets (golden, silver, and suspected sets), because this cutoff value approximately maximizes the F1 score for cell line classification (fig. S4A). At the TCGA-based precision threshold ≥ 0.7, the FDRs estimated on the original cell line labels for gene expression and DNA methylation classifiers would be 28 and 22%, respectively (fig. S4A); because the original labels are not always correct, these FDR estimates are conservative. Also, by visual inspection of the distribution densities of putatively correct predictions (originally labeled as the matching cancer type) and putatively incorrect predictions (originally labeled as another cancer type) for cell line tissue labels, one can appreciate how a TCGA_precision $\geq 70\%$ threshold separates well the two groups (fig. S4B).

Once we have a list of suspected cell lines, we have an "original" cancer type and a "suspected" cancer type. Therefore, we generated one-versus-one classifiers (original versus suspected) using TCGA dataset (balancing the classes), and for each suspected cell line, we checked whether it is predicted as original or suspected. We repeated this prediction 20 times and counted the number of times a cell line is predicted as suspected. Therefore, we defined a consistency score (range between 0 and 20) for every cell line, where 0 means never predicted as suspected and 20 always predicted as suspected. As a control, we repeated the same procedure with randomized cancer type labels, providing estimates of false discovery for different consistency score thresholds (Fig. 3B and fig. S6A). We calculated this prediction score for GE, MET, CNA, OGM, and MS96 datasets. For calculating the FDR at a score ≥ 15, we applied the following formula: FDR = FP/(FP + TP), where FP is the number of cell lines with score ≥ 15 in the randomized data and TP is the number of cell lines with score ≥ 15 in the actual data.

Independent validation

We downloaded drug sensitivity for the CL from the GDSC database (*3*). From the provided drugs, we selected trametinib and dabrafenib, U.S. Food and Drug Administration–approved drugs for melanoma treatment. We compared IC_{50} values for these two drugs for all cancer types.

We downloaded mutational signatures from cell lines available from Jarvis *et al.* (*29*) and Petljak *et al.* (*28*), and we compared the exposures of all cell lines for signature 7 (UV light). In Petljak *et al.* dataset, signature 7 is divided into signature 7a, b, c, and d. Therefore, we used the sum of exposures across all four subtypes of signature 7.

We downloaded another set of drug screening data (PRISM 19Q3) (*32*) for the CL dataset. For the suspected cell lines, we generated one-versus-one classifiers (using cv.glmnet function with alpha = 0 and family = binomial, R package glmnet 2.0.18) for predicting original versus suspected cancer type, based on the drug sensitivity data. We performed 20 runs of each case and counted how many times it is predicted as suspected (consistency score, which can vary from 0 to 20). In addition, we calculated the AUC for each classifier.

Subtype classifiers

We downloaded cancer subtypes for the TCGA samples from the R package TCGAbiolinks 2.12.6, which comprises many available molecular subtype classifications that have been described by TCGA-related publications [mRNA, DNA methylation, protein, miRNA, CNA, integrative (iCluster), and others]. From those, we selected the "subtype_selected," which is the classification that was chosen as the representative one in that cancer type (usually mRNA or integrative; see documentation of PanCancerAtlas_subtypes function for the complete list). We combined the GE and MET datasets to predict subtypes. For these data, we generated ridge regression model for predicting the subtypes in a one-versus-rest manner (using cv.glmnet function with alpha = 0 and family = binomial, R package glmnet 2.0.18) within each cancer type. We trained models in TCGA, and we predicted subtypes for the cell lines. In addition, we used cell line's subtypes for breast cancer from a previous paper (*36*) to calculate the confusion matrix and the AUPRC.

We performed a chi-square test (R package stats 3.6.0) and calculated the Cramer's *V* statistic (R package lsr 0.5) for checking whether the proportion of subtypes between TCGA and CL is maintained for each cancer type.

Drug and CRISPR screening data

We downloaded drug sensitivity and CFE data from Iorio *et al.* (*10*). CFEs are a collection of recurrent mutations, CNA, and hypermethylation events present in human tumors (*10*). We used GDSCTools (*39*) to search for associations between the drugs and the CFEs in every cancer as they did. In particular, we performed this analysis using for each tissue (i) all the cell lines, (ii) only the cell lines in the golden set (G), and (iii) only the cell lines in the golden and silver set (G&S). In addition, as controls, we included a random subset of all cell lines matching (iv) the number of cell lines in the golden set (r_G) and (v) the number in golden and silver set combined (r_G&S). We counted the number of significant hits (at FDR $\leq 25\%$) for each of the cancer types. For the controls, we repeated the subsampling 10 times and took the median of significant hits. We compared the number of hits for all the cell lines (same as in Iorio *et al.* study) with the number of hits for the different subsets of cell lines according to our grouping. In addition, we performed a sign test (R package BSDA 1.2.0) comparing the significant hits in the G/G&S subsets versus the significant hits over 10 runs in the random G/random G&S and calculated the *P* value for all cancer types (alternative = "less").

Similarly, we downloaded gene dependency data from Project Score (*42*) and Project Achilles (*52*) processed with the Project Score pipeline and combined them. From a total of 696 unique cell lines, 357 overlap with the 614 cell lines tested with our method. For those 357 tested cell lines, we repeated the same procedure as described above for the drug sensitivity data.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/27/eaba1862/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. G. Kaur, J. M. Dufour, Cell lines: Valuable tools or useless artifacts. *Spermatogenesis* **2**, 1–5 (2012).
2. A. Goodspeed, L. M. Heiser, J. W. Gray, J. C. Costello, Tumor-derived cell lines as molecular models of cancer pharmacogenomics. *Mol. Cancer Res.* **14**, 3–13 (2016).
3. W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, M. J. Garnett, Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2013).

4. J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, L. A. Garraway, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
5. R. J. Geraghty, A. Capes-Davis, J. M. Davis, J. Downward, R. I. Freshney, I. Knezevic, R. Lovell-Badge, J. R. W. Masters, J. Meredith, G. N. Stacey, P. Thraves, M. Vias; Cancer Research UK, Guidelines for the use of cell lines in biomedical research. *Br. J. Cancer* **111**, 1021–1046 (2014).
6. L. Vaughan, W. Glänzel, C. Korch, A. Capes-Davis, Widespread use of misidentified cell line KB (HeLa): Incorrect attribution and its impact revealed through mining the scientific literature. *Cancer Res.* **77**, 2784–2788 (2017).
7. J. Neimark, Line of attack. *Science* **347**, 938–940 (2015).
8. B. B. Campbell, N. Light, D. Fabrizio, M. Zatzman, F. Fuligni, R. de Borja, S. Davidson, M. Edwards, J. A. Elvin, K. P. Hodel, W. J. Zahurancik, Z. Suo, T. Lipman, K. Wimmer, C. P. Kratz, D. C. Bowers, T. W. Laetsch, G. P. Dunn, T. M. Johanns, M. R. Grimmer, I. V. Smirnov, V. Larouche, D. Samuel, A. Bronsema, M. Osborn, D. Stearns, P. Raman, K. A. Cole, P. B. Storm, M. Yalon, E. Opocher, G. Mason, G. A. Thomas, M. Sabel, B. George, D. S. Ziegler, S. Lindhorst, V. M. Issai, S. Constantini, H. Toledano, R. Elhasid, R. Farah, R. Dvir, P. Dirks, A. Huang, M. A. Galati, J. Chung, V. Ramaswamy, M. S. Irwin, M. Aronson, C. Durno, M. D. Taylor, G. Rechavi, J. M. Maris, E. Bouffet, C. Hawkins, J. F. Costello, M. S. Meyn, Z. F. Pursell, D. Malkin, U. Tabori, A. Shlien, Comprehensive analysis of hypermutation in human cancer. *Cell* **171**, 1042–1056.e10 (2017).
9. C. Virtanen, Y. Ishikawa, D. Honjoh, M. Kimura, M. Shimane, T. Miyoshi, H. Nomura, M. H. Jones, Integrated classification of lung tumors and cell lines by expression profiling. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12357–12362 (2002).
10. F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, H. Chang, H. de Silva, H. Heyn, X. Deng, R. K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang, T. Zhang, S. Moran, S. Sayols, M. Soleimani, D. Tamborero, N. Lopez-Bigas, P. Ross-Macdonald, M. Esteller, N. S. Gray, D. A. Haber, M. R. Stratton, C. H. Benes, L. F. A. Wessels, J. Saez-Rodriguez, U. McDermott, M. J. Garnett, A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
11. E. Kim, M. Dede, W. F. Lenoir, G. Wang, S. Srinivasan, M. Colic, T. Hart, A network of human functional gene interactions from knockout fitness screens in cancer cells. *Life Sci. Alliance* **2**, e201800278 (2019).
12. S. Domcke, R. Sinha, D. A. Levine, C. Sander, N. Schultz, Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat. Commun.* **4**, 2126 (2013).
13. R. Sinha, A. G. Winer, M. Chevinsky, C. Jakubowski, Y.-B. Chen, Y. Dong, S. K. Tickoo, V. E. Reuter, P. Russo, J. A. Coleman, C. Sander, J. J. Hsieh, A. A. Hakimi, Analysis of renal cancer cell lines from two major resources enables genomics-guided cell line selection. *Nat. Commun.* **8**, 15165 (2017).
14. P. Jonsson, C. Bandlamudi, M. L. Cheng, P. Srinivasan, S. S. Chavan, N. D. Friedman, E. Y. Rosen, A. L. Richards, N. Bouvier, S. D. Selcuklu, C. M. Bielski, W. Abida, D. Mandelker, O. Birsoy, L. Zhang, A. Zehir, M. T. A. Donoghue, J. Baselga, K. Offit, H. I. Scher, E. M. O'Reilly, Z. K. Stadler, N. Schultz, N. D. Socci, A. Viale, M. Ladanyi, M. E. Robson, D. M. Hyman, M. F. Berger, D. B. Solit, B. S. Taylor, Tumour lineage shapes BRCA-mediated phenotypes. *Nature* **571**, 576–579 (2019).
15. S. Kopetz, J. R. Desai, E. T. Chan, J. R. Hecht, P. J. O'dwyer, D. M. Maru, V. K. Morris, F. Janku, A. Dasari, W. Chung, J. P. J. Issa, P. Gibbs, B. P. James, G. Powis, K. B. Nolop, S. Bhattacharya, L. L. Saltz, Phase II pilot study of vemurafenib in patients with metastatic BRAF-mutated colorectal cancer. *J. Clin. Oncol.* **2015**, 4032–4038 (2015).
16. S. Park, B. Lehner, Epigenetic epistatic interactions constrain the evolution of gene expression. *Mol. Syst. Biol.* **9**, 645 (2013).
17. M. Lukk, M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, A. Brazma, A global map of human gene expression. *Nat. Biotechnol.* **28**, 322–324 (2010).
18. K. E. Varley, J. Gertz, K. M. Bowling, S. L. Parker, T. E. Reddy, F. Pauli-Behn, M. K. Cross, B. A. Williams, J. A. Stamatoyannopoulos, G. E. Crawford, D. M. Absher, B. J. Wold, R. M. Myers, Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23**, 555–567 (2013).
19. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
20. C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, C. Liu, Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLOS ONE* **6**, e17238 (2011).
21. S. P. J. M. Horbach, W. Halffman, The ghosts of HeLa: How cell line misidentification contaminates the scientific literature. *PLOS ONE* **12**, e0186281 (2017).
22. M. Salvadores, D. Mas-Ponte, F. Supek, Passenger mutations accurately classify human tumors. *PLoS Comput. Biol.* **15**, e1006953 (2019).
23. W. Jiao, G. Atwal, P. Polak, R. Karlic, E. Cuppen, A. Danyi, J. de Ridder, C. van Herpen, M. P. Lolkema, N. Steeghs, G. Getz, Q. D. Morris, L. D. Stein; PCAWG Tumor Subtypes and Clinical Translation Working Group, A deep learning system can accurately classify primary and metastatic cancers based on patterns of passenger mutations. *Nat. Commun.* **11**, 214494 (2019).
24. M. T. Furlong, C. D. Hough, C. A. Sherman-Baust, E. S. Pizer, P. J. Morin, Evidence for the colonic origin of ovarian cancer cell line SW626. *J. Natl. Cancer Inst.* **91**, 1327–1328 (1999).
25. E. Medico, M. Russo, G. Picco, C. Cancelliere, E. Valtorta, G. Corti, M. Buscarino, C. Isella, S. Lamba, B. Martinoglio, S. Veronese, S. Siena, A. Sartore-Bianchi, M. Beccuti, M. Mottolese, M. Linnebacher, F. Cordero, F. Di Nicolantonio, A. Bardelli, The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets. *Nat. Commun.* **6**, 7002 (2015).
26. L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, M. R. Stratton, Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
27. N. K. Hayward, J. S. Wilmott, N. Waddell, P. A. Johansson, M. A. Field, K. Nones, A.-M. Patch, H. Kakavand, L. B. Alexandrov, H. Burke, V. Jakrot, S. Kazakoff, O. Holmes, C. Leonard, R. Sabarinathan, L. Mularoni, S. Wood, Q. Xu, N. Waddell, V. Tembe, G. M. Pupo, R. De Paoli-Iseppi, R. E. Vilain, P. Shang, L. M. S. Lau, R. A. Dagg, S.-J. Schramm, A. Pritchard, K. Dutton-Regester, F. Newell, A. Fitzgerald, C. A. Shang, S. M. Grimmond, H. A. Pickett, J. Y. Yang, J. R. Stretch, A. Behren, R. F. Kefford, P. Hersey, G. V. Long, J. Cebon, M. Shackleton, A. J. Spillane, R. P. M. Saw, N. López-Bigas, J. V. Pearson, J. F. Thompson, R. A. Scolyer, G. J. Mann, Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).
28. M. Petljak, L. B. Alexandrov, J. S. Brammeld, S. Price, D. C. Wedge, S. Grossmann, K. J. Dawson, Y. S. Ju, F. Iorio, J. M. C. Tubio, C. C. Koh, I. Georgakopoulos-Soares, B. Rodríguez-Martín, B. Otlu, S. O'Meara, A. P. Butler, A. Menzies, S. G. Bhosle, K. Raine, D. R. Jones, J. W. Teague, K. Beal, C. Latimer, L. O'Neill, J. Zamora, E. Anderson, N. Patel, M. Maddison, B. L. Ng, J. Graham, M. J. Garnett, U. McDermott, S. Nik-Zainal, P. J. Campbell, M. R. Stratton, Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).
29. M. C. Jarvis, D. Ebrahimi, N. A. Temiz, R. S. Harris, Mutation signatures including APOBEC in cancer cell lines. *JNCI Cancer Spectr.* **2**, pky002 (2018).
30. J. Ji, X. Chen, S. Y. Leung, J.-T. A. Chi, K. M. Chu, S. T. Yuen, R. Li, A. S. Y. Chan, J. Li, N. Dunphy, S. So, Comprehensive analysis of the gene expression profiles in human gastric cancer cell lines. *Oncogene* **21**, 6549–6556 (2002).
31. R. B. Corcoran, C. E. Atreya, G. S. Falchook, E. L. Kwak, D. P. Ryan, J. C. Bendell, O. Hamid, W. A. Messersmith, A. Daud, R. Kurzrock, M. Pierobon, P. Sun, E. Cunningham, S. Little, K. Orford, M. Motwani, Y. Bai, K. Patel, A. P. Venook, S. Kopetz, Combined BRAF and MEK inhibition with dabrafenib and trametinib in BRAF V600–mutant colorectal cancer. *J. Clin. Oncol.* **33**, 4023–4031 (2015).
32. S. M. Corsello, R. T. Nagari, R. D. Spangler, J. Rossen, M. Kocak, J. G. Bryan, R. Humeidi, D. Peck, X. Wu, A. A. Tang, V. M. Wang, S. A. Bender, E. Lemire, R. Narayan, P. Montgomery, U. Ben-David, Y. Chen, M. G. Rees, N. J. Lyons, J. M. McFarland, B. T. Wong, L. Wang, N. Dumont, P. J. O'Hearn, E. Stefan, J. G. Doench, H. Greulich, M. Meyerson, F. Vazquez, A. Subramanian, J. A. Roth, J. A. Bittker, J. S. Boehm, C. C. Mader, A. Tsherniak, T. R. Golub, Non-oncology drugs are a source of previously unappreciated anti-cancer activity. bioRxiv 730119 [**Preprint**]. 9 August 2019. https://doi.org/10.1101/730119.
33. Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
34. M. Ceccarelli, F. P. Barthel, T. M. Malta, T. S. Sabedot, S. R. Salama, B. A. Murray, O. Morozova, Y. Newton, A. Radenbaugh, S. M. Pagnotta, S. Anjum, J. Wang, G. Manyam, P. Zoppoli, S. Ling, A. A. Rao, M. Grifford, A. D. Cherniack, H. Zhang, L. Poisson, C. G. Carlotti, D. P. da C. Tirapelli, A. Rao, T. Mikkelsen, C. C. Lau, W. K. A. Yung, R. Rabadan, J. Huse, D. J. Brat, N. L. Lehman, J. S. Barnholtz-Sloan, S. Zheng, K. Hess, G. Rao, M. Meyerson, R. Beroukhim, L. Cooper, R. Akbani, M. Wrensch, D. Haussler, K. D. Aldape, P. W. Laird, D. H. Gutmann, S. Anjum, H. Arachchi, J. T. Auman, M. Balasundaram, S. Balu, G. Barnett, S. Baylin, S. Bell, C. Benz, N. Bir, K. L. Black, T. Bodenheimer, L. Boice, M. S. Bootwalla, J. Bowen, C. A. Bristow, Y. S. N. Butterfield, Q.-R. Chen, L. Chin, J. Cho, E. Chuah, S. Chudamani, S. G. Coetzee, M. L. Cohen, H. Colman, M. Couce, F. D'Angelo, T. Davidsen, A. Davis, J. A. Demchok, K. Devine, L. Ding, R. Duell, J. B. Elder, J. M. Eschbacher, A. Fehrenbach, M. Ferguson, S. Frazer, G. Fuller, J. Fulop, S. B. Gabriel, L. Garofano, J. M. Gastier-Foster, N. Gehlenborg, M. Gerken, G. Getz, C. Giannini, W. J. Gibson, A. Hadjipanayis, D. N. Hayes, D. I. Heiman, B. Hermes, J. Hilty, K. A. Hoadley, A. P. Hoyle, M. Huang, S. R. Jefferys, C. D. Jones, S. J. M. Jones, Z. Ju, A. Kastl, A. Kendler, J. Kim, R. Kucherlapati, P. H. Lai, M. S. Lawrence, S. Lee, K. M. Leraas, T. M. Lichtenberg, P. Lin, Y. Liu, J. Liu, J. Y. Ljubimova, Y. Lu, Y. Ma, D. T. Maglinte, H. S. Mahadeshwar, M. A. Marra, M. McGraw, C. McPherson, S. Meng, P. A. Mieczkowski, C. R. Miller, G. B. Mills, R. A. Moore, L. E. Mose, A. J. Mungall, R. Naresh, T. Naska, L. Neder, M. S. Noble, A. Noss,

B. P. O'Neill, Q. T. Ostrom, C. Palmer, A. Pantazi, M. Parfenov, P. J. Park, J. S. Parker, C. M. Perou, C. R. Pierson, T. Pihl, A. Protopopov, A. Radenbaugh, N. C. Ramirez, W. K. Rathmell, X. Ren, J. Roach, A. G. Robertson, G. Saksena, J. E. Schein, S. E. Schumacher, J. Seidman, K. Senecal, S. Seth, H. Shen, Y. Shi, J. Shih, K. Shimmel, H. Sicotte, S. Sifri, T. Silva, J. V. Simons, R. Singh, T. Skelly, A. E. Sloan, H. J. Sofia, M. G. Soloway, X. Song, C. Sougnez, C. Souza, S. M. Staugaitis, H. Sun, C. Sun, D. Tan, J. Tang, Y. Tang, L. Thorne, F. A. Trevisan, T. Triche, D. J. Van Den Berg, U. Veluvolu, D. Voet, Y. Wan, Z. Wang, R. Warnick, J. N. Weinstein, D. J. Weisenberger, M. D. Wilkerson, F. Williams, L. Wise, Y. Wolinsky, J. Wu, A. W. Xu, L. Yang, L. Yang, T. I. Zack, J. C. Zenklusen, J. Zhang, W. Zhang, J. Zhang, E. Zmuda, H. Noushmehr, A. Iavarone, R. G. W. Verhaak, Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016).

35. Y. Liu, N. S. Sethi, T. Hinoue, B. G. Schneider, A. D. Cherniack, F. Sanchez-Vega, J. A. Seoane, F. Farshidfar, R. Bowlby, M. Islam, J. Kim, W. Chatila, R. Akbani, R. S. Kanchi, C. S. Rabkin, J. E. Willis, K. K. Wang, S. J. McCall, L. Mishra, A. I. Ojesina, S. Bullman, C. S. Pedamallu, A. J. Lazar, R. Sakai, S. J. Caesar-Johnson, J. A. Demchok, I. Felau, M. Kasapi, M. L. Ferguson, C. M. Hutter, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. Zhang, S. Chudamani, J. Liu, L. Lolla, R. Naresh, T. Pihl, Q. Sun, Y. Wan, Y. Wu, J. Cho, T. DeFreitas, S. Frazer, N. Gehlenborg, G. Getz, D. I. Heiman, J. Kim, M. S. Lawrence, P. Lin, S. Meier, M. S. Noble, G. Saksena, D. Voet, H. Zhang, B. Bernard, N. Chambwe, V. Dhankani, T. Knijnenburg, R. Kramer, K. Leinonen, Y. Liu, M. Miller, S. Reynolds, I. Shmulevich, V. Thorsson, W. Zhang, R. Akbani, B. M. Broom, A. M. Hegde, Z. Ju, R. S. Kanchi, A. Korkut, J. Li, H. Liang, S. Ling, W. Liu, Y. Lu, G. B. Mills, K.-S. Ng, A. Rao, M. Ryan, J. Wang, J. N. Weinstein, J. Zhang, A. Abeshouse, J. Armenia, D. Chakravarty, W. K. Chatila, I. Bruijn, J. Gao, B. E. Gross, Z. J. Heins, R. Kundra, K. La, M. Ladanyi, A. Luna, M. G. Nissan, A. Ochoa, S. M. Phillips, E. Reznik, F. Sanchez-Vega, C. Sander, N. Schultz, R. Sheridan, S. O. Sumer, Y. Sun, B. S. Taylor, J. Wang, H. Zhang, P. Anur, M. Peto, P. Spellman, C. Benz, J. M. Stuart, C. K. Wong, C. Yau, D. N. Hayes, J. S. Parker, M. D. Wilkerson, A. Ally, M. Balasundaram, R. Bowlby, D. Brooks, R. Carlsen, E. Chuah, N. Dhalla, R. Holt, S. J. M. Jones, K. Kasaian, D. Lee, Y. Ma, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, K. Mungall, A. G. Robertson, S. Sadeghi, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, K. Tse, T. Wong, A. C. Berger, R. Beroukhim, A. D. Cherniack, C. Cibulskis, S. B. Gabriel, G. F. Gao, G. Ha, M. Meyerson, S. E. Schumacher, J. Shih, M. H. Kucherlapati, R. S. Kucherlapati, S. Baylin, L. Cope, L. Danilova, M. S. Bootwalla, P. H. Lai, D. T. Maglinte, D. J. V. D. Berg, D. J. Weisenberger, J. T. Auman, S. Balu, T. Bodenheimer, C. Fan, K. A. Hoadley, A. P. Hoyle, S. R. Jefferys, C. D. Jones, S. Meng, P. A. Mieczkowski, L. E. Mose, A. H. Perou, C. M. Perou, J. Roach, Y. Shi, J. V. Simons, T. Skelly, M. G. Soloway, D. Tan, U. Veluvolu, H. Fan, T. Hinoue, P. W. Laird, H. Shen, W. Zhou, M. Bellair, K. Chang, K. Covington, C. J. Creighton, H. Dinh, H. Doddapaneni, L. A. Donehower, J. Drummond, R. A. Gibbs, R. Glenn, W. Hale, Y. Han, J. Hu, V. Korchina, S. Lee, L. Lewis, W. Li, X. Liu, M. Morgan, D. Morton, D. Muzny, J. Santibanez, M. Sheth, E. Shinbrot, L. Wang, M. Wang, D. A. Wheeler, L. Xi, F. Zhao, J. Hess, E. L. Appelbaum, M. Bailey, M. G. Cordes, L. Ding, C. C. Fronick, L. A. Fulton, R. S. Fulton, C. Kandoth, E. R. Mardis, M. D. McLellan, C. A. Miller, H. K. Schmidt, R. K. Wilson, D. Crain, E. Curley, J. Gardner, K. Lau, D. Mallery, S. Morris, J. Paulauskis, R. Penny, C. Shelton, T. Shelton, M. Sherman, E. Thompson, P. Yena, J. Bowen, J. M. Gastier-Foster, M. Gerken, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, N. Corcoran, T. Costello, C. Hovens, A. L. Carvalho, A. C. de Carvalho, J. H. Fregnani, A. Longatto-Filho, R. M. Reis, C. Scapulatempo-Neto, H. C. S. Silveira, D. O. Vidal, A. Burnette, J. Eschbacher, B. Hermes, A. Noss, R. Singh, M. L. Anderson, P. D. Castro, M. Ittmann, D. Huntsman, B. Kohl, X. Le, R. Thorp, C. Andry, E. R. Duffy, V. Lyadov, O. Paklina, G. Setdikova, A. Shabunin, M. Tavobilov, C. McPherson, R. Warnick, R. Berkowitz, D. Cramer, C. Feltmate, N. Horowitz, A. Kibel, M. Muto, C. P. Raut, A. Malykh, J. S. Barnholtz-Sloan, W. Barrett, K. Devine, J. Fulop, Q. T. Ostrom, K. Shimmel, Y. Wolinsky, A. E. Sloan, A. D. Rose, F. Giuliante, M. Goodman, B. Y. Karlan, C. H. Hagedorn, J. Eckman, J. Harr, J. Myers, K. Tucker, L. A. Zach, B. Deyarmin, H. Hu, L. Kvecher, C. Larson, R. J. Mural, S. Somiari, A. Vicha, T. Zelinka, J. Bennett, M. Iacocca, B. Rabeno, P. Swanson, M. Latour, L. Lacombe, B. Têtu, A. Bergeron, M. McGraw, S. M. Staugaitis, J. Chabot, H. Hibshoosh, A. Sepulveda, T. Su, T. Wang, O. Potapova, O. Voronina, L. Desjardins, O. Mariani, S. Roman-Roman, X. Sastre, M.-H. Stern, F. Cheng, S. Signoretti, A. Berchuck, D. Bigner, E. Lipp, J. Marks, S. McCall, R. McLendon, A. Secord, A. Sharp, M. Behera, D. J. Brat, A. Chen, K. Delman, S. Force, F. Khuri, K. Magliocca, S. Maithel, J. J. Olson, T. Owonikoko, A. Pickens, S. Ramalingam, D. M. Shin, G. Sica, E. G. V. Meir, H. Zhang, W. Eijckenboom, A. Gillis, E. Korpershoek, L. Looijenga, W. Oosterhuis, H. Stoop, K. E. van Kessel, E. C. Zwarthoff, C. Calatozzolo, L. Cuppini, S. Cuzzubbo, F. DiMeco, G. Finocchiaro, L. Mattei, A. Perin, B. Pollo, C. Chen, J. Houck, P. Lohavanichbutr, A. Hartmann, C. Stoehr, R. Stoehr, H. Taubert, S. Wach, B. Wullich, W. Kycler, D. Murawa, M. Wiznerowicz, K. Chung, W. J. Edenfield, J. Martin, E. Baudin, G. Bubley, R. Bueno, A. D. Rienzo, W. G. Richards, S. Kalkanis, T. Mikkelsen, H. Noushmehr, L. Scarpace, N. Girard, M. Aymerich, E. Campo, E. Giné, A. L. Guillermo, N. V. Bang, P. T. Hanh, B. D. Phu, Y. Tang, H. Colman, K. Evason, P. R. Dottino, J. A. Martignetti, H. Gabra, H. Juhl, T. Akeredolu, S. Stepa, D. Hoon, K. Ahn, K. J. Kang, F. Beuschlein, A. Breggia, M. Birrer, D. Bell, M. Borad, A. H. Bryce, E. Castle, V. Chandan, J. Cheville, J. A. Copland, M. Farnell, T. Flotte, N. Giama, T. Ho, M. Kendrick, J.-P. Kocher, K. Kopp, C. Moser, D. Nagorney, D. O'Brien, B. P. O'Neill, T. Patel, G. Petersen, F. Que, M. Rivera, L. Roberts, R. Smallridge, T. Smyrk, M. Stanton, R. H. Thompson, M. Torbenson, J. D. Yang, L. Zhang, F. Brimo, J. A. Ajani, A. M. A. Gonzalez, C. Behrens, J. Bondaruk, R. Broaddus, B. Czerniak, B. Esmaeli, J. Fujimoto, J. Gershenwald, C. Guo, A. J. Lazar, C. Logothetis, F. Meric-Bernstam, C. Moran, L. Ramondetta, D. Rice, A. Sood, P. Tamboli, T. Thompson, P. Troncoso, A. Tsao, I. Wistuba, C. Carter, L. Haydu, P. Hersey, V. Jakrot, H. Kakavand, R. Kefford, K. Lee, G. Long, G. Mann, M. Quinn, R. Saw, R. Scolyer, K. Shannon, A. Spillane, J. Stretch, M. Synott, J. Thompson, J. Wilmott, H. Al-Ahmadie, T. A. Chan, R. Ghossein, A. Gopalan, D. A. Levine, V. Reuter, S. Singer, B. Singh, N. V. Tien, T. Broudy, C. Mirsaidi, P. Nair, P. Drwiega, J. Miller, J. Smith, H. Zaren, J.-W. Park, N. P. Hung, E. Kebebew, W. M. Linehan, A. R. Metwalli, K. Pacak, P. A. Pinto, M. Schiffman, L. S. Schmidt, C. D. Vocke, N. Wentzensen, R. Worrell, H. Yang, M. Moncrieff, C. Goparaju, J. Melamed, H. Pass, N. Botnariuc, I. Caraman, M. Cernat, I. Chemencedji, A. Clipca, S. Doruc, G. Gorincioi, S. Mura, M. Pirtac, I. Stancul, D. Tcaciuc, M. Albert, I. Alexopoulou, A. Arnaout, J. Bartlett, J. Engel, S. Gilbert, J. Parfitt, H. Sekhon, G. Thomas, D. M. Rassl, R. C. Rintoul, C. Bifulco, R. Tamakawa, W. Urba, N. Hayward, H. Timmers, A. Antenucci, F. Facciolo, G. Grazi, M. Marino, R. Merola, R. de Krijger, A.-P. Gimenez-Roqueplo, A. Piché, S. Chevalier, G. McKercher, K. Birsoy, G. Barnett, C. Brewer, C. Farver, T. Naska, N. A. Pennell, D. Raymond, C. Schilero, K. Smolenski, F. Williams, C. Morrison, J. A. Borgia, M. J. Liptay, M. Pool, C. W. Seder, K. Junker, L. Omberg, M. Dinkin, G. Manikhas, D. Alvaro, M. C. Bragazzi, V. Cardinale, G. Carpino, E. Gaudio, D. Chesla, S. Cottingham, M. Dubina, F. Moiseenko, R. Dhanasekaran, K.-F. Becker, K.-P. Janssen, J. Slotta-Huspenina, M. H. Abdel-Rahman, D. Aziz, S. Bell, C. M. Cebulla, A. Davis, R. Duell, J. B. Elder, J. Hilty, B. Kumar, J. Lang, N. L. Lehman, R. Mandt, P. Nguyen, R. Pilarski, K. Rai, L. Schoenfield, K. Senecal, P. Wakely, P. Hansen, R. Lechan, J. Powers, A. Tischler, W. E. Grizzle, K. C. Sexton, A. Kastl, J. Henderson, S. Porten, J. Waldmann, M. Fassnacht, S. L. Asa, D. Schadendorf, M. Couce, M. Graefen, H. Huland, G. Sauter, T. Schlomm, R. Simon, P. Tennstedt, O. Olabode, M. Nelson, O. Bathe, P. R. Carroll, J. M. Chan, P. Disaia, P. Glenn, R. K. Kelley, C. N. Landen, J. Phillips, M. Prados, J. Simko, K. Smith-McCune, S. VandenBerg, K. Roggin, A. Fehrenbach, A. Kendler, S. Sifri, R. Steele, A. Jimeno, F. Carey, I. Forgie, M. Mannelli, M. Carney, B. Hernandez, B. Campos, C. Herold-Mende, C. Jungk, A. Unterberg, A. von Deimling, A. Bossler, J. Galbraith, L. Jacobus, M. Knudson, T. Knutson, D. Ma, M. Milhem, R. Sigmund, A. K. Godwin, R. Madan, H. G. Rosenthal, C. Adebamowo, S. N. Adebamowo, A. Boussiotas, D. Beer, T. Giordano, A.-M. Mes-Masson, F. Saad, T. Bocklage, L. Landrum, R. Mannel, K. Moore, K. Moxley, R. Postier, J. Walker, R. Zuna, M. Feldman, F. Valdivieso, R. Dhir, J. Luketich, E. M. M. Pinero, M. Quintero-Aguilo, C. G. Carlotti, J. S. D. Santos, R. Kemp, A. Sankarankuty, D. Tirapelli, J. Catto, K. Agnew, E. Swisher, J. Creaney, B. Robinson, C. S. Shelley, E. M. Godwin, S. Kendall, C. Shipman, C. Bradford, T. Carey, A. Haddad, J. Moyer, L. Peterson, M. Prince, L. Rozek, G. Wolf, R. Bowman, K. M. Fong, I. Yang, R. Korst, W. K. Rathmell, J. L. Fantacone-Campbell, J. A. Hooke, A. J. Kovatich, C. D. Shriver, J. DiPersio, B. Drake, R. Govindan, S. Heath, T. Ley, B. V. Tine, P. Westervelt, M. A. Rubin, J. I. Lee, N. D. Aredes, A. Mariamidze, V. Thorsson, A. J. Bass, P. W. Laird, Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell* **33**, 721–735.e8 (2018).

36. R. M. Neve, K. Chin, J. Fridlyand, J. Yeh, F. L. Baehner, T. Fevr, L. Clark, N. Bayani, J.-P. Coppe, F. Tong, T. Speed, P. T. Spellman, S. DeVries, A. Lapuk, N. J. Wang, W.-L. Kuo, J. L. Stilwell, D. Pinkel, D. G. Albertson, F. M. Waldman, F. McCormick, R. B. Dickson, M. D. Johnson, M. Lippman, S. Ethier, A. Gazdar, J. W. Gray, A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**, 515–527 (2006).

37. J. Ronen, S. Hayat, A. Akalin, Evaluation of colorectal cancer subtypes and cell lines using deep learning. *Life Sci. Alliance* **2**, e201900517 (2019).

38. K. Yu, B. Chen, D. Aran, J. Charalel, C. Yau, D. M. Wolf, L. J. van 't Veer, A. J. Butte, T. Goldstein, M. Sirota, Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat. Commun.* **10**, 3574 (2019).

39. T. Cokelaer, E. Chen, F. Iorio, M. P. Menden, H. Lightfoot, J. Saez-Rodriguez, M. J. Garnett, GDSCTools for mining pharmacogenomic interactions in cancer. *Bioinformatics* **34**, 1226–1228 (2018).

40. O. N. Ikediobi, M. Reimers, S. Durinck, P. E. Blower, A. P. Futreal, M. R. Stratton, J. N. Weinstein, In vitro differential sensitivity of melanomas to phenothiazines is based on the presence of codon 600 BRAF mutation. *Mol. Cancer Ther.* **7**, 1337–1346 (2008).

41. Y. Yamaguchi, T. Kasukabe, S. Kumakura, Piperlongumine rapidly induces the death of human pancreatic cancer cells mainly through the induction of ferroptosis. *Int. J. Oncol.* **52**, 1011–1022 (2018).

42. F. M. Behan, F. Iorio, G. Picco, E. Gonçalves, C. M. Beaver, G. Migliardi, R. Santos, Y. Rao, F. Sassi, M. Pinnelli, R. Ansari, S. Harper, D. A. Jackson, R. McRae, R. Pooley, P. Wilkinson, D. van der Meer, D. Dow, C. Buser-Doepner, A. Bertotti, L. Trusolino, E. A. Stronach, J. Saez-Rodriguez, K. Yusa, M. J. Garnett, Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* **568**, 511–516 (2019).

43. M. Westphal, C. L. Maire, K. Lamszus, EGFR as a target for glioblastoma treatment: An unfulfilled promise. *CNS Drugs* **31**, 723–735 (2017).

44. J. D. Campbell, A. Alexandrov, J. Kim, J. Wala, A. H. Berger, C. S. Pedamallu, S. A. Shukla, G. Guo, A. N. Brooks, B. A. Murray, M. Imielinski, X. Hu, S. Ling, R. Akbani, M. Rosenberg, C. Cibulskis, A. Ramachandran, E. A. Collisson, D. J. Kwiatkowski, M. S. Lawrence, J. N. Weinstein, R. G. W. Verhaak, C. J. Wu, P. S. Hammerman, A. D. Cherniack, G. Getz, C. G. A. R. Network, M. N. Artyomov, R. Schreiber, R. Govindan, M. Meyerson, Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).
45. R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, L. M. Staudt, Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
46. D. Aran, M. Sirota, A. J. Butte, Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, (2015).
47. Cancer Cell Line Encyclopedia Consortium; Genomics of Drug Sensitivity in Cancer Concortium, Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528**, 84–87 (2015).
48. M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, P. K.-S. Ng, K. J. Jeong, S. Cao, Z. Wang, J. Gao, Q. Gao, F. Wang, E. M. Liu, L. Mularoni, C. Rubio-Perez, N. Nagarajan, I. Cortés-Ciriano, D. C. Zhou, W.-W. Liang, J. M. Hess, V. D. Yellapantula, D. Tamborero, A. Gonzalez-Perez, C. Suphavilai, J. Y. Ko, E. Khurana, P. J. Park, E. M. Van Allen, H. Liang; MC3 Working Group; Cancer Genome Atlas Research Network, M. S. Lawrence, A. Godzik, N. Lopez-Bigas, J. Stuart, D. Wheeler, G. Getz, K. Chen, A. J. Lazar, G. B. Mills, R. Karchin, L. Ding, Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
49. K. Ellrott, M. H. Bailey, G. Saksena, K. R. Covington, C. Kandoth, C. Stewart, J. Hess, S. Ma, K. E. Chiotti, M. McLellan, H. J. Sofia, C. Hutter, G. Getz, D. Wheeler, L. Ding, S. J. Caesar-Johnson, J. A. Demchok, I. Felau, M. Kasapi, M. L. Ferguson, C. M. Hutter, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. Zhang, S. Chudamani, J. Liu, L. Lolla, R. Naresh, T. Pihl, Q. Sun, Y. Wan, Y. Wu, J. Cho, T. DeFreitas, S. Frazer, N. Gehlenborg, G. Getz, D. I. Heiman, J. Kim, M. S. Lawrence, P. Lin, S. Meier, M. S. Noble, G. Saksena, D. Voet, H. Zhang, B. Bernard, N. Chambwe, V. Dhankani, T. Knijnenburg, R. Kramer, K. Leinonen, Y. Liu, M. Miller, S. Reynolds, I. Shmulevich, V. Thorsson, W. Zhang, R. Akbani, B. M. Broom, A. M. Hegde, Z. Ju, R. S. Kanchi, A. Korkut, J. Li, H. Liang, S. Ling, W. Liu, Y. Lu, G. B. Mills, K.-S. Ng, A. Rao, M. Ryan, J. Wang, J. N. Weinstein, J. Zhang, A. Abeshouse, J. Armenia, D. Chakravarty, W. K. Chatila, I. de Bruijn, J. Gao, B. E. Gross, Z. J. Heins, R. Kundra, K. La, M. Ladanyi, A. Luna, M. G. Nissan, A. Ochoa, S. M. Phillips, E. Reznik, F. Sanchez-Vega, C. Sander, N. Schultz, R. Sheridan, S. O. Sumer, Y. Sun, B. S. Taylor, J. Wang, H. Zhang, P. Anur, M. Peto, P. Spellman, C. Benz, J. M. Stuart, C. K. Wong, C. Yau, D. N. Hayes, M. D. W. Parker, A. Ally, M. Balasundaram, R. Bowlby, D. Brooks, R. Carlsen, E. Chuah, N. Dhalla, R. Holt, S. J. M. Jones, K. Kasaian, D. Lee, Y. Ma, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, K. Mungall, A. G. Robertson, S. Sadeghi, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, K. Tse, T. Wong, A. C. Berger, R. Beroukhim, A. D. Cherniack, C. Cibulskis, S. B. Gabriel, G. F. Gao, G. Ha, M. Meyerson, S. E. Schumacher, J. Shih, M. H. Kucherlapati, R. S. Kucherlapati, S. Baylin, L. Cope, L. Danilova, M. S. Bootwalla, P. H. Lai, D. T. Maglinte, D. J. Van Den Berg, D. J. Weisenberger, J. T. Auman, S. Balu, T. Bodenheimer, C. Fan, K. A. Hoadley, A. P. Hoyle, S. R. Jefferys, C. D. Jones, S. Meng, P. A. Mieczkowski, L. E. Mose, A. H. Perou, C. M. Perou, J. Roach, Y. Shi, J. V. Simons, T. Skelly, M. G. Soloway, D. Tan, U. Veluvolu, H. Fan, T. Hinoue, P. W. Laird, H. Shen, W. Zhou, M. Bellair, K. Chang, K. Covington, C. J. Creighton, H. Dinh, H. Doddapaneni, L. A. Donehower, J. Drummond, R. A. Gibbs, R. Glenn, W. Hale, Y. Han, J. Hu, V. Korchina, S. Lee, L. Lewis, W. Li, X. Liu, M. Morgan, D. Morton, D. Muzny, J. Santibanez, M. Sheth, E. Shinbrot, L. Wang, M. Wang, D. A. Wheeler, L. Xi, F. Zhao, J. Hess, E. L. Appelbaum, M. Bailey, M. G. Cordes, L. Ding, C. C. Fronick, L. A. Fulton, R. S. Fulton, C. Kandoth, E. R. Mardis, M. D. McLellan, C. A. Miller, H. K. Schmidt, R. K. Wilson, D. Crain, E. Curley, J. Gardner, K. Lau, D. Mallery, S. Morris, J. Paulauskis, R. Penny, C. Shelton, T. Shelton, M. Sherman, E. Thompson, P. Yena, J. Bowen, J. M. Gastier-Foster, M. Gerken, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, N. Corcoran, T. Costello, C. Hovens, A. L. Carvalho, A. C. de Carvalho, J. H. Fregnani, A. Longatto-Filho, R. M. Reis, C. Scapulatempo-Neto, H. C. S. Silveira, D. O. Vidal, A. Burnette, J. Eschbacher, B. Hermes, A. Noss, R. Singh, M. L. Anderson, P. D. Castro, M. Ittmann, D. Huntsman, B. Kohl, X. Le, R. Thorp, C. Andry, E. R. Duffy, V. Lyadov, O. Paklina, G. Setdikova, A. Shabunin, M. Tavobilov, C. McPherson, R. Warnick, R. Berkowitz, D. Cramer, C. Feltmate, N. Horowitz, A. Kibel, M. Muto, C. P. Raut, A. Malykh, J. S. Barnholtz-Sloan, W. Barrett, K. Devine, J. Fulop, Q. T. Ostrom, K. Shimmel, Y. Wolinsky, A. E. Sloan, A. De Rose, F. Giuliante, M. Goodman, B. Y. Karlan, C. H. Hagedorn, J. Eckman, J. Harr, J. Myers, K. Tucker, L. A. Zach, B. Deyarmin, H. Hu, L. Kvecher, C. Larson, R. J. Mural, S. Somiari, A. Vicha, T. Zelinka, J. Bennett, M. Iacocca, B. Rabeno, P. Swanson, M. Latour, L. Lacombe, B. Têtu, A. Bergeron, M. McGraw, S. M. Staugaitis, J. Chabot, H. Hibshoosh, A. Sepulveda, T. Su, T. Wang, O. Potapova, O. Voronina, L. Desjardins, O. Mariani, S. Roman-Roman, X. Sastre, M.-H. Stern, F. Cheng, S. Signoretti, A. Berchuck, D. Bigner, E. Lipp, J. Marks, S. McCall, R. McLendon, A. Secord, A. Sharp, M. Behera, D. J. Brat, A. Chen, K. Delman, S. Force, F. Khuri, K. Magliocca, S. Maithel, J. J. Olson, T. Owonikoko, A. Pickens, S. Ramalingam, D. M. Shin, G. Sica, E. G. Van Meir, H. Zhang, W. Eijckenboom, A. Gillis, E. Korpershoek, L. Looijenga, W. Oosterhuis, H. Stoop, K. E. van Kessel, E. C. Zwarthoff, C. Calatozzolo, L. Cuppini, S. Cuzzubbo, F. DiMeco, G. Finocchiaro, L. Mattei, A. Perin, B. Pollo, C. Chen, J. Houck, P. Lohavanichbutr, A. Hartmann, C. Stoehr, R. Stoehr, H. Taubert, S. Wach, B. Wullich, W. Kycler, D. Murawa, M. Wiznerowicz, K. Chung, W. J. Edenfield, J. Martin, E. Baudin, G. Bubley, R. Bueno, A. De Rienzo, W. G. Richards, S. Kalkanis, T. Mikkelsen, H. Noushmehr, L. Scarpace, N. Girard, M. Aymerich, E. Campo, E. Giné, A. L. Guillermo, N. Van Bang, P. T. Hanh, B. D. Phu, Y. Tang, H. Colman, K. Evason, P. R. Dottino, J. A. Martignetti, H. Gabra, H. Juhl, T. Akeredolu, S. Stepa, D. Hoon, K. Ahn, K. J. Kang, F. Beuschlein, A. Breggia, M. Birrer, D. Bell, M. Borad, A. H. Bryce, E. Castle, V. Chandan, J. Cheville, J. A. Copland, M. Farnell, T. Flotte, N. Giama, T. Ho, M. Kendrick, J.-P. Kocher, K. Kopp, C. Moser, D. Nagorney, D. O'Brien, B. P. O'Neill, T. Patel, G. Petersen, F. Que, M. Rivera, L. Roberts, R. Smallridge, T. Smyrk, M. Stanton, R. H. Thompson, M. Torbenson, J. D. Yang, L. Zhang, F. Brimo, J. A. Ajani, A. M. A. Gonzalez, C. Behrens, J. Bondaruk, R. Broaddus, B. Czerniak, B. Esmaeli, J. Fujimoto, J. Gershenwald, C. Guo, A. J. Lazar, C. Logothetis, F. Meric-Bernstam, C. Moran, L. Ramondetta, D. Rice, A. Sood, P. Tamboli, T. Thompson, P. Troncoso, A. Tsao, I. Wistuba, C. Carter, L. Haydu, P. Hersey, V. Jakrot, H. Kakavand, R. Kefford, K. Lee, G. Long, G. Mann, M. Quinn, R. Saw, R. Scolyer, K. Shannon, A. Spillane, J. Stretch, M. Synott, J. Thompson, J. Wilmott, H. Al-Ahmadie, T. A. Chan, R. Ghossein, A. Gopalan, D. A. Levine, V. Reuter, S. Singer, B. Singh, N. V. Tien, T. Broudy, C. Mirsaidi, P. Nair, P. Drwiega, J. Miller, J. Smith, H. Zaren, J.-W. Park, N. P. Hung, E. Kebebew, W. M. Linehan, A. R. Metwalli, K. Pacak, P. A. Pinto, M. Schiffman, L. S. Schmidt, C. D. Vocke, N. Wentzensen, R. Worrell, H. Yang, M. Moncrieff, C. Goparaju, J. Melamed, H. Pass, N. Botnariuc, I. Caraman, M. Cernat, I. Chemencedji, A. Clipca, S. Doruc, G. Gorincioi, S. Mura, M. Pirtac, I. Stancul, D. Tcaciuc, M. Albert, I. Alexopoulou, A. Arnaout, J. Bartlett, J. Engel, S. Gilbert, J. Parfitt, H. Sekhon, G. Thomas, D. M. Rassl, R. C. Rintoul, C. Bifulco, R. Tamakawa, W. Urba, N. Hayward, H. Timmers, A. Antenucci, F. Facciolo, G. Grazi, M. Marino, R. Merola, R. de Krijger, A.-P. Gimenez-Roqueplo, A. Piché, S. Chevalier, G. McKercher, K. Birsoy, G. Barnett, C. Brewer, C. Farver, T. Naska, N. A. Pennell, D. Raymond, C. Schilero, K. Smolenski, F. Williams, C. Morrison, J. A. Borgia, M. J. Liptay, M. Pool, C. W. Seder, K. Junker, L. Omberg, M. Dinkin, G. Manikhas, D. Alvaro, M. C. Bragazzi, V. Cardinale, G. Carpino, E. Gaudio, D. Chesla, S. Cottingham, M. Dubina, F. Moiseenko, R. Dhanasekaran, K.-F. Becker, K.-P. Janssen, J. Slotta-Huspenina, M. H. Abdel-Rahman, D. Aziz, S. Bell, C. M. Cebulla, A. Davis, R. Duell, J. B. Elder, J. Hilty, B. Kumar, J. Lang, N. L. Lehman, R. Mandt, P. Nguyen, R. Pilarski, K. Rai, L. Schoenfield, K. Senecal, P. Wakely, P. Hansen, R. Lechan, J. Powers, A. Tischler, W. E. Grizzle, K. C. Sexton, A. Kastl, J. Henderson, S. Porten, J. Waldmann, M. Fassnacht, S. L. Asa, D. Schadendorf, M. Couce, M. Graefen, H. Huland, G. Sauter, T. Schlomm, R. Simon, P. Tennstedt, O. Olabode, M. Nelson, O. Bathe, P. R. Carroll, J. M. Chan, P. Disaia, P. Glenn, R. K. Kelley, C. N. Landen, J. Phillips, M. Prados, J. Simko, K. Smith-McCune, S. VandenBerg, K. Roggin, A. Fehrenbach, A. Kendler, S. Sifri, R. Steele, A. Jimeno, F. Carey, I. Forgie, M. Mannelli, M. Carney, B. Hernandez, B. Campos, C. Herold-Mende, C. Jungk, A. Unterberg, A. von Deimling, A. Bossler, J. Galbraith, L. Jacobus, M. Knudson, T. Knutson, D. Ma, M. Milhem, R. Sigmund, A. K. Godwin, R. Madan, H. G. Rosenthal, C. Adebamowo, S. N. Adebamowo, A. Boussioutas, D. Beer, T. Giordano, A.-M. Mes-Masson, F. Saad, T. Bocklage, L. Landrum, R. Mannel, K. Moore, K. Moxley, R. Postier, J. Walker, R. Zuna, M. Feldman, F. Valdivieso, R. Dhir, J. Luketich, E. M. M. Pinero, M. Quintero-Aguilo, C. G. Carlotti, J. S. D. Santos, R. Kemp, A. Sankarankuty, D. Tirapelli, J. Catto, K. Agnew, E. Swisher, J. Creaney, B. Robinson, C. S. Shelley, E. M. Godwin, S. Kendall, C. Shipman, C. Bradford, T. Carey, A. Haddad, J. Moyer, L. Peterson, M. Prince, L. Rozek, G. Wolf, R. Bowman, K. M. Fong, I. Yang, R. Korst, W. K. Rathmell, J. L. Fantacone-Campbell, J. A. Hooke, A. J. Kovatich, C. D. Shriver, J. DiPersio, B. Drake, R. Govindan, S. Heath, T. Ley, B. Van Tine, P. Westervelt, M. A. Rubin, J. I. Lee, N. D. Aredes, A. Mariamidze, Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Systems* **6**, 271–281.e7 (2018).
50. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, The Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, D. G. MacArthur, Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv 531210 [**Preprint**]. 13 August 2019. https://doi.org/10.1101/531210.
51. A. Bairoch, The Cellosaurus, a cell-line knowledge resource. *J. Biomol. Tech.* **29**, 25–38 (2018).
52. R. M. Meyers, J. G. Bryan, J. M. McFarland, B. A. Weir, A. E. Sizemore, H. Xu, N. V. Dharia, P. G. Montgomery, G. S. Cowley, S. Pantel, A. Goodale, Y. Lee, L. D. Ali, G. Jiang, R. Lubonja, W. F. Harrington, M. Strickland, T. Wu, D. C. Hawes, V. A. Zhivich, M. R. Wyatt, Z. Kalani,

J. J. Chang, M. Okamoto, K. Stegmaier, T. R. Golub, J. S. Boehm, F. Vazquez, D. E. Root, W. C. Hahn, A. Tsherniak, Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).

Acknowledgments: The results published here are, in whole or part, based on data generated by the TCGA Research Network, the Genomics of Drug Sensitivity in Cancer (GDSC) Project, the Cancer Cell Line Encyclopedia (CCLE) Project, the Cancer Dependency Map (DepMap) Project (in particular, Project Score), and the Genotype-Tissue Expression (GTEx) Project. In addition, we thank all the members from the Genome Data Science group for insightful discussions. **Funding:** F.S. is funded by the ERC StG 757700 HYPER-INSIGHT and by MINECO grant BFU2017-89833-P. M.S. is funded by the FPU2017 fellowship of the Spanish government. We acknowledge funding from the Severo Ochoa Center of Excellence award to the IRB Barcelona.
Author contributions: M.S.: conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, writing, revision, and editing; F.F.-T.: data curation and methodology; F.S.: conceptualization, funding acquisition, methodology, project administration, supervision, writing, revision, and editing. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Raw data can be downloaded from the publicly available databases as described in Materials and Methods. Additional data related to this paper may be requested from the authors.

Submitted 12 November 2019
Accepted 1 May 2020
Published 1 July 2020
10.1126/sciadv.aba1862

Citation: M. Salvadores, F. Fuster-Tormo, F. Supek, Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Sci. Adv.* **6**, eaba1862 (2020).

ScienceAdvances

Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns

Marina Salvadores, Francisco Fuster-Tormo, and Fran Supek

Sci. Adv., **6** (27), eaba1862.
DOI: 10.1126/sciadv.aba1862

Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/27/eaba1862/DC1

Supplementary Materials for

Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns

Marina Salvadores, Francisco Fuster-Tormo, Fran Supek*

*Corresponding author. Email: fran.supek@irbbarcelona.org

Published 1 July 2020, *Sci. Adv.* **6**, eaba1862 (2020)
DOI: 10.1126/sciadv.aba1862

The PDF file includes:

Figs. S1 to S9

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/27/eaba1862/DC1)

Table S1

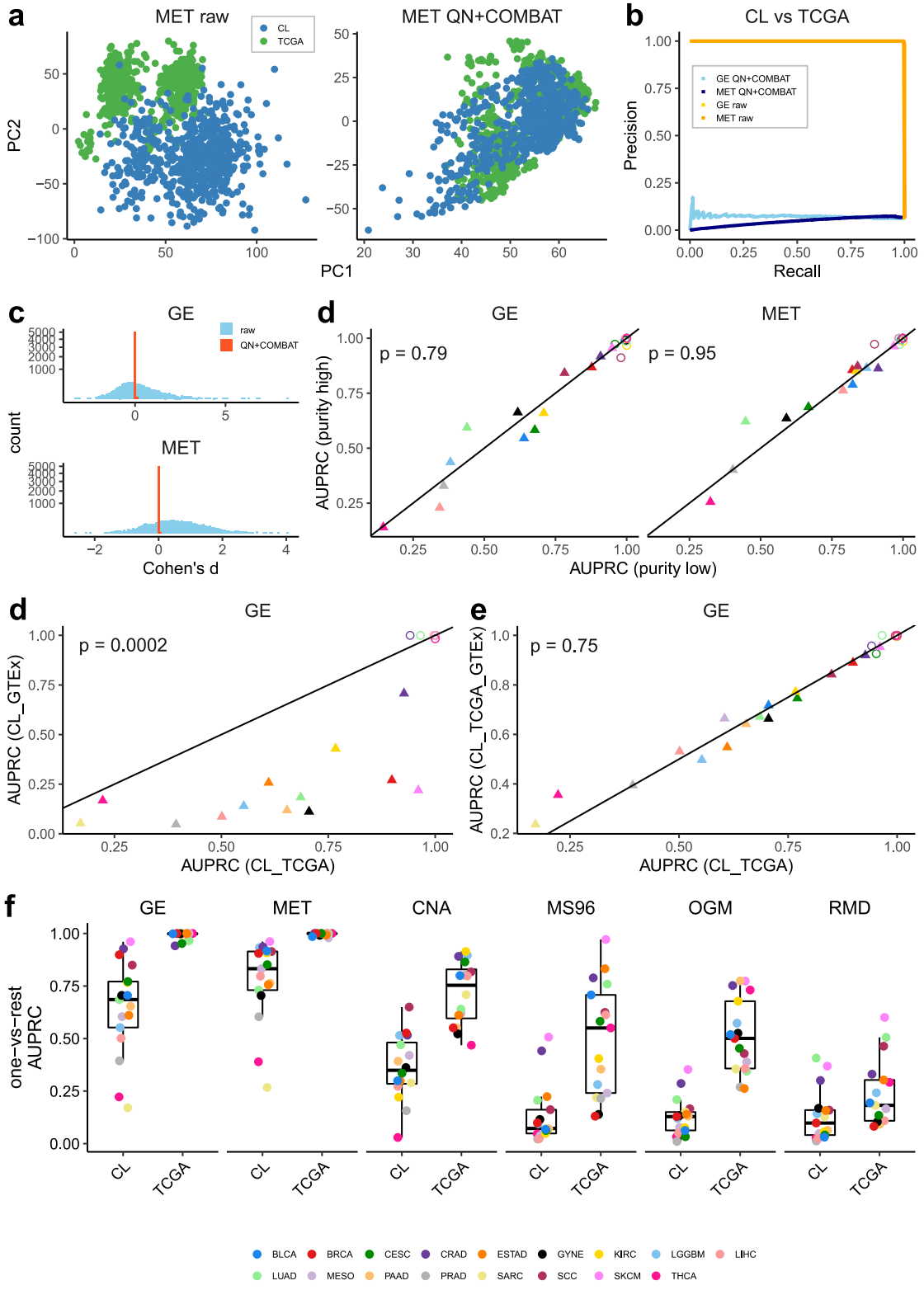


Fig S1. Quality control measures for adjustments of gene expression / DNA methylation data to remove tumor-cell line differences. ***(a)*** *Principal component (PC) 1 and PC2 of a PC analysis in the MET pre-adjustment data (raw) and in the post-adjustment data (QN+ComBat). Colors represent the batch effect labels (CL are the cell lines and TCGA are the human tumors).* ***(b)*** *PR curve for classifying tumors versus cell lines in the data pre-adjustment and post-adjustment for GE and MET.* ***(c)*** *Distribution of the cohen's d between CL and TCGA for gene expression/DNA methylation values before and after adjustment.* ***(d)*** *Comparison between the Area Under the Precision Recall Curve (AUPRC) obtained when testing in TCGA test set (circles) and cell lines (triangles) and training in high versus low purity samples from TCGA.* ***(e)*** *Comparison between the AUPRC yielded when testing in TCGA test set (circles) and cell lines (triangles) and training in healthy samples (GTEx) versus tumor samples (TCGA).* ***(f)*** *Comparison between the AUPRC yielded when testing in TCGA test set (circles) and cell lines (triangles) and training in data adjusted with the previous batch effects (GDSC/CLLE/TCGA) versus adjusted with batch effects including healthy data only (GDSC/CLLE /TCGA/GTEx).* ***(g)*** *Comparison between the AUPRC obtained when training on TCGA train set and testing on TCGA test set and on the cell lines dataset (CL) using 6 different types of features. P-values calculated for cell lines using a Wilcoxon-test (paired, two-tailed) comparing AUPRCs in x-axis versus AUPRCs in y-axis.*

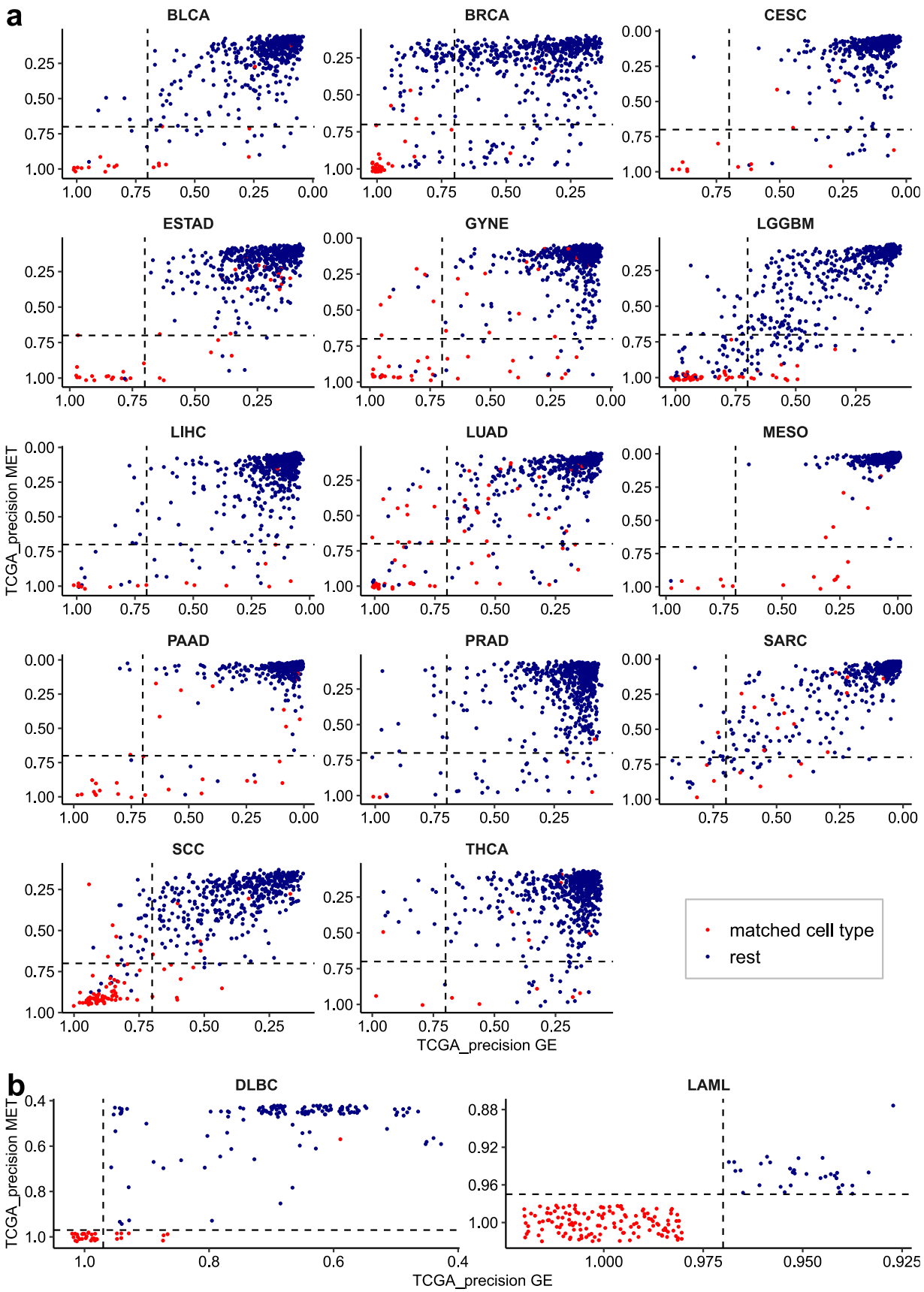


Fig S2. Reclassification of cancer cell lines in various cancer types according to classifiers based on TCGA tumor data. *TCGA-based precision scores (see Methods) for cell lines calculated in gene expression (GE, x-axis) and DNA methylation-based (MET) cancer type classifiers, using a one-vs-rest classification scheme. The higher the TCGA precision, the more likely it is that a cell line belongs to that particular cancer type (shown on top of each plot; the acronyms are from TCGA or described in Methods). The cell lines that were originally annotated as the cancer type that is being tested in each plot are shown in red in that plot, the rest of the cell lines are shown in blue. Solid tumor classifier shown in* ***(a)*** *and blood tumor classifiers in* ***(b).***

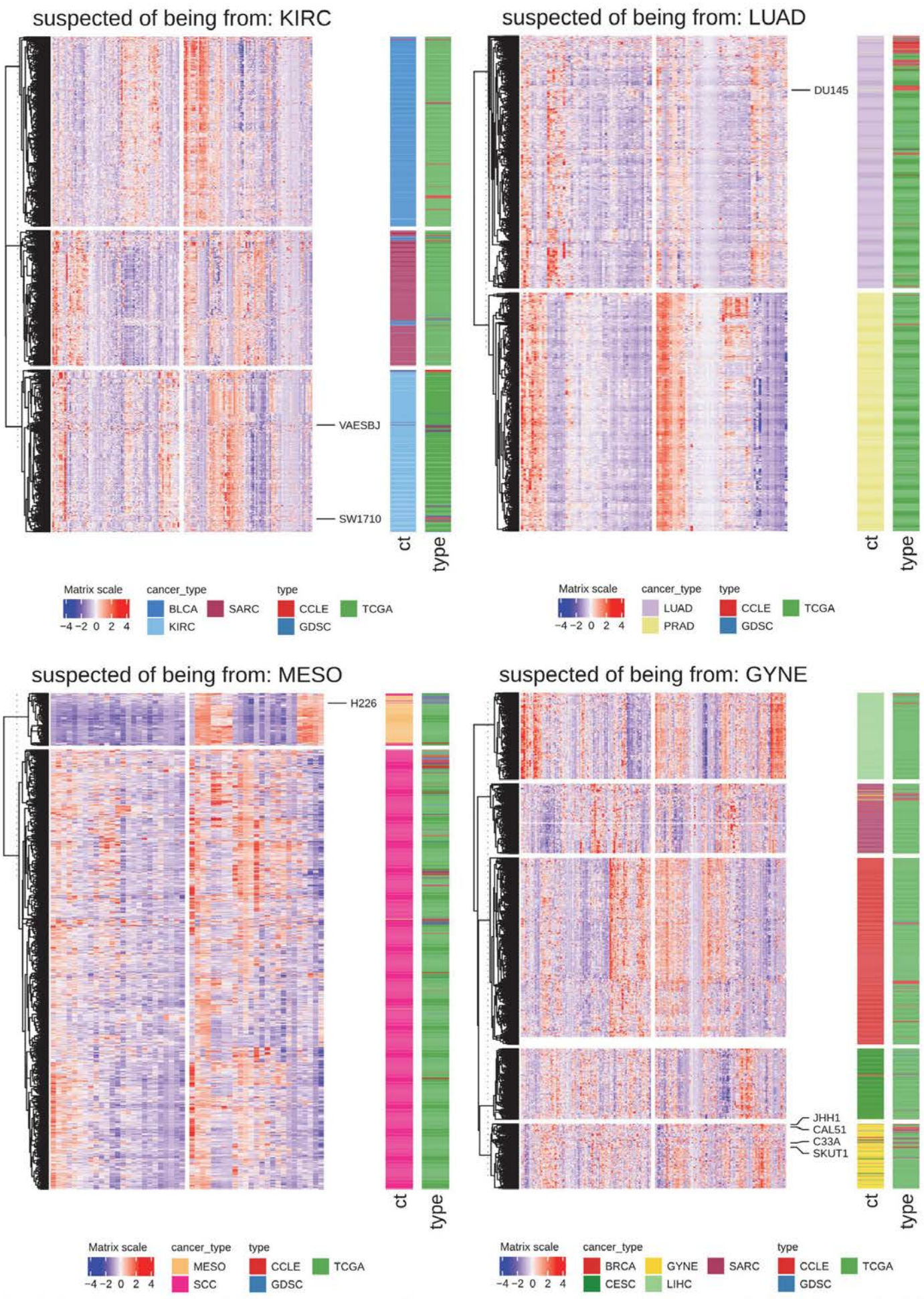


Fig S3. Visual inspection of cell lines mislabelled to a different cancer type. Heatmaps for the 25 genes (in GE) and probes (in MET) with highest absolute value of ridge regression coefficients for each of the cancer types in the plot (suspect cancer type for each case) versus rest classifiers. There is one heatmap per cancer type of those that that presented at least 1 cell line suspected of mislabeling (the suspected cancer type is written in the tittle). In each heatmap, the cell lines suspected to belong to that particular cancer type are labeled on the right-hand side of the plot. In each case, the cancer types included are the suspected cancer type, and the original cancer types of the suspect cell lines.

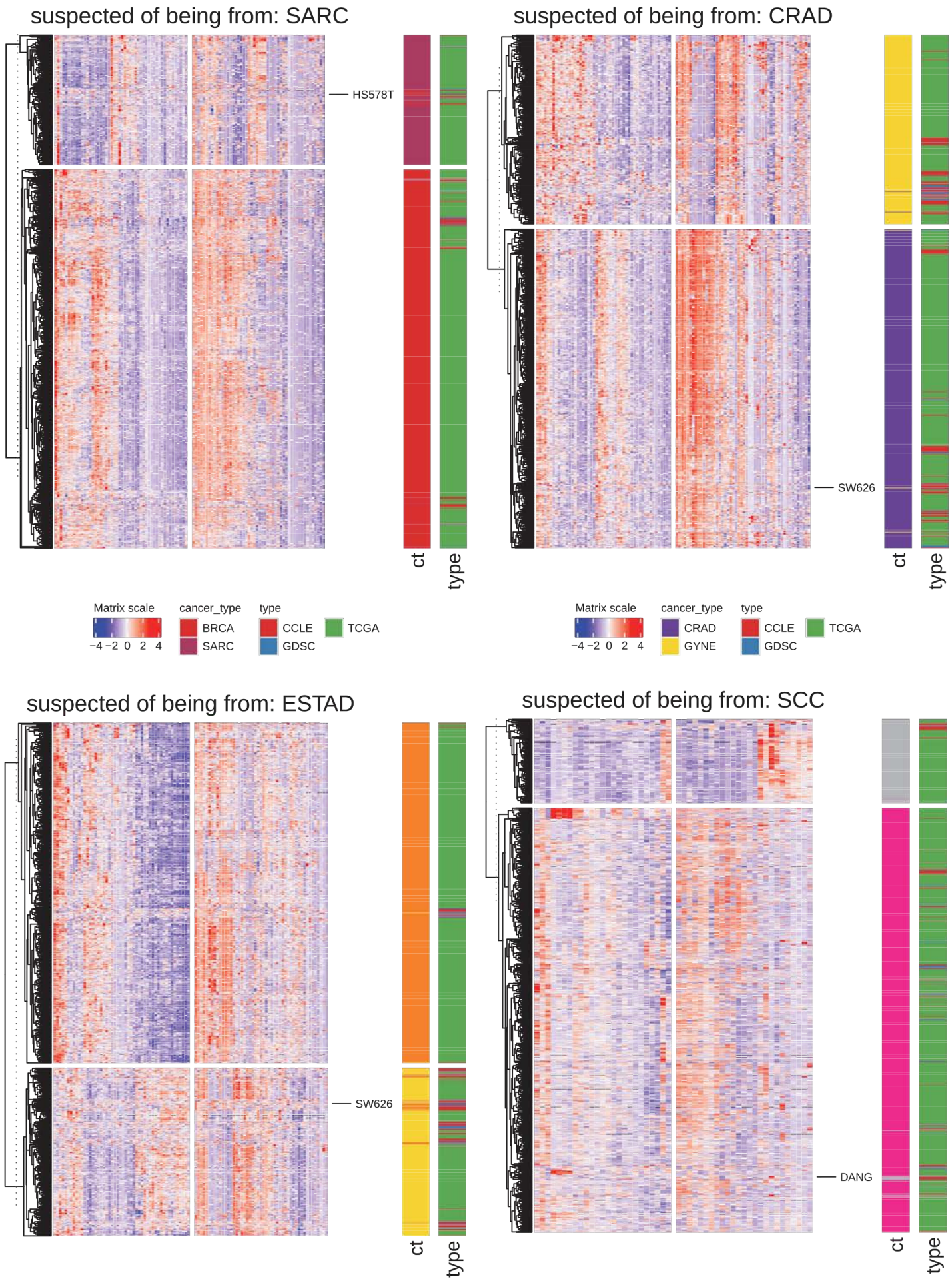


Fig S3. continuation.

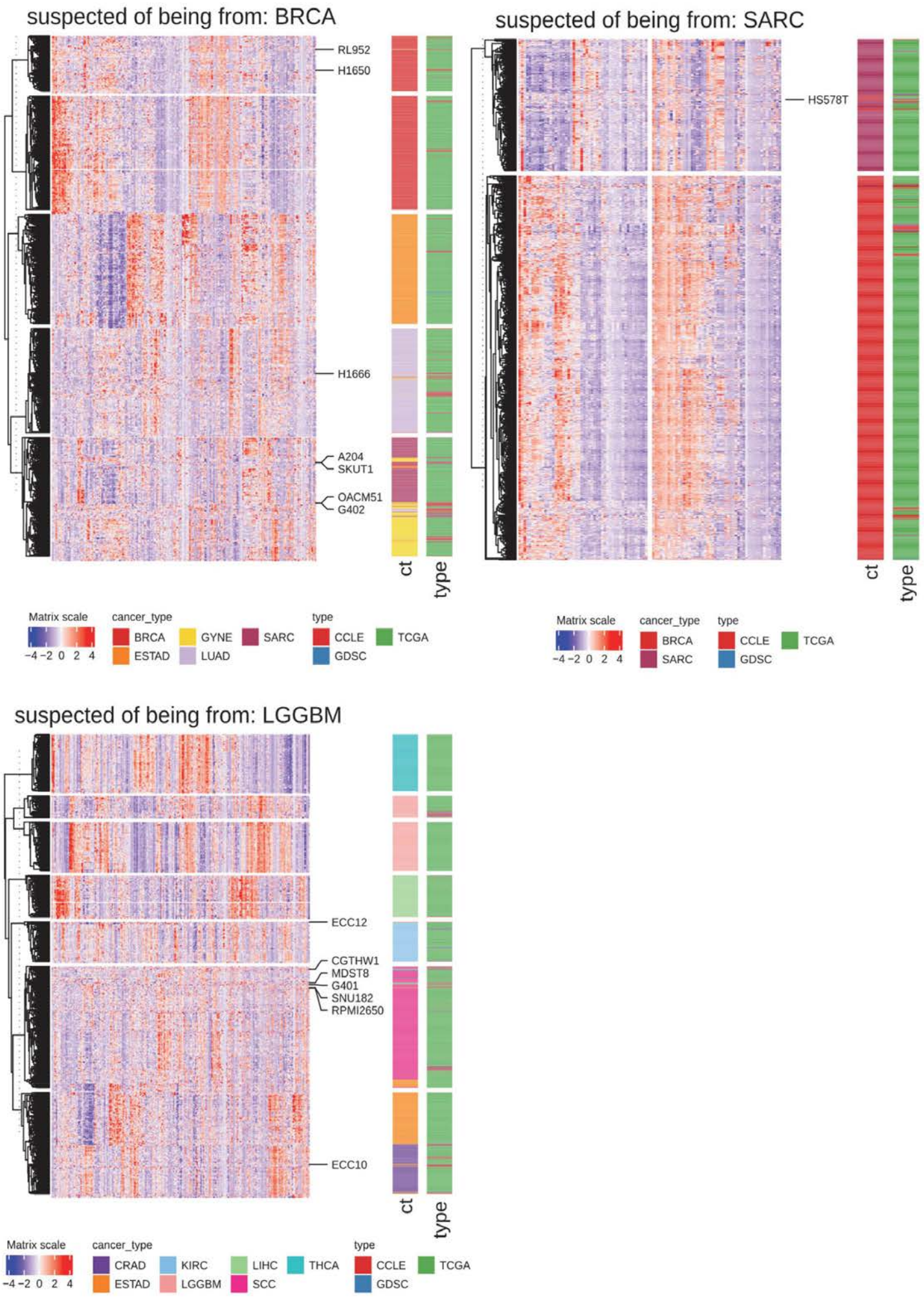


Fig S3. Continuation.

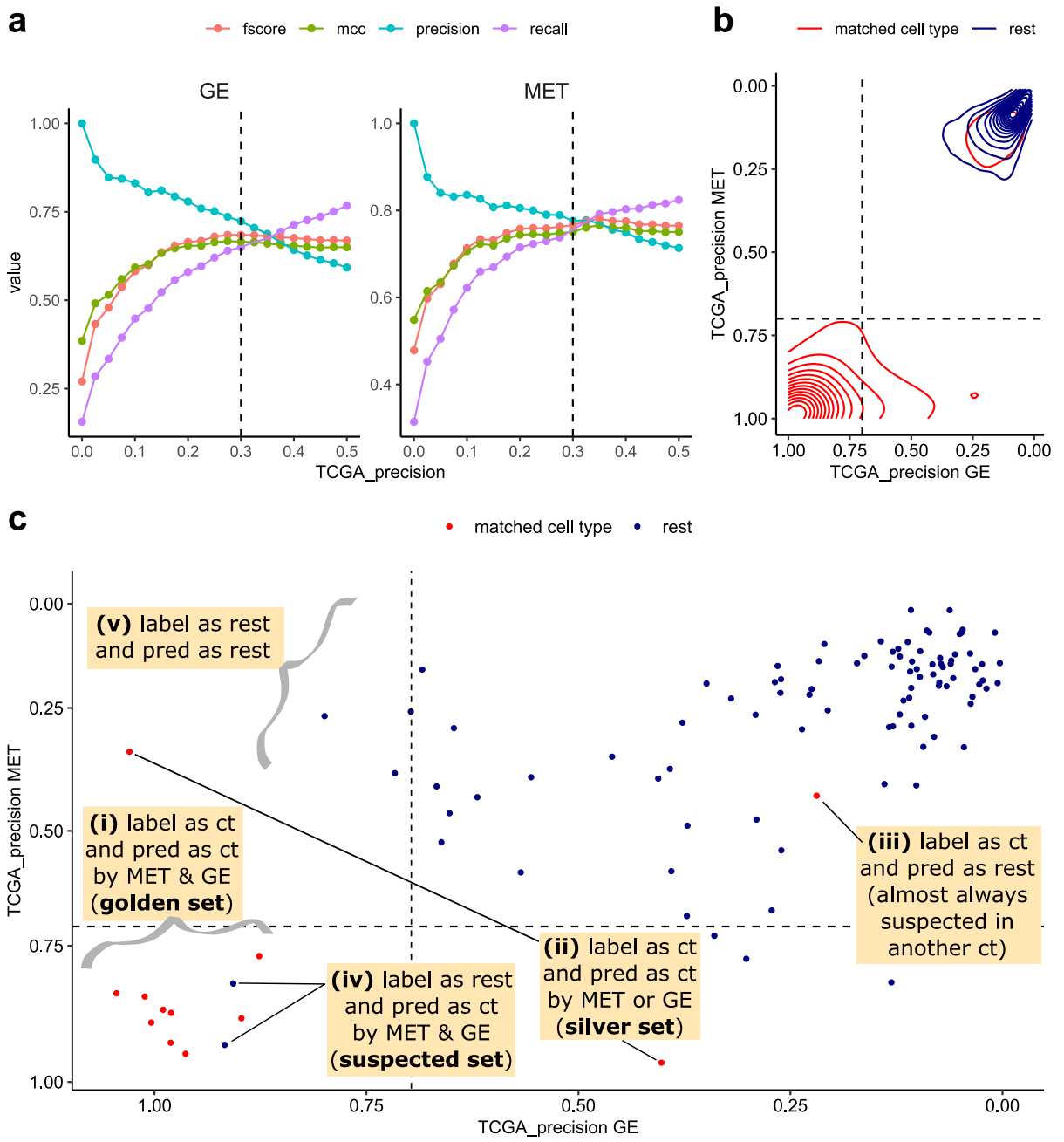


Fig S4. Decision thresholds for reclassifying a cell line to another tissue/cell type of origin. (a) *Accuracy measures (y axis) for cutting at different TCGA-based precision score thresholds (x axis) in gene expression (GE, top) and DNA methylation (MET, bottom) classifiers. "fscore" is the F1 score (harmonic mean of precision and recall), "mcc" is the Matthews correlation coefficient. Shown accuracy measures are derived only from tissue-of-origin labels on cell line data.* ***(b)*** *Distribution (two-dimensional histogram) of data points, collected across classifiers for all tissues, for matched cancer types (red) and the remainder (in blue).* ***(c)*** *A schematic diagram outlining different cases of how, for the selected threshold (TCGA-based precision score ≥0.7), a cell line is classified to the original cancer type or to an alternate cancer type. The shows shows an example of a one-versus-rest classifier, where "ct" denotes that one cancer type that the classifier is trained to recognize from the rest of the cancer types. "pred" is shorthand for "predicted". "label" implies the original label.*

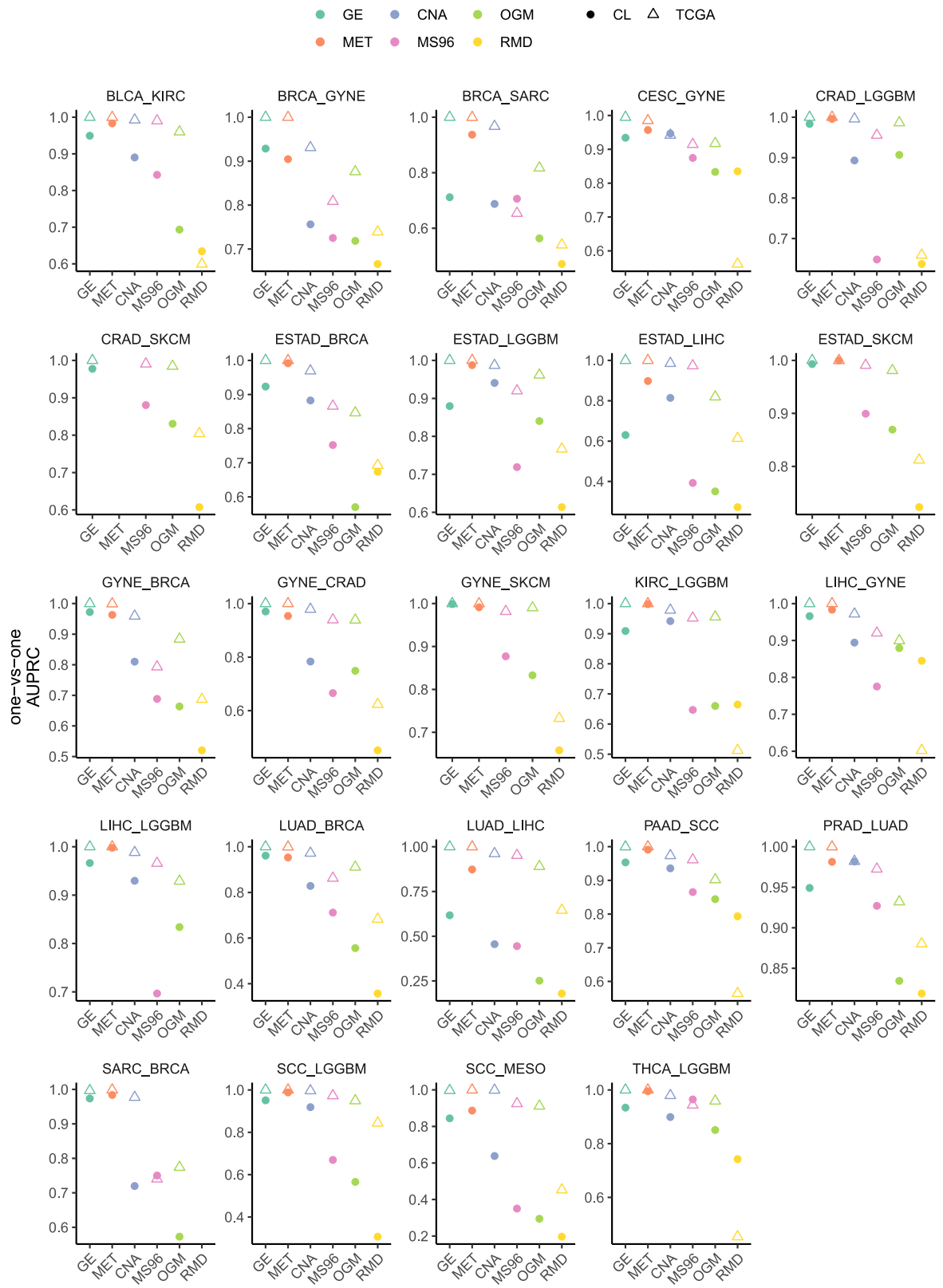


Fig S5. Accuracy of classifiers that distinguish between pairs of cancer types of interest. *Plots show the Area Under the Precision Recall curve (AUPRC) statistic for predicting cancer type in one-versus-one classifiers (the two tested cancer types are named in the top of the plot) using different types of features. GE, gene expression. MET, DNA methylation. CNA, copy number alterations. MS96, trinucleotide mutation spectrum. OGM, oncogenic mutations. RMD, regional mutation density. The classifier is trained on a training set of tumors from TCGA and the AUPRC score is calculated either on a held-out testing set of tumors (TCGA, shown with a triangle), and on all cell lines (CL, shown with a circle).*

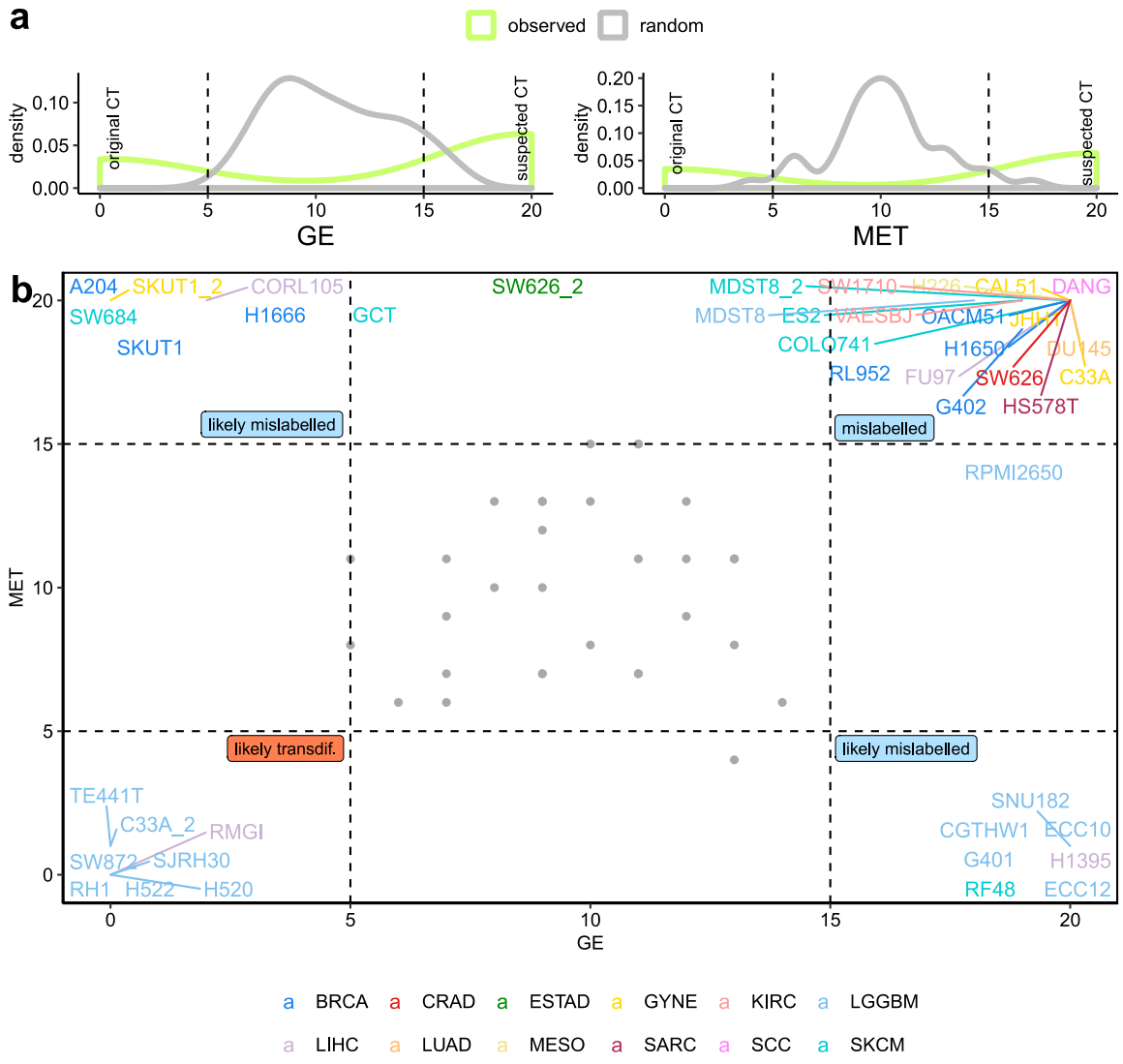


Fig S6. Additional evidence supporting tissue identity of the suspected mislabelled cell lines. (a) *Distribution of the consistency scores (outcome of classifier that predicts suspected versus the original cancer type, run 20 times), for GE and MET features on real data and on randomized data). "CT", cancer type.* ***(b)*** *Prediction consistency scores for the gene expression (GE, x axis) and DNA methylation (MET, y axis) classifiers, for the suspected mislabelled cell lines. Colors represent the suspected cancer type. Grey dots represent randomized data.*

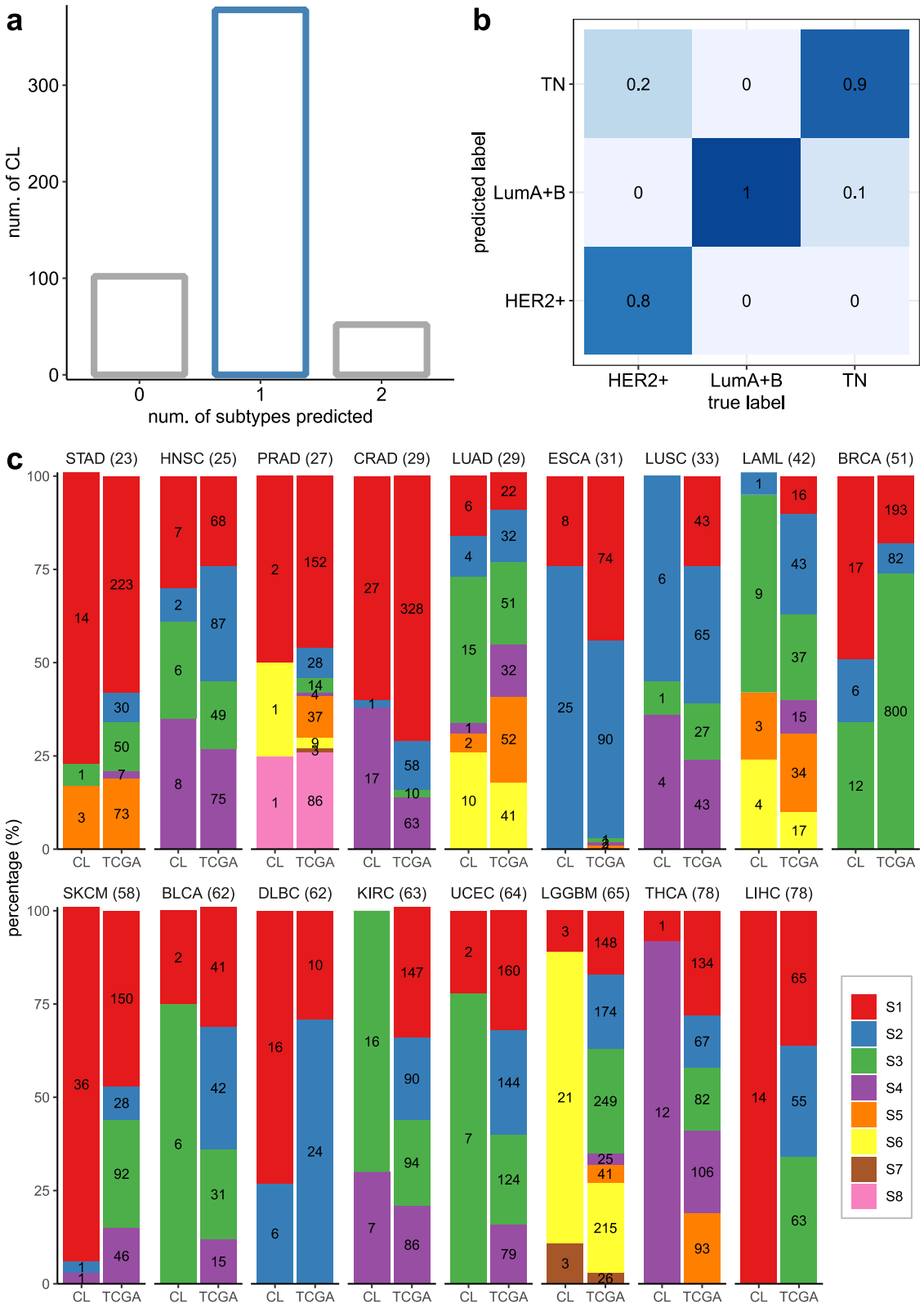


Fig S7. Assigning tumor subtypes to cancer cell lines. (a) *Number of cell lines that are assigned to 0, 1, or 2 subtypes (using one-vs-rest subtype classifiers, within each cancer type).* ***(b)*** *Normalized confusion matrix for the prediction of the breast cancer subtypes.* ***(c)*** *Subtype proportions (raw counts shown overlaid on the bars) observed in the cell lines (CL) and the tumors (TCGA) for each cancer type. The cancer types are ordered from more similar to less similar (left to right) according to concordance between subtype proportions in CL and TCGA (by Euclidian distance, shown above bars; lower scores are better, meaning that examined cell line panels are more representative of tumors in TCGA in terms of subtype representation). Chi-square statistic, p-value and Euclidean distance for the differences between CL and TCGA are provided in Supplementary Table 4.*

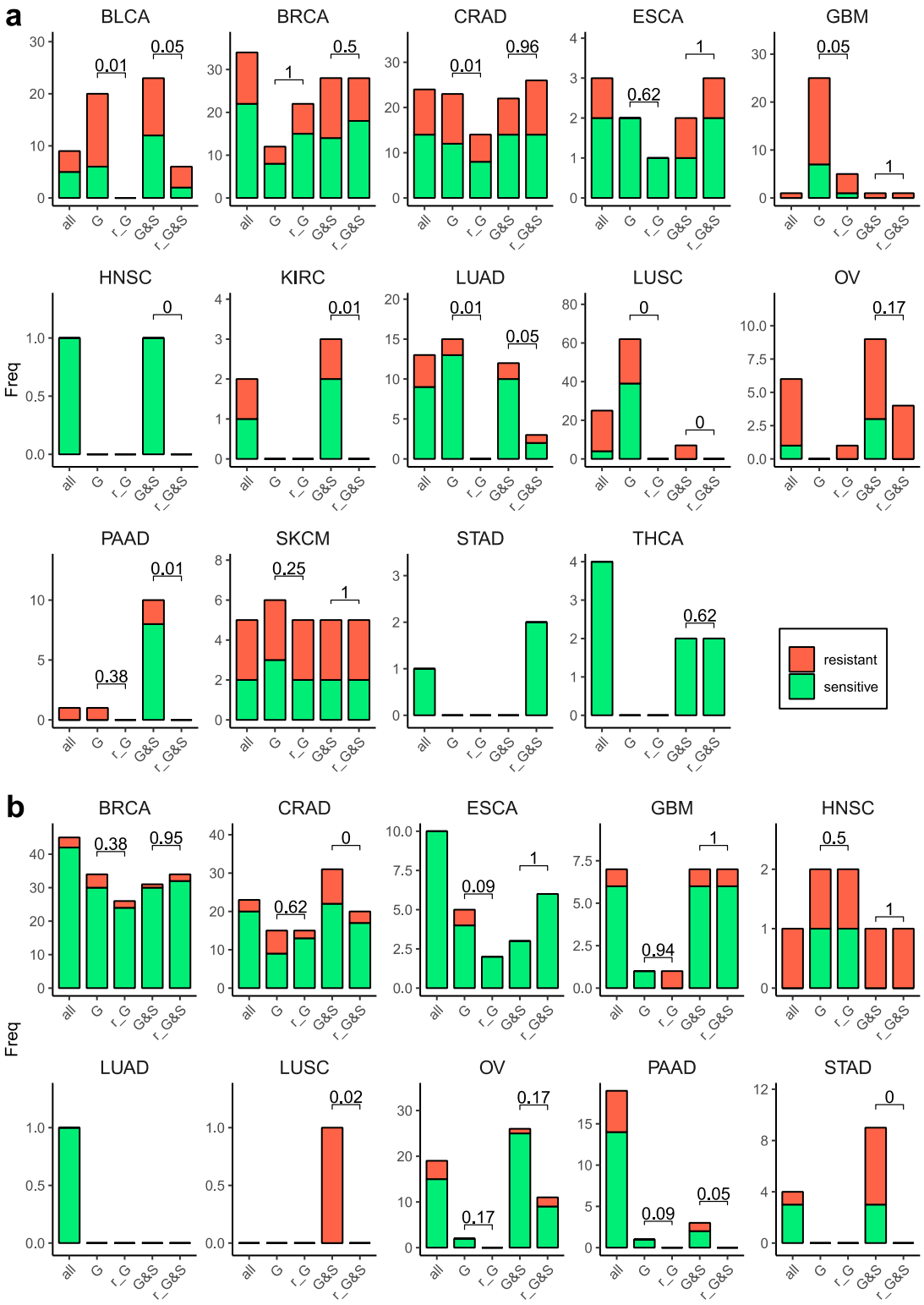


Fig S8. Tally of significant associations between cancer functional events and drug activity or gene dependency on focused panels of cell lines. (a) *Associations with drug activity were detected in an ANOVA test (at FDR 25%) for all cell lines (all), cell lines in the golden set (G), cell lines in the golden and silver sets (G&S) and a random subset of cell lines that match the number of cell lines in the golden set (r_G) and in the golden and silver set (r_G&S). For the random subsets, the number of significant associations is calculated from 10 random samplings and the median shown. P-values shown over the bars are from a sign test (one-tailed) between the associations in the G/G&S and the associations in the 10 runs of random_G/random_G&S.* ***(b)*** *Associations with gene dependency (from CRISPR/Cas9 k.o. screens) were detected in an ANOVA test (at FDR 25%). Labels and statistical tests as in (a).*

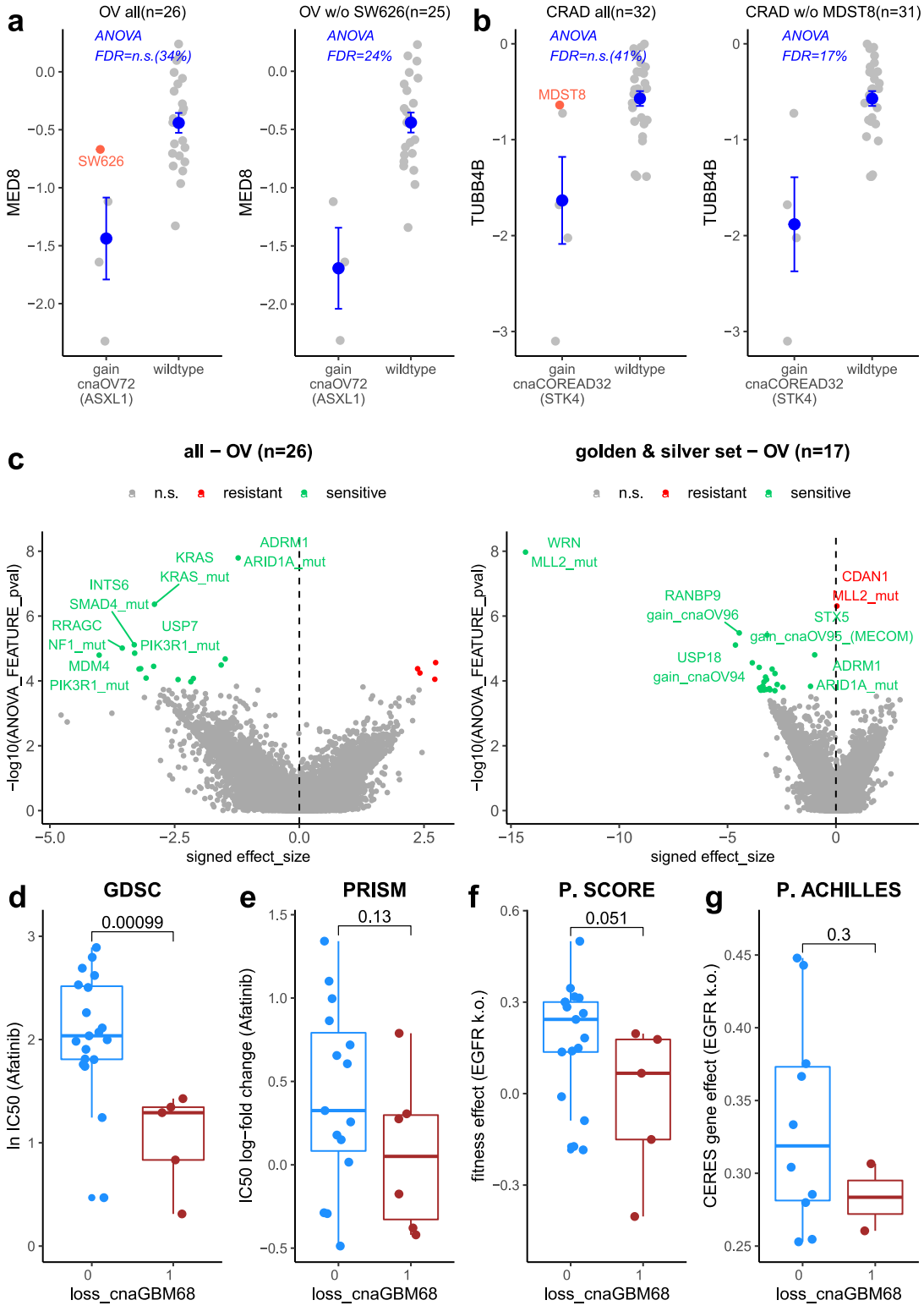


Fig S9. Analyses of genetic screening data using high-confidence cell lines and validation of an illustrative example with independent datasets. (a) *Fitness effect (fold change) for MED8 gene k.o. in all ovarian cancer (OV) cell lines (left) and all OV cell lines except SW626, which is suspected to originate from colorectal cancer (right). Cell lines with copy number gain in a region containing ASXL1 (gain_cnaOV72), and without (wild-type) are compared. ANOVA FDR for this association (MED8 k.o. and gain_cnaOV72) is shown in blue for both datasets.* ***(b)*** *Fitness effect (fold change) for TUBB4B k.o. in all CRAD cell lines (left) and all CRAD cell lines except MDST8, which is suspected to originate from skin (right). Cell lines with copy number gain in region containing STK4 ("COREAD32") and without (wild-type) are compared. ANOVA FDR for this association (TUBB4B k.o. and "COREAD32") shown in blue for both datasets.* ***(c)*** *Differential dependency biomarkers were analysed by ANOVA for all OV cell lines (left) and those in the golden and silver sets only (right). Each point is an association between the fitness effect of a gene and a genetic feature (CFE).* ***(d-g)*** *Drug sensitivity to afatinib in glioblastoma (GBM) cell lines in the 'golden set'. Two groups are compared: cell lines with a particular copy number loss (ID="cnaGBM68", coordinates=chr1:51169045-51472178 (1p32.3), genes affected are CDKN2C and FAF1) (labelled as 1) and wild-type (0). Afatinib IC50 obtained from GDSC in* ***(d)*** *and from PRISM* ***(e).*** *Gene dependency to EGFR k.o. in GBM cell lines (golden set only). Two groups are compared: cell lines with loss ofcnaGBM68 (1) and wild-type cells (0). Gene dependency data obtained from Project Score* ***(f)*** *and from Project Achilles* ***(g).*** *P-values calculated with Wilcoxon test, one-tailed.*

Chapter 6

Tissue-specific origins of replication in human associate with local chromatin remodeling

---

The following chapter is under preparation for publication.

Tissue-specific origins of replication in human associate with local chromatin remodeling

Marina Salvadores[1], Iván Galván-Femenía[1] and Fran Supek[1,2 *]

[1] Genome Data Science, Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, 08028 Barcelona, Spain.
[2] Catalan Institution for Research and Advanced Studies (ICREA), 08010 Barcelona, Spain.
* correspondence to: fran.supek@irbbarcelona.org

Abstract

DNA replication initiates at specific sites known as origins. While several experimental methods exist for mapping these origins, the potential for tissue-specific variations in origin usage has not been clearly established. To address this uncertainty, we developed a computational method, MuSAS, which leverages strand asymmetry in mutation densities across various mutational signatures to map replication origins. Applying MuSAS, we analyzed the strand bias profile switch points, indicative of origins, across roughly 300 tumors. This revealed unique tissue-specific patterns of replication origin usage, with a particularly pronounced trend on brain-specific origins. Supporting this, origins matching specific tissues were found to align with earlier replication timings in those tissues and were often situated near genes predominantly expressed in the same tissue. Additionally, these origins exhibited marked enrichment for chromatin accessibility and histone marks of enhancers and promoters in their respective tissues. Collectively, our results suggest a potential mechanistic link between tissue-specific chromatin activation and the replication origin usage in a given tissue.

Introduction

During the S-phase of the cell cycle, DNA replication initiates from thousands of DNA replication origins, specific sites in the genome from where DNA synthesis initiates and proceeds bidirectionally. These DNA replication origins are activated in a defined temporal order during each cell cycle (1,2). In yeast, the origins of replication are largely determined by sequence (3,4). However, in humans, pinpointing these origins is more complex. Even though the human origins are difficult to identify from DNA sequence alone, they associate modestly with some sequence features, such as CpG islands and G-quadruplex (G4) motifs (5,6), and epigenomic features including open chromatin, DNA hypomethylation, and H2A.Z histone variant (7,8).

Eukaryotic replication origins are established through a two-phase process: first the 'licensing' that is the marking of the site by depositing the pre-replication complex, and second the 'firing' that is the initiation of DNA replication. This sequential mechanism, with two steps separate in time, is vital to stop the same section of DNA from being replicated more than once in a single cell cycle (1,2). Notably, per cell cycle only a subset of all licensed origins are activated (1,2). Based on the frequency of their usage, the replication origins can be classified in 3 classes: (i) constitutive, used in all cell types and in all cell divisions; (ii) flexible, used in an stochastic manner in each cell cycle; and (iii) dormant, used in specific growth or differentiation conditions (1,2). The last class is particularly important for cases when replication forks stall, due to for instance bulky DNA lesions, structured DNA or replication stress conditions; the stalled forks can be rescued by firing of neighboring origins (9). However, what remains ambiguous is if there is systematic variability in origin usage across different tissues, and also across different individuals, and if so which are the genetic or epigenetic determinants of this variation. There are known examples of changes in DNA replication time (RT) between developmental stages (10), cell types and tissues (11) and also RT at some loci does vary between individuals (12,13). One hypothetical mechanism by which these RT switches might result is from locally increased density of origins and/or changed timing or frequency of their firing, raising the possibility that origin usage indeed differs in a systematic manner between tissues and individuals. This is the hypothesis that we are addressing in this study.

Various experimental methods and technologies are employed to map the DNA replication origins, including OK-seq, SNS-seq, Bubble-seq, and Ini-seq, among others (14–16). However, the outcomes from these methods differ significantly, especially in the number of origins per replication cycle (7). These disparities are often attributed to technical factors such as sensitivity, the possibility of false positives or due to resolution. Another possibility is, however, that some of these differences might genuinely mirror variation across distinct cell types or individuals. This was not able to be addressed previously, as each of the individual studies focussed only on one or a small number of cell types, meaning it was not possible to separate tissue variation from individual variation from experimental noise. Moreover the studies were mostly performed in cell line models, which are in some respects different from tissues and the physiological relevance of origins at loci detected thus needs to be established.

The inherent asymmetry in the replication process is well known to be reflected in mutation patterns. This can result from various factors, for instance the differential usage of the two replicative DNA polymerases - polymerase epsilon for the leading and delta for the lagging strand - so the mutational biases may differ (7,17). Similarly other replication strand asymmetric processes were reported to leave a mutational footprint: the leading strand is synthesized

continuously and the lagging strand discontinuously necessitating usage of primers (18) and DNA mismatch repair (MMR) is differentially active between strands (19). That these footprints might be useful as a tool to map origins was proposed by studying the case of tumors bearing pathogenic somatic mutations in DNA polymerase epsilon likely to affect proofreading (17). A recent study employed this mutational strand asymmetry to map origins in a select number of POLE-mutant colon and uterus tumors (n=20 samples in the two tissues) (7). However, it's worth noting that such POLE somatic mutations are not normally seen in most human cancer types (20). Somatic mutations and/or germline variants in DNA polymerase delta (POLD1 gene) are also rarely seen, thus precluding a systematic analysis of many individuals and tissues using this approach based on POLE or POLD1 mutational signatures.

Since we aim to investigate the variability of origin usage across different tissues, an opportunity arises because in addition to the rare POLE signatures, several other more common mutational processes were shown to have replication strand asymmetry, in particular MMR and APOBEC signatures (19,21,22). We hypothesized these, and potentially other strand-biased signatures in cancer genomes could be used to identify origins from the mutational footprint across many different somatic tissues. In this study, we pinpoint asymmetry switching loci in hundreds of human individuals bearing various tumor types, and we ascertain that they indeed correspond to replication origins that show striking differential usage between tissues.

Multiple mutational signatures show mutation asymmetry at origins of replication

Several mutational signatures display strand asymmetry at replication origins, as documented in several studies (7,19,21,22). For instance, the SBS10 signature is found in tumors where the POLE proofreading is defective. Given that POLE only replicates the leading strand, mutations resulting from its deficiency appear on opposing strands on each side of the origin (22). Other processes that exhibit mutation asymmetry at origins include MMR deficiency and APOBEC-induced mutagenesis (22).

In our study, we aim to map the DNA replication origins by leveraging the mutation strand asymmetry across the broadest range of cancer types and individual tumors, enabling us to study their variability. To achieve this, our initial step is to choose the mutational signatures that exhibit strong mutational strand asymmetry at the origins without showing transcriptional strand bias.

To measure the asymmetry at different positions we developed the mutational strand asymmetry score (MuSAS). For each position, MuSAS is calculated by tallying mutations in both the 350 kb regions upstream and downstream. Specifically, it considers the mutations of one alteration (for example, C>T) upstream, and its complementary (like G>A) downstream. This sum is then

compared to the same alteration (C>T) downstream and its complementary (G>A) upstream (see Fig 1a):

$$MuSAS = log2 \frac{\#\ muts\ alteration1\ (upstream) + \#\ muts\ complementary\ (downstream)}{\#\ muts\ alteration1\ (downstream) + \#\ muts\ complementary\ (upstream)}$$

A MuSAS score of 0 signifies the absence of asymmetry. As the score diverges from 0, whether positively or negatively, the asymmetry intensifies. When we computed the MuSAS score around replication origins, we observed a pronounced peak corresponding to the location of origins (Fig 1b). For subsequent analyses that aggregate mutational signatures and mutation types, we consider the alteration's directionality. We selected a window size of +-350 kb to maximize the number of tumor genomes with enough mutation burden within each window to identify a certain number of active origins (see Methods).

To systematically evaluate which signatures have mutation strand asymmetry, we calculated the MuSAS score both at origins and at transcription start sites (windows of +-20kb, see Methods) for every possible combination of mutational signature, mutation type and tissue (Fig 1c). From these evaluations, we picked combinations where the MuSAS at replication origins exceeded 99% of the randomized data, and the MuSAS at transcription start sites was below 99% of that same randomized data (see Fig 1c, where the threshold is indicated by dashed lines).

In total, we found 44 mutational signatures that showed mutation strand asymmetry in at least 1 tissue (Fig 1d). Some of these were expected, for instance the APOBEC signatures SBS2 and SBS13 were reported to be asymmetric (19,22) and we identified them as significant in 20 and 23 tissues (out of 21 and 23 considered). Similarly, MMR signatures SBS44, 21, 26, and 6 were significant in 90 - 100% of tissues considered.

MuSAS method maps origins of replication that agree with experimental assays

To map the origins of replication, we tiled the genome into 1 kb windows and calculated MuSAS scores for each window, after which we applied a smoothing filter. Then we applied a peak caller on the MuSAS profile to locate the replication origins. For this calculation, the higher the mutation burden, the more confidently we can detect these origins. Hence, we only analyzed samples having, on average, a minimum of 4 mutations per 350kb (see Methods). Since almost any sample (e.g. one mutational signature in one individual tumor) reached this threshold, we need to pool mutations to increase the number of MuSAS profiles.

With the goal of measuring differences in replication origin usage across tissues, we performed the pooling using two complementary approaches: (MuSAS-A) by signature and (MuSAS-B) by individual tumor. For MuSAS-A, we computed MuSAS by merging mutations from tumors with the same tissue type and mutational signature (Fig 1e). As a result, our units of analysis are pairs of tissues and mutational signatures. This is feasible because all signatures are expected to identify the same origins (22). For MuSAS-B, we computed MuSAS by combining mutations

from various signatures but from a single tumor (Fig 1f). Merging mutations from multiple signatures within a tumor retains the tissue-specific signal and enables comparisons between tissues. In total, we derived MuSAS profiles for 255 tissue-signature pairs (using MuSAS-A that pools across signatures) and 386 individual tumors (using MuSAS-B that pools across individuals).

In general, there's a strong similarity between the MuSAS profiles obtained using MuSAS-A and MuSAS-B for corresponding cancer types (Fig S1). This demonstrates the robustness of the replication origins identified, with a Pearson correlation of 0.97 when comparing the average MuSAS profiles across all cancer types between the two methods. These average MuSAS profiles represent the general patterns of universally active replication origins and additionally the commonly activated facultative origins.

To ascertain the genome-wide agreement between origins inferred through MuSAS and previously described origins, we compared the MuSAS score in windows that encompasses at least one origin against windows without any overlapping origin (+-10kb) (Fig 1g). We made this comparison using 52 sets of origins derived from various experimental techniques, including SNS.seq, Ini.seq, and OK.seq, as well as a consensus set of origins (mORI) identified in the Jaksik *et al.* study (7) that were found in multiple experimental methods (Fig 1g, see Methods).

When we compare the mean MuSAS values for windows that either overlap or do not overlap with origins, the difference can be quantified using Cohen's d value. This allows us to rank various methods based on their concordance with our MuSAS predictions (Fig 1g). Notably, the highest Cohen's d value was observed for the consensus set of origins proposed by Jaksik *et al.* (7). From this consensus, we specifically chose origins that appear in over 90%, 75%, and 50% of the OK-seq, Repli-seq, and mutORI origin datasets. A more stringent consensus (e.g., 90%) corresponds to a higher agreement with MuSAS. This indicates the effectiveness of our MuSAS score in identifying well-established origins.

Following this, there was strong concordance with MuSAS (high Cohen’s d value) for a previous computational method (Nt.comp.skew from (23)) that identifies origins based on the nucleotide compositional skew along the reference genome sequence. This compositional skew results from the summary activity of various mutational processes in the germline (in contrast with somatic mutational processes studied here) (23).

Among the origin sets from experimental methods, the Ini-seq origins (16) and the majority of OK-seq samples (7,14,24) also demonstrated good concordance with MuSAS. In contrast, the origins inferred from Repli-seq (7) showed lesser concordance by our methodology, possibly because of the lower resolution of localizing origins via Repli-seq analysis. The least agreement with MuSAS (lowest Cohen’s d values) were found for origins identified by the SNS-seq methodology.

Although some methods have a MuSAS concordance greater than others, all 52 subsets of origins from previous work (usually identified by experimental assays) show a significant MuSAS

difference between windows overlapping against non-overlapping origins (p-value< 10^-16). The consistent alignment between our MuSAS score and previously established methods robustly validates that our MuSAS score based on mutation strand bias switches can accurately pinpoint replication origins.

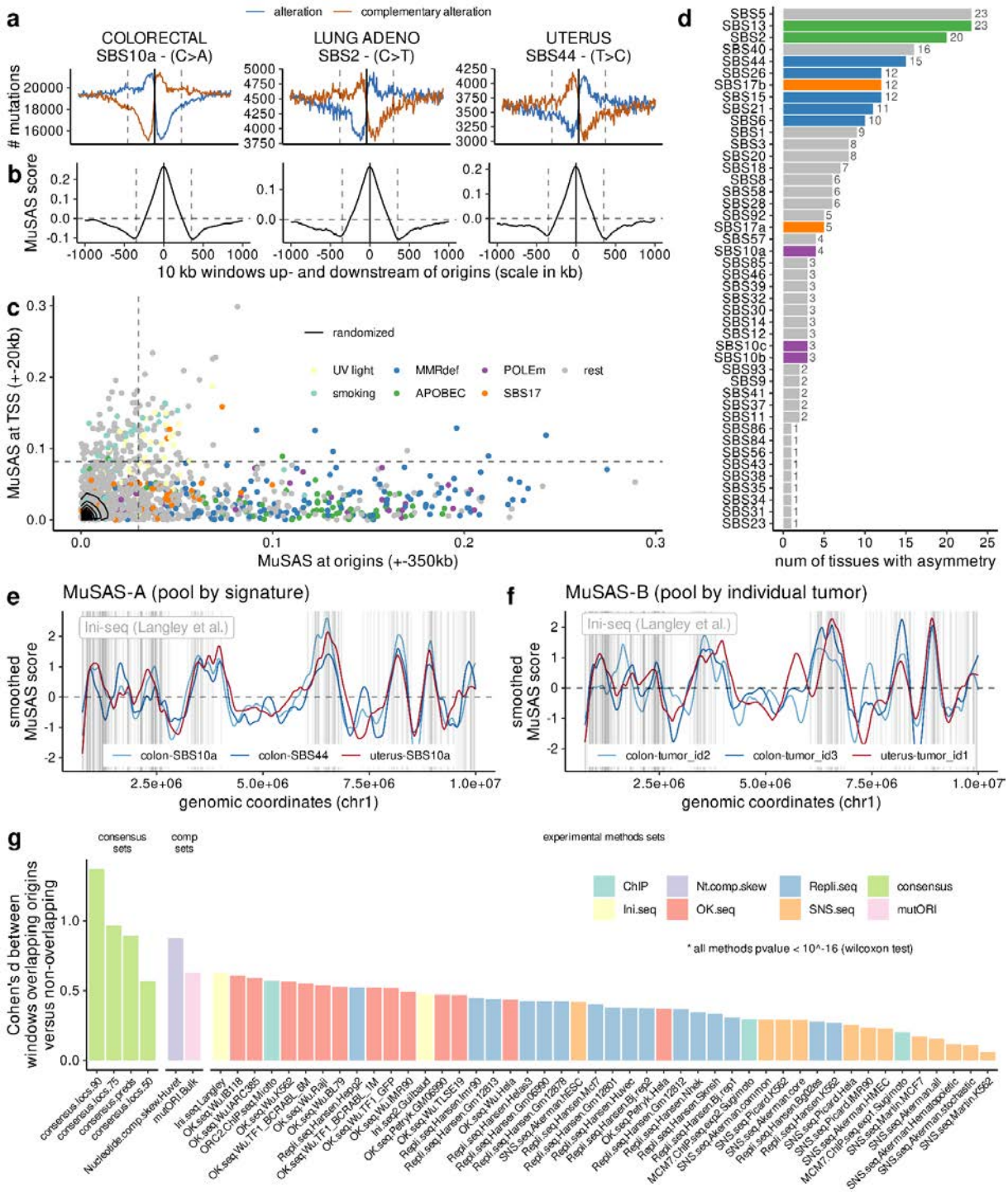


Fig 1. Detection of origins of replication through mutation strand asymmetry. a) Mutation counts within 10kb windows centered on origins (position = 0), separating by mutational signature, tissue and mutation type. **b)** Mutational strand asymmetry score (MuSAS) for the same windows and specific mutational_signature - tissue - mutation_type combinations as depicted in (a). **c)** MuSAS values at origins and transcription start sites (TSSs) with the

randomized data distribution in black and the 99th percentile thresholds marked by dashed lines. **d)** Mutational signatures displaying mutation asymmetry at origins (above the threshold in panel **c**) without showing transcription strand bias (below the threshold), showing the tissue count where each signature is significantly asymmetric. **e)** Example of smoothed MuSAS score for a segment of chromosome 1 in a tissue-signature pair from MuSAS-A. **f)** Example of smoothed MuSAS score for a segment of chromosome 1 in a individual tumor (merging all mutational signatures within the tumor) from MuSAS-B. **g)** Agreement of MuSAS score with previous methods to identify origins. Cohen's d comparison of MuSAS scores between 1kb windows overlapping at least 1 origin and windows without any origin overlap within a +-10kb range, assessed for 52 distinct origin sets identified by different methods (largely experimental, except the "Nucleotide.comp.skew" and "mutORI.bulk" method).

Mutational strand bias patterns reveal tissue-specific origins of replication

Over the last decade, numerous methods have been employed to detect origins of replication, reflecting the intricate nature of DNA replication and the significance of understanding its commencement. However, despite the array of tools at our disposal and the insights garnered, there's still limited data on how replication origins vary across different tissues. This gap in knowledge signifies a crucial area of exploration, as tissue-specific dynamics of replication origins can offer insights into cellular processes, development, and disease etiologies.

In this study, we explored the differences in origin usage across various cancer types using principal component analysis (PCA) on the MuSAS profiles. This was done separately for MuSAS-A (which pools data based on mutational signatures) and MuSAS-B (pooling data from individual tumors). For both approaches, the primary principal component (PC1) accounted for most of the variability and correlated with the average MuSAS (Fig 2ab). These values from A_PC1 and B_PC1 identify the constitutive origins, i.e. those consistently used across cancer types. Interestingly, the subsequent principal components (PCs) captured the tissue-specific variability in the mutational footprints of origins usage (Fig 2a-d).

Before further analyzing the tissue-specific origin PCs, it's crucial to differentiate between robust PCs and those that might reflect mostly noise. To discern the PCs that consistently reflect variations in origin usage, we examined their coherence across distinct genomic segments. For this, we divided the genome into three distinct bins: chr1 to 5, chr6 to 11, and chr12 to 23, each spanning roughly 1000 Mbs. We then performed a PCA on each bin. If the variations in origin usage detected by a PC were robust, the PCs would be consistent across bins, meaning that PCs with uniform origin usage patterns in these different genome sections would tend to group together, as demonstrated in Fig S2. Following this, we delved deeper into the first six PCs from both MuASAS-A and MuSAS-B.

First, we correlated the PC profiles from MuSAS-A and MuSAS-B to determine the agreement between the two complementary approaches. We found that four of these trends appeared in both methods (Fig 2e), with the remaining two exclusive to each method.

Second, to determine the cancer types relevant to our chosen tissue-specific PCs, we analyzed the sample exposure values. An example is the difference in PC2 from MuSAS-A, which separates tissue-signature pairs linked with brain tumors (Fig 2c). This means that the A_PC2 profile pinpoints origins of replication unique to brain tumors and not to other types. Similarly, we found a breast versus colon trend in MuSAS-A PC3, uterus versus squamous trend in MuSAS-A PC4 and a prostate specific trend in MuSAS-A PC6 (Fig 2cd). We conducted a systematic review of the sample exposure values for the initial 6 PCs across both methods. Ultimately, we identified six unique tissue-specific trends depicting genome-wide shifts in origin usage based on mutation patterns (Fig 2f, Fig S3-4).

As mentioned above, we found one brain specific pattern (A_PC2 and B_PC4, corr = 0.7) that strongly separates brain tumor samples, including both pediatric and adult brain cancers, from the rest of the tissues (Fig 2c). For this PC, the gastrointestinal tissues like the colorectum, intestine, and stomach exhibit high exposures in the opposite direction, possibly acting as a contrasting background due to their substantial sample size.

Next, a genome-wide pattern in replication origin usage that opposes breast *versus* colon was evident (A_PC3 and B_PC3, corr = 0.74). Both methods displayed the gastrointestinal samples, including the colon, consistently on one side of the divide, while breast along with uterus samples were on the other side (Fig 2c, Fig S3-4). This implies that the A_PC3 and B_PC3 profiles capture unique replication origins – gastrointestinal tumors are represented by peaks, while breast and uterus origins are depicted as valleys.

Additionally, we found a uterus-specific trend (A_PC4 and B_PC2, corr = 0.59) which markedly differentiates uterus samples from others (Fig 2d, Fig S3-4). For MuSAS-A, this contrast is primarily with squamous-type tissues like cervix, head and neck, and lung squamous (Fig 2d, Fig S3) while for MuSAS-B, breast samples predominantly formed the background (Fig S4). Similarly, the prostate specific pattern (A_PC6 and B_PC6, corr = 0.43) sets prostate samples apart from squamous samples in MuSAS-A (Fig 2d, Fig S3) and brain samples in MuSAS-B (Fig S4).

Two PCs were exclusive to individual methods and did not replicate between our MuSAS-A and MuSAS-B, possibly because of statistical power issues. A bladder-associated pattern of origin usage (B_PC5) distinguishes bladder from colorectal and breast in MuSAS-B (Fig S4). Meanwhile, a lymphoid-specific trend (A_PC5) distinguishes lymphoid samples from others in MuSAS-A (Fig S3). For these examples of tissue-specific origin usages in bladder and lymphocytes, there is less confidence than for the other tissue-specific patterns described above.

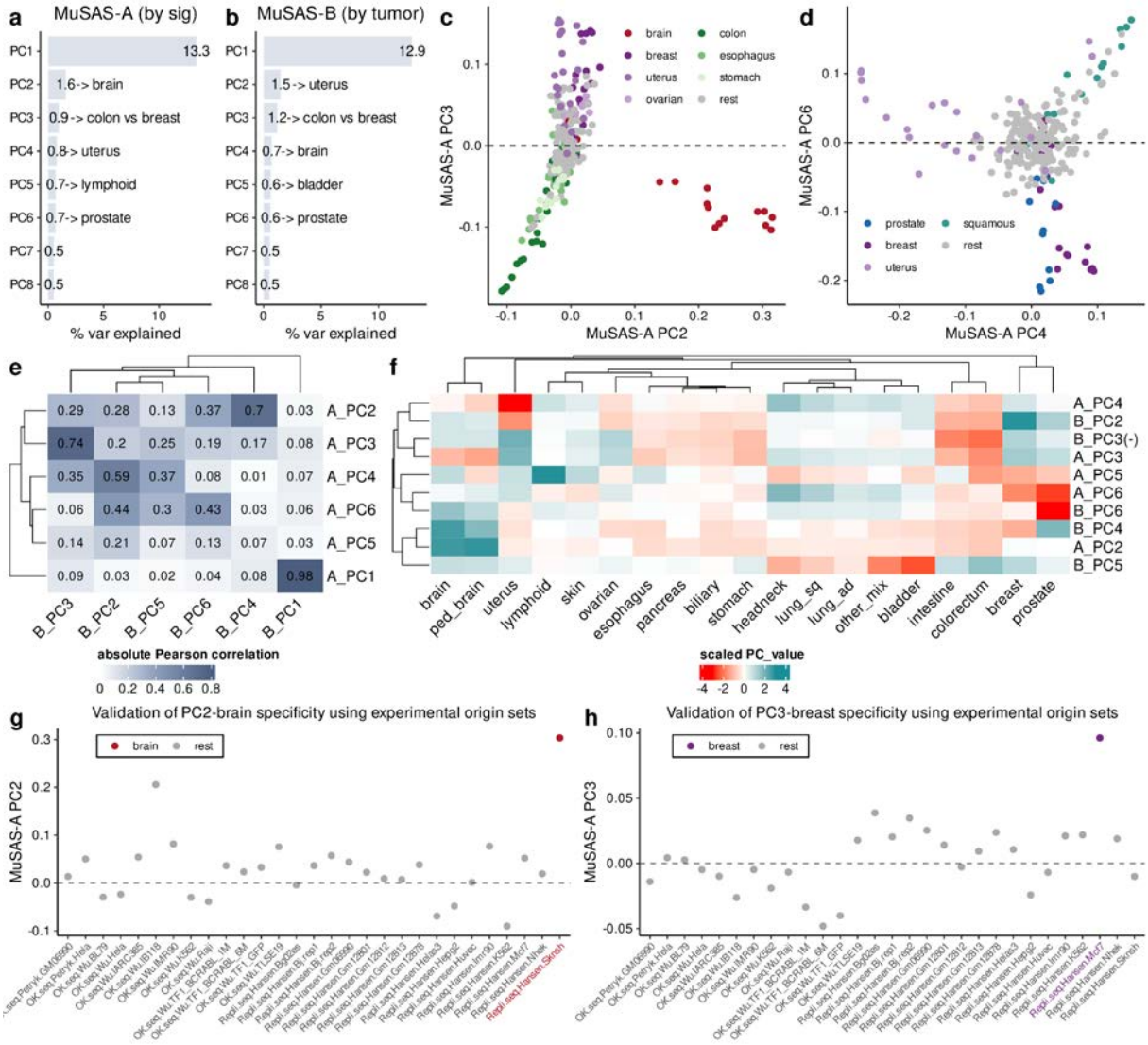


Fig 2. Tissue-specific origins of replication identified by PCA on MuSAS scores. a) Variance explained of the first origin PCs for a PCA onto a matrix of MuSAS score for 255 tissue-signature pairs (MuSAS-A) x 2678657 1-kb windows. **b)** Variance explained of the first origin PCs for a PCA into a matrix of MuSAS score for 386 individual tumors (MuSAS-B) x 2752410 1-kb windows. **c)** Origin usage PC2 and PC3 activities inferred from different mutational signatures using MuSAS-A (pooling across individual tumor samples), stratified by cancer type. Each dot is a tissue-signature pair. **d)** Origin usage PC4 and PC6 activities inferred from different mutational signatures using MuSAS-A (pooling across individual tumor samples), stratified by cancer type. Each dot is a tissue-signature pair. **e)** Robustly recovered origin usage PCs that replicate between methods MuSAS-A (y axis) and MuSAS-B (x axis). **f)** Mean PC exposure per cancer type for the tissue-specific PCs by methods MuSAS-A and MuSAS-B. **g-h)** Origin usage PC2 (in **g**) and PC3 (in **h**) values for the windows that contain at least one origin identified by experimental assays.

Validating tissue-specificity of replication origin usage in experimental data sets

For validation, we analyzed the overlap between PC2 (specific to the brain) and PC3 (breast versus colon) with cell type specificity in experimental readings using the OK-seq assay (7,14,24), and in the origin zones inferred from Repli-seq experimental measurements (7,25), which we combined into one dataset. For performing this cell type association with experimental data, we compared the average PC value across various experimental origin subsets (Fig 2g-h, Fig S5). The highest average of PC2 (pertaining to brain-specific origin activity) appears in the origin subset specific to the Sk-n-sh cell line (neuroblastoma), while these values are lower in the other cell lines (PC2 = 0.30 in neuroblastoma versus -0.09 to 0.20 in the other n=29 cell lines considered). Likewise, the highest PC3 average (breast-specific origin usage, contrasted vs colon) is detected in the origin subset from the Mcf7 cell line, originating from the breast (Fig 2g-h, Fig S5). In both instances, the origin-usage PC value is notably elevated in tissues with the closest match to the cell line, offering further validation when compared to the non-matched tissues.

Next, we aim to validate the identified tissue-specific origins of replication by comparing the changes in DNA replication timing (RT) at the origin loci across tissues. This is based on the reasonable assumption that RT will be, overall, earlier at an origin in the tissue(s) where that origin is more frequently used, than in the tissues where the origin is less frequently used, where the replication forks would need some time to reach that locus starting from other flanking origins. The RT profiles used here were inferred from accessible chromatin patterns (ATAC-Seq datasets for 410 tumors (26) occurring in various human tissues including 2 brain cancer types), using the Replicon tool (27); this was shown to be a highly accurate method for predicting domain-scale RT profiles (28,29).

First, for every PC profile of origin usage, we applied a peak caller to pinpoint the identify the local peaks and valleys (Fig 3a). Since the direction of the PCs is arbitrary, tissue-specific origins will be at the peaks for cancer types with a positive PC exposure (e.g. A_PC2 brain), while they will be at the valleys for those with a negative PC exposure (e.g. A_PC4 uterus). In this analysis, we compared the RT at the origin-PC peaks versus at the PC valleys for each tissue-specific origin usage PC (Fig 3b). A difference in RT – calculated as the RT at the peaks subtracted from the RT at the valleys – will be <0 if the replication timing is earlier at the peaks, and >0 if it is later (Fig 3c). We summarized these comparisons in Fig 3d (see legend for description of color coding and direction of PCs).

Upon inspection, we observed that the variation in replication timing robustly supports the six tissue-specific patterns we identified (Fig 3d, Fig S6). Strikingly, the difference in RT at peaks versus valleys for the PC2-brain-specific origins distinctly differentiates both glioblastoma (GBM) and lower-grade glioma (LGG) cancer types, meaning that these cancer types have earlier replication at the PC2-peaks (brain-specific origins) in comparison to RT in other analyzed cancer types (as are shown in Fig 3c, median dif RT at peak-valley of -0.014 and -0.014 in brain, while ranging from -0.008 to 0.023 in other tissues). Similarly, for PC3-breast-specific origins, replication is earlier in breast (BRCA) and uterus (UCEC) tumors than in other cancer

types (Fig 3d, Fig S6). In contrast, for PC3-colon-specific origins, replication occurs earlier in colon (COAD) and stomach (STAD) adenocarcinoma, compared to other tumor types where RT was inferred (Fig 3d, Fig S6). The RT patterns also corroborate the tissue specificity of the uterus-specific, prostate-specific and bladder-specific trends (Fig 3d, Fig S6).

Tissue-specific genes co-localize with their matching tissue-specific origins

Given the tissue-specific nature of the detected origins and their shift towards earlier RT, we wondered if genes with tissue-specific expression might be located near these origins. To investigate this, for each PC, we ranked genes by the origin activity-PC value of their encompassing window and conducted a gene set enrichment analysis (GSEA). We analyzed two distinct gene sets: (i) from the GTEX database, we chose the top 50 genes that showed the most significant expression variation in one tissue relative to others (Fig 3e); and (ii) from the TiGER database (30), we used gene sets that emphasize tissue-specific expression and regulation (Fig S7).

The results from GSEA indicate that tissue-specific highly expressed genes are located in areas with high tissue-specific origin PC values (Fig 3e, Fig S7), meaning that higher local gene expression predicts higher origin usage. For the brain-specific origin patterns, we found an enrichment of brain specific expressed genes: FDR for association of origin-PC2 with 10 brain-specific expression gene sets range from 0.05-5%, with other tissue-specific expression gene sets median FDR=52%, Q1-Q3=16-72%) (Fig 3ef). An illustration of this is the genes OLIG1 and OLIG2. These two transcription factors are expressed in the spinal cord and play a sequential role in the formation of motoneurons and oligodendrocytes. In the GTEx data, the expression levels of OLIG1 and OLIG2 in the brain cortex are 21 and 106 TPM, respectively. However, in other tissues, the next highest expression levels are only 0.23 and 4 TPM, respectively. These genes, predominantly expressed in the brain, are found at a brain-specific origin identified using our MuSAS method (Fig 3g). In this region, replication timing shifts to earlier in brain cancers compared to other tumors such as the colon or stomach (Fig 3g).

For the origin-PC3 representing the breast versus colon origin usage pattern, there was a positive enrichment on mammary tissue associated genes (FDR = 2%) versus a negative enrichment of colon specific genes (FDR = 1%) (Fig 3e). One example is the gene ELOVL5 (150 TPM in mammary tissue versus 127 TPM in the next highest expressed tissue), which is located at a breast-specific origin of replication and where there is a RT switch to earlier in breast cancers (Fig 3h). We additionally observed an abundance of genes distinctive to the cervix in the uterus pattern (FDR = 25%) and genes unique to the prostate in the prostate-specific pattern (FDR = 15%) (Fig 3e).

In summary, the above analyses support that MuSAS reliably identifies different tissue-specific replication origins usage patterns, which were validated using tissue-specific DNA replication timing, chromatin accessibility at regulatory elements, and highly expressed gene set enrichment. We observed that tissue-specific origins exhibit earlier replication timing in their

respective tissues compared to others and that genes specifically expressed in these tissues are found to co-localize at these origins (Fig 3f-g). This supports a mechanistic link between tissue-specific activation of genes and/or of regulatory elements, and tissue-specific origin usage.

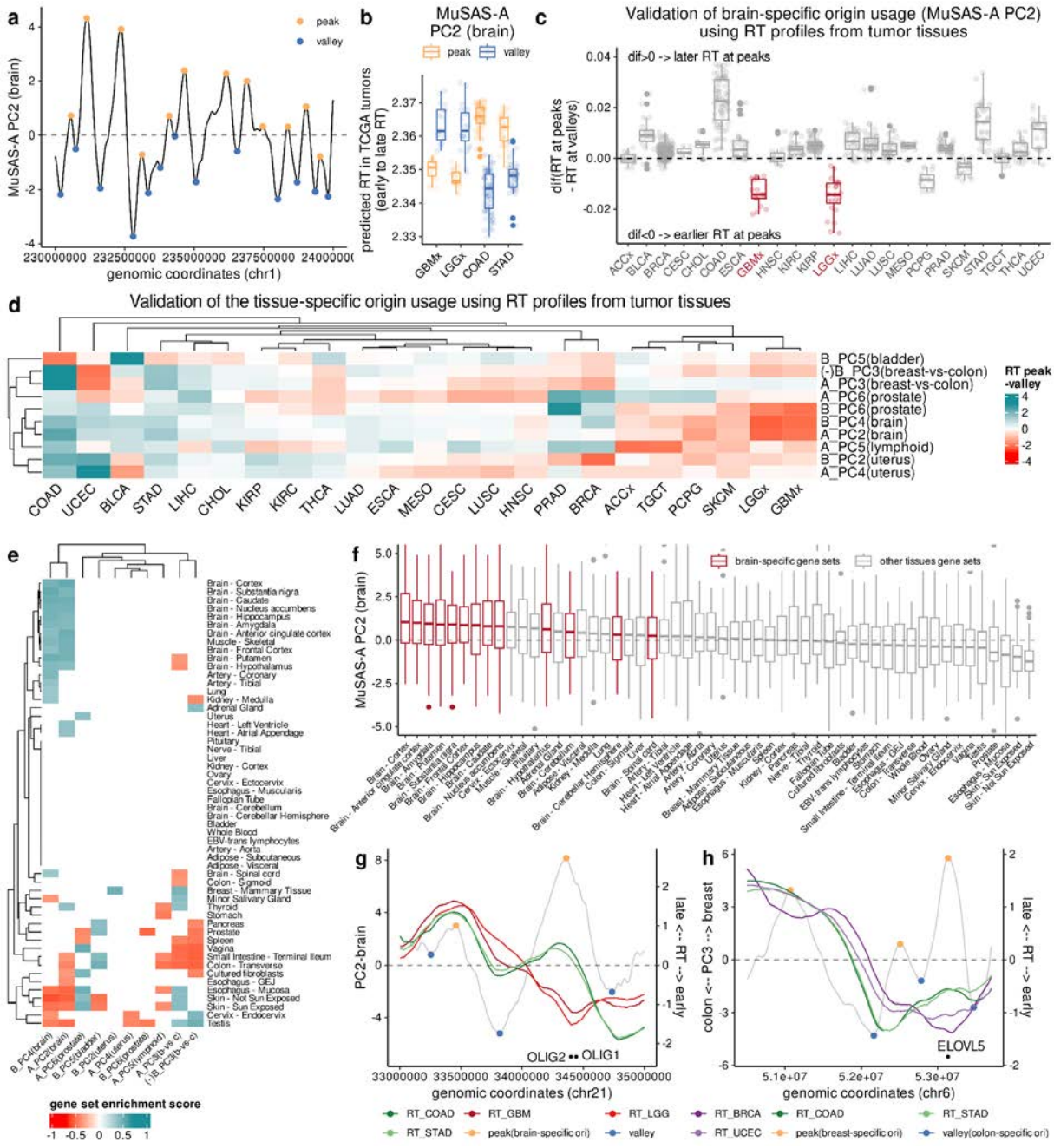


Fig 3. Tissue-specific origins of replication co-localize with earlier replication timing and the tissue-specifically expressed genes of the matched tissues. a) Example of origin PC2

(by MuSAS-A) window weights for a segment of chromosome 1. The identified peaks and valleys are marked with dots, corresponding to brain-enriched origin usage and brain-depleted origin usage, respectively. **b)** Replication timing at the peaks and valleys of PC2 (MuSAS-A) for 4 tissues. The RT is predicted from ATAC-seq data in 410 TCGA tumors using the Replicon software (27). **c)** Validation of the tissue-specific origin usage PC2 using RT profiles from different tumor types. Difference between the mean RT at peaks and the mean RT at valleys of PC2 (MuSAS-A) for each TCGA tumor grouped by cancer type. **d)** Difference between the mean RT at peaks and the mean RT at valleys for each tissue-specific PCs of method MuSAS-A and MuSAS-B for each TCGA tumor type. The color coding mirrors the direction from Fig 2d: a pronounced blue hue in Fig 2d indicates that the origins are at the peaks, and if the same tissue(s) display a strong blue shade in Fig 2i, it signifies an earlier RT at those peaks. **e)** Tissue-specific gene set enrichment scores using GTEx data (y axis) for the tissue-specific origin PCs (x axis) by method MuSAS-A and MuSAS-B. The gene sets with an FDR > 25% are coloured in white. **f)** The brain-specific origin usage PC2 (method MuSAS-A) weights in the 1 kb windows where the tissue-specific expressed genes are located, stratified by tissue (x axis). **g)** Brain-specific origin PC2 profile, where the peaks correspond to the brain-specific origin usage increase, and RT profiles are shown for brain tissues (GBM, LGG tumor types) versus digestive tissues (COAD, STAD tumor types). The black dots marked the position of two brain specific expressed genes. **h)** Breast-associated origin usage PC3- profile where the peaks correspond to the breast-specific origins, and the valleys to the colon-specific origins; overlaid are shown RT profiles for breast (BRCA) and uterus (UCEC) versus colon (COAD) and stomach (STAD). The black dots marked the position of an example breast specifically expressed gene.

Tissue-specific chromatin accessibility is enriched at the corresponding tissue-specific origins of replication

Different histone marks and epigenomic features were reported to be associated with origins of replication. Jaksik *et al.* (7) found that several DNA features co-localized with the consensus set of ORI (identified by various assays from earlier studies (15,23,31–34)), including chromatin loop anchors, G-quadruplexes, S/MARs (scaffold/nuclear matrix attached regions), and CpG islands (often found at promoters and enhancers). The H2A.Z histone marked showed the most significant association among all the features (7). In another recent study that considered epigenomic features in the K562 cell line, Dao *et al.* (8) reported that various histone modification signals are enriched in ORI-dense compared to ORI-sparse regions, with the strongest enrichment observed in the enhancer mark H3K4me1 and the regulatory region-associated DNase I hypersensitive sites (DHS), a mark of locally accessible chromatin.

In our study we identified both constitutive and tissue-specific replication origins. This led us to ask whether tissue-specific origins exhibited any distinct chromatin features in the matched cell types, which would distinguish them from the non-matched cell types. The tissue-specific origins were further assessed in how they differ in the chromatin associations from the constitutive origins. In a sense, the tissue-specific origin usage (as identified in the brain, colon, breast etc. by our MuSAS scores) could serve as a natural experiment to identify which of the many

proposed origin-associated chromatin features are more likely to be the causal features for origins.

To investigate this, we analyzed each origin subset separately, such as brain or GI specific origins. For each subset, we grouped all regions containing origins (centering the origin within a 10-kb window) together with their 10-kb neighboring areas (extending up to 1Mb both upstream and downstream from the origins). Next, we calculated the density of each DNA feature within the origin regions and contrasted this with the flanking regions across various biosample groups (broadly corresponding to tissues or cell types) (Fig 4a, Fig S8).

For evaluating chromatin accessibility, we employed the cell-type-specific DHS vocabulary provided by the global analysis of 733 ENCODE biosamples by Meuleman *et al.* (35), where they categorized the DHS into different signatures or components and associated these to cell types (e.g. neural or lymphoid). We assessed the density of each DHS component within the origin areas in relation to the neighboring regions (Fig 4a). As expected, in the case of constitutive origins, there was a pronounced overlap of DHS at the origins as opposed to the adjacent areas for the majority of components (Fig S8a). Notably, when considering tissue-specific origins, there was a more pronounced enrichment for the component closely aligned with the tissue of the origins (Fig 4b-c).

Specifically, we saw a co-localization of the neural, embryonic and musculoskeletal DHS components at brain origins (log2 FC = 1.28 to 0.75, compared to origin flanks) but not so for the other DHS components (log2 FC = 0.17 to -2.09). Similarly, for colon-specific origins the strongest enrichment is for the DHS digestive component (1.51 log2 FC; other components 0.80 to -0.42 log2 FC) and for lymphoid-specific origins the strongest enrichment is for lymphoid, myeloid DHS components (1.42 and 0.69 log2 FC), and also the tissue invariant component (Fig 4c). These results further support the tissue specificity property of the origins we identified via MuSAS mutational analysis, because the origin usage is mirrored in the tissue-specific DHS signal. Moreover, these suggest that chromatin accessibility is causally involved in origin definition and/or activation, since the DHS strongly covaries with origin usage in a tissue-specific manner.

Tissue-specific enhancer and promoter histone marks are enriched at their corresponding tissue-specific origins of replication

We additionally examined the enrichment of thirteen histone modifications available from ENCODE for diverse human tissues (Fig 4d, Fig S8b) (36). We separated the samples based on tissue type to broadly match with our tissue-specific origins. The groups are blood (for lymphoid origins), central nervous system (CNS) (for brain origins), breast (for breast origins), gastrointestinal (GI) (for GI origins) and prostate (for prostate origins).

As before, for each tissue type, we calculated the mean density of every histone mark within our origin areas and compared it to the adjacent areas of the origin (Fig S8b). For each set of

tissue-specific origins and each histone mark, we calculated the fold change between the origin regions (ori +-10kb) and the flanking (+-50kb to 1Mb). Notably, the matched combination (same tissue in the mark and in the origins) clustered together and showed a greater enrichment compared to other tissues in several histone marks (Fig 4d).

In particular, we found the two active enhancer marks H3K4me1 and H3K27ac to be specifically enriched in the matched tissue in the origins compared to flanking regions, for all tissue-specific origin sets (0.61 and 0.46 mean log2 FC in matched combinations versus -0.55 and -0.06 log2 FC in the rest, in H3K4me1 and H3K27ac respectively) (Fig 4d). Both these marks, H3K4me1 and H3K27ac, are enhancer-associated modifications and are linked with activated transcription of target genes (37), with some association with super-enhancers (38). Of note, the H3K27ac is found also at active promoters.

Similarly, H3K4me3 and H3K4me2 marks were also enriched at origins compared to flaking in the tissue-associated marks of the corresponding tissue-specific origins (0.52 and 0.63 mean log2 FC in matched combinations versus 0.38 and 0.27 log2 FC in the rest, in H3K4me3 and H3K4me2 respectively) (Fig 4d). H3K4me3 is a hallmark of the promoters of actively transcribing and poised genes. The H3K4me2 is highest toward the 5′ end of transcribing genes (39). The other promoter mark (H3K9ac) did not cluster with the other enriched marks but showed a similar enrichment (0.46 log2 FC in the matched combinations versus -0.06 log2 FC in the rest). These findings indicate that tissue-specific origins are closely associated with enhancers and promoters that are selectively activated in the respective tissue.

Additionally, tissue-specific patterns of histone variants H2A.Z (H2AFZ) and H3.3A (H3F3A) are also associated with tissue-specific origin usage in most cases (Fig 4d). The H2A.Z variant was reported to regulate the licensing and activation of early replication origins and to maintain the replication timing (40). The H3.3 histone A was reported to associate with activated genes by occupying the intronic regions. The H3.3 tends to bind regions that show characteristics of regulatory DNA elements (41). Association of these histone variants with origins in a tissue-specific manner supports that activation of genes or regulatory elements is closely intertwined with tissue-specific replication origin usage.

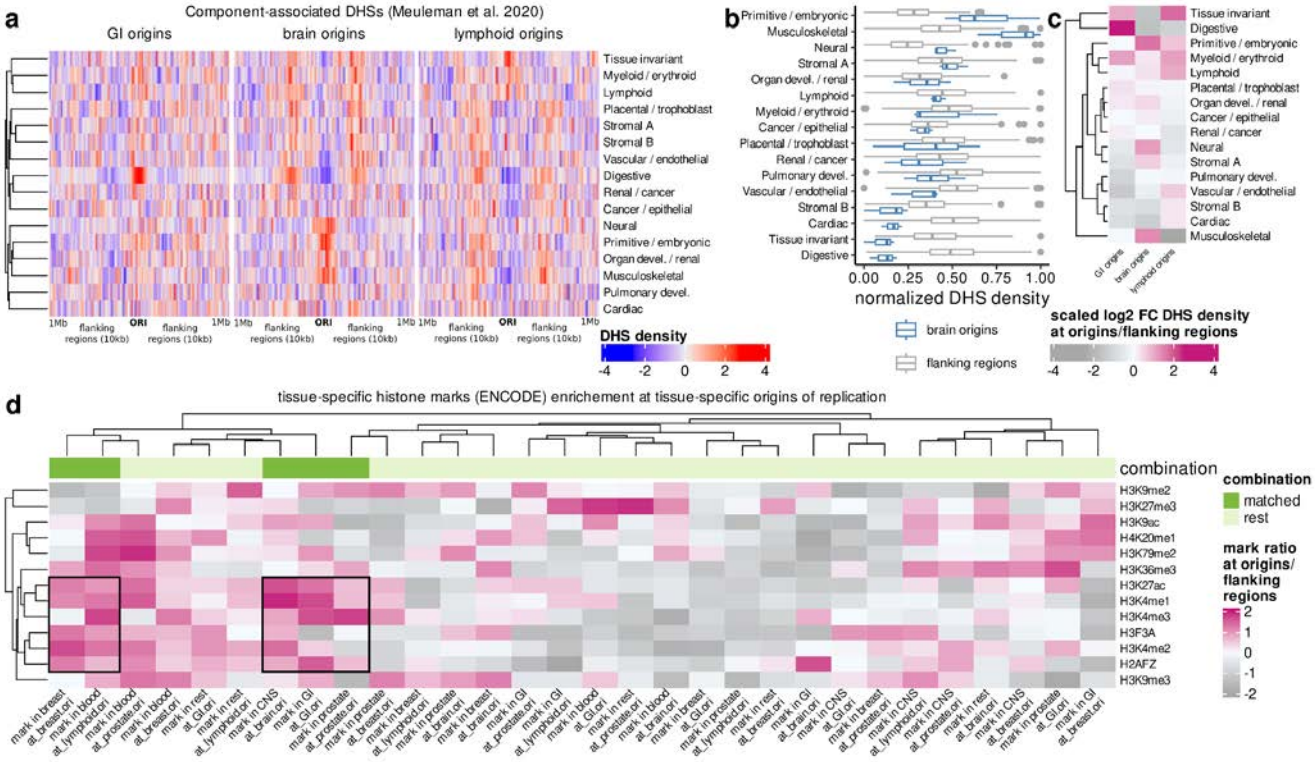


Fig 4. Tissue-specific chromatin accessibility and histone marks are enriched at tissue-specific origins. a) Density of chromatin accessibility (DHS) across 10-kb windows centered at tissue-specific origins of replication, stratified by components defined by Meuleman et al. (35) **b)** Analysis contrasting DHS density between the origin position (±10kb) and its flanking regions (ranging from ±50kb to 1Mb) in brain-specific origins. This contrast helps identify the components that show enrichment at the origins. **c)** Assessment of component-specific DHS density enrichment at tissue-specific origins. **d)** Analysis of enrichment for 13 tissue-specific histone marks densities at tissue-specific origins. The combinations where the tissue of the mark matches the tissue of the origin set are highlighted in dark green (matched), with all other instances in light green.

Tissue-specific transcription factor bindings are enriched at their corresponding tissue-specific origins of replication

To further narrow down the possible mechanistic underpinnings of the tissue-specificity of the origins reflected in gene activation, we considered that transcription factors (TFs) might mediate this link between active enhancers/promoters and origins. We thus systematically evaluated the enrichment of various bound TFs available as ChIP-Seq from ENCODE in select cell lines that corresponded to the tissues of our replication origins (36). These cell lines included GM12878 (lymphoblastoid cell), HCT116 (colon adenocarcinoma), K562 (myeloid leukemia), MCF-7

(breast carcinoma) and SK-N-SH (neuroblastoma). In these cell lines, we tested binding in a total of 419 distinct TFs for association to replication origins in a tissue-specific manner.

For this particular analysis, considering not every cell line has data for all the TFs, we evaluated each tissue-specific origin set individually, linking it with the TF set for that particular cell line. To jointly analyze the coordinated variation of TF binding, we performed a PCA on the TFs binding density profiles across the origin and its adjacent regions. Next, we selected those TF binding-PC (corresponding to a summary of multiple TFs) that exhibited a peak within the replication origin zone (Fig S9). We then investigated the activities for that selected TF-PC, testing for association with a tissue-matched cell line, compared to the other cell lines.

For the brain, breast, colon, and lymphoid tissue-specific replication origin sets, we observed that the cell line matching the tissue showed more pronounced TF-PC activity than other cell lines (Fig S9), supporting that differential activation of regulatory elements underlies differential replication origin usage. In all four cases, the higher TF-PC activity spanned across many TFs, rather than having a clearly defined subgroup of TFs driving the change (Fig S9).

We further check this by splitting the TFs by categories, however we still observed the gradient in all the categories (Fig S10). Therefore, this analysis did not identify any particular group of TFs distinctly linked to the tissue-specific activation of origins, but rather supports the model where binding of many distinct TFs may promote origin activity by activating the enhancers and promoters. Of note, our analysis does not formally exclude that the ordering of events in time might be the opposite meaning that origin activation precedes the activity of the co-located enhancer.

Discussion

In eukaryotic cells, thousands of potential replication origins are present during each cell division. Yet, the initiation at these origins is stochastic, so not all are activated in each replication cycle. When comparing origins reported by different experimental methods, there is notable variability in both their number and location (7). While these differences are often attributed to technical factors, the genuine biological variability of origin usage across different tissues and individuals remains largely uninvestigated. Our work addresses this question: We developed a genomics method to map origins of replication across different somatic tissues in human and uncovered several tissue-specific patterns on the origin usage.

Our method, named MuSAS, leverages the mutational strand asymmetry bias associated with specific mutational signatures at replication origins to map their location. Numerous signatures, including those linked to POLE mutations, MMR deficiency, and APOBEC mutagenesis, were reported to exhibit pronounced mutational strand biases at origins (7,19,21). Our findings suggest that indeed those and possibly additional mutational signatures can be applied to detect origins. A recent study has shown this to be feasible by means of POLE mutational signature (known as Signature 10 or SBS10) (7), which largely restricts the analysis to colorectal and

uterus cancers, and also limits statistical power as even in these organs very few tumors have this signature SBS10. In contrast, our MuSAS can combine these diverse mutational signatures instead of relying only on SBS10. We can thus notably enhance statistical power, increasing the number of individual tumors and importantly we can cover many distinct tissues in our analysis.

Interestingly, we found that SBS5 exhibited sufficient mutational strand bias across 23 different cancer types to be able to detect origins. Two prior studies did not report any replication strand bias at origins for SBS5 (19,21). A possible explanation is that our methodology calculates the mutational strand asymmetry in a specific manner: instead of merging all mutation types within a single mutational signature after determining their directionality (19), we assess the asymmetry for each mutation type (e.g. C>T versus C>A) individually, later merging only the combinations with asymmetry; this might be more robust to noise. We cannot rule out the explanation that mutations from other mutational processes (possibly MMR deficiency) are 'leaking' into the SBS5, meaning that some part of mutations labeled as SBS5 originate from another signature. Despite this potential overlap, we retained these potentially confounded SBS5 mutations in the method on empirical grounds. Their presence benefits our method if they indeed have inherent replication strand asymmetry, while their inclusion in case of lacking asymmetry is not expected to adversely impact the score.

A parameter in our MuSAS method is the window width used on both sides of a given location to determine the strand asymmetry score. For this study, we chose a window size of 350kb to maximize our sample size in terms of number of mutations. This window width permits the inclusion of more samples but trading off the resolution somewhat (we note, importantly, that the effective resolution is considerably more precise than 350 kb; this interval is so wide such that smoothing over multiple mutations can mitigate noise). Therefore, unlike high-resolution experimental techniques for identifying origins, such as SNS-seq, our approach addresses the issue of identifying regions where origins are typically activated. Yet, within these regions, multiple sites could potentially serve as origins, as we see in Fig 1e-f, where numerous Ini-seq origins fall within a single MuSAS peak.

Importantly, the window width parameter could be adapted to the study's goals and additional resolution can be gained by restricting window size but thus reducing the number of available samples. In our case to explore tissue-specific variability in origin usage, we sought to maximize the number of individual samples trading off resolution. In a hypothetical analysis that would focus on universal origins across all tumors, pooling samples would allow for a narrower window, thereby increasing resolution. In a similar vein, we defined two complementary approaches in this study: MuSAS-A and MuSAS-B, which pool mutations (to increase statistical power) according to different criteria. Similarly to the window width parameter choice, these approaches are designed to maximize sample size and encompass a diverse range of cancer types. Depending on the specific research objectives, alternative combinations of mutations, signatures, and samples might be utilized. Finally, a limitation of our approach is that our method relies on mutation accumulation; we acknowledge that our method does not have power for tissues and/or samples with low mutation burdens. Moreover it is unlikely to recover origins that are rarely activated.

Our results from the global MuSAS analysis (using the described parameters above) reveals clear variability in replication origin usage across diverse tissues. Here, the brain tissues in specific are identified as the ones with the most prominent tissue-specific variation in origin usage, according to the PC analysis. This is in accord with the known particularities of the brain, compared to other tissue, in terms of gene expression and regulation (35). Other tissue specificities were also noted, which we identified in seven tissues: brain, breast, colorectal, uterus, lymphoid, bladder, and prostate. Consistent, we found four of these tissue-specific patterns replicating across two complementary approaches (MuSAS-A and -B). Further reinforcing the reliability of our findings. These tissue-specific origins align with tissue-specific earlier replication timing, and moreover co-localize with numerous tissue-specifically expressed genes in the matched tissues. In other words, increased origin density in a chromosomal domain in a certain tissue associates with earlier replication and higher gene activity in that tissue.

A possible mechanistic basis for this is suggested by analysis of histone modifications. In particular, these origins are enriched in tissue-specific promoters and more so enhancer histone marks, prominently the H3K4me1 and H3K27ac. Consistently, we note an accumulation of tissue-specific TF bindings, observed across different TFs suggesting that general activation of tissue-specific enhancers, rather than binding of one particular TF, associates with promoter usage. Prior studies have explored the general relationship between various chromatin marks and presence or absence of origins at the locus (7,8,14,15). Our results further reveal the tissue-specific nature of these associations. Specifically, we observed a higher enrichment of enhancer and promoter marks from one tissue in the origins corresponding to that same tissue, suggesting a causal role of enhancer/promoter activation in origin usage. A caveat in interpretation is that we are not able to distinguish the direction of causality; experiments measuring a time-course, or analyses considering developmental time would be able to suggest if enhancer activation precedes or follows origin activation.

Altogether, our analyses suggest a tight association interplay between tissue-associated expressed genes and the firing of co-localized replication origins, and switches to earlier replication timing in the corresponding genomic domain. These changes are likely to have consequences on mutation accumulation in two ways: firstly, at the origins themselves certain mutational processes can be hyperactivated (42); secondly, the switch of the whole domain to earlier replication time in a tissue predicts a lowering of mutation rate in that domain (43,44).

In summary, this work provides the first overview of tissue-specificity origin usage, a phenomenon observed across various cancer types. Moreover, our findings hint at a deep interplay between the origin usage with other tissue-specific processes, including gene expression and chromatin markings.

Methods

WGS somatin mutation collection and processing

We collected whole genome sequencing (WGS) somatic mutations from 7 different cohorts as described in (29). We gather 1950 WGS from the PCAWG study; 724 WGS from the TCGA study; 2932 WGS from the ICGC consortium (non-PCAWG, non-TCGA, non-TARGET); 4823 WGS from the Hartwig Medical Foundation (HMF) project; 1147 WGS from the CPTAC project; 552 WGS from the MMRF COMMPASS project; and 105 WGS from the DECIDER project.

We collected the samples' metadata (MSI status, purity, ploidy, smoking history, gender) from data portals and/or from the supplementary data of the corresponding publications. We harmonized the cancer type labels across cohorts.

Fitting mutational signatures to mutations

SNV calls with low mappability were filtered out according to the CRG 75 alignability track (45). SigprofilerMatrixGenerator was used to generate the SBS-96 count matrix for all the samples (46). Then, we used SigProfilerAssignment tool (47) to assign mutational signatures to all the mutations by using the COSMIC v3.3 mutational signatures as reference (48). Iteratively, we classified each mutation to the most likely causal mutational signature by sampling a multinomial distribution of the vector of probabilities of each mutational signature per sample and mutation type (Decomposed_MutationType_Probabilities.txt output file from SigProfilerAssignment).

Calculating mutational strand asymmetry score (MuSAS)

We developed a mutational strand asymmetry that can be calculated at any position. For each position, MuSAS is calculated by counting the number of mutations (stratified by mutation type) in both the flanking regions upstream and downstream of the position. For origins the flanking regions considered are 350kb windows on each side, while for TSS the flaking regions are 20kb windows on each side. For each mutation type, the MuSAS score is calculated as the sum of mutations from one alteration (for example, C>T) upstream, and its complementary (like G>A) downstream. This sum is then compared to the same alteration (C>T) downstream and its complementary (G>A) upstream. The formula is as follows:

$$MuSAS = log2 \frac{\#\ muts\ alteration1\ (upstream) + \#\ muts\ complementary\ (downstream)}{\#\ muts\ alteration1\ (downstream) + \#\ muts\ complementary\ (upstream)}$$

A MuSAS score of 0 means that there is not asymmetry and the furthest the score diverges from 0, both positively or negatively, the more the asymmetry. The direction of the score (positive or negative) is arbitrary as it will depend on which mutation type we consider as alteration 1 and which as complementary.

To align them, we calculated the MuSAS score for each tissue, mutational signature and mutation types separately. Based on the MuSAS score, we assign for each tissue, signature and mutation type which change is set as alteration1 and which change is set as complementary to

obtain a positive MuSAS score. Once we do this we can merge the mutations across different mutational signatures and mutation types for the subsequent analyses.

Selecting the signatures with mutational strand asymmetry

We calculated the MuSAS score for every combination of tissue, mutational signature and mutation type at (i) origins and (ii) at transcription start sites (TSSs). For the origins, we selected the origins that are represented in at least 50% of samples in the consensus from Jaksik et al. (7), remove the mutORI.Bulk and Nucleotide.comp.skew.Huvet and calculate the origin position as the median across the rest of the methods. For the origins the flanking window used is of +-350kb. For the TSSs, we downloaded them from the Table browser and used flanking windows of +-20kb. For the TSS, we orient the flanking windows in relation to the gene being in the positive or negative strand.

As a control, we shuffle the position across mutation and calculate the MuSAS score for each signature, tissue and alteration. We performed this 100 times for both origins and TSSs. The selected thresholds (99% of randomized data below) are 0.03 for origins and 0.08 for TSSs.

Determining the flanking windows size

From Jaksik *et al.* (7), we downloaded a comparison across origins detected with several experimental methods. We calculated the distance between origin locations across each method. Mean difference between origin position is 2468016 (Nucleotide.comp.skew.Huvet), 281967 (OK.seq.Petryk), 314691 (OK.seq.Wu), 288764 (Repli.seq.Hansen) and 539611 (mutORI.Bulk). The mean distance between the 5 methods is 356258.2. Therefore, we select 350kb and the window size.

Merging mutations to increase the mutation burden

<u>MuSAS-A: by signature</u>
We computed MuSAS by merging mutations from tumors with the same tissue type and mutational signature. This is feasible because in the previous section all the mutations are assigned the correct strand to have a positive MuSAS score when there is asymmetry. As a result, our units of analysis are pairs of tissue and mutational signatures. In total, we calculated 255 tissue-signature MuSAS profiles.

<u>MuSAS-B: by individual tumor</u>
We computed MuSAS by combining mutations from different mutational signatures but from the same individual tumor. The MuSAS peaks correspond to where there is higher asymmetry (origins of replication). In total, we derived MuSAS profiles for 386 individual tumors.

Determining origins of replication

To identify the origins of replication, we divided the genome into 1 kb windows and calculated the MuSAS score for each of them, after which we applied a smoothing filter. Then, we applied a peak caller on the MuSAS profile to locate the replication origins.

Overlapping of MuSAS with experimental methods

We collected 52 sets of origins of replication from distinct experimental methods. Names, source and description of the different sets in Table S1.

To calculate the genome-wide agreement between our MuSAS score and previously described origins, we compared the MuSAS score in windows that encompasses at least one origin against windows without any overlapping origin (+-10kb). We used the wilcoxon test to ascertain the significance of a greater MuSAS score in the origin overlapping windows versus the non-overlapping (alternative= greater).

Calculating variability on MuSAS profiles

To calculate the variability on the MuSAS profiles we applied a Principal component analysis (PCA) on the matrix of samples x 1-kb windows along the whole genome. We center and scale the MuSAS values per sample.

To differentiate robust PCs, we examined their coherence across distinct genomic segments. For this, we divided the genome into three distinct parts: chr1 to 5, chr6 to 11, and chr12 to 23, each spanning roughly 1000 Mbs. We then conducted a PCA on each segment. We performed a hierarchical clustering on the PC exposures to find the PCs that cluster.

Replication timing data

We predicted RT using the Replicon software (27) from human tumors (n = 410 samples, most of them with technical replicates) using ATAC-seq data of TCGA tumors downloaded from (26). We used the Replicon tool with the default settings.

Tissue-specific expressed genes

We downloaded the median gene-level TPM by tissue from GTEx Analysis V8 (dbGaP Accession phs000424.v8.p2) [https://www.gtexportal.org/home/datasets]. We selected the top 50 genes with the highest difference in expression for 1 tissue compared to the rest (removing from the rest those tissues that are similar, e.g. all brain categories when considering 1 brain subgroup).

DNA features data

Histone marks. We downloaded from ENCODE (https://www.encodeproject.org/ (36)) all data available for Homo sapiens in the genome assembly hg19 for H3F3A, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H2AFZ, H3K27ac, H3K36me3, H3K79me2,

H3K9me2 and H4K20me1 marks. Data is described in Table S2. For each of these features, we downloaded the narrow peaks file.

TFs chip-seq data. We downloaded all available TFs chip-seq data for GM12878, HCT116, K562, MCF-7 and SK-N-SH cell lines from ENCODE. We downloaded the files from GRCh38 assembly and IDR thresholded peaks. We performed a liftover to hg19.

HiC data. We downloaded chromatin domain hierarchies and compartment scores generated by the Calder method from Hi-C data from 114 cell lines (49).

DHS data. We downloaded the DHS position stratified by component from Meuleman et al. (35). We downloaded the file DHS_Index_and_Vocabulary_hg38_WM20190703.txt.gz from https://zenodo.org/record/3838751 and performed a liftover to hg19.

References

1. Méchali M. Eukaryotic DNA replication origins: many choices for appropriate answers. Nat Rev Mol Cell Biol. 2010 Oct;11(10):728–38.
2. Fragkos M, Ganier O, Coulombe P, Méchali M. DNA replication origin activation in space and time. Nat Rev Mol Cell Biol. 2015 Jun;16(6):360–74.
3. Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, Conway A, et al. Replication Dynamics of the Yeast Genome. Science. 2001 Oct 5;294(5540):115–21.
4. Méchali M. DNA replication origins: from sequence specificity to epigenetics. Nat Rev Genet. 2001 Aug;2(8):640–5.
5. Cayrou C, Coulombe P, Vigneron A, Stanojcic S, Ganier O, Peiffer I, et al. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res. 2011 Sep;21(9):1438–49.
6. Cayrou C, Coulombe P, Puy A, Rialle S, Kaplan N, Segal E, et al. New insights into replication origin characteristics in metazoans. Cell Cycle. 2012 Feb 15;11(4):658–67.
7. Jaksik R, Wheeler DA, Kimmel M. Detection and characterization of constitutive replication origins defined by DNA polymerase epsilon. BMC Biol. 2023 Feb 24;21(1):41.
8. Dao FY, Lv H, Fullwood MJ, Lin H. Accurate Identification of DNA Replication Origin by Fusing Epigenomics and Chromatin Interaction Information. Research [Internet]. 2022 Oct 30 [cited 2023 Sep 1];2022. Available from: https://spj.science.org/doi/10.34133/2022/9780293
9. Courtot L, Hoffmann JS, Bergoglio V. The Protective Role of Dormant Origins in Response to Replicative Stress. Int J Mol Sci. 2018 Nov 12;19(11):3569.
10. Rivera-Mulia JC, Buckley Q, Sasaki T, Zimmerman J, Didier RA, Nazor K, et al. Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. Genome Res. 2015 Aug 1;25(8):1091–103.

11. Poulet A, Li B, Dubos T, Rivera-Mulia JC, Gilbert DM, Qin ZS. RT States: systematic annotation of the human genome using cell type-specific replication timing programs. Bioinformatics. 2019 Jul 1;35(13):2167–76.
12. Koren A, Handsaker RE, Kamitaki N, Karlić R, Ghosh S, Polak P, et al. Genetic Variation in Human DNA Replication Timing. Cell. 2014 Nov 20;159(5):1015–26.
13. Ding Q, Edwards MM, Hulke ML, Bracci AN, Hu Y, Tong Y, et al. The Genetic Architecture of DNA Replication Timing in Human Pluripotent Stem Cells. bioRxiv. 2020 May 10;2020.05.08.085324.
14. Wu X, Kabalane H, Kahli M, Petryk N, Laperrousaz B, Jaszczyszyn Y, et al. Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-replicating regions. Nucleic Acids Res. 2018 Nov 2;46(19):10157–72.
15. Akerman I, Kasaai B, Bazarova A, Sang PB, Peiffer I, Artufel M, et al. A predictable conserved DNA base composition signature defines human core DNA replication origins. Nat Commun. 2020 Sep 21;11(1):4826.
16. Langley AR, Gräf S, Smith JC, Krude T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). Nucleic Acids Res. 2016 Dec 1;44(21):10230–47.
17. Shinbrot E, Henninger EE, Weinhold N, Covington KR, Göksenin AY, Schultz N, et al. Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. Genome Res. 2014 Nov 1;24(11):1740–50.
18. Reijns MAM, Kemp H, Ding J, de Procé SM, Jackson AP, Taylor MS. Lagging-strand replication shapes the mutational landscape of the genome. Nature. 2015 Feb 26;518(7540):502–6.
19. Otlu B, Díaz-Gay M, Vermes I, Bergstrom EN, Zhivagui M, Barnes M, et al. Topography of mutational signatures in human cancer. Cell Rep [Internet]. 2023 Aug 29 [cited 2023 Sep 5];42(8). Available from: https://www.cell.com/cell-reports/abstract/S2211-1247(23)00941-5
20. Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, Borja R de, et al. Comprehensive Analysis of Hypermutation in Human Cancer. Cell. 2017 Nov 16;171(5):1042-1056.e10.
21. Tomkova M, Tomek J, Kriaucionis S, Schuster-Böckler B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. 2018 Sep 10;19(1):129.
22. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell. 2016 Jan 28;164(3):538–49.
23. Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A, et al. Human gene organization driven by the coordination of replication and transcription. Genome Res. 2007 Sep 1;17(9):1278–85.
24. Petryk N, Kahli M, d'Aubenton-Carafa Y, Jaszczyszyn Y, Shen Y, Silvain M, et al. Replication

landscape of the human genome. Nat Commun. 2016 Jan 11;7(1):10208.
25. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci. 2010 Jan 5;107(1):139–44.
26. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018 Oct 26;362(6413):eaav1898.
27. Gindin Y, Meltzer PS, Bilke S. Replicon: a software to accurately predict DNA replication timing in metazoan cells. Front Genet [Internet]. 2014 [cited 2022 Oct 4];5. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2014.00378
28. Gnan S, Josephides JM, Wu X, Spagnuolo M, Saulebekova D, Bohec M, et al. Kronos scRT: a uniform framework for single-cell replication timing analysis. Nat Commun. 2022 Apr 28;13(1):2329.
29. Salvadores M, Supek F. Cell cycle alterations associate with a redistribution of mutation rates across chromosomal domains in human cancers [Internet]. bioRxiv; 2022 [cited 2023 Jul 5]. p. 2022.10.24.513586. Available from: https://www.biorxiv.org/content/10.1101/2022.10.24.513586v2
30. Liu X, Yu X, Zack DJ, Zhu H, Qian J. TiGER: A database for tissue-specific gene expression and regulation. BMC Bioinformatics. 2008 Jun 9;9(1):271.
31. Besnard E, Babled A, Lapasset L, Milhavet O, Parrinello H, Dantec C, et al. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat Struct Mol Biol. 2012 Aug;19(8):837–44.
32. Wilson RHC, Coverley D. Relationship between DNA replication and the nuclear matrix. Genes Cells Devoted Mol Cell Mech. 2013 Jan;18(1):17–31.
33. Sugimoto N, Maehara K, Yoshida K, Ohkawa Y, Fujita M. Genome-wide analysis of the spatiotemporal regulation of firing and dormant replication origins in human cells. Nucleic Acids Res. 2018 Jul 27;46(13):6683–96.
34. Courbet S, Gay S, Arnoult N, Wronka G, Anglana M, Brison O, et al. Replication fork movement sets chromatin loop size and origin choice in mammalian cells. Nature. 2008 Sep;455(7212):557–60.
35. Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature. 2020 Aug;584(7820):244–51.
36. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57–74.
37. Kang Y, Kim YW, Kang J, Kim A. Histone H3K4me1 and H3K27ac play roles in nucleosome eviction and eRNA transcription, respectively, at enhancers. FASEB J Off Publ Fed Am Soc Exp Biol. 2021 Aug;35(8):e21781.
38. Kang Y, Kang J, Kim A. Histone H3K4me1 strongly activates the DNase I hypersensitive sites in super-enhancers than those in typical enhancers. Biosci Rep. 2021 Jul

30;41(7):BSR20210691.
39. Hyun K, Jeon J, Park K, Kim J. Writing, erasing and reading histone lysine methylations. Exp Mol Med. 2017 Apr;49(4):e324–e324.
40. Long H, Zhang L, Lv M, Wen Z, Zhang W, Chen X, et al. H2A.Z facilitates licensing and activation of early replication origins. Nature. 2020 Jan;577(7791):576–81.
41. Park SM, Choi EY, Bae M, Kim S, Park JB, Yoo H, et al. Histone variant H3F3A promotes lung cancer cell migration through intronic regulation. Nat Commun. 2016 Oct 3;7:12914.
42. Murat P, Perez C, Crisp A, van Eijk P, Reed SH, Guilbaud G, et al. DNA replication initiation shapes the mutational landscape and expression of the human genome. Sci Adv. 2022 Nov 9;8(45):eadd3686.
43. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015 Feb;518(7539):360–4.
44. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015 May;521(7550):81–4.
45. Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, et al. Fast Computation and Applications of Genome Mappability. PLOS ONE. 2012 Jan 19;7(1):e30377.
46. Bergstrom EN, Huang MN, Mahto U, Barnes M, Stratton MR, Rozen SG, et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genomics. 2019 Aug 30;20(1):685.
47. Díaz-Gay M, Vangara R, Barnes M, Wang X, Islam SMA, Vermes I, et al. Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment [Internet]. bioRxiv; 2023 [cited 2023 Sep 7]. p. 2023.07.10.548264. Available from: https://www.biorxiv.org/content/10.1101/2023.07.10.548264v1
48. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019 Jan 8;47(D1):D941–7.
49. Liu Y, Nanni L, Sungalee S, Zufferey M, Tavernari D, Mina M, et al. Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. Nat Commun. 2021 May 10;12:2439.

Supplementary figures

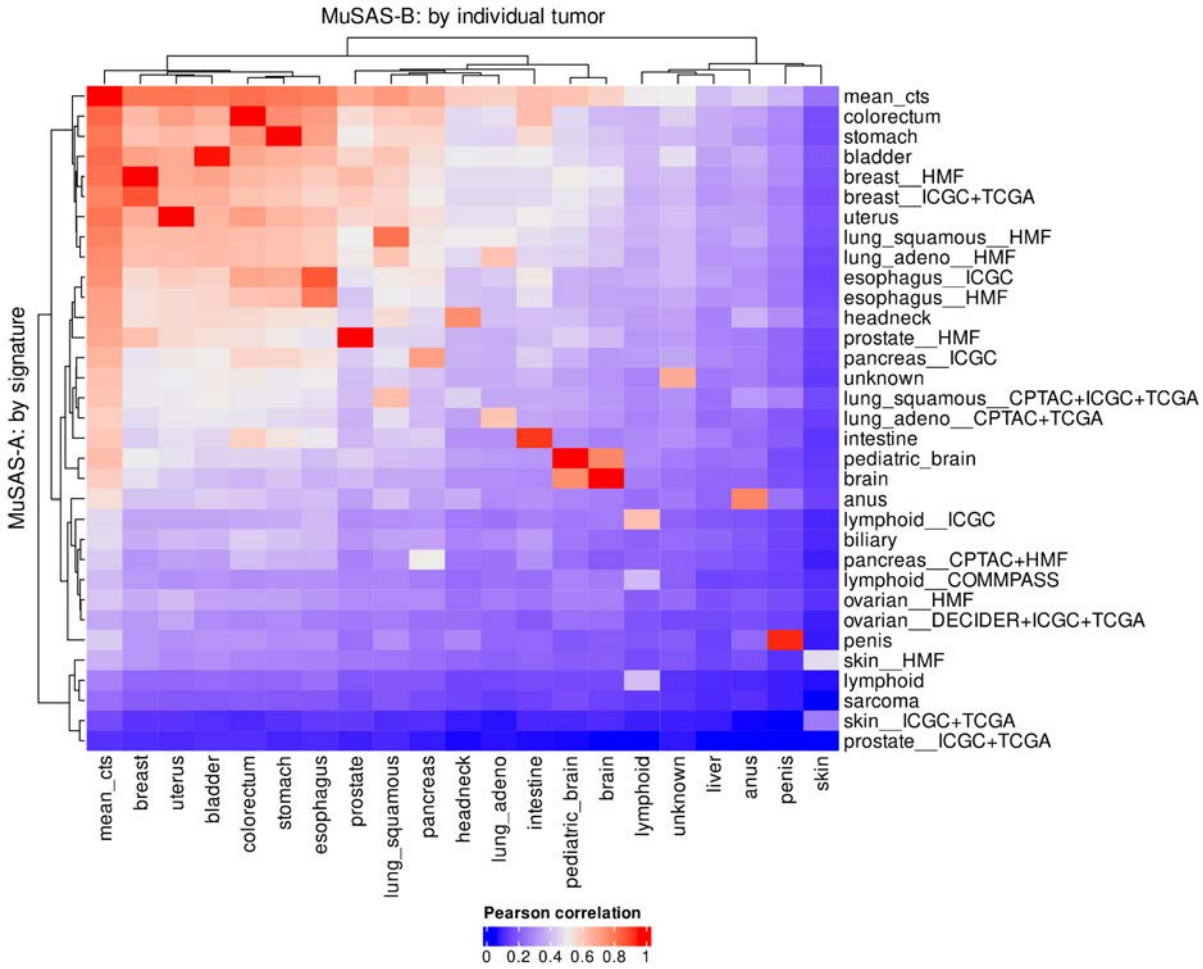


Fig S1. Comparison of MuSAS profiles between methods MuSAS-A and MuSAS-B stratified by cancer type. The strongest correlations are observed when the profiles correspond to the same tissue in both methods. This suggests that our methodology effectively captures tissue-specific variability.

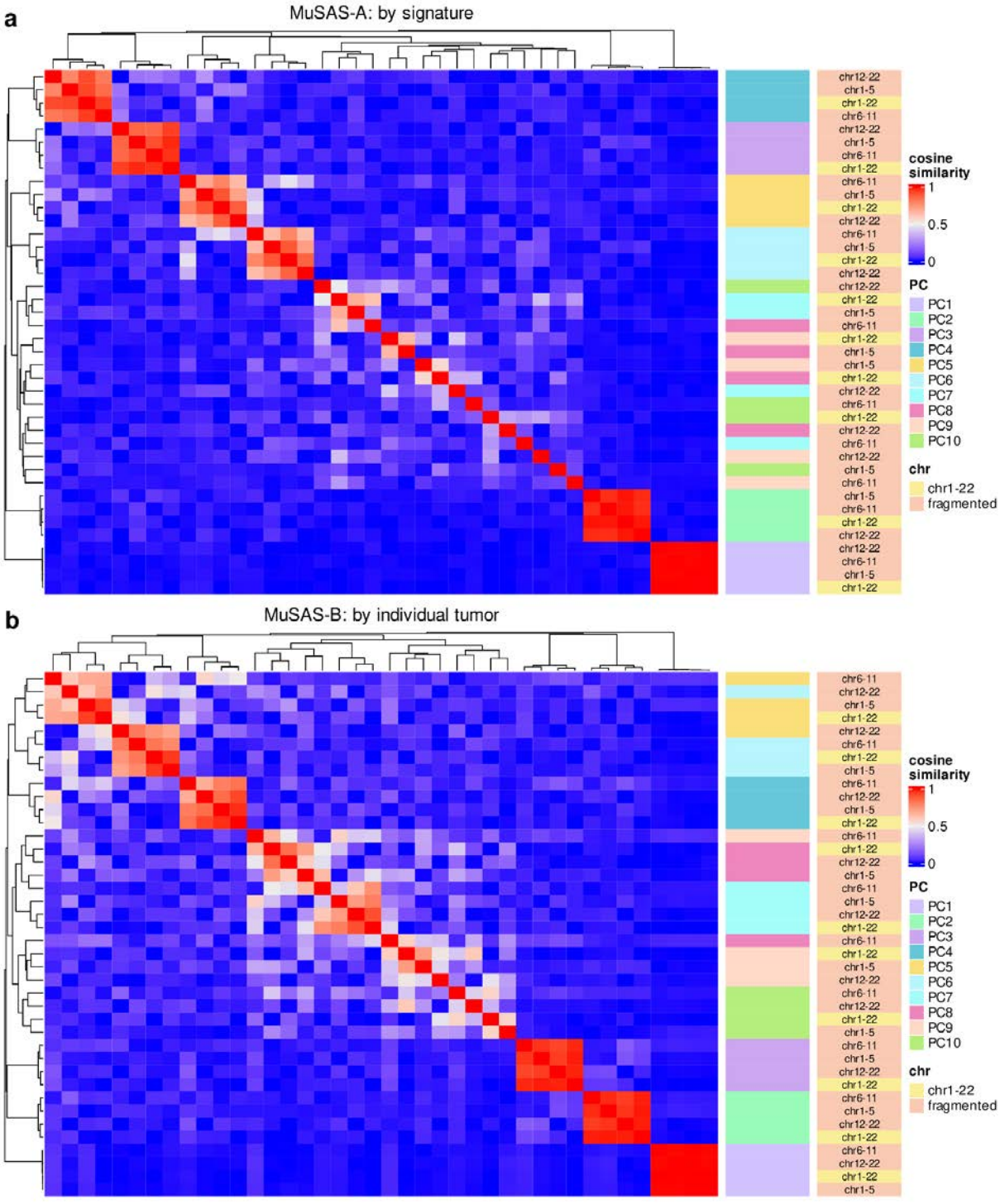


Fig S2. **Evaluation of the robustness of the extracted PCA signatures.** Hierarchical clustering of the cosine similarity between the first 10 PC signatures obtained when applying a PCA into the MuSAS profiles from method MuSAS-A **(a)** and MuSAS-B **(b)**. This PCA was repeated splitting the genome into 3 segments of approximately 10 Mb (chr1 to chr5, chr6 to

chr11 and chr12 to chr22). If the signal is consistent across these genome segments, their respective PCs will cluster together. This approach enables us to determine the number of reliable PCs.

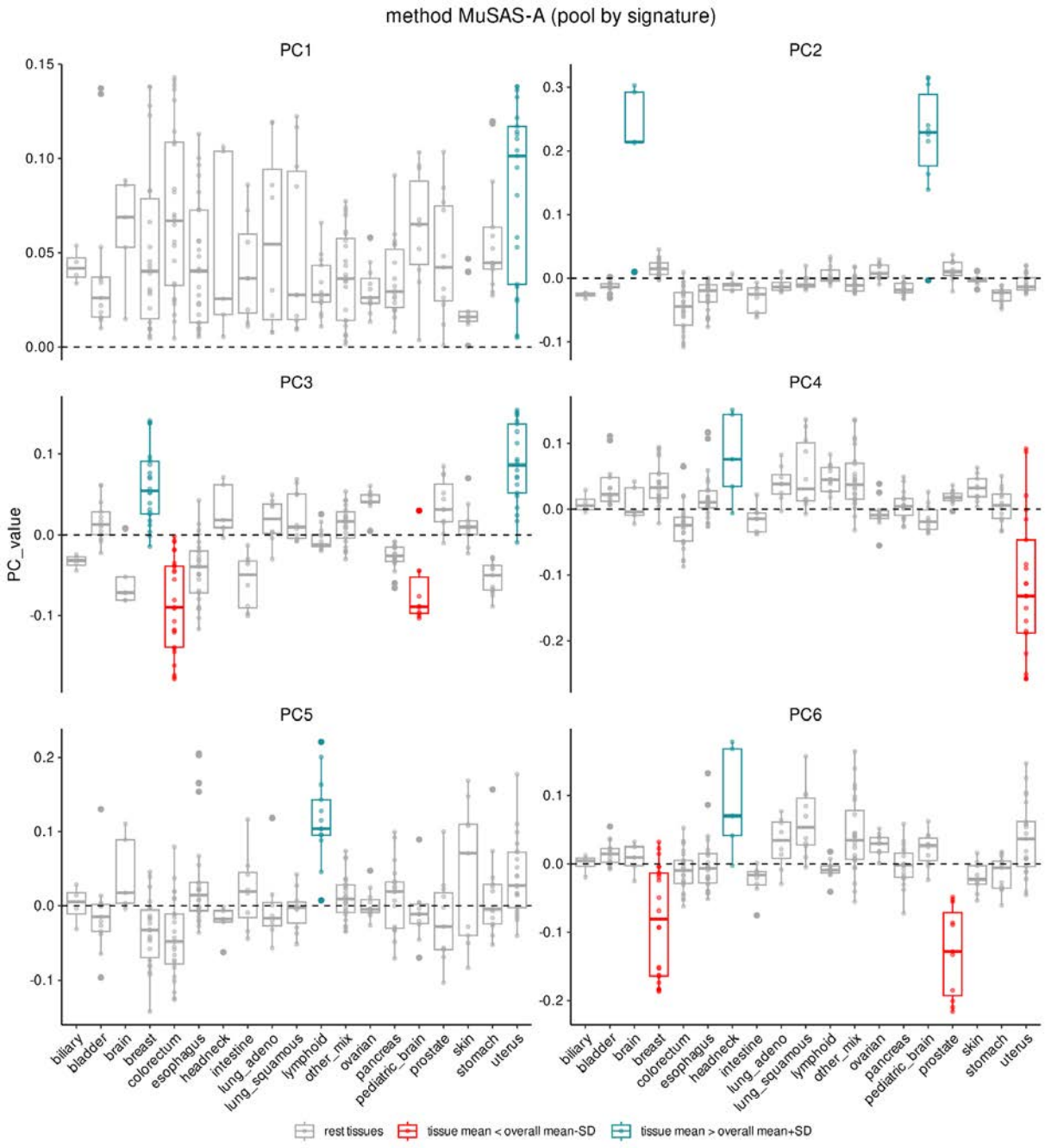


Fig S3. Identification of the tissue-specificity for the MuSAS-A PC origin's signatures. PC activity for the first 6 PCs on the MuSAS-A profiles stratified by cancer type. We marked the tissues that stick out on the positive side (blue) and on the negative side (red). The color code corresponds to the heatmap in Fig 2f.

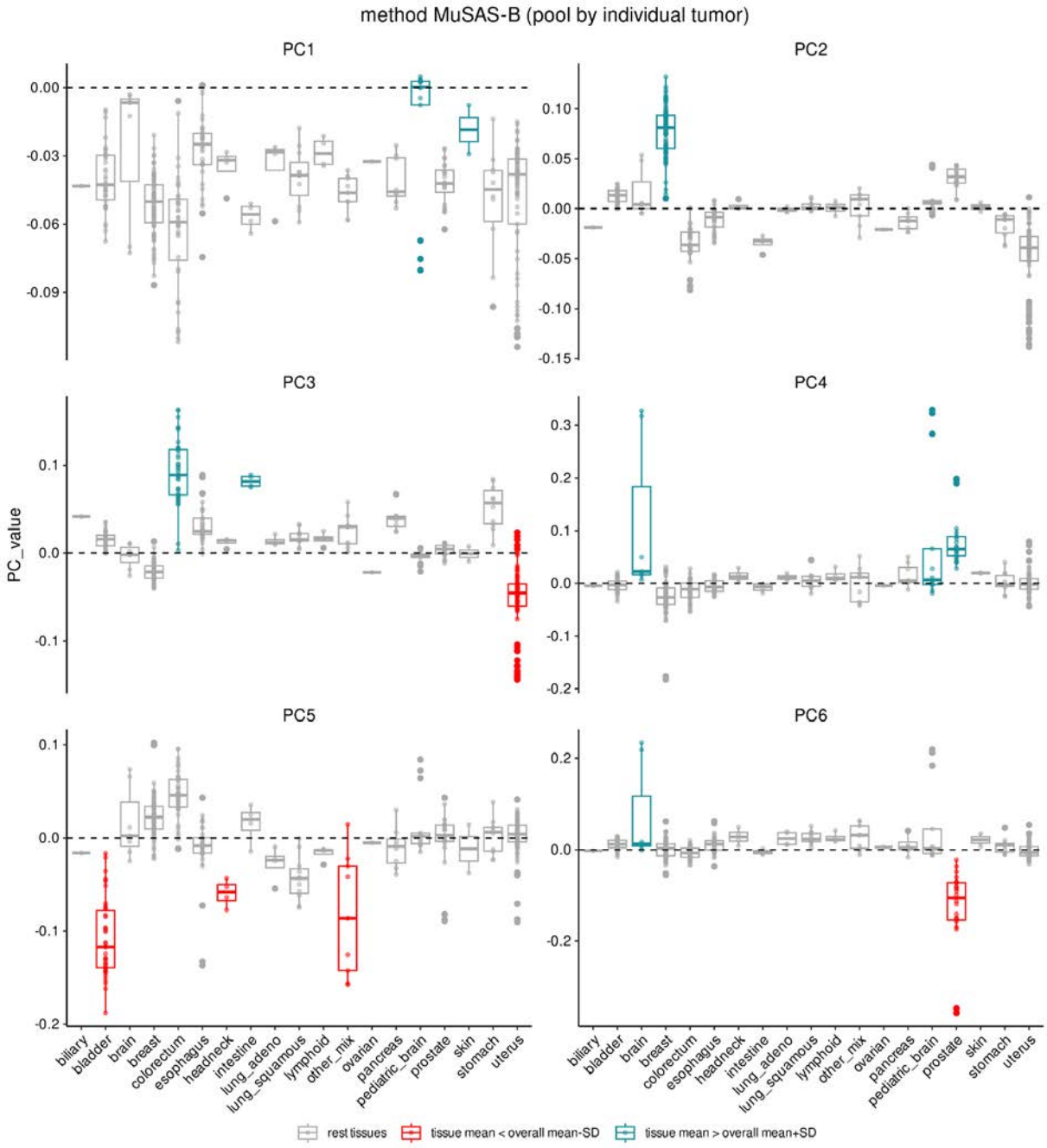


Fig S4. Identification of the tissue-specificity for the MuSAS-B PC origin's signatures. PC activity for the first 6 PCs on the MuSAS-B profiles stratified by cancer type. We marked the tissues that stick out on the positive side (blue) and on the negative side (red). The color code corresponds to the heatmap in Fig 2f.

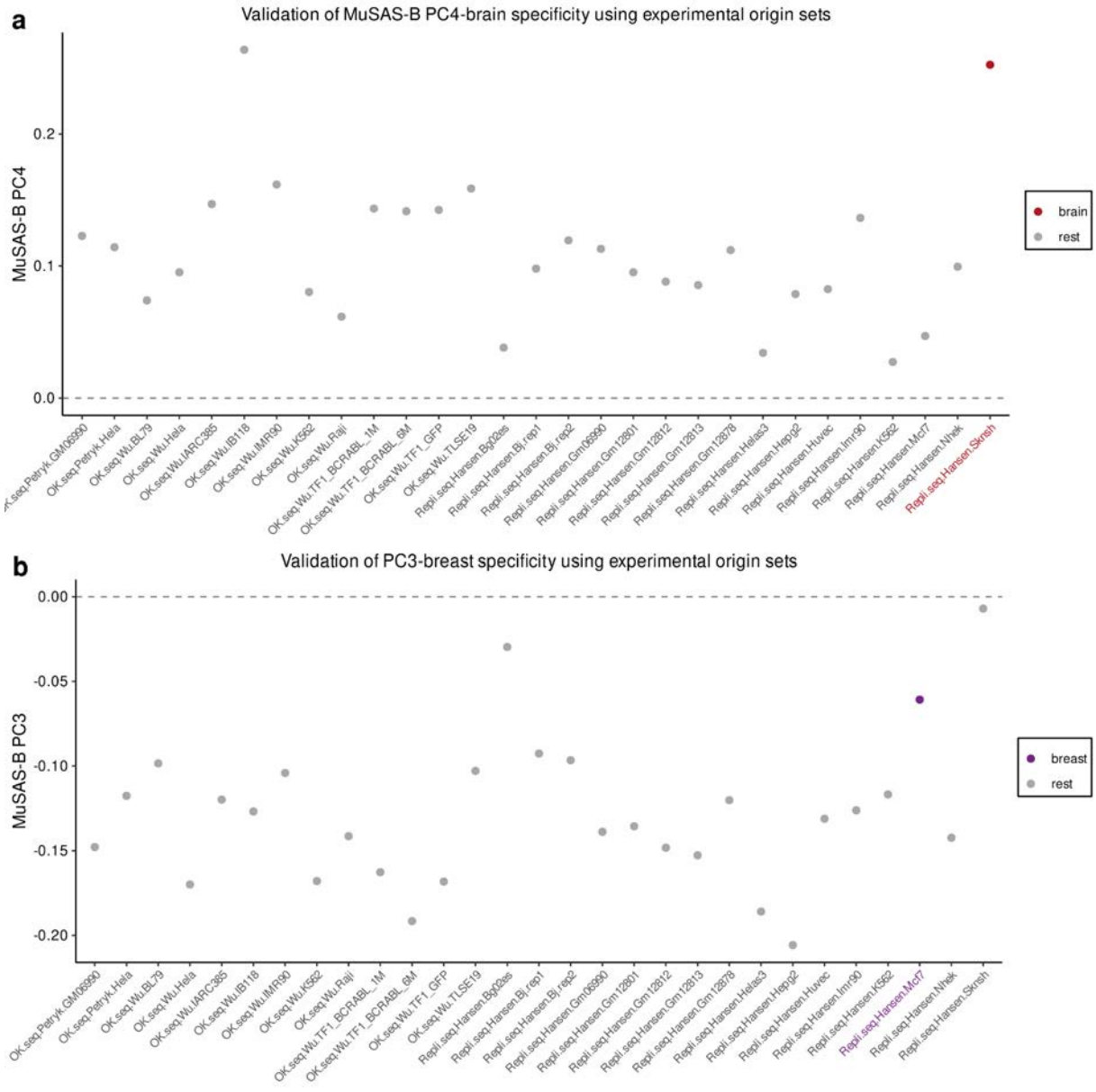


Fig S5. Validation of tissue specific origins using experimental data. a-b) Origin usage PC4-brain (in **a**) and PC3-breast (in **b**) values for the windows that contain at least one origin identified by experimental assays.

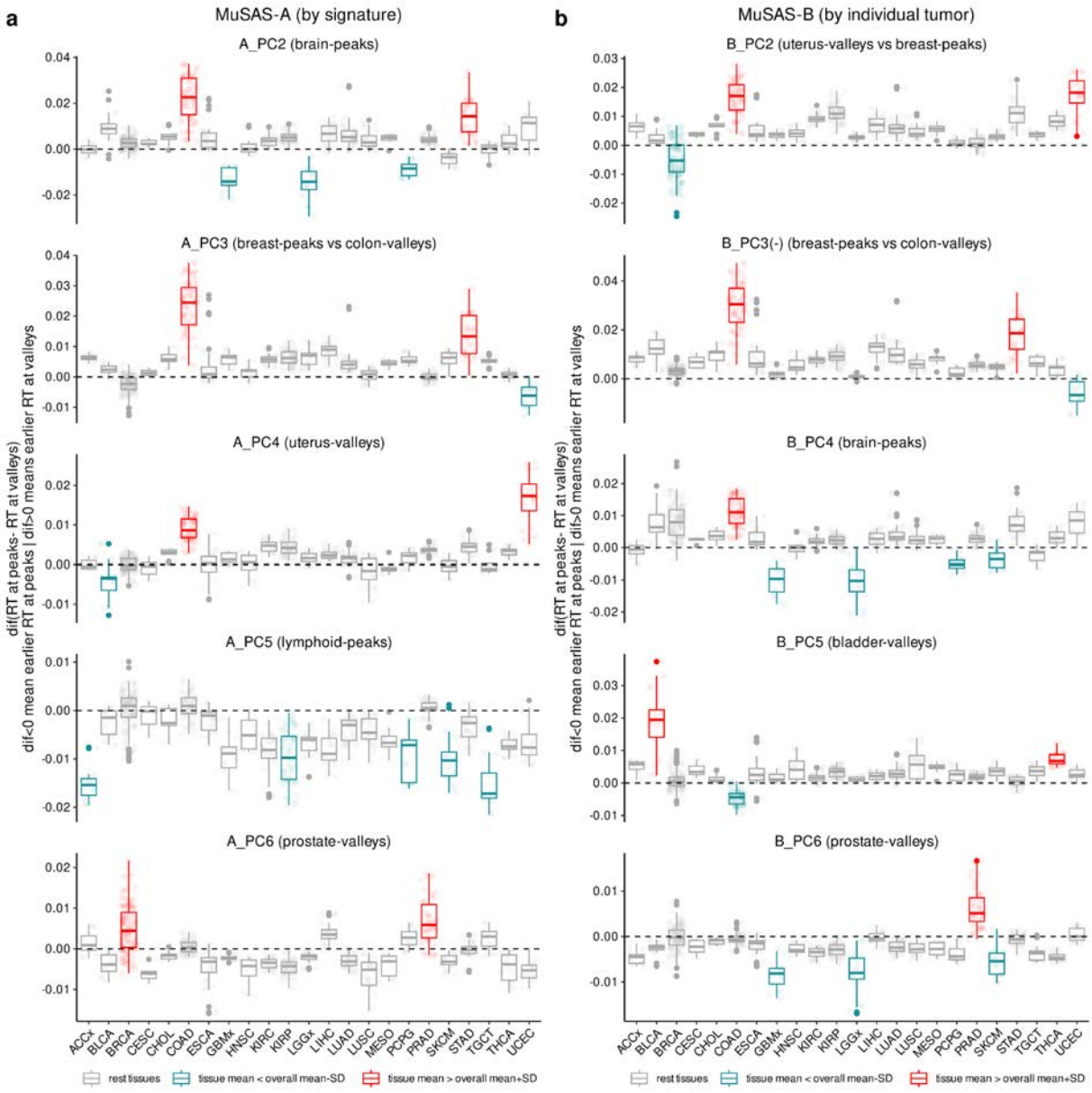


Fig S6. Validation of the tissue-specific origin usage using RT profiles from different tumor types. Difference between the mean RT at peaks and the mean RT at valleys of the MuSAS-A **(a)** and MuSAS-B **(b)** PCs for each TCGA tumor grouped by cancer type. The color code corresponds to the heatmap in Fig 2f, where red means that the origin is in the valleys and blue that is in the peaks. The difference between RT at peaks and RT at valleys when it is <0 means that that cancer type replicates earlier in peaks and when it is >0 means that that cancer type replicates earlier in valleys. Therefore, the tissue-specificity of one PC validates when the cancer type that is in a peak has a dif<0 and a cancer type that is in a valley has a dif>0.

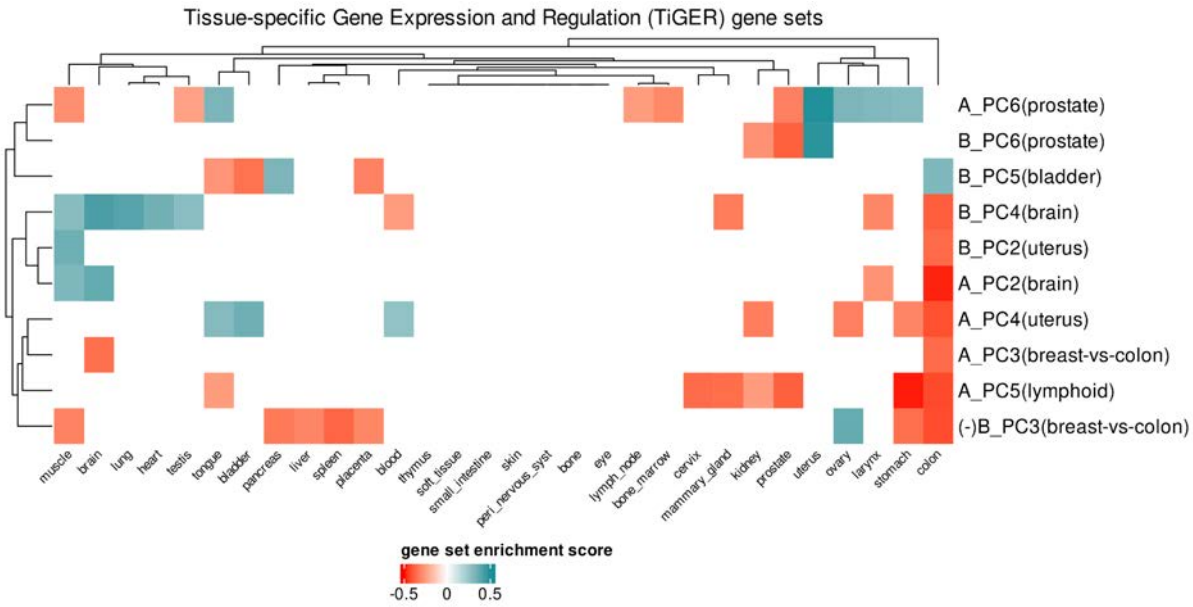


Fig S7. Tissue-specific origins of replication co-localize with the tissue-specifically expressed genes of the matched tissues. Tissue-specific gene set enrichment scores using the TiGER database (30) gene sets (x axis) for the tissue-specific origin PCs (y axis) by method MuSAS-A and MuSAS-B. The gene sets with an FDR > 25% are coloured in white. The color code corresponds to the heatmap in Fig 2f, where red means that the origin is in the valleys and blue that is in the peaks.

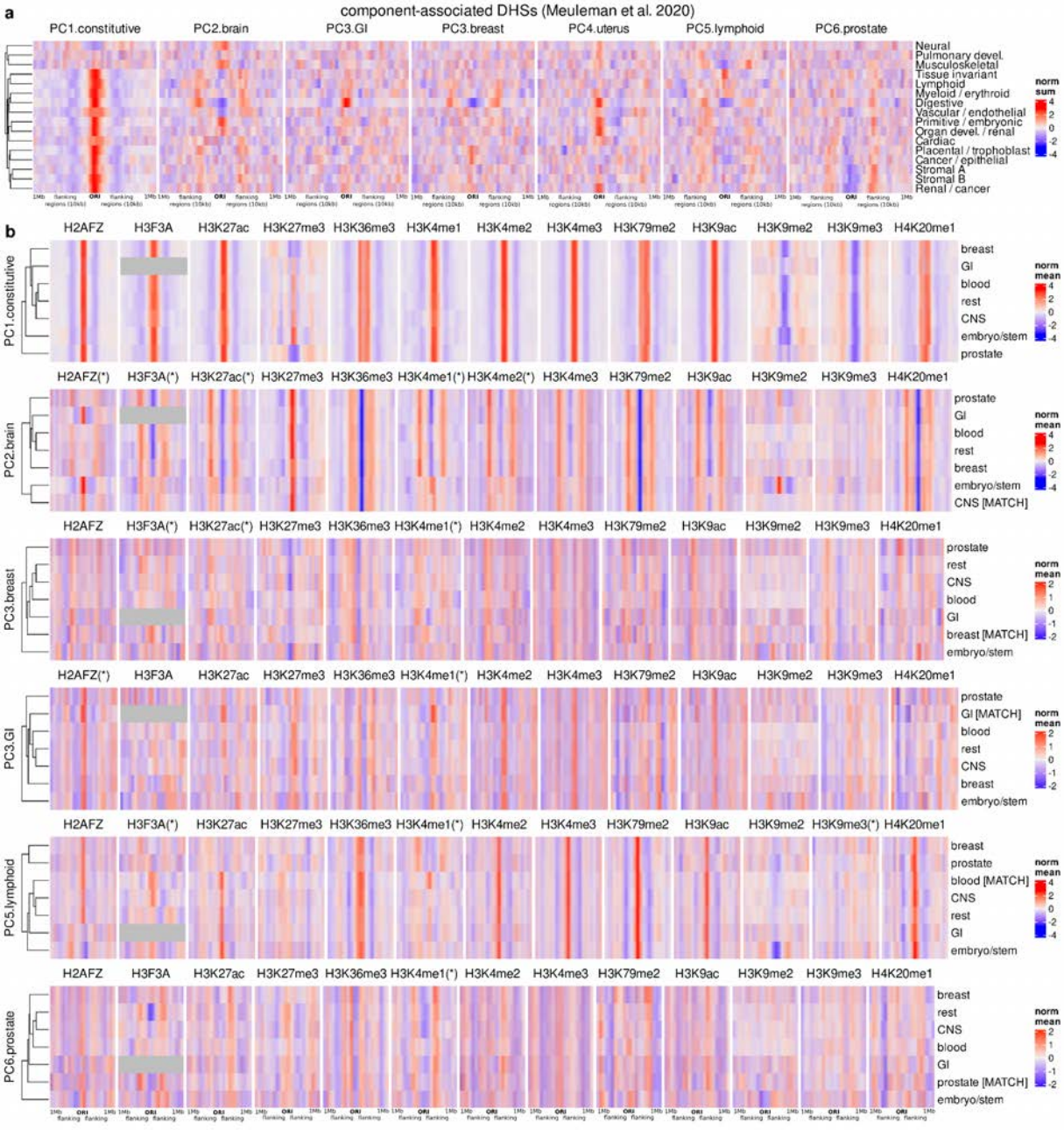


Fig S8. Chromatin accessibility and histone marks density at origins of replication. a) Density of chromatin accessibility (DHS) across 10-kb windows centered at tissue-specific origins of replication, stratified by tissue/cell-type components. **b)** Density of 13 histone marks across 10-kb windows centered at tissue-specific origins of replication, stratified by tissue groups.

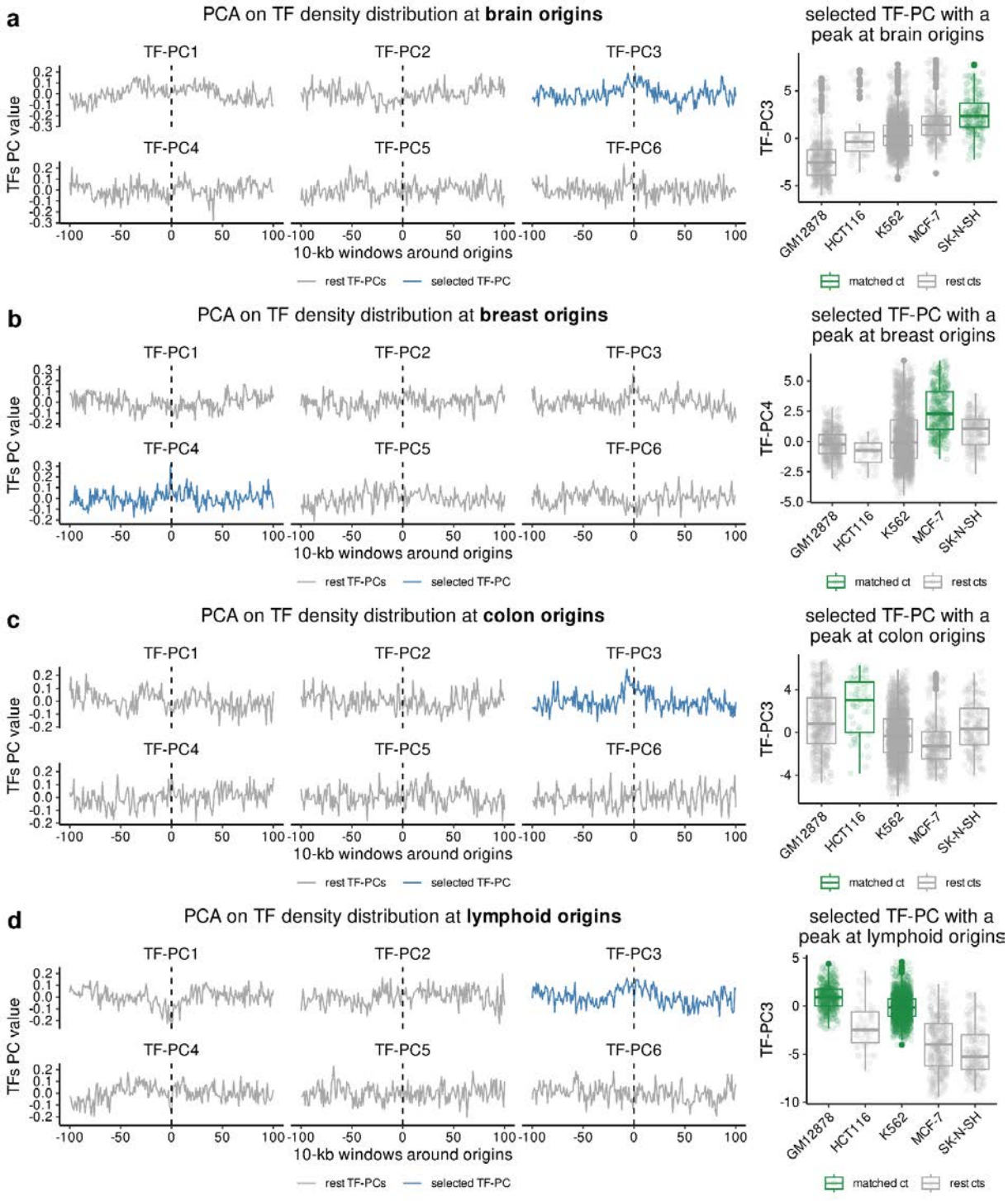


Fig S9. Distribution of TF density around tissue-specific origins. **(left panels)** PCA profiles of TF density around tissue specific origins from brain (in **a**), breast (in **b**), colon (in **c**) and lymphoid (in **d**). **(right panels)** Activities of the selected TF-PCs that exhibit a peak at the origin's position in left panels. The higher the activity the more enrichment in TF density at

origins in a specific cell line. Cell lines that have a cancer type corresponding to the origin tissues are highlighted in green.

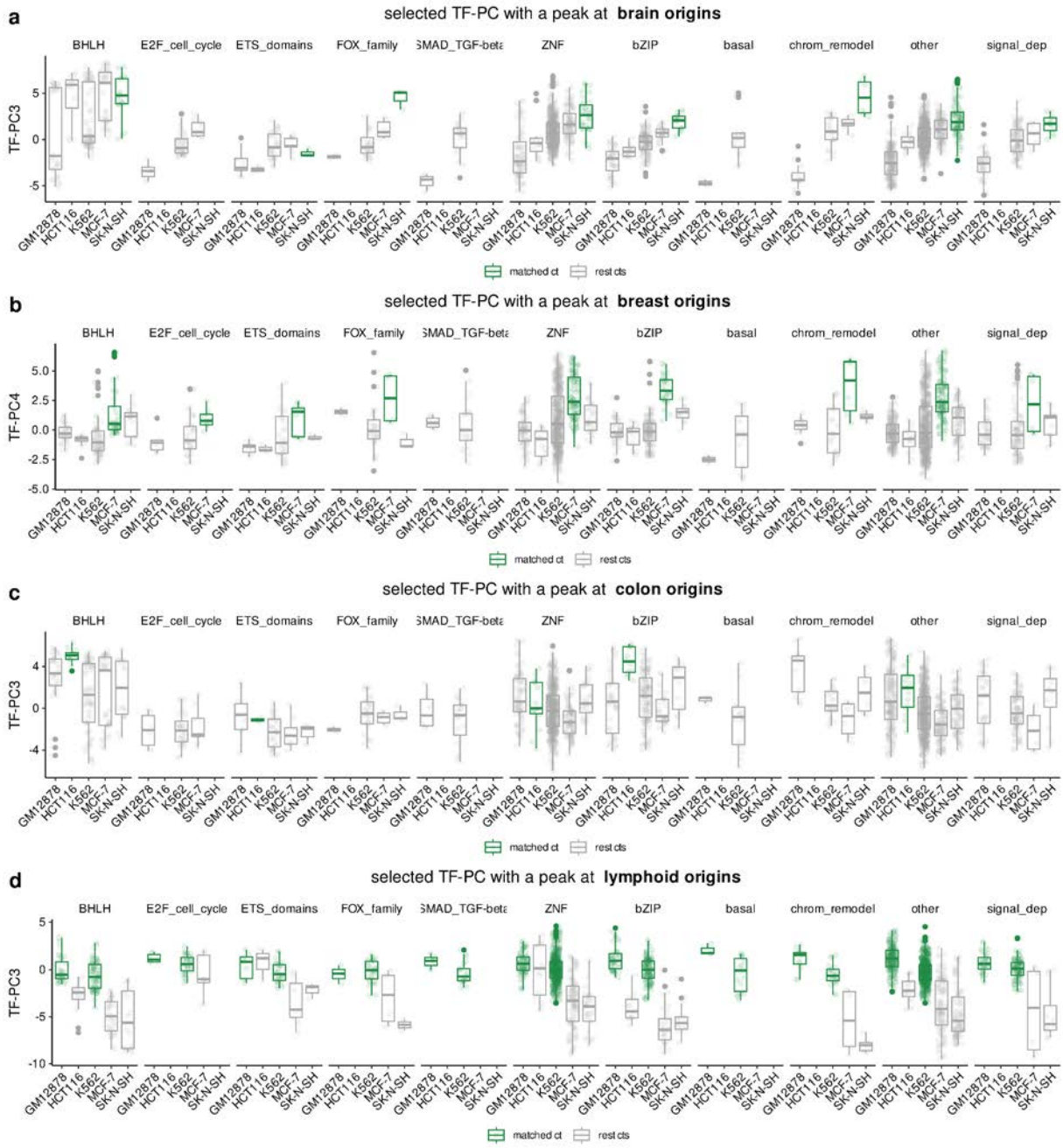


Fig S10. Activities of the selected TF-PCs that exhibit a peak at the origin's position in Fig S9 for brain (a), breast (b), colon (c) and lymphoid (d) specific origins. The transcription factors are stratified by groups. The higher the activity the more enrichment in TF density at origins in a specific cell line and TF group. Cell lines that have a cancer type corresponding to the origin tissues are highlighted in green.

Chapter 7

Summary of results

Somatic mutations are distributed highly unevenly across the genome. This variability in the regional mutation density strongly correlates with the RT and chromatin landscapes in a cell type-specific manner. However, little is known about inter-individual variability shaping the mutational profile. In this thesis, we explored and quantified the variability in regional mutation density (RMD) at the domain scale between tissues and individuals. Additionally, we applied the RMD tissue-specificity to predict the tissue of origin for two different scenarios. Finally, we used the regional mutational strand asymmetry to detect origins of replication and study the variability in origin usage across tissues.

In **Chapter 3**, we performed a systematic, genome-wide study of the inter-individual variation in mutation rate distribution across chromosomal domains (here measured as 1 Mb windows) in humans. Our study was performed on >4,000 whole-genome sequences of various cancer types. We used a newly-developed statistical methodology based on NMF for extracting "regional mutation density (RMD) signatures" (Fig 3). These RMD signatures are robust patterns that describe how mutation risk varies in a coordinated manner across large chromosomal segments in groups of tumors.

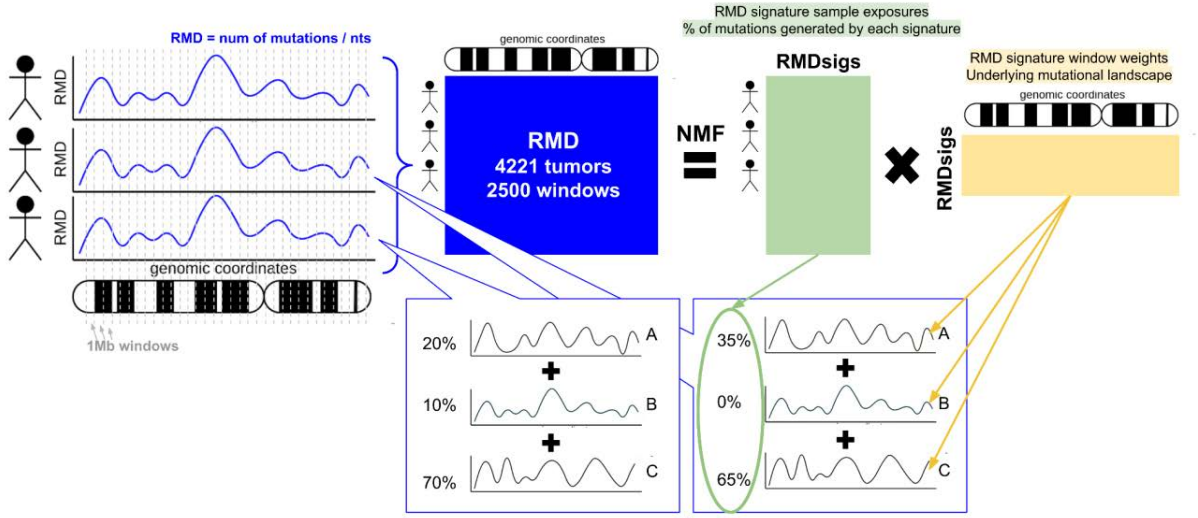


Fig 3. **Schematic representation of the method to extract RMD signatures.** We applied the NMF algorithm to a matrix of RMD patterns for 4221 tumor WGS. The output of the NMF are two matrices: (i) the RMD signature "exposures" per tumor sample i.e. the % of mutations generated by each RMD signature in that tumor; and (ii) the RMD signature window weights, which are the coefficient of increase or decrease in mutation rates upon the signature activity in each specific window.

In total, we identified 13 RMD robust signatures that describe a global genome-wide mutation redistribution in the 4000 tumors. Out of the 13, we identified 10 that are tissue-specific and 3 that are tissue-independant ("global") variability in RMD between individual tumors. One of the three global signatures captures the known flatter distribution associated with MSI tumors (RMDflat),

thus serving as a validation of the method. The other two global signatures are to our knowledge novel and we named them RMDglobal1 and RMDglobal2.

The spectrum of tissue-related RMD signatures ranges from highly specific patterns unique to individual tissues (e.g RMD signature from skin tumors), to signatures that encompass multiple tissues with similar molecular attributes (e.g. there is a shared RMD signature from several gastrointestinal tumor types). This suggests that those tumors shared some biology in their cell-of-origin with its characteristic RT/chromatin landscape.

The RMDflat signature recovered the global mutation pattern where the heterogeneity in mutation rates between domains gets overall reduced, resulting in a more 'flat' appearing landscape. Previously reported causes of this are the loss of MMR activity or also NER activity (48,60), thus removing the preferential DNA repair bias towards early-replicating domains. APOBEC mutagenesis also was previously reported to increase mutations in early replicating gene-rich regions (8). In this work, we found an additional novel mechanism that can underlie this global 'flat' pattern of mutations, which is the deficiency in the homologous recombination (HR) repair pathway. Our study indicates that these three diverse DNA repair disturbances (MMR, NER, HR) converge onto a common global pattern of regional mutagenesis. We evaluated the consequences of this global redistribution, and found that the RMDflat mechanism causes mutation risk to be altered for ~3/4 of all known cancer genes.

The RMDglobal1 signature is a widespread redistribution of mutation rates that occurs mainly the facultative heterochromatin domains (Polycomb mark-enriched; centrally located in the nucleus i.e. not lamina-associated as is typical for heterochromatin). The RMDglobal1 signature strongly correlated with chromatin accessibility, heterochromatin marks and RT remodeling signatures we identified across many human cell types (cancerous or healthy) in the ENCODE resource. These epigenome remodeling signatures correlate with each other, suggesting that they are a mirror-image of the same process, plausibly whole chromosomal domains switching to active or inactive state. We found that these global, multi-domain coordinated RT and heterochromatin shifts are associated with differential expression of cell cycle and E2F target genes, suggesting more proliferative cells. Further, the RMDglobal1 mutation risk redistribution signature was robustly associated with the deletions of the *RB1* tumor suppressor gene, converging onto cell cycle disturbances as the likely cause.

Together, our analyses suggest a mechanistic model where due to more rapid cycling in some tumors (which is commonly caused by oncogenic events such as RB1 loss), heterochromatin and RT program are remodeled in specific nuclear compartments (enriched in Polycomb regions). As a consequence, this changes the local mutation risk across chromosomal domains and so alters the mutational opportunity for acquiring pathogenic changes in particular genes. For instance, RMDglobal1 signature increases the risk of acquiring a downstream mutation in the loci harboring *ATM*, *BAP1* or *KMT2C* driver genes.

The RMDglobal2 signature occurs independently of the above and alters the relative mutation risk between the late-replicating, constitutive heterochromatin, and the rest of the genome. This

global pattern is strongly associated with the mutations in the *TP53* gene and, consistently with the amplifications of oncogenes that phenocopy *TP53* loss (*MDM2* and *MDM4*). Via this RMDglobal2 mechanism of redistribution of mutation rates, the *TP53* loss-of-function may be able to commonly 'redirect' downstream tumor evolution, by accelerating the local mutation supply towards some driver genes and away from other driver genes. For example, the region containing the common tumor suppressor gene *ARID1A* is affected by this mutation redistribution mechanism, resulting in a mutual exclusivity between *ARID1A* and *TP53* causal mutations.

In summary, in this Chapter 3 we developed a statistical methodology to extract RMD signatures of domain-scale mutation risk. Our analyses showed that there are at least 3 RMD signatures universally present across many human cell types, resulting from distinct mechanisms associated with alterations in cell cycle control genes. The redistribution of mutations caused by these RMD signatures have a direct impact in the mutation supply of various cancer driver genes, possibly affecting the subsequent activation or inactivation of relevant cancer genes during the cancer evolution.

In **Chapter 4,** we used the tissue-specific property of the regional mutation density pattern to predict the cancer type of origin in human tumors. An accurate classification of the cancer type is important for approximately 3% of metastatic cancers, called cancers of unknown primary (CUP), for which the standard diagnostic procedure cannot identify the site of origin (64), hampering therapy.

First, we created machine learning classifiers trained on RMD features and evaluated if this pattern alone is able to distinguish the different cancer types and subtypes. Here, the RMD profile is the normalized mutation counts across 2655 megabase-sized chromosomal domains. In the task of predicting 18 tumor types, the RMD models obtained very high accuracies (median AUC=0.99, testing in crossvalidation). Furthermore, the RMD was also able to discriminate between molecular subtypes and/or anatomical sites of six major cancers (breast cancer, liver cancer, stomach cancer, colorectal cancer, head-and-neck squamous cell carcinoma and melanoma). We believe that the proposed classifiers based on genomic features are a promising approach to prioritize the tumor type and subtype for CUP, and upon adaptation, can be used for liquid biopsy data if WGS is applied. Our analyses however suggest caution in the latter case, because very low-coverage sequencing may not yield enough power to discriminate tissues; similarly WES mutation patterns are not promising for tissue classification.

Second, we examined the predictive ability of causal (driver) somatic mutations in this task, comparing it against global patterns of non-selected (passenger) mutations, including features based on regional mutation density and trinucleotide mutation frequencies. In the task of distinguishing 18 cancer types, the driver mutations – mutated oncogenes or tumor suppressors, pathways and hotspots – classified only 36% of the patients to the correct cancer type. In contrast, the features based on passenger mutations did so at 92% accuracy, with similar contributions from the RMD features and the trinucleotide mutation spectra. The RMD and the spectra covered distinct sets of patients with predictions. In particular, introducing the RMD features into a

combined classification model increased the fraction of diagnosed patients by 50 percentage points (at 20% FDR).

In sum, we propose that WGS is valuable for classifying tumors because it captures global patterns emanating from mutational processes, which are informative of the underlying tumor biology. In addition, despite direct applications of these classifiers, this work showed that the mutational profile of passenger mutations across chromosomes varies consistently across tissues.

In **Chapter 5**, we extended the cancer type classifiers to predict the tissue-of-origin in human cancer cell lines. Cell lines are commonly used models of cancer biology because they are cost effective, convenient and amenable to high-throughput screening (116). However, not all the cell lines are equally suitable as tumor models, and those that better resemble the biological and molecular characteristics of the original cancer should be prioritized. In particular, the tissue (or cell type) of origin of a tumor is key because it provides important context for understanding the mechanisms of cancer and is one of the main determinants for selecting the drug treatment for a patient (117). Therefore, having accurate information on the tissue of origin of a cancer cell line is important for interpreting the experimental results obtained using these cell lines.

The misidentification of the cancer type of a cell line may happen for different reasons: that the tumor thought to originate in a certain tissue may in fact be an unrecognized metastatic lesion from a distal site; that the tumor presents ambiguous histological/anatomical features which may cause the cancer type to be misdiagnosed; or that the process of establishing the cell culture might select a rare cell type which is not representative of the tumor. In addition, the changes harbored due to the establishment of the cell culture might cause a cell line to diverge completely from the characteristics of the originating tissue.

In this chapter, we used transcriptomic and epigenomic profiles of human tumors and cell lines to systematically identify (i) cell lines that might be mislabeled with respect to cancer type or that might have diverged from their original tissue identity and (ii) cell lines whose molecular profiles are highly consistent with human tumors of their declared cancer type of origin. For this, we aligned mRNA expression and DNA methylation data between ~9,000 solid tumors and ~600 cell lines to enable joint analysis. Next, we created classification models for predicting cancer type and subtype using tumor data and applied them to cell line data.

Of 614 cell lines, we cataloged 60% as “golden set” cell lines because both transcriptomic and epigenomic profiles strongly resembled the known tumor type of origin. We suggest these cell lines are better suited for experimental work and we designated tentative cancer subtypes to them. In contrast, we detected 35 cell lines (6%) which better matched a different cancer type than the one the cell line was originally annotated with. We recommend caution in using these cell lines in experimental work. Furthermore, we applied genomic classifiers that corroborated as mislabelled 22 of these suspect cases, drawing upon mutational signatures and copy number alteration profiles. For instance, 5 cell lines detected as suspected of being melanoma were corroborated by the presence of UV radiation-like mutational signatures in them.

Finally, we corrected the cancer type labels (i.e. removing misclassified-suspected cell lines or using only the cell lines from the "golden set" list) and re-calculated associations between various genetic biomarkers (e.g. mutated oncogenes) and drug sensitivity / CRISPR genetic screening data. We showed that including misclassified cell lines may lead to false associations. Thus focussing on the cell lines that are best matched to the cancer type of interest allows to more precisely identify active drugs and/or specific markers for stratification.

In **Chapter 6**, we developed a method to identify origins of replication by leveraging the mutational strand asymmetry of various mutational signatures present at these origins. For this, we first calculated a Mutational Strand Asymmetry Score (MuSAS) at the set of known origins of replication to identify the mutational signatures useful for further analyses. Next, for these various signatures, including the known strand-biased signatures of MMR deficiencies and APOBEC mutagenesis and others, we calculated the MuSAS score across 1 kb windows along the genome, aiming to identify potential origins of replication using two complementary approaches. In the first, MuSAS-A, we calculated the MuSAS profiles by merging all the mutations from the same mutational signature from different tumors, while stratifying by cancer type. In this context, the "samples" for analysis are the combinations of tissue and mutational signature. In the second approach, MuSAS-B, we merged all mutations from one individual tumor irrespective of their mutational signatures. Here, the "samples" are individual tumor genomes, enabling us to study inter-individual variation in origin usage, in the tumors with sufficient number of mutations to afford statistical power.

In total, we obtained MuSAS profiles (scores at every 1 kb genomic window) for 255 tissue-signature pairs using MuSAS-A, and 386 individual tumors using MuSAS-B. We applied a Principal Component Analysis (PCA) in the MuSAS-A and MuSAS-B datasets separately to extract the main trends in variability in origin usage across tissues and individuals. Interestingly, we found four trends of tissue-specific origin usage in both MuSAS-A and -B, associated with: (i) brain; (ii) breast *versus* colon contrast; (ii) uterus; and (iv) prostate. We additionally found a lymphoid-specific origin usage trend in MuSAS-A, and a bladder-specific origin usage trend in MuSAS-B method.

To corroborate the tissue specificity of these origins, we compared replication timing (RT) at origin positions against non-origin loci, across various cancer types from TCGA (using RT inferred from ATAC-seq patterns as per the Replicon method). Notably, in most of the sets of tissue-specific origins, the matched cancer type exhibited earlier replication timing at origins than non-matched cancer types (e.g., earlier RT in LGG and GBM at the loci of brain-specific origins). Moreover, at the tissue-specific origins there was a co-localization of genes specifically expressed in the corresponding tissue. This implies a strong link between tissue-specific gene or regulatory element activation and origin usage.

Prior studies showed that specific markers on chromatin (histone marks and other epigenomic features) are associated with origins of replication. We, therefore, explored whether tissue-

specific origins possessed unique markers differentiating them from general origins. For this, we analyzed the density of the marks at these origins compared to their flanking areas, separating each mark by tissue. Using ENCODE data, we found that origin regions are enriched in chromatin accessibility (DHS), enhancer (H3K4me1, H3K27ac) and promoter (H3K4me3, H3K9ac) marks in a tissue-specific manner. For instance, brain-specific origins predominantly featured promoter and enhancer markers related to brain or neural tissues. Additionally, we found tissue-specific enrichment in H2A.Z and H3.3A histone marks, consistent with their known roles in influencing replication origin licensing and activation (118,119). In summary, we discovered a greater abundance of enhancer and promoter marks in origins from the same tissue, implying that the activation of enhancers/promoters may play a pivotal role in origin utilization.

In conclusion, our study underscores a strong link between tissue-specific replication origins and tissue-specific gene expression patterns, an earlier replication timing at the origin locations, and a tissue-specific enrichment of promoter and enhancer histone marks at origins. Overall, this research offers convergent evidence about tissue-specific origin use observed across multiple somatic cell types and suggests a mechanistic connection between origin usage and enhancer or promoter activation in different cell types.

Collectively, our studies focused on understanding the heterogeneity of mutation risk along the human genome at the domain-scale, and how this varies across tissues and individuals. Within this thesis, we identified two novel mutation risk redistribution patterns that affect most cancer types and we provided mechanistic insight into their etiologies, linked with alterations in cell cycle genes. Additionally, we used the properties of the somatic mutations to address three distinct applications. First, we demonstrated the viability of the RMD as an informative feature for predicting the tissue of origin in cancers of unknown primary. Moreover, we established a scoring system based on transcriptomes and epigenomic to assess the appropriateness of a cell line as a model for a specific tumor type, applying to a panel of 600 cell lines, supporting this with a mutation pattern classifier. Lastly, we analyzed regional mutation strand biases in RMD to identify the loci harboring origins of replication, uncovering their tissue-specific variation, associated with tissue-specific activity of promoters and enhancers.

Chapter 8

Discussion

As in all cells, so in human cells, the mutations result from an interaction of DNA damage and DNA repair mechanisms. When analyzing tumor genome sequences, we encounter a mixture of somatic mutations originating from various mutational processes. The distribution of these mutations across the genome is influenced by DNA damaging processes, DNA repair mechanisms and the genome organization, all of which can vary over time. Consequently, within a single tumor's regional mutation density profile, using statistical analyses we may identify multiple mutational landscapes superimposed on each other. The overall aim of this thesis is to study mutation risk at the chromosomal domain scale, and to distinguish these superimposed profiles to understand the mechanisms affecting mutation distribution across the tumor genomes (Fig 4).

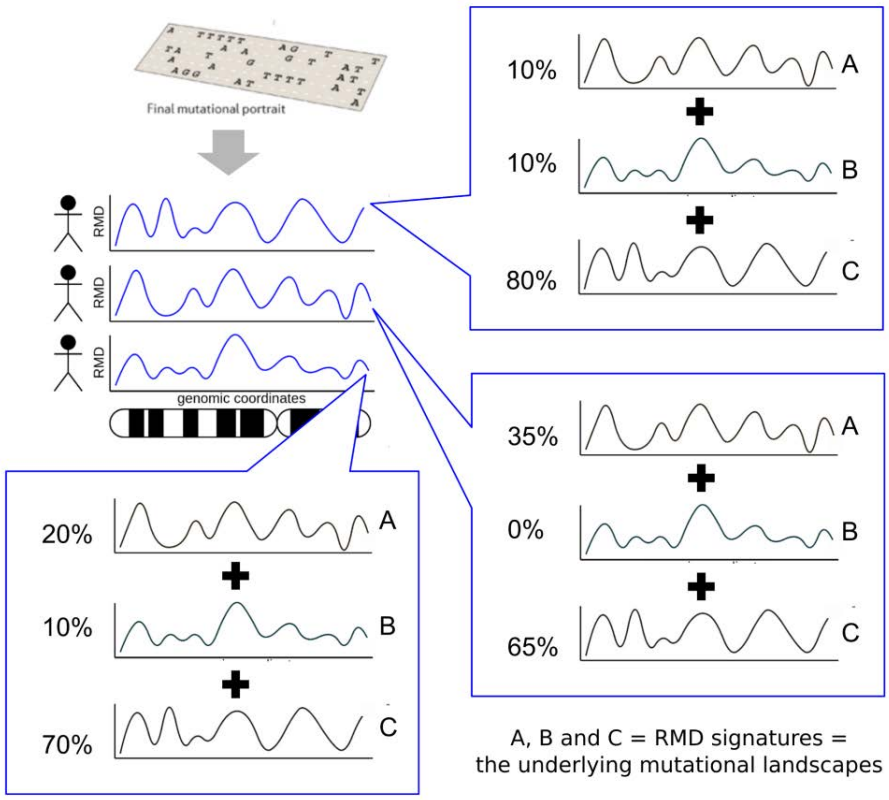


Fig 4. Illustration of the RMD profiles in human tumors and their deconvolution into the underlying mutational landscapes that composes them. Each tumor has a different percentage of contribution from each of the mutational landscapes.

An NMF-based method to detect inter-individual variability in RMD

To deconvolute the RMD patterns, we **developed a method** (Fig 3) based on non-negative matrix factorization (NMF) algorithm to disentangle the distinct mutation landscapes from the RMD profiles across approximately 4000 human tumors. Although the NMF algorithm is at the core of the state-of-the-art method to extract trinucleotide mutational signatures (28,35,37), we used it here on the megabase domain-scale signatures. Therefore addressing a completely different biological question, for which the underlying mechanisms could be different.

One important feature of our methodology is its ability to account for various potential confounding factors, including mutation rates at finer resolutions (e.g. hypermutation at CTCF binding sites), importantly the different trinucleotide composition within each genomic window examined (meaning the patterns we detect cannot be confounded by mutational signatures), and arm-level copy number alterations. A limitation of the current implementation of our method is that low mutation burden tumors are underrepresented as we imposed a minimum threshold of 3 muts/Mb per tumor. To address this limitation in future studies we could relax the mutation burden threshold, and increased WGS sequencing depths would also help by virtue of detecting subclonal mutations. Additionally, include more WGS of tumors does help extract robust mutational signatures so conceivably it will also help in analyses like ours.

To test if our statistical methodology was capable of robustly extracting the inter-individual variability in RMD, we performed simulations using synthetic cancer genomes with artificially "spiked-in" RMD signatures. The most relevant features of these synthetic signatures influencing the power of the method were: the intensity of the RMD signature (how much the mutation count increases or decreases in the windows) and the exposure of the RMD signature i.e. its contribution to the total mutation burden. Notably, our capacity to discern more subtle RMD signatures improved when a sufficient number of tumors exhibited a particular signature. This observation suggests that with a higher number of WGS tumors included in the analysis new RMD signatures might emerge, similarly to what happened in the analyses of the trinucleotide mutational signatures (28).

The current setting of our method extracts the RMD signatures from the entire cohort of tumors available (pan-cancer analysis) because we aimed to maximize the power to identify inter-individual signatures that occur across a diverse array of cancer types. It is important to note that the same methodology can be adapted for application on a per-cancer type basis, or for specific subsets of tumor samples, depending on the specific objectives of the analysis (e.g. if we are interested in subtyping a specific tumor type). Under the current settings, we identified a total of 13 distinctive RMD signatures. Among these, 10 are associated with specific tissue types, while the remaining 3 are global (tissue-independent), capturing the sought-after inter-individual variability in mutation risk of chromosomal domains.

Tissue-related RMD signatures can help identify cell of origin and subtype of the tumor

We found that some of the **tissue-related RMD signatures** combined various cancer types, usually reflecting shared biology of these tumors. For instance, RMD_upper-GI signature is present in most esophagus, stomach, pancreas and biliary tumor samples, and some intestine

tumors. The RMD_lower-GI, in turn, contains mainly the colorectal and most of the intestinal tumors, broadly consistent with the subdivision by developmental origin into the foregut (RMD_upper-GI) and the midgut/hindgut (RMD_lower-GI). These results support the proposed uses of RMD profiles for elucidating cell-of-origin and cancer development trajectories (e.g. metaplasia and/or invasion) by matching to chromatin profiles (120). Moreover, the tissue-associated signatures provide a valuable tool for gaining biological insights into both the shared characteristics and unique distinctions among diverse cancer types.

The RMD signatures can also be useful for subtyping. For example, we observed that the RMD_squamous signature spans various squamous lung cancers, head-and-neck cancers, some bladder cancers (consistent with reports based on gene expression data (121)), also expectedly some cervical and esophageal tumors, and surprisingly some sarcomas and uterus cancers. These results suggest the existence of a squamous-like subtype, possibly originating from squamous cells or cells that changed identity to a squamous-like one, within tumor samples exhibiting high exposure to this signature across cancer types where such a subtype hasn't been previously described. It is possible that a more detailed investigation of this phenomenon, such as applying the NMF methodology in each cancer type separately, could potentially uncover subtypes with clinical relevance.

Interestingly, we obtained one signature that we provisionally named “RMD_B.O.P.S.” and spans brain (B), ovarian (O), prostate (P), sarcomas (S), and some uterus cancers. Since these 4 cancer types do not have their own specific tissue-RMD signature this can indicate that RMD “B.O.P.S.” is a residual, “catch-all” RMD signature that collects diverse RMD patterns that current NMF methodology does not resolve well, possibly due to lack of statistical power because of small sample sized. However, this RMD B.O.P.S. pattern was similarly statistically robust as the other tissue-specific patterns we observed, arguing against this being an artefactual mix of signatures and suggesting this RMD may reflect some commonalities in chromatin organization and gene regulation connecting those cancer types. For instance, an analysis of transcriptome-based cell states across tumor types, based on single-cell gene expression data (122), suggested a module of cilium/cytoskeleton-related genes common to some ovarian cancers, glioblastoma, uterine cancer and lung adenocarcinoma, thus the tissue spectrum approximately corresponds to RMD B.O.P.S. We acknowledge the possibility that, as has occurred with the trinucleotide mutational signatures (28), with arrival of more data some of the initially proposed signatures such as RMD_B.O.P.S. may, be able to be ‘split’ into component RMD signatures that may more precisely match tissue identity.

In summary, the RMD features provide an important tool for understanding the relatedness of cancer types and the chromatin organization in the cell-of-origin of these cancers (120,123). The RMD profiles hold the potential to be a novel methodology for subtyping tumors, reflecting the chromatin organization in the tumor cell of origin, and clinical implications of this should be further studied.

RMD signatures reveal three distinct patterns of inter-individual variability across tumors

Our work uncovered three **global RMD signatures** that capture the redistribution of mutation risks, changing the canonical RMD landscape (as determined by the RT/heterochromatin landscape of each tissue type). These three signatures significantly contribute to the overall mutation burden of the affected tumors, with certain cases exhibiting up to 50% contribution. The 3 global RMD signatures are consistently observed across cancer types originating from most tissues, indicating their widespread relevance.

Interestingly, our association studies found that the three RMD signatures are driven by distinct events: DNA repair failure (MMR, NER, and HR) or APOBEC mutagenesis for RMDflat signature, loss-of-function in the *RB1* tumor suppressor gene for RMDglobal1, and *TP53* mutation for RMDglobal2. In accord with being caused by somatic alterations, these RMD signatures were enriched in subclonal mutations rather than clonal, in contrast to the tissue-related RMD signatures that show an enrichment of clonal mutations. As more WGS data becomes available, it would be interesting to explore the possibility of separating the mutations that occurred before and after the specific causal events, such as RB1 loss, and examine their respective RMD profiles. This analysis could provide further insights into the temporal dynamics and functional consequences of these events on the regional mutation risk landscape.

An intriguing observation emerges from the **RMDflat signature,** where three distinct DNA repair mechanisms and one DNA damaging process (APOBEC mutagenesis) converge onto the same phenotype (a flatter distribution of mutations across chromosomal domains). For MMR and NER repair mechanisms, prior studies described that their activity is enriched in early replicating, euchromatic, gene-rich regions. Consequently, when these DNA repair mechanisms fail a relative increase in mutation rates occurs within these regions compared to the canonical landscape, which is precisely the pattern described by the RMDflat signature (48) (Fig 2). For APOBEC signature, there are prior reports of APOBEC mechanisms being enriched in early-replicating DNA, possibly via their association with DNA repair activity providing ssDNA substrate for APOBECs (8,124,125). If DNA repair mechanisms, particularly DNA MMR, provide the substrate for APOBEC-induced mutagenesis, this would explain the observed co-enrichment of mutations in the same regions as in case of tumors with DNA MMR deficiency.

In addition to the known associations involving MMR, NER, and APOBEC activity, in this work we found that tumors with HRD mutational signatures (both BRCA1 and BRCA2 subtypes (126)) showed the same RMDflat mutation risk phenotype. When HR is deficient, there is an increase in the spectrum of the trinucleotide mutational signature SBS3 (35,127) that likely results from the activity of error-prone DNA polymerases (128). Intriguingly, we observed that in HRD tumors the accumulation of the SBS3-like mutational spectrum tends to cluster within early-replicating DNA regions, opposite to SBS3-like mutational spectrum of HR proficient tumors. The SBS3 pattern in HRD tumors opposes the canonical RMD landscape, thus contributing to the “flatness” of the RMD landscape. This contrasts with previous studies in which they did not find an association between SBS3 and early replication timing (68–70) or they found the opposite, an association of SBS3 with late replication (129). However, none of these studies stratified the tumors into HRD and HR-proficient.

Once we take into account the three known DNA repair failures mentioned above, and APOBEC mutagenesis there is a remaining 28% of tumor samples with the RMDflat signature that remain unexplained. We suggest this is unlikely to be due to false-negatives in the tests for MMR or HR deficiency, since these tumors had, on average, an indel spectrum (in microsatellite loci and elsewhere) not substantially different from the general indel spectrum of same cancer types. Thus we suggest there are other mechanism(s) involved. For instance, in kidney cancer there were many unexplained RMDflat signature samples, however this cancer type very rarely has known MMR, HR deficiencies or APOBEC mutagenesis. A possible explanation is a particular mutational process in kidney (130) that may evade DNA repair mechanisms operative in early-RT domains (MMR and NER) and thus accumulate uniformly across the genome.

Analogous to the scenario observed with the RMDflat signature, we propose that the emergence of the **RMDglobal1 signature** can be acquired by different processes or events that converge onto the same phenotype. Here, this would imply the disruption of the cell cycle controls and/or increased proliferation. The gene expression patterns linked to the RMDglobal1 signature exhibit a notable enrichment in E2F target genes, MYC target genes, and cell cycle regulation pathways. While in the causal analyses, tumors with high RMDglobal1 exposure show a positive association with *RB1* deletions, and independently show a negative association with *CDK6* deletions (negative regulator upstream of RB1). Additionally there is a positive association with *KRAS* mutations. KRAS was reported to act downstream of RB1 in developmental and in tumor mouse phenotypes (131,132)). All the three events lead to the higher proliferation and activation of the cell cycle, findings that are mirrored in the observed gene expression patterns.

Our study suggests, based on analysis of mutation patterns, that these tumors exhibiting a higher proliferative phenotype, undergo changes in the overall epigenome organization. In particular, we found that the specific genome domains that increase mutation risk upon RMDglobal1 signature in tumors, tend to be the same domains that become later replicated and more heterochromatic in cells with a higher proliferation-like gene expression. This observation was consistently supported by multiple independent datasets: the pattern across domains was seen in RT, heterochromatin (both H3K9me3 marked constitutive and to some extent H3K27me3 marked facultative heterochromatin). Intriguingly, our analysis of mutation risk across tumors suggested a broader trend of proliferation-associated chromatin remodeling at the domain level among various biological samples (including many types of healthy cells) that was not previously described to our knowledge.

In our study , we calculated the variability across samples for several features independently such as replication timing, heterochromatin marks and chromatin accessibility at the megabase scale. An initial surprising finding was that indeed there was a significant variability not only in RT, which is known to be variable on a scale of 400-800 kb (74), but also in chromatin accessibility and heterochromatin marks, which are typically distributed at a much finer resolution, encompassing peaks spanning from a few nucleotides to kilobases. Notably, even when assessing the density of these heterochromatin marks and chromatin accessibility peaks across 1-megabase windows

and employing principal component analysis (PCA), we are able to separate several signatures of heterochromatin remodeling at this coarser scale.

A second notable finding was that we found a strong agreement in the RT, heterochromatin, DHS and HiC remodeling signatures that all correlate with each other and with the RMDglobal1 mutagenesis signature. In other words, remarkably, all these distinct features experience remodeling within the same genomic windows where the RMDglobal1 signature increases mutation rates. This observation suggests that the heterochromatin, DHS, RT, and HiC remodeling signatures likely represent facets of a singular process, which results in the activation (euchromatinization) or inactivation (heterochromatinization) of entire chromosomal domains.

As described in the Introduction section, the variability in epigenome organization landscapes has been extensively studied in various prior studies. It was shown that RT varies across cell types (reflecting diverse functions and gene expression programs (73,75)), during differentiation, due to genetic factors, epigenetic modifications and carcinogenesis (73,75,76,79,80). Even though in our analysis we did observe cell type-specific RT changes and developmental-like RT remodeling signatures in ENCODE samples, the RT remodelling signature signature that best correlates with RMDglobal1 mutagenesis is a proliferation-associated RT signature, independent of those RT programs previously reported.

For the rest of the chromatin features mentioned, their local variability across biological samples has only been studied at the finer genomic resolutions, typically on the kilobase scale. For instance for chromatin accessibility, cell-specific regulatory patterns were extracted from 3.6 million accessible chromatin (DHS) sites (83), and similarly cancer-, tissue- and subtype-specific variability was extracted from ATAC-seq variation in human tumors from TCGA (84). However, there are studies that show that epigenetic deregulation in cancer can indeed span large domains of long-range epigenetic silencing (LRES) (88) and activation (LREA) (89). In particular, these regions showed coordinated gene expression, histone modification, DNA methylation changes and disruption of topologically associated domains (TADs) between several kilobases to megabases (80,90). However, our global RMD signatures did not exhibit enrichment within these previously-reported LRES and LREA regions. This is consistent with our study because those epigenomic organization remodeling signatures that correlated with RMDglobal1 mutagenesis did not separate cancer versus healthy cells (as the LRES and LREA would be expected to), but instead -- judging from gene expression -- they separated lowly versus highly proliferative cells or tissues.

An additional intriguing finding emerged as we observed the RT remodeling signatures in a single-cell RT dataset (133). Notably, in the single-cell RT data, the timeline larger does not allow for specific driver events (e.g. RB1 deletion) to occur in many cells. This suggests the presence of stochastic switches in gene expression programs, epigenetic events or potential microenvironmental influences that selectively impact a subset of cells, thus generating the RT signature. In this scRT dataset, lack of associated gene expression data prevented us from testing whether the cells exhibiting the remodeling display a proliferative gene expression phenotype as we would expect. This finding suggests that the genome organization remodeling that we uncovered is very plastic and malleable, potentially indicating its adaptability over time.

To summarize, we have described the epigenomic variability at the megabase domain resolution in human cells, encompassing RT reprogramming, DNA accessibility changes, Hi-C changes and remodeling in heterochromatin mark profiles, uncovering robust intercorrelations among them. Our findings lay the foundation for future investigations into genomic organization variability at the domain scale, a topic that has remained relatively unexplored until now.

For **RMDglobal2 mutation risk signature** we found a strong association with TP53 mutation, which is further independently validated by a positive correlation of RMDglobal2 with amplification in genes that phenocopy TP53 loss (*MDM2* and *MDM4* oncogenes). However, for RMDglobal2 we also found a negative correlation with the mutational signature SBS1: a very abundant signature of C>T changes at CpG dinucleotide, with clock-like properties, likely stemming at least in part from spontaneous 5-methylcytosine deamination. In our dataset, the tumors with *TP53* loss showed reduced exposures to SBS1. These findings suggest that the relationship between these two events may either involve converging onto the same phenotype (similarly to how multiple DNA repair deficiencies converge onto RMDflat) or that there is a link between SBS1 and RMDglobal2 that is not understood yet.

In Yaacov *et al.* study, a modest bias towards late RT was reported for SBS1 in cancer samples (68), while a more pronounced late RT bias was found in healthy samples, implying the loss of this bias in cancer (69). In our study, RMDglobal2 signature shows a reduction of mutation rates in certain windows that are very late replicating. Therefore, if the tumor samples with high RMDglobal2 signature have low exposure to SBS1 this could explain why we found less mutation accumulation in our tumors with high RMDglobal2 activity. However, it remains unexplained why some of the windows in very late replication undergo the mutation risk decrease but not the others, as the prior study pooled all late replicating windows (68,69).

To investigate this, we performed an additional analysis comparing mutation density of the four main trinucleotide SBS1 mutations (C>T) across our 1-megabase windows. As anticipated, we observed increased SBS1 mutation density in late replicating versus early replicating regions. Intriguingly, within late RT windows, those experiencing a reduction in the RMDglobal2 signature displayed elevated SBS1 mutation accumulation. This explains why in samples without SBS1 mutations those windows will relatively acquire less mutations that the other late replication windows. This provides an explanation for the relatively lower mutation acquired in these windows in samples with reduced SBS1 mutations, yet the underlying reasons for differing SBS1 mutation rates between the two groups of late RT windows (affected by RMDglobal2, or not affected) remain an open question, as well as the link of TP53 with this process.

In summary, in this work we deconvoluted the regional mutation density of human tumors at the megabase domain scale into the different mutational landscapes that shape mutation risk. While our analysis focused on tumors, exploring RMD profiles in healthy samples would provide valuable insights into whether normal cell mutational processes differ. We speculate that beyond tissue-specific patterns, other global mutation redistribution processes may exist in healthy cell

samples. For instance, individuals with germline BRCA1 or BRCA2 pathogenic variants (HRD) may exhibit a flatter redistribution of mutations. Additionally, QTL variants in specific RT distributions (rtQTLs) (75,76), likely influencing mutation rates in corresponding regions.

RMD signatures can divert the mutation supply towards domains harboring cancer genes

We consider the study of these mutagenic risk redistributions to be very relevant for modelling cancer evolution, as they impact the mutation supply to different domains in the subsequent stages of cancer development. These redistributions of incoming mutations may play an important role in shaping the likelihood of cancer genes acquiring driver mutations that influence the trajectory of cancer evolution. Future studies about these RMD patterns and their changes upon time, using multiregion sequencing or longitudinal sequencing data would give an interesting perspective into the regional mutation supply dynamics.

Moreover, as the RMD profile of a tumor, coupled with its superimposed mutational processes (e.g., RMDflat signature), directly influences mutation supply, we speculate that these profiles could serve as predictors of cancer development. They have the potential to predict the most probable forthcoming oncogenic or tumor suppressor mutations, and therefore could aid in treatment decisions to prevent these anticipated driver mutations from taking hold in the tumor. Further studies on how RMD profiles can predict prognosis and response to treatment would need to be pursued. One promising example is predicting response to immunotherapy. Currently, that a tumor is MSI tumors (i.e. has an MMR failure) serve as a marker for immunotherapy response (134). Yet, only a small fraction of tumors exhibit MSI. We propose that cases beyond MSI (HRD, NER loss, APOBEC mutagenesis) with RMDflat signatures may yield similar outcomes because mutations get redirected towards gene-rich domains thus increasing potential for generating neoantigens, thus representing a promising area for further exploration.

Applications of cancer type and subtype classifiers in the context of cancer genomics

As earlier indicated, regional mutation density exhibits strong tissue specificity. In this thesis, we use this tissue specificity to predict the cancer type in human tumors, aiming to extend its application to cases of Cancer of Unknown Primary. However, the main limitation arises from its reliance on costly whole-genome sequencing data. Our study showed a reduced accuracy when using whole-exome sequencing and testing on simulated liquid biopsy data. To address this, it would be interesting to explore combining classifiers with the tissue-specific RMD signatures described in Chapter 3 in an effort to enhance accuracy when mutation data is sparse, for instance for liquid biopsy cases where the tumor DNA fraction can be very low.

In Chapter 5, we expanded cancer type classifiers to predict tissue-of-origin in cell lines, as accurate cell line representation is essential for experimental outcomes and drug and genetic screenings. We hypothesized that the mislabeling can stem from metastatic origins, ambiguous histology, or cell culture-induced selection or changes. Unexpectedly, our analysis unveiled a striking 6% of cell lines that appear to be mislabelled, exhibiting molecular profiles inconsistent

with their annotated cancer type but closely resembling another tissue type. This was supported in mutation pattern analysis (in this case trinucleotide signatures, as the RMD profiles are less reliable with exome sequencing; availability of WGS for the cell lines will enable use of RMD as an additional data source). Additionally, we identified a further 12% of cell lines that had seemingly diverged from their initial cancer type and no longer were consistent with their original molecular characteristics in terms of methylome and transcriptome.

Our study highlights the critical role of selecting appropriate experimental models for testing cancer drugs and gene therapies. It underscores the need for cautious consideration when employing cell line panels, as we reveal that focusing on the 60% of cell lines closely mirroring their annotated cancer type significantly improves the efficacy of drug target identification, despite the overall lower number of cell lines. Future efforts that establish new cell line panels by selecting cell lines guided by their resemblance to actual tumor genomes (also epigenomes and transcriptomes) would be justified.

Mutational strand asymmetry as a tool to detect origins of replication

In human cells, thousands of potential replication origins are present during each cell division. Despite the existence of several experimental methods to detect them, the biological diversity of origin usage across tissues and individuals remains largely unexplored. In **Chapter 6**, we perform a study to tackle this, revealing the existence of unique patterns of tissue-specific replication origin usage.

We developed a genomics method, MuSAS, which identifies replication origins using the mutational strand asymmetry associated found in specific mutational signatures. A recent approach demonstrated the feasibility of this principle, relying on the individual example of the POLE mutational signature (SBS10) (106). However, the MuSAS method stands out by combining a considerably broader set of mutational signatures, and extends to tissues beyond that in which POLE mutants are present (normally just colorectum and uterus). This pan-signature analysis via MuSAS vastly boosts statistical power by considerably increasing the number of mutations and allows for the consideration of many tissues, enabling a comparison between tissues. We also designed two approaches for using the MuSAS method, MuSAS-A and MuSAS-B. Each strategy aggregates mutations based on distinct criteria, aiming to encompass a range of cancer types. The MuSAS method's primary limitation lies in its reliance on mutation accumulation, making it less potent for samples with low mutation burdens, and we presume it would not easily detect rarely-used origins.

Interestingly, our study identified substantial variability in replication origin activity across seven tissues: in addition to the very prominent brain-specific pattern of origin usage, there was also the breast, colon, uterus, lymphoid, prostate, and bladder. Confirming our methodology's reliability, many of these patterns remained consistent between both MuSAS methods. In addition, we observed that these origins align with tissue-specific early replication timings and are situated near tissue-specific genes. A deeper exploration revealed these origins to be abundant in tissue-specific enhancer and promoter histone marks, notably H3K4me1 and H3K27ac. This suggests

that the activation of these enhancers and promoters might influence origin usage, although we cannot rule out that the direction of causality can also be the converse, where in principle the activated origins could activate nearby enhancers and promoters in a tissue. While past studies have delved into the relationship between chromatin marks and origins, this was performed using global sets of origins, or on results of experiments on individual cell types. Our work goes beyond that by systematically highlighting the tissue-specific nuances of these associations between origin usage and chromatin switching to active chromatin.

Here, we addressed between-tissue variation but MuSAS-B provides a great resource for future work to address the existence of inter-individual variability on origin usage. Plausibly, since there exist germline SNPs associated with locally earlier or later RT (75,76), there might also exist SNPs that affect origin usage, which could act via abolishing or establishing DNA motifs relevant to origin usage. In our current study, when analyzing the variability of MuSAS-B (by individual tumor), we found a few PCs (with a lower amount of variance explained than the tissue-specific ones) that separate out a small number of individual tumors. This suggests that there might indeed be some inter-individual variability in origin usage. However, these PCs were not robustly supported by our consistency analysis, when we divided the genome into three parts and looked for recurrent trends. This could plausibly happen because the sample size was too small. Further research on this with more whole-genome sequencing data to boost power, or perhaps some alternative algorithmic approaches, such as NMF or ICA, will offer more clarity.

To sum up, our work offers insight into tissue-specific replication origin usage across diverse human somatic cell type, inferred via mutation analysis in hundreds of cancers. It reveals the close relationship between tissue-associated gene expression, replication origin activation, and shifts to earlier replication timing in the related genomic regions. Additionally it highlights the intricate tissue-specific coupling between origin usage and chromatin marks and histone variants, which emerges as a pivotal area for further exploration of mechanisms.

Finally, our research on the origins of replication highlights its direct link with the replication timing, with the (expected) relationship that a higher density of active origins implies an earlier RT time in a given tissue. These changes in RT are likely to have consequences on mutation accumulation in two ways: firstly, certain mutational processes can become more pronounced at the narrow region of the origins themselves (135); and secondly, switching an entire domain to an earlier replication time in a specific tissue is well expected to result in a reduced mutation rate in that region (48,54). Despite the distinct resolutions affected by these two mechanisms -- origins at base pair (bp) resolution and regional mutation density (RMD) at the 1 Mb scale -- our findings pave the way for understanding how RMD patterns result from the tissue or individual-specific origin usage and tissue or individual-specific RT.

Chapter 9

Conclusions

- We developed a new statistical methodology based on NMF that identified 13 RMD signatures, which describe a genome-wide mutation risk redistribution across many megabase-sized domains in >4000 tumors
- Out of the 13 RMD signatures, we identified 10 that are tissue-specific at least to some degree and 3 that are universal, tissue-independant global variability in RMD observed between individual tumors.
- The tissue-related RMD signatures range from highly specific patterns unique to individual tissues, to signatures that encompass multiple tissues with apparently similar physiology. RMD signatures may help identify the subtype of a tumor.
- The RMDflat global signature indicates a “flatter”, less heterogeneous distribution of mutation rates across domains, which can be caused by deficiencies in the MMR, NER, or HR DNA repair pathways.
- The RMDglobal1 signature reflects regional plasticity in DNA replication time and in heterochromatin, which was linked with a higher expression of cell cycle genes and E2F target genes. Its occurrence is associated with altered cell cycle control via loss of the RB1 tumor suppressor gene.
- RMDglobal2 signature is associated with loss-of-function of the TP53 pathway, mainly affecting the redistribution of relative mutation rates away from the late RT regions.
- The redistribution of mutations by these RMD signatures affects the mutation supply towards domains harboring various cancer genes. This has the potential to influence the acquisition of driver mutations, steering cancer evolution.
- Machine learning classifiers trained on RMD patterns achieved high accuracy (median AUC of 0.99) in predicting 18 tumor types and can distinguish major cancer subtypes, suggesting their potential in diagnosing cancers of unknown primary (CUP).
- While driver mutations classified only 36% of patients correctly at high precision, features based on passenger mutations, including RMD, achieved a 92% accuracy, emphasizing the significance of whole-genome sequencing for tumor classification.
- We extended the machine learning classifiers to predict the tissue-of-origin, generalizing from tumors to a panel of 600 cancer cell lines, based on transcriptomic and epigenomic profiles as features.
- We found 60% of the analyzed cell lines closely matched their known tumor type of origin thus making a good cancer model in terms of transcriptome and epigenome, while 6% were misidentified to the wrong tissue, potentially affecting downstream experimental results.

- Focussing analyses on correctly classified cancer cell lines led to more precise identification of active drugs and markers, underscoring the utility of epigenomic analysis to rank cancer cell lines as good experimental models for cancer research.
- We developed a genomics method, named MuSAS, to identify replication origin loci via switches in mutational strand asymmetry, using two approaches: MuSAS-A (pooling mutational signatures) and MuSAS-B (pooling individual tumors).
- We identified several tissue-specific replication origin usage trends using an analysis of tumor WGS, including distinct origin patterns for brain, breast, colon, uterus, prostate, lymphoid and bladder.
- Tissue-specific origins showed correlations with earlier replication timing and had a higher prevalence of genes specifically expressed in their tissue type, indicating an association between gene activation and local origin usage.
- Tissue-specific origins presented distinct chromatin marks, with noticeable enrichments in chromatin accessibility, enhancer and promoter marks of the corresponding tissue, suggesting a possible mechanistic significance of these features origin utilization.

Chapter 10

References

1. Definition of mutation - NCI Dictionary of Cancer Terms - NCI [Internet]. 2011 [cited 2023 Jun 1]. Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/mutation
2. Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet. 2010 Oct 15;19(R2):R131–6.
3. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 2015 Sep 1;14(5):305–14.
4. Chatterjee N, Walker GC. Mechanisms of DNA damage, repair, and mutagenesis. Environ Mol Mutagen. 2017;58(5):235–63.
5. Arana ME, Kunkel TA. Mutator phenotypes due to DNA replication infidelity. Semin Cancer Biol. 2010 Oct;20(5):304–11.
6. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci. 2010 Jan 19;107(3):961–8.
7. Ray PD, Huang BW, Tsuji Y. Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. Cell Signal. 2012 May;24(5):981–90.
8. Mas-Ponte D, Supek F. DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. Nat Genet. 2020 Aug 3;1–11.
9. Wang Y, Robinson PS, Coorens THH, Moore L, Lee-Six H, Noorani A, et al. APOBEC mutagenesis is a common process in normal human small intestine. Nat Genet. 2023 Feb;55(2):246–54.
10. Pfeifer GP. Mechanisms of UV-induced mutations and skin cancer. Genome Instab Dis. 2020;1(3):99–113.
11. Hoeijmakers JHJ. DNA Damage, Aging, and Cancer. N Engl J Med. 2009 Oct 8;361(15):1475–85.
12. Krokan HE, Bjørås M. Base Excision Repair. Cold Spring Harb Perspect Biol. 2013 Apr;5(4):a012583.
13. Robertson AB, Klungland A, Rognes T, Leiros I. DNA repair in mammalian cells: Base excision repair: the long and short of it. Cell Mol Life Sci CMLS. 2009 Mar;66(6):981–93.
14. Spivak G. Nucleotide excision repair in humans. DNA Repair. 2015 Dec 1;36:13–8.
15. Mas-Ponte D, McCullough M, Supek F. Spectrum of DNA mismatch repair failures viewed through the lens of cancer genomics and implications for therapy. Clin Sci. 2022 Mar 11;136(5):383–404.
16. Li GM. Mechanisms and functions of DNA mismatch repair. Cell Res. 2008 Jan;18(1):85–98.
17. Zhao L, Washington MT. Translesion Synthesis: Insights into the Selection and Switching of DNA Polymerases. Genes. 2017 Jan;8(1):24.
18. Powers KT, Washington MT. Eukaryotic translesion synthesis: Choosing the right tool for the job. DNA Repair. 2018 Nov 1;71:127–34.
19. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. 2015 Sep 25;349(6255):1483–9.
20. Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, et al. The mutational landscape of human somatic and germline cells. Nature. 2021 Sep;597(7876):381–6.
21. What Is Cancer? - NCI [Internet]. 2007 [cited 2023 Jun 7]. Available from: https://www.cancer.gov/about-cancer/understanding/what-is-cancer
22. Global Cancer Observatory: Cancer Today. Lyon Int Agency Res Cancer 2020 [Internet]. [cited 2023 Jun 7]; Available from: http://gco.iarc.fr/today/home
23. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000 Jan 7;100(1):57–70.
24. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011 Mar 4;144(5):646–74.

25. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009 Apr 9;458(7239):719–24.
26. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Loo PV, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017 Nov 16;171(5):1029-1041.e21.
27. Hudson (Chairperson) TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature. 2010 Apr;464(7291):993–8.
28. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. 2020 Feb;578(7793):82–93.
29. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature. 2019 Nov;575(7781):210–6.
30. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat Rev Genet. 2018 May;19(5):269–85.
31. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009 Oct;461(7265):747–53.
32. Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. 2013 Nov 7;155(4):948–62.
33. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 Jul;499(7457):214–8.
34. Boström M, Larsson E. Somatic mutation distribution across tumour cohorts provides a signal for positive selection in cancer. Nat Commun. 2022 Nov 17;13(1):7023.
35. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug 22;500(7463):415–21.
36. Alexandrov LB, Nik-Zainal S, Siu HC, Leung SY, Stratton MR. A mutational signature in gastric cancer suggests therapeutic strategies. Nat Commun. 2015 Oct 29;6(1):8683.
37. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb;578(7793):94–101.
38. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013 Sep;45(9):970–6.
39. Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, et al. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell. 2019 Mar 7;176(6):1282-1294.e20.
40. Supek F, Lehner B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. Cell. 2017 Jul 27;170(3):534-547.e23.
41. Harris K, Nielsen R. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Res. 2014 Sep 1;24(9):1445–54.
42. Steele CD, Abbasi A, Islam SMA, Bowes AL, Khandekar A, Haase K, et al. Signatures of copy number alterations in human cancer. Nature. 2022 Jun;606(7916):984–91.
43. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019 May 2;177(4):821-836.e16.
44. Meier B, Volkova NV, Hong Y, Schofield P, Campbell PJ, Gerstung M, et al. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. Genome Res [Internet]. 2018 Apr 10 [cited 2023 Sep 12]; Available from: https://genome.cshlp.org/content/early/2018/04/10/gr.226845.117
45. Supek F, Lehner B. Scales and mechanisms of somatic mutation rate variation across the human genome. DNA Repair. 2019;81:102647.
46. Kaiser VB, Taylor MS, Semple CA. Mutational Biases Drive Elevated Rates of Substitution

at Regulatory Sites across Cancer Types. PLOS Genet. 2016 Aug 4;12(8):e1006207.
47. Chen X, Chen Z, Chen H, Su Z, Yang J, Lin F, et al. Nucleosomes Suppress Spontaneous Mutations Base-Specifically in Eukaryotes. Science. 2012 Mar 9;335(6073):1235–8.
48. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015 May;521(7550):81–4.
49. Mao P, Brown AJ, Esaki S, Lockwood S, Poon GMK, Smerdon MJ, et al. ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. Nat Commun. 2018 Jul 6;9(1):2626.
50. Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science [Internet]. 2019 Jun 28 [cited 2020 Jun 24];364(6447). Available from: https://science.sciencemag.org/content/364/6447/eaaw2872
51. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010 Jan;463(7278):191–6.
52. Li F, Mao G, Tong D, Huang J, Gu L, Yang W, et al. The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. Cell. 2013 Apr 25;153(3):590–600.
53. Polak P, Lawrence MS, Haugen E, Stoletzki N, Stojanov P, Thurman RE, et al. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. Nat Biotechnol. 2014 Jan;32(1):71–5.
54. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015 Feb;518(7539):360–4.
55. Hodgkinson A, Chen Y, Eyre-Walker A. The large-scale distribution of somatic mutations in cancer genomes. Hum Mutat. 2012 Jan;33(1):136–43.
56. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012 Aug;488(7412):504–7.
57. Woo YH, Li WH. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nat Commun. 2012 Aug 14;3(1):1004.
58. Liu L, De S, Michor F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. Nat Commun. 2013 Feb 19;4(1):1502.
59. Akdemir KC, Le VT, Kim JM, Killcoyne S, King DA, Lin YP, et al. Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. Nat Genet. 2020 Nov;52(11):1178–88.
60. Zheng CL, Wang NJ, Chung J, Moslehi H, Sanborn JZ, Hur JS, et al. Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. Cell Rep. 2014 Nov 20;9(4):1228–34.
61. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science. 2017 13;358(6360):234–8.
62. Zou X, Owusu M, Harris R, Jackson SP, Loizou JI, Nik-Zainal S. Validating the concept of mutational signatures with isogenic cell models. Nat Commun. 2018 May 1;9(1):1–16.
63. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci. 2010 Jan 5;107(1):139–44.
64. Krämer A, Bochtler T, Pauli C, Baciarello G, Delorme S, Hemminki K, et al. Cancer of unknown primary: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up☆. Ann Oncol. 2023 Mar 1;34(3):228–46.
65. Weiss LM, Chu P, Schroeder BE, Singh V, Zhang Y, Erlander MG, et al. Blinded comparator

study of immunohistochemical analysis versus a 92-gene cancer classifier in the diagnosis of the primary site in metastatic tumors. J Mol Diagn JMD. 2013 Mar;15(2):263–9.
66. Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, et al. MicroRNAs accurately identify cancer tissue origin. Nat Biotechnol. 2008 Apr;26(4):462–9.
67. Moran S, Martínez-Cardús A, Sayols S, Musulén E, Balañá C, Estival-Gonzalez A, et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. Lancet Oncol. 2016 Oct;17(10):1386–95.
68. Yaacov A, Vardi O, Blumenfeld B, Greenberg A, Massey DJ, Koren A, et al. Cancer Mutational Processes Vary in Their Association with Replication Timing and Chromatin Accessibility. Cancer Res. 2021 Dec 15;81(24):6106–16.
69. Yaacov A, Rosenberg S, Simon I. Mutational signatures association with replication timing in normal cells reveals similarities and differences with matched cancer tissues. Sci Rep. 2023 May 15;13(1):7833.
70. Tomkova M, Tomek J, Kriaucionis S, Schuster-Böckler B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. 2018 Sep 10;19(1):129.
71. Caballero M, Ge T, Rebelo AR, Seo S, Kim S, Brooks K, et al. Comprehensive analysis of DNA replication timing across 184 cell lines suggests a role for MCM10 in replication timing regulation. Hum Mol Genet. 2022 Sep 1;31(17):2899–917.
72. Zhao PA, Rivera-Mulia JC, Gilbert DM. Replication Domains: Genome Compartmentalization into Functional Replication Units. In: Masai H, Foiani M, editors. DNA Replication: From Old Principles to New Discoveries [Internet]. Singapore: Springer; 2017 [cited 2023 Jul 12]. p. 229–57. (Advances in Experimental Medicine and Biology). Available from: https://doi.org/10.1007/978-981-10-6955-0_11
73. Rivera-Mulia JC, Buckley Q, Sasaki T, Zimmerman J, Didier RA, Nazor K, et al. Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. Genome Res. 2015 Aug 1;25(8):1091–103.
74. Marchal C, Sima J, Gilbert DM. Control of DNA replication timing in the 3D genome. Nat Rev Mol Cell Biol. 2019 Dec;20(12):721–37.
75. Koren A, Handsaker RE, Kamitaki N, Karlić R, Ghosh S, Polak P, et al. Genetic Variation in Human DNA Replication Timing. Cell. 2014 Nov 20;159(5):1015–26.
76. Ding Q, Edwards MM, Hulke ML, Bracci AN, Hu Y, Tong Y, et al. The Genetic Architecture of DNA Replication Timing in Human Pluripotent Stem Cells. bioRxiv. 2020 May 10;2020.05.08.085324.
77. Klein KN, Zhao PA, Lyu X, Sasaki T, Bartlett DA, Singh AM, et al. Replication timing maintains the global epigenetic state in human cells. Science. 2021 Apr 23;372(6540):371–8.
78. Pratto F, Brick K, Cheng G, Lam KWG, Cloutier JM, Dahiya D, et al. Meiotic recombination mirrors patterns of germline replication in mice and humans. Cell. 2021 Aug 5;184(16):4251-4267.e20.
79. Ryba T, Battaglia D, Chang BH, Shirley JW, Buckley Q, Pope BD, et al. Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. Genome Res. 2012 Oct;22(10):1833–44.
80. Du Q, Bert SA, Armstrong NJ, Caldon CE, Song JZ, Nair SS, et al. Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. Nat Commun. 2019 Jan 24;10(1):416.
81. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. 2019 Apr;20(4):207–20.
82. Klein DC, Hainer SJ. Genomic methods in profiling DNA accessibility and factor localization. Chromosome Res. 2020 Mar 1;28(1):69–85.
83. Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, et al. Index and biological

spectrum of human DNase I hypersensitive sites. Nature. 2020 Aug;584(7820):244–51.
84. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018 Oct 26;362(6413):eaav1898.
85. Siggens L, Ekwall K. Epigenetics, chromatin and genome organization: recent advances from the ENCODE project. J Intern Med. 2014;276(3):201–14.
86. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020 Jul;583(7818):699–710.
87. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017 Dec;12(12):2478–92.
88. Coolen MW, Stirzaker C, Song JZ, Statham AL, Kassir Z, Moreno CS, et al. Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity. Nat Cell Biol. 2010 Mar;12(3):235–46.
89. Bert SA, Robinson MD, Strbenac D, Statham AL, Song JZ, Hulf T, et al. Regional activation of the cancer genome by long-range epigenetic remodeling. Cancer Cell. 2013 Jan 14;23(1):9–22.
90. Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir K, Gould CM, et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. Genome Res. 2016 Jun;26(6):719–31.
91. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, et al. Global Epigenomic Reconfiguration During Mammalian Brain Development. Science. 2013 Aug 9;341(6146):1237905.
92. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012 Jul;13(7):484–92.
93. Luo C, Hajkova P, Ecker JR. Dynamic DNA methylation: in the right place at the right time. Science. 2018 Sep 28;361(6409):1336–40.
94. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol. 2019 Oct;20(10):590–607.
95. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet. 2018 Apr;50(4):591–602.
96. Zhang X, Jeong M, Huang X, Wang XQ, Wang X, Zhou W, et al. Large DNA Methylation Nadirs Anchor Chromatin Loops Maintaining Hematopoietic Stem Cell Identity. Mol Cell. 2020 May 7;78(3):506-521.e6.
97. Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina–associated domains. Nat Genet. 2012 Jan;44(1):40–6.
98. Brinkman AB, Nik-Zainal S, Simmer F, Rodríguez-González FG, Smid M, Alexandrov LB, et al. Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation. Nat Commun. 2019 Apr 15;10(1):1749.
99. Wei SH, Balch C, Paik HH, Kim YS, Baldwin RL, Liyanarachchi S, et al. Prognostic DNA methylation biomarkers in ovarian cancer. Clin Cancer Res Off J Am Assoc Cancer Res. 2006 May 1;12(9):2788–94.
100. Méchali M. Eukaryotic DNA replication origins: many choices for appropriate answers. Nat Rev Mol Cell Biol. 2010 Oct;11(10):728–38.
101. Fragkos M, Ganier O, Coulombe P, Méchali M. DNA replication origin activation in space and time. Nat Rev Mol Cell Biol. 2015 Jun;16(6):360–74.
102. Courtot L, Hoffmann JS, Bergoglio V. The Protective Role of Dormant Origins in Response to Replicative Stress. Int J Mol Sci. 2018 Nov 12;19(11):3569.
103. Shinbrot E, Henninger EE, Weinhold N, Covington KR, Göksenin AY, Schultz N, et al.

Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. Genome Res. 2014 Nov 1;24(11):1740–50.
104. Cayrou C, Coulombe P, Vigneron A, Stanojcic S, Ganier O, Peiffer I, et al. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res. 2011 Sep;21(9):1438–49.
105. Akerman I, Kasaai B, Bazarova A, Sang PB, Peiffer I, Artufel M, et al. A predictable conserved DNA base composition signature defines human core DNA replication origins. Nat Commun. 2020 Sep 21;11(1):4826.
106. Jaksik R, Wheeler DA, Kimmel M. Detection and characterization of constitutive replication origins defined by DNA polymerase epsilon. BMC Biol. 2023 Feb 24;21(1):41.
107. Dao FY, Lv H, Fullwood MJ, Lin H. Accurate Identification of DNA Replication Origin by Fusing Epigenomics and Chromatin Interaction Information. Research [Internet]. 2022 Oct 30 [cited 2023 Sep 1];2022. Available from: https://spj.science.org/doi/10.34133/2022/9780293
108. Mesner LD, Valsakumar V, Cieslik M, Pickin R, Hamlin JL, Bekiranov S. Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. Genome Res. 2013 Nov;23(11):1774–88.
109. Langley AR, Gräf S, Smith JC, Krude T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). Nucleic Acids Res. 2016 Dec 1;44(21):10230–47.
110. Petryk N, Kahli M, d'Aubenton-Carafa Y, Jaszczyszyn Y, Shen Y, Silvain M, et al. Replication landscape of the human genome. Nat Commun. 2016 Jan 11;7(1):10208.
111. Wu X, Kabalane H, Kahli M, Petryk N, Laperrousaz B, Jaszczyszyn Y, et al. Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-replicating regions. Nucleic Acids Res. 2018 Nov 2;46(19):10157–72.
112. Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A, et al. Human gene organization driven by the coordination of replication and transcription. Genome Res. 2007 Sep 1;17(9):1278–85.
113. Dellino GI, Cittaro D, Piccioni R, Luzi L, Banfi S, Segalla S, et al. Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. Genome Res. 2013 Jan;23(1):1–11.
114. Miotto B, Ji Z, Struhl K. Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. Proc Natl Acad Sci. 2016 Aug 16;113(33):E4810–9.
115. Sugimoto N, Maehara K, Yoshida K, Ohkawa Y, Fujita M. Genome-wide analysis of the spatiotemporal regulation of firing and dormant replication origins in human cells. Nucleic Acids Res. 2018 Jul 27;46(13):6683–96.
116. Goodspeed A, Heiser LM, Gray JW, Costello JC. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. Mol Cancer Res MCR. 2016 Jan;14(1):3–13.
117. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell. 2016 Jul 28;166(3):740–54.
118. Long H, Zhang L, Lv M, Wen Z, Zhang W, Chen X, et al. H2A.Z facilitates licensing and activation of early replication origins. Nature. 2020 Jan;577(7791):576–81.
119. Park SM, Choi EY, Bae M, Kim S, Park JB, Yoo H, et al. Histone variant H3F3A promotes lung cancer cell migration through intronic regulation. Nat Commun. 2016 Oct 3;7:12914.
120. Kübler K, Karlić R, Haradhvala NJ, Ha K, Kim J, Kuzman M, et al. Tumor mutational landscape is a record of the pre-malignant state [Internet]. bioRxiv; 2019 [cited 2022 Oct 4]. p. 517565. Available from: https://www.biorxiv.org/content/10.1101/517565v1

121. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014 Aug 14;158(4):929–44.
122. Barkley D, Moncada R, Pour M, Liberman DA, Dryg I, Werba G, et al. Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. Nat Genet. 2022 Aug;54(8):1192–201.
123. Nguyen L, Van Hoeck A, Cuppen E. Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. Nat Commun. 2022 Jul 11;13(1):4013.
124. Seplyarskiy VB, Soldatov RA, Popadin KY, Antonarakis SE, Bazykin GA, Nikolaev SI. APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. Genome Res. 2016 Feb 1;26(2):174–82.
125. Chen J, Miller BF, Furano AV. Repair of naturally occurring mismatches can induce mutations in flanking DNA. eLife. 2014 Apr 29;3:e02001.
126. Pothuri B. BRCA1- and BRCA2-related mutations: therapeutic implications in ovarian cancer. Ann Oncol. 2013 Nov 1;24:viii22–7.
127. Nguyen L, W. M. Martens J, Van Hoeck A, Cuppen E. Pan-cancer landscape of homologous recombination deficiency. Nat Commun. 2020 Nov 4;11(1):5584.
128. Chen D, Gervai JZ, Póti Á, Németh E, Szeltner Z, Szikriszt B, et al. BRCA1 deficiency specific base substitution mutagenesis is dependent on translesion synthesis and regulated by 53BP1. Nat Commun. 2022 Jan 11;13(1):226.
129. Otlu B, Díaz-Gay M, Vermes I, Bergstrom EN, Zhivagui M, Barnes M, et al. Topography of mutational signatures in human cancer. Cell Rep [Internet]. 2023 Aug 29 [cited 2023 Sep 5];42(8). Available from: https://www.cell.com/cell-reports/abstract/S2211-1247(23)00941-5
130. Franco I, Helgadottir HT, Moggio A, Larsson M, Vrtačnik P, Johansson A, et al. Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. Genome Biol. 2019 Dec 18;20(1):285.
131. Takahashi C, Contreras B, Bronson RT, Loda M, Ewen ME. Genetic Interaction between Rb and K-ras in the Control of Differentiation and Tumor Suppression. Mol Cell Biol. 2004 Dec;24(23):10406–15.
132. Lee KY, Ladha MH, McMahon C, Ewen ME. The Retinoblastoma Protein Is Linked to the Activation of Ras. Mol Cell Biol. 1999 Nov;19(11):7724–32.
133. Du Q, Smith GC, Luu PL, Ferguson JM, Armstrong NJ, Caldon CE, et al. DNA methylation is required to maintain both DNA replication timing precision and 3D genome organization integrity. Cell Rep. 2021 Sep 21;36(12):109722.
134. Guan J, Li GM. DNA mismatch repair in cancer immunotherapy. NAR Cancer. 2023 Sep 1;5(3):zcad031.
135. Murat P, Perez C, Crisp A, van Eijk P, Reed SH, Guilbaud G, et al. DNA replication initiation shapes the mutational landscape and expression of the human genome. Sci Adv. 2022 Nov 9;8(45):eadd3686.

Annex: Collaborations in other publications

In addition to the published manuscripts included in this text, this section lists other scientific publications contributed by the author during the doctoral thesis period:

- B Fito-Lopez, M Salvadores, MM Alvarez and F Supek (2023). Prevalence, causes and impact of TP53-loss phenocopying events in human tumors. BMC biology 21 (1), 92
- J Levatić, M Salvadores, F Fuster-Tormo and F Supek (2022). Mutational signatures are markers of drug sensitivity of cancer cells. Nature Communications 13 (1), 2926
- J Biayna, I Garcia-Cao, MM Alvarez, M Salvadores, J Espinosa-Carrasco, M McCullough, F Supek and TH Stracker (2021) Loss of the abasic site sensor HMCES is synthetic lethal with the activity of the APOBEC3A cytosine deaminase in cancer cells. PLoS Biology 19 (3), e3001176