Redistribution of mutation rates across chromosomal domains in human cancer genomes

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 have a heterogeneous genomic distribution, with increased numbers of mutations in late-replication time (RT), heterochromatic domains of chromosomes. While this regional mutation rate density (RMD) landscape is known to vary between tissues and due to deficiencies in DNA repair, we asked whether it varies between individual tumors and what would be the mechanisms underlying such variation. Here, we identified 13 RMD signatures that describe mutation redistribution across megabase-scale domains in ~4200 tumors. Of those, 10 RMD signatures corresponded to groupings or subdivisions of cancerous tissues and cell types. We further identified 3 global RMD signatures of somatic mutation landscapes that transcended cancer types. One is a known general loss of RMD variation, previously associated with DNA mismatch repair failures, and was here additionally linked with homologous recombination (HR) repair deficiencies. Next, we identified a global RMD signature affecting facultative heterochromatin domains. This RMD signature strongly reflects regional variation in DNA replication time and in heterochromatin across state tumor samples, and is associated with altered cell cycle control. Finally we identified a global RMD signature associated with *TP53* loss-of-function, mainly affecting the very late RT regions. The local mutation rates in 26%-75% of cancer genes are notably changed in the tumors affected by these three global RMD signatures of mutation redistribution. Our study highlights how the plasticity of chromatin states and the RT program in cancers bears upon the regional somatic mutation rate landscape, and the downstream consequences on mutation supply to disease genes.

Introduction

During cancer evolution, somatic cells accumulate a number of mutations, most of them non-selected “passengers”. These somatic mutations are caused by different mutagenic processes, many of which generate higher mutation rates in late DNA replication time (RT), inactive, heterochromatic DNA. This is likely due to higher activity and/or accuracy of DNA repair in early-replicating, active chromosomal domains [1,2].

These chromosomal segments are defined roughly at the megabase scale, and 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 domain RT, local gene expression levels, chromatin accessibility (as DNAse hypersensitive sites (DHS)), density of inactive histone marks such as H3K9me3 and inversely with density of active marks such as H3K4me3 [1,6–8]. The RMD signatures have been shown to be tissue-specific, and can be used to predict cancer type, and potentially subtype at high accuracy [9,10]. The tissue-specificity of RMD is paralleled in the tissue-specificity of active or inactive domains. 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 may help identify the cell-of-origin of some cancers [11].

Apart from variation in active chromatin and gene expressions between tissues, 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 [12,13]. Such 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 in some chromosomal domains with cell cycling, as well as a generalized loss of heterochromatin upon transformation [14–18]. These changes in RT, DNA methylation and heterochromatin occuring in cancer cells may plausibly affect chromosomal stability, given the links of various DNA damage and repair processes and chromatin organization [1,2,16,19–21].

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

We study the RMD profiles at the megabase scale of somatic mutations from tumor whole-genome sequences, modeling this mutational portrait as a mixture of several underlying regional distributions, which may correspond to different mechanisms that produce or prevent mutations preferentially in some genomic domains. To disentangle these distributions, we apply an unsupervised factorization approach and extract RMD signatures from ~4200 whole genome sequenced human tumors. Some of these RMD signatures represent the expected differences between tissues/cell types, or they may represent consequences of common DNA repair failures. However others are novel and are associated with RT variation and with chromatin remodeling upon cell cycle disturbances. We characterize the differences between individuals in the usage of these different RMD distributions of mutations, suggesting that the chromatin remodeling RMD signatures are ubiquitous amongst human cancers. They reflect wide-spread mutation redistribution across domains and affect mutation supply to regions harboring cancer genes.

Results

Inter-individual variability in megabase-scale regional mutation density in human tumors

We hypothesize that, in addition to the variability between cancer types, the RMD patterns encompass variability between individuals that is observed across many tissues. To test this, we performed a global unsupervised analysis of diversity in one-megabase (1 Mb) RMD patterns across 4221 whole-genome sequenced tumors that had a mutation burden >3 single-nucleotide variants (SNV) per Mb. To prevent confounding by the variable SNV mutational signatures across tumors [22] we controlled for trinucleotide composition across the 1 Mb windows (Methods). We additionally normalized the RMDs at chromosome arm-level to control for possible confounding of large-scale copy-number alterations (CNA) on mutation rates. Finally we removed known mutation hotspots (e.g. CTCF binding sites, see Methods), and also exons of all protein-coding genes to reduce effects of selection.

To quantify the systematic variability contained within tumor RMD landscapes, we applied a Principal Component (PC) analysis on the RMD profiles across all tumor samples (n=4221). Expectedly, most of the 22 relevant PCs (those with a % of variance explained higher than a random baseline (Fig 1a)), separated different tissues (Fig 1b, Fig S1a). However, we found some PCs that captured variability between individuals but not between the known tissue-of-origin of tumors (Fig S1a). Serving as a positive control, the PC1 separated the canonical RMD landscape with increased mutation rates in late-replicating DNA versus the known "flat" landscape of tumors with failed DNA mismatch repair (MMR) [1] (Fig 1c). Next, we observed that PC7 separates lymphoid tumors with higher somatic hypermutation (SHM) activity (Fig 1c), using the exposure of mutational signature SBS9 as a proxy for prior activity of SHM in that lymphoma sample [22]. Reassuringly, we observed that the PC7 1 Mb window weights are strongest in known SHM regions containing antibody genes (Fig 1d). In summary, our RMD features were able to capture two known examples of regional redistribution of mutations: one affects specific sites (SHM regions in B-lymphocytes) and the other causes a global 'flattening' of mutation rate landscape along the genome in MMR-deficient samples, supporting the utility of our RMD profiling method.

Next, we asked if clustering the tumor samples by their RMD feature vectors would reveal, in addition to an expected grouping by tissue, also other sources of inter-individual variability in RMD (Fig 1e). For clustering, we selected first 22 RMD PCs based on the amount of variance explained (Fig 1a), based on the PC window weights' autocorrelation with neighboring windows (Fig S1b) indicating a nonrandom organization of the RMD pattern along the chromosomes; and based on additional criteria (Fig S1c-d). This revealed there exist 3 different types of clusters. On the one extreme, for example the RMD_cluster2 contained samples from almost all cancer types, and was very enriched with MSI (MMR-deficient) samples. On the other extreme, there were tissue-specific clusters that contained only a single cancer type (e.g. the liver RMD_cluster8) (Fig 1e). Interestingly, there was also the third, intermediate case with clusters that spanned several, apparently similar cancer types (e.g. RMD_cluster3 with various digestive tract cancers, or the squamous-like RMD_cluster11, with head-and-neck cancers, the non-

melanoma skin cancers and some esophagus and lung cancers) (Fig 1e). Therefore, there is information in the RMD feature vector that can transcend the tissue-of-origin, in this case uniting similar tissues or cell types.

In addition to RMDs bridging cancer types, conversely RMD profiles can be used to subdivide some cancer types such as breast cancer. Breast cancers in RMD_cluster2 have high APOBEC mutagenesis (Fig 1f), thus sharing cluster with MSI samples; APOBEC mutagenesis has been reported to change the regional mutational landscape by preferring early replicating regions [23,24]. similarly as in MSI tumors [1]. Furthermore, most breast cancer samples in RMD_cluster6 (ovarian-like), which have visually distinct RMD profiles from the typical breast-like RMD_cluster9 (Fig 1f-g), are from the triple negative breast subtype, which was reported to be more similar to ovarian cancer by gene expression [25]. Another example of how RMD profiles can be used for subtyping is the head-and-neck cancer, which is split into RMD_cluster11 (squamous-like, includes non-melanoma skin cancers) and RMD_cluster13 (also contains some lung cancers) (Fig 1h).

Overall, even though the RMD profiles are tissue specific, there is systematic RMD variability in certain tumor genomes observed apparently independently of the tissue. This motivated us to devise a method that is able to extract this inter-individual variability from genomic RMD profiles, while robustly accounting for the strong tissue-specific signal in RMD.

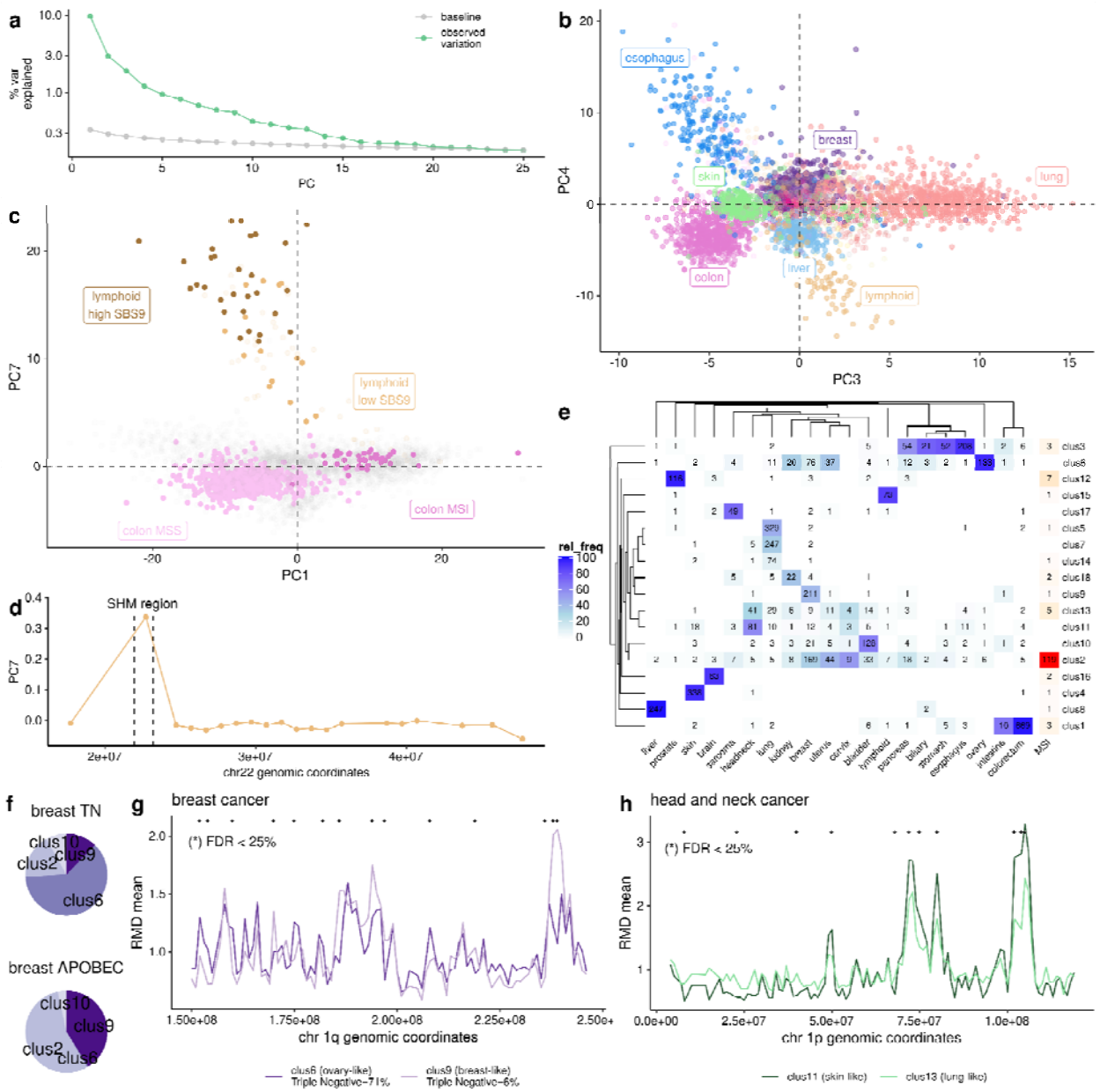

Figure 1. Chromosomal domain RMD variability across tissues and individuals. **a)** Variance explained for the first 25 PCs of a PCA on the RMD matrix (4221 samples x 2542 one-megabase windows), and a baseline (by the broken stick rule). **b)** PC3 and 4 separate various cancer types. **c)** As controls, the PC1 separates MSI versus MSS tumors, and PC7 separates lymphoid samples according to their level of the SHM mutational signature (SBS9). **d)** PC7 window weights for chromosome 22 agree with the known SHM region. **e)** 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)** Mean RMD profiles for breast cancer samples in cluster 6 (n = 76) and cluster 9 (n= 211), shown for chr 1q. **h)** Mean RMD profiles for head and neck squamous samples in cluster 11 (n = 81) and cluster 13 (n= 41), for chr 1p.

A methodology to detect inter-individual variation in regional mutation density

To separate the inter-individual RMD variability from the tissue-specific variability we applied a methodology analogous to that recently used for extracting trinucleotide SNV mutational signatures [22,26,27] however here applied to megabase-sized domains. In brief, non-negative matrix factorization (NMF) is repeatedly applied to bootstrapped mutational data, to find solutions (sets of factors) that are consistent across bootstrap runs. These solutions contain multiple RMD signatures (factors), each with RMD window weights (all 1 Mb windows with varying contributions) and RMD sample 'exposures' or activities (the weight of each tumor for that signature).

To test whether our NMF method is sufficiently powered to capture RMD inter-individual variability, we simulated cancer genomes containing known, ground-truth patterns of RMD that affected a variable number of windows, being present in variable number of tumor samples, and at variable intensity (fold-increase over canonical mutation rates) (Fig S2a, see detailed description in Methods). We ran our NMF methodology for these different scenarios independently. We selected the number of factors and clusters based on a clustering quality measure, the silhouette index (SI), over multiple runs of NMF (Fig S2b) and matching the known ground-truth signatures (Methods, Fig S3). We show an example of an extracted RMD signature compared to its matching ground-truth signature in Fig 2a.

By comparing the different scenarios (Fig S4), encouragingly, we observed that even with a small fraction of samples affected (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 high (>=20%). In addition, we observed that the NMF setup is very robust to the number of windows affected and is usually able to recover RMD signatures that affect as little as 10% of all windows. Out of other characteristics that may affect power to recover RMD signatures, we identified the signature strength/exposure (fold-enrichment) as showing the highest effect, thus the signatures with subtle effects on RMD might not be recovered (Fig S4). In summary, our simulations support that our NMF-based methodology can recover the genome-wide RMD signatures in a wide variety of tested scenarios.

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

We applied the NMF methodology to the somatic RMD profiles of 4221 tumor WGS, here requiring a minimum of 3 mutations/Mb per sample thus restricting to tumors with less noisy RMD profile (as a limitation, we note that this may exclude samples from some cancer types preferentially). In total, we extracted a total of 13 RMD signatures based on the silhouette index that scores the reproducibility of solutions upon 100 bootstraps (Fig 2bc, Fig S5).

In accordance with the above RMD clustering analysis (Fig. 1e), we observed that the RMD signatures from NMF span a continuum from very tissue specific (high Gini index, Fig 2c), to

global signatures (low Gini index). We named ten signatures according to the tissue or tissues they affect (e.g. RMD_upper-GI, RMD_liver), while the three global signatures that affect many cancer types were named RMDglobal1, RMDglobal2 and RMDflat (Fig 2c, Fig S5) (the latter is named by the visually recognizable pattern, and also has in part known mechanisms; see below).

In one extreme, there are tissue-specific signatures (e.g. RMD_skin, RMD_liver) which capture the genomic regions with an increase of mutations only in that particular cancer type (e.g. skin in RMD_skin, or liver and some biliary and some kidney cancers in RMD_liver) (Fig 2c, Fig S5). Windows in these RMD signatures could be used to improve cancer-type classification of tumors based on regional mutation density from WGS data [9].

Between the two extremes, there are signatures present in several cancer types which are apparently similar (Fig 2c, Fig S5). 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; Fig 2c). The RMD_squamous signature spans some squamous lung cancers, head-and-neck cancers, some bladder cancers (consistent with reports based on gene expression data [28], also expectedly some cervical and esophageal tumors, and surprisingly some sarcomas and uterus cancers. Interestingly, one signature, provisionally named “RMD_B.O.P.S.”, spans brain (B), ovarian (O), prostate (P), sarcomas (S), and uterus cancers and so probably reflects a convergent phenotype rather than a common cell-of-origin. These examples suggest that there are commonalities in mutation rates, probably reflecting chromatin organization in the cell-of-origin of tumor types. These commonalities usually reflect anatomical subdivisions or cell type similarity, and shape the RMD profiles of those samples. Our RMD signatures support the proposed uses of RMD profiles for elucidating cell-of-origin and cancer development trajectories (e.g. metaplasia and/or invasion) [11] by matching to chromatin profiles.

In the other extreme, we identified 3 global RMD signatures, which capture the inter-individual RMD variability within most cancer types (Fig 2c, Fig S5). While the profile of RMDflat captures the known “flat” RMD landscape (i.e. a low variation in mutation rates between segments) profile associated with MMR and NER failures [1,2], RMDglobal1 and RMDglobal2 profiles have an apparently complex pattern with their peaks appearing distributed throughout the chromosomes. We can rule out that RMDglobal1 and 2 are due to random noise, because (a) the silhouette index of RMDglobal1 and 2 (measuring robustness of their profile to noise that is introduced in repeated NMF runs) is comparable to the other RMD signatures, and (b) the autocorrelation of their profiles (measuring similarity in weights of consecutive 1 Mb windows) is comparable to the other, tissue-associated RMD signatures (Fig S6a-b).

In addition to the pan-cancer analysis, we ran NMF for each cancer type independently, for the 12 cancer types with more than 100 genomes meeting criteria (Fig S7). All three global signatures can be found also in the per-cancer-type NMF runs (Fig S8). We found signatures in

breast, lung and esophagus with a cosine similarity > 0.84 with RMDglobal1, and in colon, uterus and breast with a cosine similarity > 0.89 with RMDglobal2, supporting that the global RMD signatures capture inter-individual RMD variation recurrently observed in various somatic tissues.

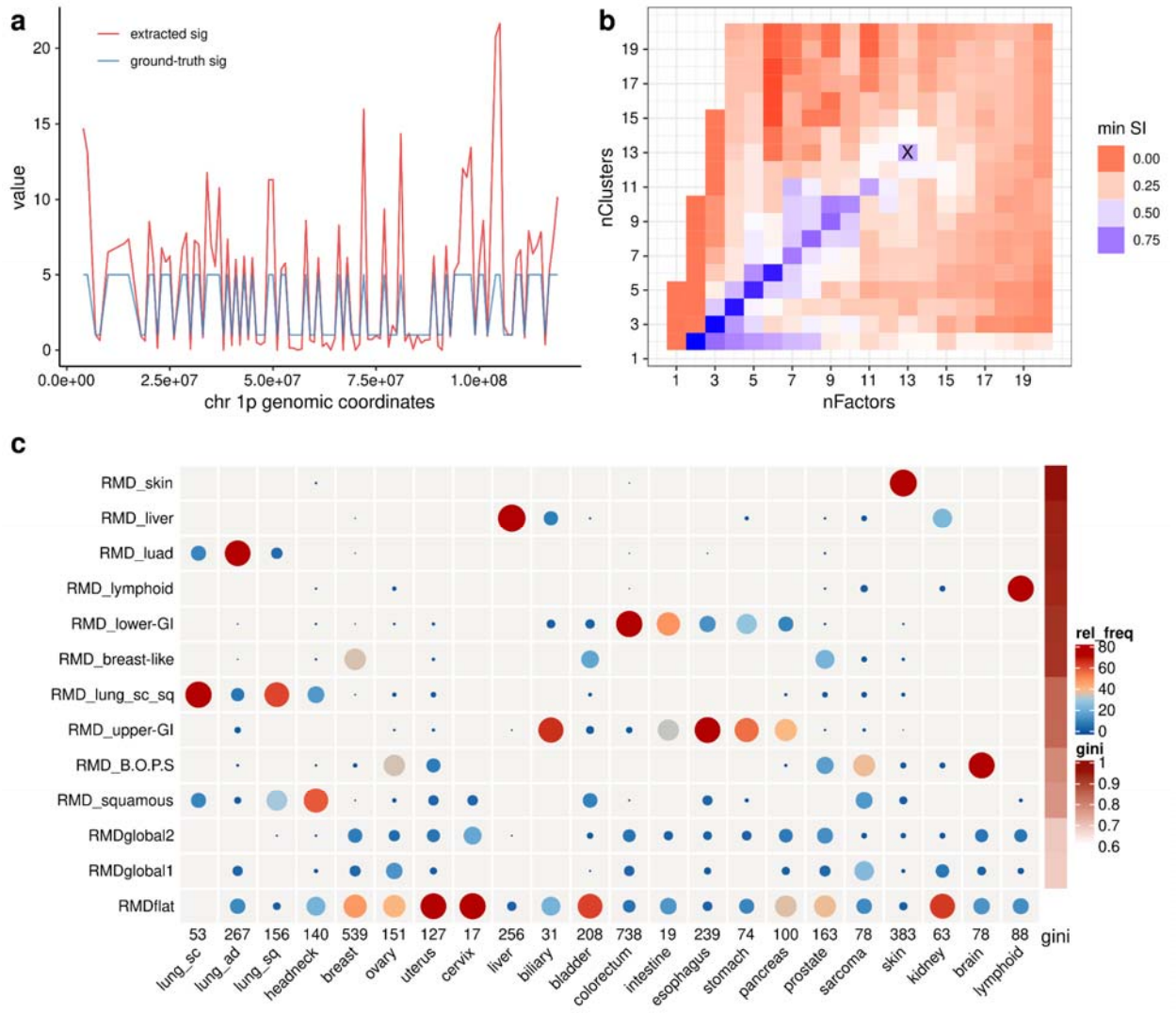

Figure 2. Identifying RMD signatures by an application of a NMF-based methodology to WGS of human tumors. a) Example signature from a simulation study, comparing window weights for an extracted NMF signature and its matching simulated ground-truth signature along chr 1p. See Supplementary Figs 2-4 for additional simulation data. **b)** NMF run on data from 4221 human tumors. Minimum silhouette index (SI) across clusters (RMD signatures) for different numbers of NMF factors and clusters. Selected case (nFactor=13, nCluster=13) is marked with a cross. **c)** Overview of the 13 RMD signatures extracted (rows) and their distribution across different cancer types (columns). The circle size and, equivalently, color corresponds to the fraction of samples from a specific cancer type exhibiting a specific signature (signature exposure >= 0.177). Total number of samples per cancer type written beneath table. The Gini index quantifies the distribution of the signature across different cancer types; higher index means more specificity to few cancer types.

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 [1]. The window weights of RMDflat correlated with the average RT (Fig 3a), 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 3b), as well as bladder samples with mutations in the *ERCC2* gene, participating in the NER pathway (Fig 3b), consistent with previous reports [1,29]. In addition, we observed that tumor samples with high APOBEC signature mutagenesis also showed high exposures to the RMDflat signature (Fig 3b), 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 [23,24,30].

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 [31] The HR deficient samples also presented higher RMDflat exposures, both for BRCA1 and BRCA2 subtypes (Fig 3b). When HR is deficient, there is an increase in the spectrum of the trinucleotide mutational signature SBS3 [22,31] may result from activity of error-prone DNA polymerases [32]. 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) (Fig S9), 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 3c).

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 (Fig S10). 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 [33] that may evades DNA repair mechanisms operative in early-RT domains.

Mutation supply towards major cancer genes is altered by global RMD signatures

Tumors with RMDflat undergo an increase in mutation rates in early replicating, euchromatic regions[23,34]. These regions also have a higher gene density, so we quantified how RMDflat affects the mutation supply to cancer driver genes. In particular, we tested whether there is a difference in mutation density in cancer genes (considering intronic mutations, to avoid effects of selection, and further normalizing to the mutation burden of that chromosome arm to avoid effects of gross CNA; Methods) , between tumor samples with a high RMDflat exposure (top tertile) versus low RMDflat exposure tumors (bottom tertile). 75% of the 460 tested cancer genes [35] undergo an increase in mutation supply from RMDflat-low to RMDflat-high tumors, when compared to the 95th percentile of a randomized distribution (Fig 3d). Conversely, few cancer genes decreased in mutation supply in RMDflat-high tumors (9% are below the 5th percentile of the random distribution). We considered the mutation supply density for 5 examples of common driver genes, for which mutation supply is increased 1.8-2.5 fold between RMDflat-high and RMDflat-low tumors (Fig 3e). Considering for instance the *ARID1A* tumor suppressor gene, located in a lowly-mutated region in chromosome 1p, its mutation supply increased 1.8-fold, 2.1-fold and 2.4-fold in MSI, HRD and APOBEC tumors (all RMDflat-high), respectively, compared to the *ARID1A* baseline mutation supply in tumors without DNA repair deficiencies (Fig 3f). Similarly, the *BRAF* oncogene (where causal mutations are known to be highly enriched in MSI compared to MSS colorectal tumors [36]) has considerably increased mutation supply in the RMDflat-high tumors (Fig 3f).

In summary, we detected the three known mechanisms that cause RMDflat (APOBEC mutagenesis, MMR and *ERCC2* deficiency) and we found an additional cause (HR deficiency) of this phenotype (Fig 3d). The consequence for tumors with RMDflat is an increase in the mutation supply for three-quarters of all cancer genes.

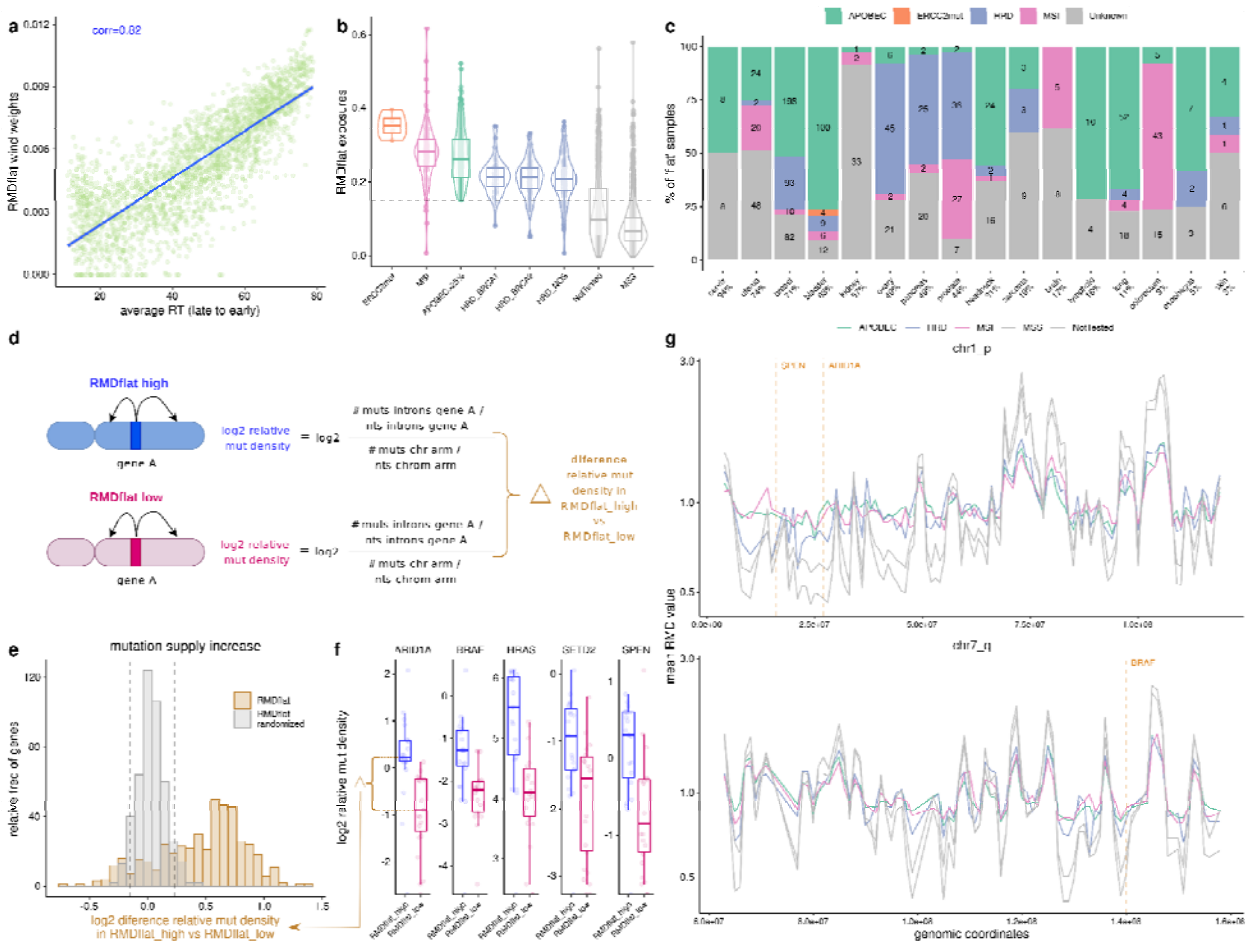

Figure 3. Characterization of the RMDflat RMD signature, which represents a loss of mutation rate heterogeneity. a) Correlation between RMDflat signature NMF window weights and the DNA replication timing (RT) (Repli-Seq average across 10 cell lines). **b)** RMDflat signature exposures (i.e. activities) for groups of tumor samples with various DNA repair failures (MSI, microsatellite instable, indiciating MMR failure; HRD, deficient homologous recombination, by the BRCA1 type or BRCA2 type [31] or not otherwise specified), or high levels of APOBEC mutation signatures. **c)** Percentage of tumor samples with 'flat' mutation rate landscapes (RMDflat exposure>0.177, a threshold that recovers 95% of MSI samples) belonging to each of the DNA repair categories, stratified by cancer type. The percentage of 'flat' samples in each cancer type is indicated in x-axis labels. **d)** Schematic of the mutation supply analysis in panels e-g. **e)** Distribution of the log2 difference in the relative mutation density (intronic) for 460 cancer genes, comparing between RMDflat high tumors and RMDflat low tumors, using the actual values ("RMDflat" histogram) and randomized values ("RMDflat randomized" histogram). **f)** Log2 relative mutation density (normalized to flanking DNA in same chromosome arm, see panel **d**) for RMDflat-high versus RMDflat-low for 5 example genes (common drivers across >=4 cancer types and with highest effect sizes in this test) . Each dot is a 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 example genes from panel **f**.

RMDglobal1 signature increases mutation rate in regions with variable replication timing

We were interested in the mechanism behind the RMDglobal1 signature. To elucidate this, we first tried to predict RMDglobal1 signature (the megabase window weights) from epigenomic features previously reported to associate with megabase mutation rates (reviewed in [34]): replication timing (RT), density of accessible chromatin (DNAse hypersensitive sites, DHS) and ChipSeq data from a variety of histone marks (Fig 4a). We first tried to predict the chromosome-wide profile of the RMDglobal1 signature using the average of each feature across many epigenomic datasets, which failed to predict (Fig 4a). Predicting RMDglobal1 from each RT/DHS/ChipSeq dataset individually fared slightly better, with moderate associations (R^2 ~= 0.2) for certain datasets with regional density of facultative heterochromatin (H3K27me3) and constitutive heterochromatin (H3K9me3) marks (Fig 4a), suggesting a role of heterochromatin organization in determining RMDglobal1.

Remarkably, we observed that RMDglobal1 signature can be highly accurately predicted (R^2 up to 0.7) from certain features most prominently RT, DHS, and the two heterochromatin marks above, however only when predicting using multiple samples jointly (but not when predicting from the averaged feature across the samples (Fig 4a)). This suggests that RMDgloblal1 signature is explained by the variation between the samples for one feature, e.g. differences between the individual RT profiles. We observed the same trend using regional density of chromHMM segmentation states (Fig S11).

The features that best predicted RMDglobal1 were the three RT datasets (Fig 4a): (i) a collection of RT profiles from experiments [RepliChip or RepliSeq] in multiple cell types (expRT, n = 158 samples), (ii) predicted RT in a collection of noncancerous tissues, cultured primary cells and cell lines including cancer and stem cell lines (predRT, n = 597 samples), and (iii) predicted RT in human tumors (predRT-TCGA, n = 410 samples, majority measured in technical duplicate). For the latter two RT datasets, we predicted RT from DHS [37] or ATAC-seq data [38], respectively, using the Replicon software, which predicts RT profiles from local distributions in chromatin accessibility at very high accuracy [39] (see Methods).

Next, we aimed to characterize the variability in RT across individuals that predicts RMDglobal1. By calculating the difference in window-wise RT for each pair of RT samples, and correlating this difference with RMDglobal1 window weights (Fig 4b, Fig S12), we observed that often only two RT profiles can be enough to predict RMDglobal1 using either expRT (max R=0.47), predRT (max R=0.49) and predRT-TCGA (max R=0.62). In predRT, the best correlations are obtained when the difference in RT is when contrasting a pair that consist of one RT from (noncancerous) intact tissue *versus* one RT from primary cultured cells (Fig 4b), This suggests that selection for proliferation-capable stem-like cells when introducing cells into culture may alter RT, and that this altered RT is reflected in mutation rates in RMDglobal1 (see below for further discussion). As an illustrative example in a classification analysis using two selected RT profiles, one from a primary cell culture (ENCFF145RIZ) and one from an intact tissue (ENCFF315RKI), we observed that while the RT profile of each sample alone does not accurately identify RMDglobal1-high windows (Fig 4c), the difference in RT of windows between

these two RT samples can classify the windows with high RMDglobal1 weights from those windows with low RMDglobal1 weights (AUC = 0.82) (Fig 4d).

Cell cycling gene expression-associated changes in RT are relevant for RMDglobal1

To further characterize the source of variability within RT profiles that explains RMDglobal1 signature, we applied a PCA with the predRT-TCGA dataset of RT in TCGA tumors (Fig 4e), and correlated each RT-PC with the profile of RMDglobal1 signature across megabase windows. We observed that the strongest PCs, RT-PC3 and RT-PC4 are either tissue-associated, separating breast from kidney and brain tumors, or represent the average RT profile (RT-PC1, RT-PC2) (Fig S13). However, the RT-PC5 does not have a strong tissue bias but correlates strongly with RMDglobal1 (R=-0.49) (Fig 4e, Fig S13c). Indeed, when we checked the RT profiles for the top RT-PC5 positive and RT-PC5 negative tumor samples, we observed RT differences in the RMDglobal1-relevant windows (Fig S14). The next best correlation was with RT-PC6 (R=0.35).

These RT-PCs summarize the global variation in the RT program between tumors of the same cancer type, and also predict RMDglobal1 global variation in mutation distribution. To interpret the RT-PCs, we asked how gene expression changed between the TCGA tumor samples with high values of a RT-PC versus tumors with low values. We considered the RHP gene sets, representing gene expression programs that are variable in a coordinated manner between individual cells, and that were recurrently observed across different cancer cell lines [12].

In particular, RT-PC5 correlates strongly with gene expression of cell cycle genes from the RHP gene sets [12] (there are two RHP sets of cell cycle genes, the G2/M and G1/S, and both correlate at p= 9e-40 and 1e-14, respectively) (Fig 4f, Fig S13d), while the other RHP gene sets correlate less (next strongest is p-value=5e-05) Consistently, also the RT-PC6, which has a more subtle correlation to RMDglobal1, also correlates with the cell-cycle RHP gene expression programs (Fig 4f, Fig S13d). This suggests that the RMDglobal1 regional mutability pattern reflects the RT program alterations associated with variable speed of cell cycling across different tumors.

To further understand the biology of the systematic variation in RT captured by the RT-PCs relevant to mutation rates, we projected the predRT and expRT data into the existing predRT-TCGA PC coordinate system. RT-PC5 separated tissues *versus* cultured primary cells in predRT samples (Fig 4g). One possible interpretation is that RT-PC5 captures the effect of tissue culture conditions on RT profiles, however this is unlikely because there is a considerable spread within the cultured cells group, which span across the tissue-side of RT-PC5 on the one extreme of RT-PC5 and cell line side on the other extreme of RT-PC5 (Fig 4g). The other interpretation is that RT-PC5 captures the RT program of proliferation-capable, stem-like cells, which are normally a minority in an intact tissue, but are selected during establishment of cell culture; this is consistent with the above-mentioned cell cycling RHP gene expression program association with RT-PC5 and RT-PC6 and so we favor this interpretation. Next, the RT-PC5 also separated healthy *versus* cancerous cells in the expRT samples (considered for blood cells, where both healthy and tumor was available (Fig 4h)). This suggests that this property

captured in RT-PC5 is more prominent in cancerous cells than in normal cells, again consistent with the property being related with cell cycling, which is often unchecked and accelerated in cancer cells.

In summary, RT-PC5 separates intact tissue samples or tumors with lower cell cycle gene expression on one side, and cultured primary cells or tumors with higher cell cycle gene expression on the other side. This suggests that the windows with high RMDglobal1 weights are those that undergo changes in RT in more proliferative/stem cell-like samples compared to less proliferative/differentiated cell-like samples.

Within a subset of the expRT data, changes in RT were studied previously[40], reporting late-to-early (LtoE) and early-to-late (EtoL) RT changes between noncancerous samples (lymphoblastoid cell lines) and cancers (leukemias and cell lines). Interestingly, their pre-calculated ratio of LtoE/EtoL strongly correlate with our RT-PC6 (R=-0.72) and RT-PC5 (R=0.63) (Fig 4i), adding evidence that RMDglobal1 is linked to the genome-wide changes in RT that occur during cancerous transformation.

As a validation, we saw the same trends when we performed the PCA in predRT initially (i.e. using a mix of tissues and cell types, rather than only tumors in predRT-TCGA), and then projected the expRT into it (Fig S15). Of note, the expRT-PC that reflects developmental changes as reported earlier [41] does not correlate with RMDglobal1 (Fig S16), meaning that RMDglobal1 mutagenesis pattern does not relate to embryonal-characteristic patterns of RT.

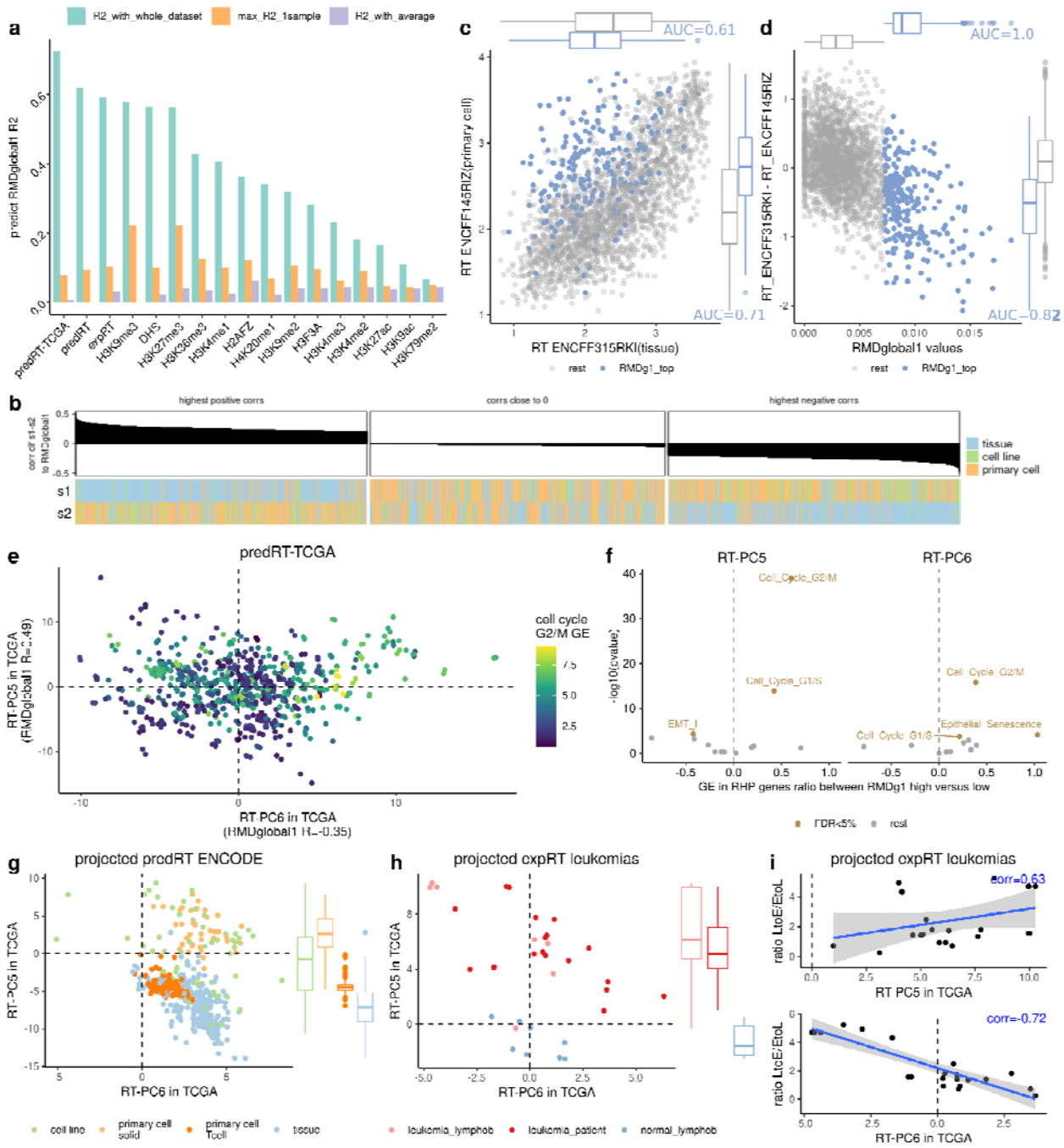

Figure 4. RMDglobal1 signature is linked to regional variability in replication timing. a) Adjusted R^2 of a regression predicting RMDglobal1 window weights from various epigenomic features (x axis) using either the whole dataset jointly, or selecting the maximum R2 of each sample in the dataset individually, or using the average values of the feature across the samples in the dataset. **b)** Correlation between RMDglobal1 signature, and the difference between each pair of RT profiles (all combinations tested). Panel shows 1st decile (highest positive R), 5th decile (R close to 0) and 10th decile (highest negative R) deciles ordered by correlation. **c)** RT profiles for two selected samples, where dots are megabase windows, colored by their weight in RMDglobal1 signature (top decile in blue). RT of each sample individually is modestly predictive of RMDglobal1 (AUCs for discriminating top-decile windows are listed

next to boxplots of RTs). **d)** Difference between the two RT profiles in panel c (on y axis) is predictive of the RMDglobal1 signature (see AUC for discriminating top-decile windows). **e)** A PCA on a matrix of predicted RT from TCGA tumor samples. RT-PC5 and RT-PC6 are shown because of their correlation with RMDglobal1 (R=0.49 and -0.35, respectively). Points are colored by the mean gene expression in the cell cycle G2/M RHP module, in each TCGA tumor sample. **f)** Association between RT-PC5-high (top tertile) versus RT-PC5-low (bottom tertile) with the expression of genes in various RHP programs [12], and same for RT-PC6. **g)** Predicted RT from ENCODE data with tissues, primary cells and cell lines (predRT-ENCODE) was projected into PCs of the tumor predRT-TCGA data. **h)** Projection of experimentally determined RT data for leukemias and normal blood cells into the same PCs of predRT-TCGA data. **i)** Correlation between the projection of expRT leukemia samples in RT-PC5 and RT-PC6, and the ratio of late-to-early and early-to-late regional RT changes reported previously [40].

RMDglobal1 signature associates with RB1 loss and affects regions that undergo chromatin remodeling

To identify events that may drive the changes in RT we found linked with cell cycling, we performed an analysis to detect somatic copy number alteration (CNA) events and deleterious point mutations are associated with RMDglobal1 exposure, while adjusting for cancer type and for confounding between linked CNAs (qq-plots in Fig S17; Methods for details). Here, we considered 1543 chromatin modifier genes, cell cycle genes, DNA replication and repair genes and cancer genes, compared against a background of 1000 control genes (Methods).

For CNA, we found a strong positive association of RMDglobal1 with *RB1* deletion (FDR=0.05%, and better p-value than all control genes) (Fig 5a-b, Fig S18a). Because CNA often affects large segments, we also checked associations with *RB1* neighbors (Fig 5c), noting that *RB1* is at the CNA peak (by mean estimated copy-number across tumor samples), meaning it is the likely causal gene. RMDglobal1 association with *RB1* is gene dosage dependent (Fig S18b). Consistently, we see that the effect of *RB1* mutations shows a trend in the same direction as *RB1* deletions, even though it is nonsignificant (*RB1* mutations are rarer) (Fig S18c). As independent supporting evidence, we identified deletions in *CDK6*, a negative regulator upstream of *RB1*, negatively associated with RMDglobal1, also having a stronger p-value than any of the control genes considered (Fig 5a).

In addition to its effects on cell cycle regulation, RB1 has additional important roles in chromatin organization [19,42–44]. In specific, *RB1* deletions were reported to change heterochromatin marks H3K9me3 and H3K27me3 in regions enriched at subtelomeres, and this associates with regional propensity to DNA damage [19]. These same two marks we found to be more highly correlated to RMDglobal1 than other tested marks (Fig 4a), and interestingly we find that RMDglobal1 window weights are also strongly enriched approximately 5 Mb nearby telomeres (Fig 6h). Notably, the changes in regional H3K9me3 profile when *RB1* is wild-type *versus* in isogenic *RB1* k.o. cells [19] predicted RMDglobal1 signature (adjusted R^2=0.29), and so did changes in regional H3K27me3 albeit subtly (adjR^2=0.18) (Fig 5d, Fig S19ab). The genome regions with 10% highest weights in RMDglobal1 are the ones where the level of H3K9me3 heterochromatin mark is more likely to be asymmetrically altered upon RB1 disruption [19] (off-diagonal dots in Fig 5d). This indicates that loci where heterochromatin is remodeled upon *RB1*

loss-of-function [19] significantly overlap with loci where RMDglobal1 mutation rates change in many tumors, further implicating RB1 in shaping the mutation rate landscape.

As for CNA, we also tested associations between the presence of deleterious point mutations in cancer and chromatin and DNA repair genes and the exposure to the RMDglobal1 mutagenic pattern. Here, we found the *KRAS* mutation to strongly positively associate with RMDglobal1, at FDR=0.1% (Fig 5e, Fig S19c), and this is observed consistently across individual cancer types (Fig S19c) and significantly in colon, uterus and bladder (see Fig S19d-e legend for comment on lung adenocarcinoma). Of note, the *KRAS* gene was reported to act downstream of *RB1* loss-of-function with *RB1* in developmental and in tumor mouse phenotypes [45,46]. 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, RMDglobal1.

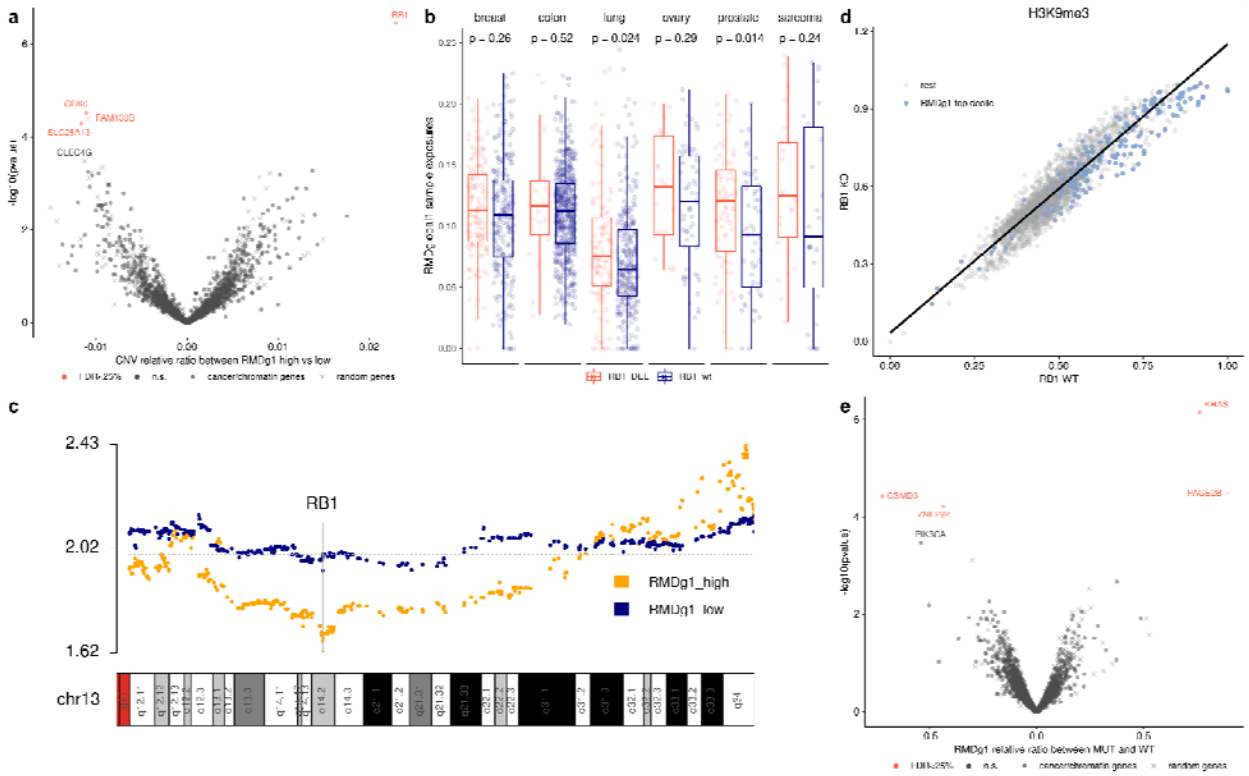

Figure 5. Genetic alterations associated with the activity RMDglobal1 signature. a) Associations between CNA deletions and tumors with higher RMDglobal1 exposures in a pan-cancer analysis, adjusting for cancer type and for global CNA patterns (Methods). N=1543 cancer genes and chromatin-related genes are shown (dots), as well as a 1000 set of randomly chosen genes (crosses). **b)** Differences in RMDglobal1 exposures between RB1 deletion (-1 or -2 deletion) and wt for several cancer types (those with the highest number of samples with *RB1* deletion); remainder in Fig S18. **c)** Mean local CN profile in groups of tumors, grouped by RMDglobal1 high and low, of the segment of chromosome 13 containing the gene *RB1*. Each dot represents one gene. **d)** Correlation between the H3K9me3 heterochromatin profiles for samples with *RB1* knock-out ("KO") versus wild-type ("WT"). Each dot represents a window, colored by RMDglobal1 window weight top decile versus the rest of the windows. **e)**

Associations between deleterious SNV and indel mutations in the same sets of genes as in panel **a**, and the RMDglobal1-high versus RMDglobal1-low activity of tumor samples, in a pan-cancer analysis.

Motivated by these associations between RB1-caused regional heterochromatin changes [19] and the RMDglobal1 regional mutation rates, we further investigated the variation in the H3K27me3 and H3K9me3 marks across datasets in ENCODE (Fig 6a). To characterize the regional heterochromatin variability, we performed a PCA on the profiles of the two marks and the predRT together. The resulting heterochomatin-PC4 (het-PC4) correlated best with RMDglobal1 window weights (R=0.53). As above, the difference in the three features (H3K9me, H3K27me3, RT) separated the proliferative, putatively stem-like samples (het-PC4 positive) *versus* the rest of the samples (het-PC4 negative) (Fig 6b). The stem-like samples (het-PC4 positive) are later replicating (relative increase in RT [het-PC4 high vs low] = 61%) and have higher H3K27me3 and H3K9me3 in RMDglobal1 top windows (relative increase of 55% and 78% respectively) (Fig 6c). In summary, our analyses suggest that the chromosomal domains with highest RMDglobal1 weights become later-replicating and more heterochromatic in more stem-like cells (cell lines/primary cells), associated with an increase of relative mutation rates in these domains.

Gene regulation and chromatin compartments associated with the RMDglobal1 mutation rate phenotype

The regional variability in RT and heterochromatin marks, as reflected in variable somatic RMD in tumors, suggests the existence of concomitant changes of regional gene expression, because early RT was reported to be broadly associated with higher gene expression [1]. Therefore we asked if there are coordinated changes in gene expression levels in certain windows between the RMDg1-high and RMDg1-low cancers. Indeed, we found several windows with gene expression upregulation and downregulation between RMDg1-high and low cancers (FDR < 25%). The windows with coordinated gene expression downregulation are enriched in higher values of RMDglobal1 window weights, compared to the windows with non-coordinated gene expression changes (Wilcoxon test, greater; downregulation p-value = 0.03; there is a nonsignificant trend for coordinated upregulation) (Fig S20a). These regional changes in gene expression are consistent with chromatin remodeling affecting various chromosomal domains, which is also mirrored in regional mutation rates (Fig S20b).

To additionally characterize the regions affected by the chromatin remodelling, we analyzed data from diverse types of genomic assays from various studies (Table S3) that reported some correlations with RT. We compared the regional density of these various features with our RMDglobal1 window weights (Table S3). Correlations were noted with Hi-C subcompartments (Fig 6d), inferred from long-range chromatin interactions at fine resolution (25 kb) [47]. In particular, the B1 subcompartment was associated with RMDglobal1; this subcompartment replicates during middle S phase, and correlates positively with the Polycomb H3K27me3 mark (Fig S21) and negatively with H3K36me3 suggesting that it represents facultative heterochromatin [47]. Next, we observed a correlation with two SPIN states (Fig 6e), derived by integrating nuclear compartment mapping assays and chromatin interaction data [48].

RMDglobal1 signature regions are enriched in the two “Interior repressed” SPIN states, marking regions that are inactive, however unlike other inactive heterochromatic regions they are located centrally in the nucleus, rather than peripherally (next to the lamina). Additionally, RMDglobal1 important windows are enriched in subtelomeric regions (Fig 6f). In sum, the windows with higher weights in RMDglobal1 signature are enriched in subtelomeric regions, the B1 facultative heterochromatin subcompartment, and nuclear interior located, repressed chromatin states.

Since RMDglobal1 captures a redistribution of mutation rates genome-wide, we predicted that this will affect the supply of mutations to some cancer genes. To quantify this, we performed a similar analysis as for the MSI-associated RMDflat above; for RMDglobal1 shown in (Fig 6g-i). When compared to a randomized baseline (95th percentile of the random distribution used as cutoff), 28% of the 460 cancer genes suffer a significant increase of mutation supply when comparing RMDglobal1-low (bottom tertile) and -high (top tertile) tumor samples. Regarding the effect size of increase, these genes increase mutation rates on average by 1.21-fold between the RMDglobal1-low *versus* high tertile tumors. The mutation rate density is shown for 5 example genes with high fold-difference in Fig 6h, where for instance the median mutation rate for the *ATM* tumor suppressor increases by 1.18-fold, and for the *KMT2C* tumor suppressor by 1.79-fold, in the top tertile by RMDglobal1 signature of mutation redistribution.

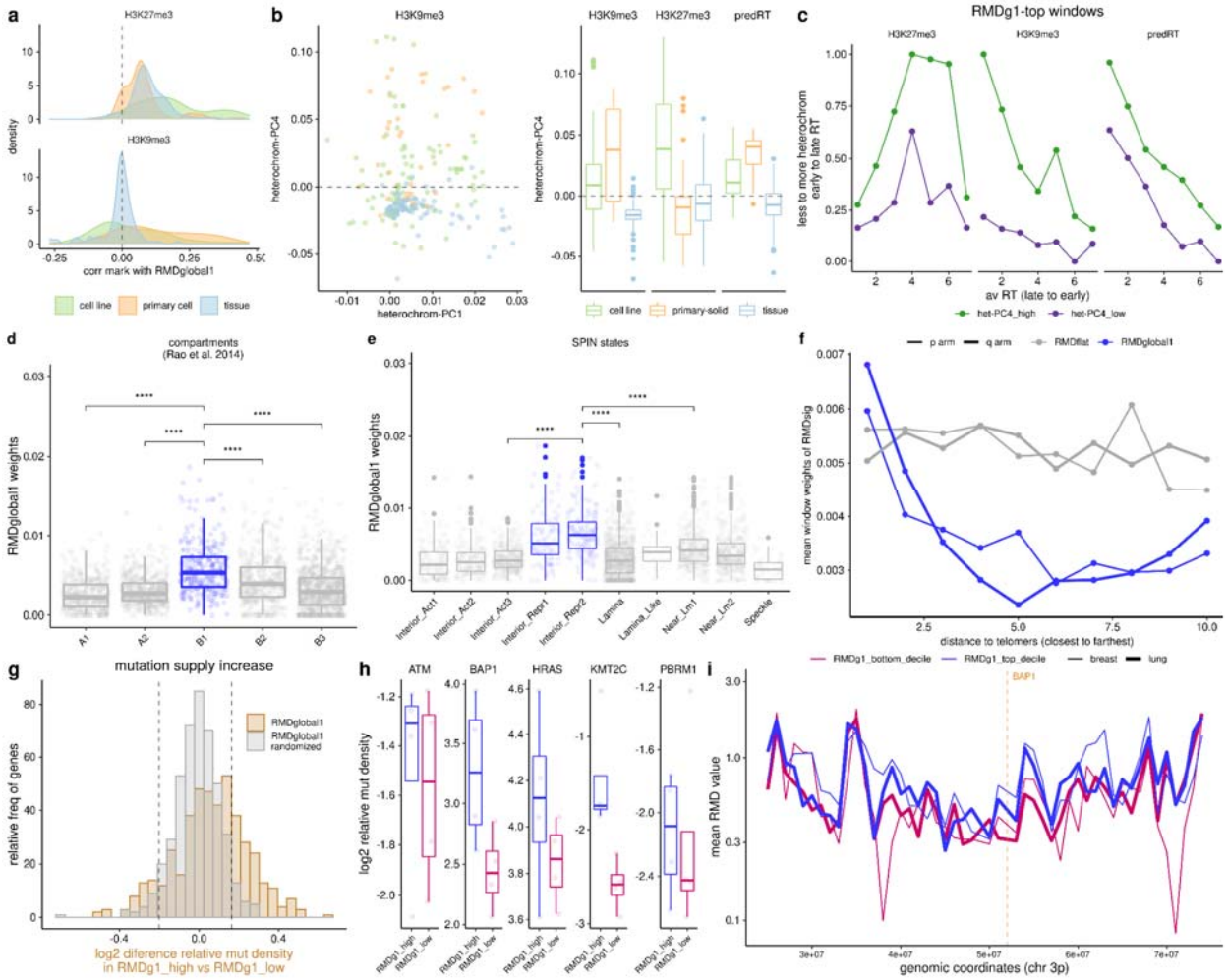

Figure 6. RMDglobal1 mutation rate redistribution is linked with chromatin remodeling. a) Heterochromatin marks correlation with RMDglobal1. Distribution of the correlations between H3K9me3 and H3K27me3 profiles in various ENCODE datasets of healthy tissues/primary cells/cell lines, and the RMDglobal1 signature extracted from tumor mutations. This shows a wide spread of correlations, with some examples of cell lines or primary cells with high correlations of heterochromatin profiles with RMDglobal1. **b)** A PCA was performed on the predicted RT and the heterochromatin marks from ENCODE data. Left panel shows the heterochromatin PC4 (het-PC4) selected for its high correlation with RMDglobal1 (R=0.52) and het-PC1 shown for highest amount of variance explained for H3K9me3. Right panel shows the het-PC4 distribution across different cell types for the 3 features. **c)** Mean predicted RT, H3K27me3 and H3K9me3 across PC4-high versus PC4-low groups in ENCODE data, split by RT bins. **d)** RMDglobal1 signature across different Hi-C nuclear subcompartments from reference [47]. **e)** RMDglobal1 signature across different SPIN nuclear compartmentalization states from [48]. **f)** RMDglobal1 and RMDflat signature window weights compared to distance to telomeres. **g)** Distribution for the difference in mutation density (see Fig. 2d), shown for 460 cancer genes, comparing between RMDglobal1-high and low tumors, using the actual values of RMDglobal1 and as a baseline randomized of RMDglobal1**.** Vertical lines show 5th and 95th percentile of the randomized distribution. **h)** Mutation density for RMDglobal1-high versus low tumor samples (here, top tertile versus bottom tertile) for 5 example genes (drivers in >=4 cancer types and with the highest effect size); dots are cancer types. **i)** Mean RMD profile on chromosome 3p 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 (example gene in panel **h**).

A TP53-associated RMDglobal2 signature reduces relative mutation rates in late replicating regions

RMDglobal2 signature mutations follow a distribution similar to the canonical RMD landscape, increasing mutation density in late replication, except for a set of very late RT windows , which acquire fewer mutations than expected from RT (Fig 7ab). Mutation density increases near 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 fit (Fig 7c, Fig S22). Therefore, qualitatively the canonical RT-associated RMD landscape is preserved regardless of RMDglobal2 being low or high. However RMDglobal2 changes the shape of association to RT, by exaggerating (or suppressing) the more prominent peaks in regional mutation rates, but not affecting the minor peaks.

We aimed to identify driver event behind this redistribution of mutations by testing for associations of RMDglobal2 high (top tertile) versus low (bottom tertile) samples, and genetic events (CNAs, deleterious mutations) in cancer driver, DNA repair and chromatin modifier genes. Strikingly, we found *TP53* mutation to be uniquely strongly associated with RMDglobal2 signature (effect size = 1.27, FDR = 9e-10) (Fig 7d). As supporting evidence, we found that *TP53* deletions also positively associated (Fig 7e) and, independently, the known amplifications that phenocopy TP53 loss (*MDM2*, *MDM4* and *PPM1D* oncogenes) are also positively associated with RMDglobal2 RMD signature exposures (Fig 7e, Fig S23). This rules out that the *TP53* driver mutation is merely the consequence of RMDglobal2 redistribution, and provides evidence for a causal effect of *TP53* inactivation.

Since *TP53* mutations were reported to be associated with increased burdens of CNA events [49], we tested whether RMDglobal2 RMD signature could be due to confounding from a multiplicity of focal CNA events (we note our method for RMD analysis does control for confounding by arm-level CNAs, Methods), which can 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 likely does not simply reflect changes in local DNA copy number (Fig S24).

RMDglobal2 signature describes variation in certain genome regions, which may affect mutation supply to genes 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 7f). When compared to randomized data (5th percentile), 26% of cancer genes exhibited decreased mutation supply; only 6% genes exhibit an increased mutation supply with high RMDglobal2 (Fig 7f-g). As an example, we show the mutation density of *ARID1A* and *GATA3*, which decreased in mutation supply (as above, measured using intronic rates; the decrease implies they are below the 5th percentile of the randomized distribution) with high RMDglobal2 (Fig 7h).

We hypothesized that apparent genetic interactions like this mutual exclusivity example might arise due to redistribution of mutations altering local mutation supply to genes. We thus considered 13 genes mutually exclusive with *TP53* mutations [50] (note that *TP53* loss is strongly associated with RMDglobal2) and found that nearly half (6/13) of these genes were below the 5th percentile of the random distribution (Fig 7f-g). Upon inspection of the raw RMD profiles for RMDglobal2 high and low tumors for several cancer types we noted a difference in the region where *ARID1A* resides (Fig 7i). Overall, this illustrates how a global redistribution of mutation rates, here mediated by TP53 loss, can create apparent genetic interactions that may not indicate selection on functional effects of the genetic interaction. Thus, regional mutation rates, which vary extensively between tumors, should be explicitly controlled for in statistical studies of epistasis in cancer genomes.

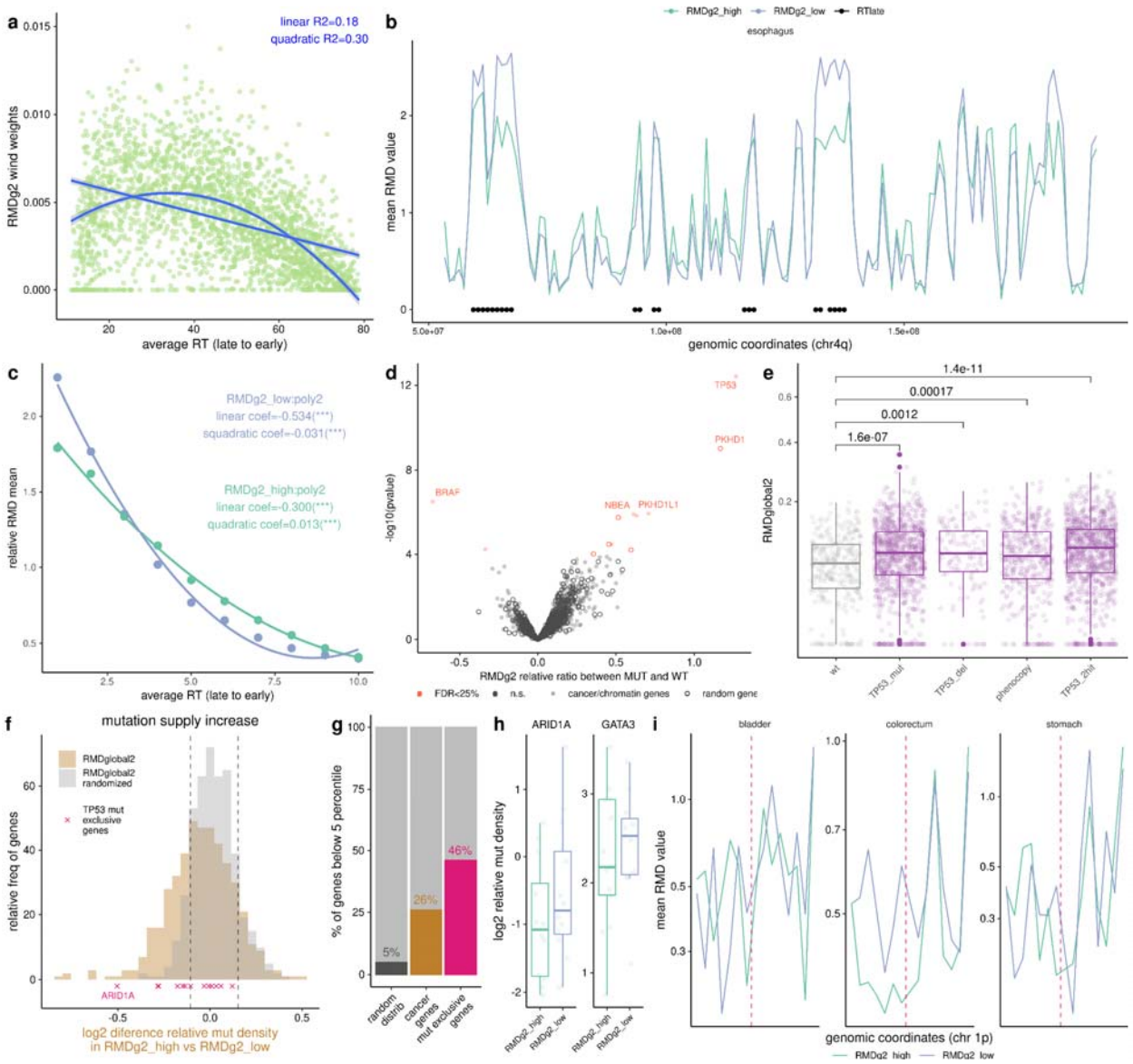

Figure 7. A TP53-associated mechanism underlies the RMDglobal2 mutation rate redistribution pattern. a) A quadratic association of RMDglobal2 signature with the average replication timing. **b)** Mean RMD profiles in chromosome 4q for the RMDglobal2-high versus low tumor samples in esophagus cancer. Latest RT windows (avRT<20) marked with black dots. **c)** Relative RMD mean profile across 10 RT bins for tumors that are RMDglobal2-high (RMDglobal2 exposures > 0.17) versus RMDglobal2-low (RMDglobal2 exposures < 0.01), showing a linearization of the link between RT and mutation rates in RMDglobal2-high. **d)** Associations between deleterious mutations in known cancer genes and chromatin-related genes (dots) and a control set of randomly chosen genes (hollow circles), and RMDglobal2 exposures in samples (p-values from Z-test on regression coefficient). **e)** RMDglobal2 signature exposures of tumor samples stratified by: wild-type for *TP53* (wt), TP53 with 1 mutation (TP53_mut), TP53 with 1 deletion (TP53_del), TP53 loss phenocopy via a amplification in *MDM2*, *MDM4* or *PPM1D* (TP53_pheno), or *TP53* with any two hits of the previously mentioned alteration (TP53_2hit). **f)** Distribution of the log2 difference in the relative mutation density (intronic) for 460 cancer genes, comparing between RMDglobal2 high tumors and RMDglobal2 low tumors, using the actual values

("RMDglobal2" histogram) and randomized values ("RMDglobal2 randomized" histogram). Position of the genes mutually exclusive with TP53 marked with crosses. **g)** Percentage of genes above 95 percentile of a random distribution for the random distribution, cancer genes and TP53 mutually exclusive genes. **h)** Log2 relative mutation density (normalized to flanking DNA in same chromosome arm, see Fig 3d) for RMDglobal2-high versus RMDglobal2-low for 2 example genes (TP53 mutually exclusive genes above the 95 percentile). Each dot is a cancer type. **i)** Mean RMD profile across theRMDglobal2-high versus low groups in a region of chr 1p. Vertical lines mark the position for the *ARID1A* gene.

Discussion

Even though there clearly exists a common, canonical regional mutation rate landscape shared across human cells, there are RMD patterns superimposed that are tissue-specific. This is consistent with the fact that tissues have different RT programs and chromatin landscapes. Here we systematically characterized patterns of regional redistribution of somatic mutations independent of tissue, identifying two new global RMD patterns and possible underlying mechanisms.

We demonstrate how an NMF-based approach can deconvolute the RMD mutational patterns that compose the final regional mutagenesis 'portrait' of each individual tumor. The method is related for those applied to trinucleotide mutational signatures, however it also rigorously adjusts for the confounding by these trinucleotide signatures. Of 13 RMD signatures, expectedly the majority were tissue-associated.

Some of the tissue-related RMD signatures may bridge various cancer types, usually reflecting known biology, however the surprising RMD signature ("B.O.P.S.") has high activity in many brain samples but also in ovary, uterus, prostate and sarcoma 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 RMD signature that collects diverse RMD patterns that current NMF methodology does not resolve well, possibly due to lack of power. However, this RMD B.O.P.S. pattern was similarly robust (by autocorrelation across windows) as the other tissue-specific patterns (Fig S6) and so 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 [13] suggested a module of cilium/cytoskeleton-related genes common to some ovarian cancers, glioblastoma, uterine cancer and lung adenocarcinoma, thus the tissue spectrum corresponds to B.O.P.S. (we do note the lung tissue is separate in our RMD analysis). We acknowledge that, as has recently occurred with the trinucleotide mutational signatures [51], some of the initially proposed signatures such as RMD_B.O.P.S. may, with arrival of more data, be able to be 'split' into component RMD signatures that more precisely match tissue identity. Overall, the RMD features provide an important tool for understanding the relatedness of cancer types and the chromatin organization in the cell-of-origin of cancers [11,52].

Here, we identified 3 robust 'global' (i.e. largely independent of tissue) RMD signatures of redistribution of mutation rates occurring in human cancer. As expected, we recovered the

redistribution of mutations towards a flatter mutational landscape ("RMDflat") that was previously described for DNA MMR failures in tumors, cell lines and xenografts [1,53,54], validating our methodology. A simulation study further supports the broad adequacy of the NMF method, with the number of mutations generated by a RMD signature being a limiting element in identifying the signature. With a higher number of tumor samples, additional RMD patterns may become sufficiently represented to be recognized by NMF, and our RMD catalog will likely be extended with rarer RMD patterns, as was the case for trinucleotide mutational signatures [27,51]. A limitation of the current implementation of our method is that low mutation burden tumors are not analyzed; increasing WGS sequencing depth and so power to detect subclonal variants may alleviate this constraint.

Of the widespread global RMD signatures, RMDglobal1 causes a genome-wide redistribution of mutations predominantly in inactive regions enriched in Polycomb facultative heterochromatin mark (H3K27me3), and centrally located in the nucleus (i.e. not lamina associated heterochromatin). These regions showed variable RT programs and variable heterochromatin state, comparing a more proliferative/stem-like group of samples versus a less stem-like group of samples. Consistently, the corresponding RMDglobal1 mutational pattern was associated with genetic alterations implicated in cell cycle disturbances. Of note, *RB1* is involved in cell cycle but also has important roles in chromatin organization [19] and so may affect this RMD pattern in multiple ways. Together, our analyses converge onto a model where due to more rapid cycling in tumor cells (e.g. caused by oncogenic drivers) and/or loss of cell cycle control (e.g. caused by *RB1* loss-of-function), chromatin is remodeled in facultative heterochromatin locations and RT program changed, and as a consequence the mutation rates in those regions are altered.

One question that arises is why those specific regions undergo the chromatin remodeling and RT change. We found these RMDglobal1 regions to be enriched in the B1 Hi-C subcompartment, which is the most dynamic (less conserved) subcompartment across cell lines [55]. Additionally, chromatin remodelling and increase of risk to DNA damage was reported to affect subtelomeric regions upon *RB1* disruption [19]. Collectively, this suggests that chromatin state in those facultative heterochromatin regions are likely more malleable and prone to change upon different processes, either developmental or cancerous in nature.

The second global change in the mutational patterns we identified (RMDglobal2) occurs independently of the above and can be described as a sharp relative reduction of mutations in latest RT regions, associated with loss of TP53 activity via mutation, CNA or phenocopying events. Since *TP53* mutations are very common in tumors, the impact of this redistribution of mutations to many other cancer genes may be widespread. This study provides examples of how an alteration in one gene -- here, deletions in RB1 or mutations in *TP53* -- can affect future evolutionary scenarios: by 'redirecting' the regional mutation supply away from one set of genes and towards another.

In conclusion, our large-scale analysis recovered the known differences in mutation density between tissues and identified three robust global RMD signatures of mutation rate variability

across chromosomal domains. The global redistribution of mutations can have an important impact in mutation supply on cancer genes at their affected regions, increasing their likelihood to acquire a deleterious mutation.

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 [56] Data portal (https://dcc.icgc.org/pcawg). Second, we obtained 4823 WGS somatic SNVs from the Hartwig Medical Foundation (HMF) project [57] (https://www.hartwigmedicalfoundation.nl/en/). Third, we downloaded 570 WGS somatic SNVs from the Personal Oncogenomics (POG) project [58] 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 [59,60] and bam files for 758 tumor samples from the MMRF COMMPASS project [61] from the GDC data portal (https://portal.gdc.cancer.gov/). Somatic variants were called using Illumina's Strelka2 caller [62], 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 [63]. 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 [63]. 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 [64] with alignability <1.0. In addition, we removed the regions that are unstable when converting between GRCh37 and GRCh38 [65] and the ENCODE blacklist of problematic regions of the genome [66].

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 [67]. 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 tos 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 <u>RMD</u>, 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 <u>RMD applied to NMF analysis</u>, 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 [35] (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 [35] and Cancer Gene Census list [68], and furthermore we included genes associated with chromatin and DNA damage [69]. As control, we used a subset of 1000 random genes selected as in [69].

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.

ChromHMM chromatin states. We downloaded the 25 ChromHMM states segmented files ("imputed12marks_segments") for the 129 cell types available from Roadmap epigenomics [70](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 [71], Constitutive and Developmental RT domains [72], RT changes upon overexpression of the oncogene *KDM4A* [73], RT signatures of replication stress [74], RT signatures of tissues [41], RT states [75], changes in RT upon *RIF1* knock-out [76] and RT changes due to RT QTLs [77]. In addition, we downloaded data for variability in DNA methylation [15,78], HMD and PMD regions [16], CpG density, gene density, lamina associated domains (LADs), asynchronous replication domains [79], early replicating fragile sites [80], SPIN states [48], A/B subcompartments [47], DHS signatures [81] and H3K27me3 and H3K9me profiles for *RB1* wild-type and knock-out [19]. 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) [72] in multiple human cell types (n = 158 samples). In addition, we predicted RT using the Replicon software [39] 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 [38]. We used Replicon tool with the default settings.

Analysis of coordinated gene expression changes

For the genomes from the HMF data set, we downloaded gene expression data (as adjusted TPM values) from Hartwig [57], 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. We tested whether the gene expression values of the genes within one window show an increase or decrease compared to their flanking windows in RMDglobal1 high (exposure >= 0.13) versus RMDglobal1 low (exposure < 0.06) tumor samples using a linear regression model:
gene_expression (adjTPM) ~ is_RMDglobal1 + is_window + is_RMDglobal1:is_window + tissue

In this analysis, we removed from the datasets samples with high RMDflat or with high RMDglobal2 value (exposure > 0.15). We used samples from breast, colorectum, lung, ovary and skin because they had >=5 samples in both categories (RMDglobal1 high and low). To analyze the coordinated changes in gene expression we checked the coefficient and p-values of the interaction term *is_RMDglobal1:is_window.*

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. We compared the gene expression between RT-PC5 (and RT-PC6) high and low for a group of pathways which has been reported to be related with recurrent heterogeneity across cell types [12] using a regression model. We binarized RT-PC5 (and RT-PC6) by dividing each into tertiles and keeping the samples in the 1st tertile to be compared versus the samples in the 3rd tertile. We applied a regression for all the genes in each RHP gene set separately. We controlled for tissue by including it as covariate. The regression formula is:

gene_expression (TPM) ~ is_RT-PC5 + tissue

We considered the regression coefficient and its p-value of the variable is_RT-PC5. We applied the same analysis for RT-PC6.

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.

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. Kinker, G. S. *et al.* Pan-cancer single cell RNA-seq uncovers recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
13. 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).
14. Du, Q. *et al.* Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat. Commun.* **10**, 416 (2019).
15. 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).
16. Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).
17. Brinkman, A. B. *et al.* Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation. *Nat. Commun.* **10**, 1749 (2019).
18. 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).
19. 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).
20. 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).

21. 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).
22. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
23. 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.
24. Seplyarskiy, V. B. *et al.* APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res.* **26**, 174–182 (2016).
25. Jönsson, J.-M. *et al.* Molecular Subtyping of Serous Ovarian Tumors Reveals Multiple Connections to Intrinsic Breast Cancer Subtypes. *PLOS ONE* **9**, e107643 (2014).
26. Supek, F. & Lehner, B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell* **170**, 534-547.e23 (2017).
27. Degasperi, A. *et al.* Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. *Science* **376**, abl9283 (2022).
28. 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).
29. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
30. Chen, J., Miller, B. F. & Furano, A. V. Repair of naturally occurring mismatches can induce mutations in flanking DNA. *eLife* **3**, e02001 (2014).
31. Nguyen, L., W. M. Martens, J., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. *Nat. Commun.* **11**, 5584 (2020).
32. Chen, D. *et al.* BRCA1 deficiency specific base substitution mutagenesis is dependent on translesion synthesis and regulated by 53BP1. *Nat. Commun.* **13**, 226 (2022).
33. Franco, I. *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.* **20**, 285 (2019).
34. Supek, F. & Lehner, B. Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair* **81**, 102647 (2019).
35. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).
36. Yaeger, R. *et al.* Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer. *Cancer Cell* **33**, 125-136.e3 (2018).
37. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
38. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
39. Gindin, Y., Meltzer, P. S. & Bilke, S. Replicon: a software to accurately predict DNA replication timing in metazoan cells. *Front. Genet.* **5**, (2014).
40. Ryba, T. *et al.* Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. *Genome Res.* **22**, 1833–1844 (2012).
41. 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).
42. Gonzalo, S. *et al.* Role of the RB1 family in stabilizing histone methylation at constitutive heterochromatin. *Nat. Cell Biol.* **7**, 420–428 (2005).
43. Krishnan, B. *et al.* Active RB causes visible changes in nuclear organization. *J. Cell Biol.* **221**, e202102144 (2022).
44. 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).
45. 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).
46. 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).
47. 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).
48. SPIN reveals genome-wide landscape of nuclear compartmentalization | Genome Biology | Full Text.

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02253-3.
49. 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).
50. 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).
51. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
52. 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.* **13**, 4013 (2022).
53. Zou, X. *et al.* Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, 1–16 (2018).
54. Drost, J. *et al.* Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234–238 (2017).
55. Xiong, K. & Ma, J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat. Commun.* **10**, 5069 (2019).
56. Hudson (Chairperson), T. J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
57. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
58. Pleasance, E. *et al.* Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* **1**, 452–468 (2020).
59. 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).
60. Edwards, N. J. *et al.* The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J. Proteome Res.* **14**, 2707–2713 (2015).
61. Walker, B. A. *et al.* A high-risk, Double-Hit, group of newly diagnosed myeloma identified by genomic analysis. *Leukemia* **33**, 159–170 (2019).
62. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
63. 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).
64. Derrien, T. *et al.* Fast Computation and Applications of Genome Mappability. *PLOS ONE* **7**, e30377 (2012).
65. 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).
66. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).
67. Buisson, R. *et al.* Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, (2019).
68. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
69. Vali-Pour, M., Lehner, B. & Supek, F. The impact of rare germline variants on human somatic mutation processes. *Nat. Commun.* **13**, 3724 (2022).
70. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
71. 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).
72. Sima, J. *et al.* Identifying cis Elements for Spatiotemporal Control of Mammalian DNA Replication. *Cell* **176**, 816-830.e18 (2019).
73. Van Rechem, C. *et al.* Collective regulation of chromatin modifications predicts replication timing during cell cycle. *Cell Rep.* **37**, 109799 (2021).
74. 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).
75. Poulet, A. *et al.* RT States: systematic annotation of the human genome using cell type-specific replication timing programs. *Bioinformatics* **35**, 2167–2176 (2019).

76. Klein, K. N. *et al.* Replication timing maintains the global epigenetic state in human cells. *Science* **372**, 371–378 (2021).
77. Ding, Q. *et al.* The genetic architecture of DNA replication timing in human pluripotent stem cells. *Nat. Commun.* **12**, 6746 (2021).
78. Gunasekara, C. J. *et al.* A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol.* **20**, 105–105 (2019).
79. Mukhopadhyay, R. *et al.* Allele-Specific Genome-wide Profiling in Human Primary Erythroblasts Reveal Replication Program Organization. *PLoS Genet.* **10**, e1004319 (2014).
80. Barlow, J. H. *et al.* Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. *Cell* **152**, 620–632 (2013).
81. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).