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

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

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

Abstract

Mutations in human cells exhibit increased burden in heterochromatic, late DNA replication time (RT) chromosomal domains, with variation in mutation rates between tissues mirroring variation in heterochromatin and RT. We observed that regional mutation risk further varies between individual tumors in a manner independent of cell type, identifying three signatures of regional mutagenesis in >4000 tumor genomes. The major signature reflects domain-scale remodeling of heterochromatin and of the RT program seen across tumors, tissues and cultured cells, and is robustly linked with higher expression of cell proliferation genes. Regional mutagenesis is associated with loss-of-activity of the tumor suppressor genes *RB1* and *TP53*, consistent with their roles in cell cycle control, with distinct mutational patterns for the two genes. Loss of regional heterogeneity in mutagenesis associates with deficiencies in various DNA repair pathways. These mutation risk redistribution processes modify the mutation supply towards important genes, diverting the course of somatic evolution.

Introduction

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

These segments with variable mutation rates tend to correspond to topologically associating domains (TADs) and, similarly, to RT domains [3–5]. Regional mutation density (RMD) strongly correlates with the later RT of the domain, as well as with lower gene expression levels, lower chromatin accessibility (e.g. as measured by DNAse hypersensitive sites (DHS)), higher levels of inactive histone marks such as H3K9me3 and, in the opposite direction, with active chromatin marks such as H3K4me3 [1,6–8]. The global RMD landscape in somatic cells is to some extent tissue-specific, sufficiently so that it can be used to predict cancer type at high accuracy [9,10]. The tissue-specificity of RMD in a domain is paralleled in the tissue-specificity of RT and heterochromatin in the domain. For instance, the domain that switches from late-RT to early-RT,

or where genes increase in expression levels, or that gets more accessible chromatin in a particular tissue, also exhibits a reduced rate of somatic mutations in that tissue [1,6]; this property can identify the cell-of-origin of some cancers [11].

Apart from variation in active chromatin and RT between tissues, there is variation in chromatin within the same tissue/cell types, but across individuals or across cells. For instance, studies of quantitative trait loci (QTL) have demonstrated genetically determined local changes in RT, chromatin accessibility and DNA methylation [12–15]. Further, gene expression programs exist that are variably active between tumors originating from the same tissue (and also between individual cells), but are recurrently seen across many different tissues [16,17]. Such recurrent expression programs may conceivably drive, or be driven by chromatin remodeling that activates or silences chromosomal domains.

Indeed, chromatin remodeling can occur during tumor evolution, and this can manifest as changes in RT between normal and cancerous cells, loss of regional DNA methylation, or changes in heterochromatin marks in some chromosomal domains upon cancerous transformation [18–22]. This large-scale, global heterochromatin remodeling occurring across various chromosomal domains of cancer cells may plausibly affect local and regional DNA damage and repair processes, which are linked with different features of chromatin organization [1,2,20,23–25].

Here, we hypothesized that chromatin remodeling that occurs variably between tumors may generate variation in regional mutation rates, beyond the known cell-of-origin identity effects on mutagenesis. We study the domain-scale regional profiles of somatic mutations from ~4200 tumor whole-genome sequences, modeling them as a mixture of several underlying regional distributions, which correspond to different mutation risk mechanisms acting preferentially in some genomic domains. Some of these RMD signatures represent the expected differences between tissues/cell types stemming from chromatin organization in the cell-of-origin, or consequences of common DNA repair failures. However other commonly-occurring RMD signature patterns were associated with large-scale chromatin remodeling and with changes in RT programs. This chromatin and RT remodeling, in turn, associated with activity of cell cycling gene expression programs. The resulting wide-spread mutation risk redistribution across chromosomal domains can increase or decrease mutation supply towards regions harboring cancer genes, potentially generating driver events and genetic interactions.

Results and Discussion

A statistical method to quantify variation in megabase-scale regional mutation density

We performed an exploratory unsupervised analysis of diversity in one megabase (1 Mb) mutation density patterns across 4221 whole-genome sequenced tumors, controlling for confounding factors such as arm-level CNA, possibly selected regions, and trinucleotide mutational signatures [26] (Methods).

A Principal Component (PC) analysis on the RMD profiles yielded clusters largely reflecting identity of tissues and subtypes, as expected [1,6] (Fig. 1ab; Extended Data Fig. 1f).We note a similarity of RMD profiles between related cancer types, and conversely RMD profiles may subdivide some cancer types into plausible subtypes, as exemplified in breast (Fig 1b) and head-and-neck cancers (Fig. 1c; Extended Data Fig. 1). As a control, we captured two known examples of redistribution of mutation rates: one affects somatic hypermutation regions in B-lymphocytes (Extended Data Fig. 1gh), and the other causes a global homogenization ('flattening') of the RMDs in MMR-deficient tumors [1] (Extended Data Fig. 1g). While the RMD profiles expectedly reflect cell type-specific signals, the PCA suggests additional systematic RMD variability (Extended Data Fig. 1b).

Aiming to separate the tissue-specific RMD variability from the mutation patterns independent of tissue-of-origin, we devised a methodology based on non-negative matrix factorization (NMF), analogous to that used for extracting trinucleotide SNV mutational signatures [26–28], however here applied to 2540 megabase-sized domains instead of the typical 96-channel trinucleotide SNV spectrum. Each of robustly extracted NMF factors corresponds to one "RMD signature" of mutation risk redistribution across domains, with a spectrum consisting of RMD window weights (for all 2540 windows). Additionally the RMD signature has data on 'exposure' or activity for each tumor genome.

To test whether our NMF method is sufficiently powered to capture RMD inter-individual variability, we simulated cancer genomes containing ground-truth patterns of RMD. These were generated to affect a variable number of domains, to be present in a variable number of tumor samples, and to be present at variable intensity (fold-increase over default mutation rate at each window) (Extended Data Fig. 2a, see detailed description in Methods). Upon running our NMF methodology, we selected the number of factors and clusters based on the silhouette index (SI), estimating reproducibility over repeated bootstraps and runs of NMF (Table S1), and assessing accuracy of match to the ground-truth signatures (Methods, Table S2, example in Fig 1d).

Encouragingly, we observed that even with a small fraction of tumor samples affected by a signature (5%), the ground-truth RMD signatures can be identified reliably, as long as the contribution of the RMD signature to the total mutation burden is reasonably high (>=20%) (Extended Data Fig. 2b). In addition, the NMF setup is usually able to recover RMD signatures that affect as little as 10% of all 1 Mb windows. The signature exposure strength (fold-enrichment over baseline mutation rate) does affect power to recover RMD signatures (Extended Data Fig. 2b).

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

We applied the NMF methodology to the somatic RMD profiles of 4221 tumor WGS with ≥ 3 mutations/Mb, allaying noise in RMD profiles (as a limitation, this depletes low mutation burden cancer types). A simulation indicated that 3 mutations/Mb are well sufficient for 1 Mb analysis

(Extended Data Fig. 2c-d). In total, we extracted 13 RMD signatures based on optimizing SI, which scores the reproducibility of solutions (Fig. 1ef, Extended Data Fig. 3).

The RMD signatures span a continuum from very tissue-specific (high Gini index, Fig. 1f), to global signatures seen across many cancer types (low Gini index). We named the 10 tissue-specific RMD signatures according to the tissue or tissues they affect (e.g. RMD_upper-GI, RMD_liver), while the three global signatures were named RMDglobal1, RMDglobal2 and RMDflat (Fig. 1f, Extended Data Fig. 3); the latter has in part known mechanisms (see below).

While some RMD signatures are tissue-specific and capture the genomic regions with high mutation risk only in that particular organ, many RMD signatures are observed in several related cancer types (Fig. 1f, Extended Data Fig. 3). For instance, tissue activity spectra of RMD_upper-GI and RMD_lower-GI signatures are broadly consistent with the subdivision by developmental origin into the foregut and the midgut/hindgut (Fig. 1f). The RMD_squamous signature unites some squamous lung cancers, head-and-neck cancers, some bladder cancers (consistent with reports based on gene expression data [29]), also expectedly some cervical and esophageal tumors, however surprisingly includes some sarcomas and uterus cancers, suggesting similarity of chromatin organization in these samples. Thus our 2540-channel RMD signatures support the proposed uses of mutational profiles for elucidating cell-of-origin and cancer development trajectories (metaplasia and/or invasion) [6,9,11].

Three patterns of megabase-scale mutation risk observed across most somatic tissues

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

Firstly, the profile of the RMDflat signature captures the known pattern of reduced variation in mutation rates, previously associated with MMR and NER deficiencies [1,2] and with high APOBEC3A mutagenic activity [30] . These known associations explained 52% RMDflat-high tumors in our data (Extended Data Fig. 4) , and we hypothesized that the remainder may also result from DNA repair deficiencies. Indeed we found that homologous recombination deficiencies (HRd) were commonly associated with RMDflat, and this was the case for both the BRCA1 and the BRCA2-subtypes of HRd ascertained by the CHORD method [31] (Extended Data Fig. 4b). This is consistent with reported enrichment of HRd trinucleotide signature SBS3 towards early-replicating domains [32], opposite of the canonical RMD distribution; same distribution of the SBS3-like mutation spectrum, particularly C>G changes, is observed in our data (Extended Data Fig. 4h)Thus different DNA repair defects converge onto the RMDflat mutational phenotype, with varying prevalence depending on the cancer type: in colorectal tumors the main mechanism is the MMR deficiency, while in ovary and pancreas it is the HR deficiency, and the main mechanism in bladder and lung is APOBEC mutagenesis (Extended Data Fig. 4c). For the remaining 28% of RMDflat tumor samples that are not explained by the above mechanisms, we find they are unlikely to be caused by false-negative calls for MMR or HR deficiency (based on indel spectra in

Extended Data Fig. 4i), therefore additional mutational mechanisms involving DNA repair deficiency or evasion are likely to be relevant in those tumors.

Unlike the homogeneous pattern resulting when the RMDflat signature profile is superimposed onto the canonical RMD landscape, the RMDglobal1 and RMDglobal2 profiles have a complex pattern with peaks scattered throughout the chromosomes. We can rule out that RMDglobal1 and 2 resulted from random noise, because their SI and autocorrelation are comparable to the tissue-RMD signatures, which have a known biological basis (Extended Data Fig. 5ab). The RMD signatures were determined using single-nucleotide variant (SNV) mutations; as a validation, their regional biases were similarly observed in indel and SV mutation distributions (Extended Data Fig. 5c-f).

Tissue-specific NMF analyses also robustly recovered the three RMDglobal signatures (Table S3), which thus capture inter-tumoral RMD heterogeneity in mutation risk of chromosomal domains that is recurrently observed in various human somatic tissues.

RMDglobal1 mutation risk in regions with plastic replication timing and heterochromatin

We were interested in the mechanism underlying the widespread RMDglobal1 signature, which was significant in ~25% of the tumor genomes (Fig. 1f; using a conservative threshold, see Extended Data Fig. 3), and which contributed variable mutation burden across individual tumors (Fig. 1g). Because tissue-specific RMD patterns reflect tissue-specific chromatin organization [1,6,9], we hypothesized that other, tissue-independent variation in chromatin across domains may underlie the tissue-independent RMDglobal1. We tried to predict the RMDglobal1 signature spectrum (1 Mb window weights) from epigenomic features relevant to megabase-scale mutation rates (reviewed in [33]): replication timing (RT), density of accessible chromatin (DNAse hypersensitive sites, DHS) and ChipSeq data for histone modifications including the heterochromatin marks H3K9me3 and H3K27me3. The average of each feature (either RT, or DHS, or heterochromatin mark per 1 Mb window) across many epigenomic datasets did not predict (Fig. 2a), and predicting from each sample individually identified only moderate associations (R^2 ~= 0.2) for certain datasets with regional density heterochromatin (H3K27me3, H3K9me3 marks) (Fig. 2a).

In stark contrast, RMDglobal1 spectrum can be accurately predicted (R^2 up to 0.7) from either RT, or DHS, or either of the two heterochromatin marks, if predicting using multiple tissue samples jointly (Fig. 2a). This suggests that RMDglobal1 spectrum is explained by some pattern in the variation across the cell/tissue samples for a chromatin feature (individual examples shown in Fig. 2b). As a validation we observed the same trend using regional density of chromHMM segmentation states (Extended Data Fig. 5g).

Heterochromatin restructuring at the domain scale across human cell types

Next, we quantified the systematic variation in heterochromatin states at the domain level, recurrently observed across diverse human tissues and cell types including tumor cells in the ENCODE repository, with the goal of identifying heterochromatin variation that predicts RMDglobal1 mutation risk. We performed PCAs on the megabase-window signal of the H3K9me3 and H3K27me3 marks, RT, DHS, and Hi-C compartments [34]. While each feature was analyzed independently, the PCs that resulted -- representing chromatin restructuring across domains -- were often recurrently observed across the analyses (Fig. 2c). In particular the RMDglobal1 mutagenesis pattern was in a tight cluster of chromatin restructuring PCs, most correlated with the H3K9me3_PC3, Hi-C_PC2 and DHS_PC3 (R=0.53, -0.53 and 0.43, respectively), and additionally also with RT_PC4 and H3K27me3_PC2 (Fig. 2c). These chromatin PCs exhibited a regional distribution whose peaks collocated with the peaks in RMDglobal1 regional mutagenesis (example shown in Fig. 2d).

Next, to understand the mechanism driving the H3K9me3_PC3 heterochromatin restructuring program, we analyzed gene expression levels of ENCODE samples with higher *versus* lower H3K9me3_PC3. Interrogating the MSigDB hallmarks [35] revealed associations with expression of MYC target genes, E2F target genes and G2M checkpoint genes (all GSEA FDRs ≤ 10^{-33}; Fig. 2e), implicating increased anabolism, DNA replication, and mitotic processes, respectively. Similarly, enrichment of cell cycling-associated genes was observed using the single cell-derived RHP gene programs [16] (Extended Data Fig. 6a). These enrichments were consistent with gene expression analysis from contrasting ENCODE samples by the chromatin accessibility (DHS_PC3) or by Polycomb repressive mark shifts (H3K27me3_PC2) across domains (all FDR<1% by GSEA, Fig. 2e), thus converging onto a model of genome-wide chromatin restructuring program linked with rapid cell proliferation.

Chromatin restructuring program PCs were differentially active between ENCODE intact tissues (lower scores) and cultured primary cells (higher score) ($p=10^{-12}$ and 10^{-23} for H3K9me3 and DHS, respectively, by Mann-Whitney test Fig. 2f); cell culture selects for proliferation-competent cells and is expected to have a higher proportion of cycling cells than intact tissues. Consistently, immortal cell lines were more similar to the primary cell cultures than to tissues (Fig. 2f). Comparing the cancerous cell lines to noncancerous cell lines however reveals considerable overlap, both in the H3K9me3 mark PC and also in DHS density PC (Fig. 2f), suggesting these chromatin restructuring programs do not reflect cancerous transformation *per se* (Fig. 2f).

We further asked if these particular chromatin remodeling programs linked with RMDglobal1 mutation risk reflect tissue specificity, but did not find evidence thereof (Extended Data Fig. 6de. Table S4). Samples from different tissues overlapped each other in the intensity of the chromatin remodeling signatures, and not even the nervous system nor the muscle cells (known to have distinctive tissue-specific patterns of active chromatin [36]) showed a notable difference..

Differences in levels of some selected cell proliferation genes were striking when comparing the ENCODE samples on the two ends of either the H3K9me3_PC3 constitutive heterochromatin reorganization, or DHS accessible chromatin domain-level reorganization (Fig. 2g; H3K27me3_PC3 Polycomb reorganization, Extended Data Fig. 6c). Taken together, this data

suggests that likely the cell proliferation itself, rather than the tissue/cell type identity or the oncogenic transformation status, predicts the domain-scale heterochromatin reorganization that is mirrored in RMDglobal1 mutation risk.

RT profiles of tumors and cells link chromatin changes with mutation risk signature

The above analyses of chromatin features at the domain-scale was performed over data from ENCODE, which does include some cancer cell lines (74 out of 256 H3K9me3, and 153 out of 676 DHS datasets), but does not include tumor tissue per se. Therefore we turned to examine chromatin domain restructuring directly in tumors, drawing on accessible chromatin (via ATAC-Seq) measurements in 410 TCGA tumors [37]. The local distributions of accessible chromatin sites can be used to accurately infer RT programs [38], as applied to large-scale studies of RT in various nontumoral tissues [39,40]. Here we perform a large-scale analysis of RT in tumors ("predRT-TCGA"); the tumoral RT predictions using Replicon tool [38] were deemed accurate and relevant to mutation risk modeling (), based on several observations. The tool was accurate in our tests (mean R between predicted RT and RepliSeq = 0.87; Extended Data Fig. 6f) (see Methods), and the modeling of RMDglobal1 mutagenesis is similarly accurate using predicted RT (from ENCODE DHS) as is with experimental RT (from various datasets) (Fig. 3a). Additionally, the predicted RT profiles are relevant to analysis of tumor mutation data: the predRT-TCGA better models the RMDglobal1 mutagenesis than predicted RT profiles from ENCODE diverse cell types do (predRT-ENCODE Fig. 3a). This suggests that our global RT analysis of TCGA tumors may capture tumor-relevant RT switching programs that reflect in RMDglobal1 mutagenesis.

To understand the mechanism underlying variability in RT that predicts RMDglobal1 mutagenesis, we systematized trends in RT profile variation of tumors using a PCA with the predRT-TCGA dataset. Expectedly, the TCGA RT PCs with most variance explained represent the average RT profile (predRT-TCGA-PC1 and 2), or the cancer type-associated RT programs (3 and 4, separating breast from kidney and brain tumors, Extended Data Fig. 6g). However, the following pattern of systematic RT variation (e.g. TCGA-RT_PC5) did not exhibit a tissue signal (Extended Data Fig. 6h).

Independently, we also performed PCAs on experimentally measured RT data (expRT, n=158), and on predicted RT in nontumor tissues and primary cells and varied cell lines (predRT-ENCODE, n=597). Certain RT-PCs from varied datasets converge onto the same global pattern of alteration in the RT program, exemplified in the predRT-TCGA_PC5 (Fig. 3b). They also correlated with the heterochromatin remodeling PCs highlighted above (DHS_PC3, H3K27me3_PC2, H3K9me_PC3 and HiC_PC2) (Fig. 3b; median pairwise correlation R = 0.43); thus these RT PCs represent the remodeling of the RT program and heterochromatin that can be observed across tumors, cultured cells (either cancerous or not), and healthy tissues.

Additionally, a data set of RT measured in single cells [19] (scRT) generated a scRT-PC5 which correlates moderately (R=-0.36) with the TCGA RT program (Fig. 3b), thus the variation of RT programs between individual cells [40] -- presumably indicating those chromosomal domains that

have more labile RT -- may predispose these domains to the systemic RT switches between tumors.

Next, we asked whether shifts in RT observed across tumors can explain the shifts in domain mutation risk across tumor WGS as per profile of RMDglobal1. The TCGA-RT_PC5, which did not exhibit a tissue signal (Extended Data Fig. 6gh), correlated strongly with RMDglobal1 mutation risk redistribution (R=-0.49) (Fig. 3b). and affected the RMDglobal1-relevant domains where mutation rate changes notably (example cancer types in (Fig. 3c)), compared with next best PC6 at R=0.35, and other RT-PCs up to PC10 at R<0.2.

Importantly in a multiple regression test not only RT but other chromatin features were independently predictive of RMDglobal1 mutagenesis (with exception of H3K27me3, which is dispensable; Extended Data Fig. 6i). Upon “orienting” these various PCs from chromatin or RT remodelling analysis (Fig. 2c, Fig. 3b), we infer that chromosomal domains with highest RMDglobal1 mutation rate increase become later-replicating and heterochromatinized in cells exhibiting a stronger signal of proliferation in gene expression.

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

To understand the biology underlying this tumoral RT-PC5 that mirrors RMDglobal1 mutagenesis, we asked how gene expression changes between the TCGA tumors with high values of a RT-PC versus tumors with low values. As with heterochromatin analysis, both the RT-TCGA_PC5 and the independently derived RT-ENCODE_PC3 strongly associate with gene expression of E2F target, MYC target and G2M checkpoint genes in Hallmark sets (all FDR<0.1%, Fig. 3d). Thus, this RT switching pattern represents a RT program characteristic of tumors bearing programs of rapid cell cycling. This was supported in an independent analysis of single-cell derived RHP gene sets [16], where the TCGA RT-PC5 correlated strongly with expression of cell cycle genes [16] (G2/M and G1/S genes, correlated at GSEA FDR=10^{-20} and 10^{-10}, respectively) (Extended Data Fig. 6j), while the expression of other RHP gene sets correlated less well with TCGA RT-PC5 (next strongest $5 \cdot 10^{-5}$).

To more directly support the association of the RMDglobal1 mutation redistribution with cell cycle gene expression in the same tumor samples, we next considered the subset of WGS where matched RNA-Seq was available. Expression of E2F target genes and of G2M checkpoint genes in tumors associated with their RMDglobal1 mutagenesis status (at FDR<1%; Fig. 3d), and expression of MYC target genes had a positive trend (FDR=31%). Prominent genes linked to cell division and with potential uses as cell proliferation markers were associated with mutagenesis-relevant RT programs, both in TCGA tumors with predicted RT and in ENCODE tissues (Fig. 3e) and this also independently validated in their direct association with RMDglobal1 mutagenesis pattern in our set of tumors with RNA-Seq and WGS (Fig. 3f).

Overall, the chromosomal domains with higher mutation rate changes in RMDglobal1 are those domains that undergo changes in RT in tumor samples with more proliferative-like transcriptomes, compared to tumor samples with less proliferative-like transcriptomes.

Spatial chromatin compartments that are plastic are prone to mutation risk changes

We further asked what characterizes these domains where heterochromatin and RT are more malleable (and which mirror the RMDglobal1 mutation rates). To this end, we analyzed data from diverse epigenomic assays (studies listed in (Table S5)) with reported correlations with RT. We compared the regional density of each epigenomic feature with our RMDglobal1 spectrum window weights (Table S5). Consistently with the chromatin/RT restructuring analyses above, we noted strong correlations with Hi-C subcompartments (Fig. 3g)[41]. In particular, the B1 inactive subcompartment, rather than B2 and B3 inactive heterochromatin, was most associated with RMDglobal1. B1 replicates during middle S phase, and correlates positively with the Polycomb H3K27me3 mark suggesting that it represents facultative heterochromatin [41]. Next, we observed a positive correlation with two SPIN states (Fig. 3h) intranuclear territories [42]. classified as "Interior repressed" [42], marking inactive regions that are (unlike other heterochromatin) located centrally in the nucleus, rather than lamina-associated [42]. Additionally RMDglobal1 windows are enriched in subtelomeric parts of chromosomes (Fig. 3j).

In addition to RMDglobal1 mutagenesis, also the chromatin and RT restructuring programs that we identified (Fig 2, H3K9me3_PC3, DHS_PC3, H3K27me3_PC2; Fig 3, predRT-TCGA_PC5 and predRT-ENCODE_PC3) were enriched in the same nuclear territories (Extended Data Fig. 7a-d)., suggesting they contain chromatin prone to restructuring that is associated with expression of cell proliferation genes (Fig 2, Fig 3). We additionally found a correspondence between the "CORES" regions [43] i.e. domains that change chromatin conformation upon whole-genome duplication, and our RMDglobal1 mutation redistribution domains (Fig 3j) and also the chromatin restructuring PC signatures (Extended Data Fig. 7a-d).

In summary, this analysis suggests that certain heterochromatin compartments may be intrinsically more malleable, undergoing remodeling that determines mutation rates.

RMDglobal1 mutagenesis associates with *RB1* pathway alterations

We further hypothesized that genetic alterations may drive changes in RT/heterochromatin accompanied by cell cycling gene expression and the RMDglobal1 mutation risk redistribution. We thus performed a genome-wide association analysis, linking somatic copy number alteration (CNA) events and deleterious point mutations with RMDglobal1 mutation risk (adjusting for cancer type and for confounding between linked neighboring CNAs; qq plots in Extended Data Fig. 8a; Methods for details). Here, we considered a set of 1543 chromatin modifiers, cell cycle, DNA replication and repair genes and cancer driver genes, compared against a background of 1000 random genes (Methods).

For CNA, we found a strong positive association of RMDglobal1 with deletions of the *RB1* tumor suppressor, which has key roles in cell cycle control and also in chromatin organization (FDR=0.05%, better p-value than all 1000 control genes) (Fig. 4abc, Extended Data Fig. 7e). Because CNA often affects large chromosomal segments, we also tested associations with *RB1* neighboring genes (Fig 4d), noting that *RB1* is at the CNA frequency peak, meaning it is the likely causal gene in the CNA segment. Strength of RMDglobal1 association with *RB1* alterations is gene dosage dependent (Extended Data Fig. 7f), and moreover the effect of (rarer) *RB1* point mutations shows a (nonsignificant) supporting trend in the same direction as the *RB1* deletions (Extended Data Fig. 7g). As independent evidence, we identified deletions in *CDK6*, a negative regulator upstream of *RB1*, as the CNA event negatively associated with RMDglobal1 with the strongest p-value, exceeding all control genes (Fig. 4b).

The *RB1* alterations were anticipated to associate with certain gene expression patterns; indeed in two tumor datasets, we observe *RB1* deletions associated with higher expression of E2F target genes (as expected from pRb function in inhibiting the E2F transcription factors), G2M checkpoint genes and MYC target genes (GSEA all FDR<1%; Fig. 4i; additionally the mitotic spindle genes and DNA repair genes were upregulated here). Thus gene expression signatures of *RB1* deletion are consistent with gene expression signatures of RT/heterochromatin restructuring above (H3K9me3_PC3 or predRT-TCGA_PC5, Fig. 2e, 3d).

In addition to cell cycle regulation, pRb has additional roles in chromatin organization [23,44–46], and *RB1* deletions were reported to change H3K9me3 and H3K27me3 marks, affecting more the regions enriched at subtelomeres and associating with propensity to DNA damage in those regions[23]. We found these same two heterochromatin marks more highly correlated to RMDglobal1 than other tested marks (Fig. 2a), and the RMDglobal1 domain weights were strongly enriched in the approximately ¼ of chromosome arm proximal to telomere (Fig. 3i).

Prompted by the above, we asked if the location of heterochromatin restructuring upon *RB1* perturbation by experiment [23] matches the locations of the RMDglobal1 mutation risk changes in cancer genomes. Indeed, overlap with H3K9me3-switching regions in *RB1* k.o. cells [23] is substantial (OR=3.25, 95% C.I [2.3-4.5], $p<10^{-13}$ for overlap in the top-10% regions), however there is no enrichment with H3K27me3-switching regions (Fig. 4e). Next, we asked if the RMDglobal1 mutation risk results, at least in part, from the increase in DNA damage in these regions upon *RB1* perturbation (measured as CPD lesions after UV exposure [23]). The overlap between the top-10% RMDglobal1 mutation risk domains and the top-10% DNA damage-sensitized domains (upon RB1 KO) was strong (OR=5.05, $p<10^{-29}$), and similarly the overlap in bottom-10% RMDglobal1 and the top-10% damage-protected domains (OR=8.17, $p<10^{-54}$; Fig. 4f). The overlap was seen in regional mutation risk both in skin cancers, which are UV mutagenized, but similarly so in lung cancers, which are tobacco smoking chemical-mutagenized (Extended Data Fig. 8c), suggesting the link extends to multiple types of DNA damage. Further, the telomere-proximal enrichment of RMDglobal1 mutation risk (Fig. 3i) associated with the DNA damage-increase upon RB1 KO [23], with a clear gradual increase in RMDglobal1 mutation risk towards the telomere spanning ~10 Mb telomere-proximal DNA on average (Fig. 4g). Overall, this

overlap of heterochromatin remodelling loci as well as DNA damage sensitive loci upon *RB1* loss-of-function [23] with the RMDglobal1 mutation risk loci in tumors underscores *RB1*'s role in shaping the somatic mutation rate landscape.

In addition to the CNA analysis above, we also tested associations with the presence of deleterious somatic point mutations and identified the *KRAS* mutation to positively associate with RMDglobal1, at FDR=1% (Fig. 4h), consistently across individual cancer types (Extended Data Fig. 8d); Extended Data Fig. 8e-f). he *KRAS* gene is known to act downstream of *RB1* in developmental and in tumor mouse phenotypes [47,48]. Consistently, *KRAS* mutation and *RB1* loss (either deletion or mutation) are mutually exclusive in our tumor dataset (chi-square $p < 2.2e-16$), supporting that the driver alterations in *RB1* and *KRAS* may converge onto the same mutation rate redistribution phenotype, the RMDglobal1. Consistently, we found the subclonal, later-occurring mutations are enriched in the RMDglobal1 pattern compared to the clonal mutations in various cancer types (Extended Data Fig. 8g-h).

Mutation supply towards cancer genes is altered by RMD signatures

Since RMDglobal1 captures a redistribution of mutation rates genome-wide, it follows that this will affect the local supply of mutations towards loci harboring some cancer genes. To test this, we considered 460 cancer genes and the intronic mutation rate thereof (to avoid effects of selection), contrasting tumors with a high RMDglobal1 activity (top tertile) versus low RMDglobal1 activity (bottom tertile) (Fig. 4jk). When compared to a randomized baseline (95th percentile of the random distribution used as cutoff; Methods), the mutation supply was significantly increased towards 28% of the 460 cancer genes in RMDglobal1-high tumors (Fig. 4k). These genes increase mutation rates on average by 1.21-fold in the RMDglobal1-high tertile tumors, with bigger increases for some genes (example in Fig. 4l) such as *BAP1* 1.78-fold, *KMT2C* 1.79-fold, and *ATM* 1.18-fold increase in median mutation supply. Importantly, the mutation supply measured does not reflect selected mutations but instead only the relative risk of mutations appearing in the given region.

Next, we similarly considered the redistribution effects of the RMDflat signature, associated with DNA repair deficiencies (see above), increasing relative mutation rates in early replicating, euchromatic regions [30,33] (Extended Data Fig. 4d). These early RT regions also have a higher gene density. Indeed, RMDflat commonly affects the mutation supply to many cancer driver genes, with 75% of the cancer genes [49] exhibiting an increased supply comparing RMDflat-low to RMDflat-high tumors (Extended Data Fig. 4e). The converse case was rare, with 9% cancer genes decreased in relative mutation supply. As one example driver gene (Extended Data Fig. 4f), the *ARID1A* tumor suppressor gene, located in a lowly-mutated region in chromosome 1p, has mutation supply increased 1.8-fold, 2.1-fold and 2.4-fold in MSI (i.e. MMR-deficient), HR-deficient and APOBEC tumors (all cases of RMDflat-high), respectively (Extended Data Fig. 4f-g). As another example, the *BRAF* oncogene (where driver mutations are highly enriched in MSI compared to MSS colorectal tumors [50]) has a considerably increased mutation supply in the RMDflat-high tumors (Extended Data Fig. 4f-g).

TP53 disruption reduces mutation supply towards late replicating regions

In addition to RMDglobal1 and RMDflat, there exists a third, commonly occurring mutation rate redistribution signature observed across 21% of tumor genomes across multiple tissues (Fig. 1f, Extended Data Fig. 3), the RMDglobal2. Its 1 Mb domain mutation rates do follow a distribution increasing mutation density in later RT overall, except for latest RT windows, which acquire fewer mutations than expected in the RMDglobal2 pattern (Fig. 5ab). As a consequence, mutation rates increase approximately linearly with RT bins in tumors with high RMDglobal2, while in tumors with a low RMDglobal2 exposure the RT relationship to mutation rates is better described by a quadratic function (Fig. 5c, Extended Data Fig. 9a). In other words, the RMDglobal2 redistribution "linearizes" the association of RMD to RT, by suppressing the prominent RMD peaks.

We aimed to identify the causal event behind RMDglobal2 redistribution, again testing for associations of RMDglobal2-high (top tertile) versus low (bottom tertile) tumor samples with CNAs and deleterious coding mutations. Strikingly, we found *TP53* mutation to be uniquely strongly associated with RMDglobal2 signature (FDR = $9 \cdot 10^{-10}$; next strongest positive association is *PKHD1* FDR=$2.2 \cdot 10^{-6}$) (Fig. 5d). *TP53* deletions were also positively associated (Fig. 5e) and these trends were observed consistently across many cancer types (Extended Data Fig. 9b). As independent supporting evidence, the amplifications in oncogenes that phenocopy TP53 loss (*MDM2*, *MDM4* and *PPM1D*) are all also positively associated with the activity of the RMDglobal2 mutation redistribution signature (Fig 5e, Extended Data Fig. 9b). This rules out that the *TP53* driver mutation occurrence is the consequence of the RMDglobal2 redistribution redirecting local mutation supply, but rather provides evidence for a causal effect of *TP53* inactivation in RMDglobal2 mutation risk redistribution.

Since *TP53* mutations are associated with increased burdens of CNA events [51], we tested whether RMDglobal2 RMD signature could be due to confounding from a multiplicity of focal CNAs. However, there is only a weak correlation between the CNA burden and RMDglobal2 signature levels upon stratifying for *TP53* status (R<=0.11) (Extended Data Fig. 9c-d). As with RMDglobal1, also RMDglobal2 is enriched in subclonal, late-occurring mutations (Extended Data Fig. 8g-h), consistent with it being triggered by *TP53* alterations, which are unlikely to be present in noncancerous cells while they accumulate mutations. The activity of the RMDglobal2 signature, inferred from SNV mutations, is also mirrored in the regional pattern of indel and SV mutations (Extended Data Fig. 5c-f).

Overall, the above convergent genetic associations strongly implicated the deficiencies in *TP53* pathway in the RMDglobal2 mutation risk redistribution, similarly as the deficiencies in *RB1* pathway -- another cancer gene controlling the cell cycle -- were implicated in the RMDglobal1 mutation redistribution (Fig. 4). Interestingly, RMDglobal2 was negatively correlated with the clock-like trinucleotide mutational signature SBS1 (C>T changes at CpG dinucleotides), consistently across cancer types (Extended Data Fig. 10ab), and there was also a positive association with the SBS93 trinucleotide signature (Extended Data Fig. 10ac). We did not identify

associations of similar magnitude and consistency between the RMDglobal1 redistribution and trinucleotide SBS signatures (Extended Data Fig. 10a; some tentative associations are in Extended Data Fig. 10de).

Interestingly, changes to local mutation supply because of risk redistribution can result in epistasis-like phenomena**.** For instance, 26% of cancer genes including *ARID1A* and *GATA3* exhibited a decreased relative mutation supply in high-RMDglobal2 tumors (which are often *TP53* mutant). (Fig. 5fg).

Apparent genetic interactions -- for example mutual exclusivity with *TP53* mutations that are drivers of RMDglobal2 -- might arise therefore. Indeed, considering13 genes known to bear coding mutations mutually exclusive with *TP53* mutations [52] , nearly half (6/13) were below the 5th percentile of the random distribution of local mutation rates, implicating RMDglobal2. What appears to be epistatic interaction is in fact commonly just a diversion of the mutational supply by a TP53-dependant redistribution (Fig. 5fh), resulting in RMD profiles with a difference in the local mutation supply towards the *ARID1A* locus (Fig. 5i). Overall, this illustrates how a change in local mutation risk, here mediated by *TP53* loss, can create apparent genetic interactions that may not indicate selection on functional effects, and should be explicitly controlled for in statistical studies selection and epistasis in cancer genomes.

Concluding remarks

Mutation rates are lower in early-replicating, euchromatic DNA compared to late-replicating heterochromatic DNA [8,53–56]. If either RT or heterochromatin (or both) are causal to mutation rates, which is likely the case and is often mediated by differential DNA repair [1,2,6,57,58] and/or differential DNA damage [25,59] then local changes in RT or in heterochromatin status would change local mutation risk. We provides robust evidence this is commonly the case, plausibly reflecting various molecular consequences of accelerated and/or dysregulated cell cycling on RT and heterochromatin organization, with downstream effects on mutation risk in chromosomal domains that affects mutation supply towards disease genes and steers the course of somatic evolution.

Methods

WGS mutation data collection and processing

Our research complies with all relevant ethical regulations. We collected whole genome sequencing (WGS) somatic mutation data from 6 different studies (Table S6). First, we downloaded somatic single-nucleotide variants (SNVs) from 1950 WGS from the Pan-cancer Analysis of Whole Genomes (PCAWG) study at the International Cancer Genome Consortium [60] Data portal (https://dcc.icgc.org/pcawg). Second, we obtained somatic SNVs for 4823 WGS from the Hartwig Medical Foundation (HMF) study [61] (https://www.hartwigmedicalfoundation.nl/en/). Third, we downloaded somatic SNVs from 570 WGS from the Personal Oncogenomics (POG) project [62] 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 applied QSS_NT>=12 mutation calling threshold in this study.

Finally, we downloaded alignments (BAM files) for 781 WGS samples from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) project [63,64] and BAM files for 758 tumor samples from the MMRF COMMPASS project [65] from the GDC data portal (https://portal.gdc.cancer.gov/). Somatic variants were called using Illumina's Strelka2 caller [66], using the variant calling threshold SomaticEVS >=6. Additionally, for these samples we performed a liftOver from GRCh38 to the hg19 reference genome.

Subtype assignment

We collected the sample 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 studies. 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 studies using ComBat as described in our recent work [67]. We trained a Ridge regression model with TCGA data to discriminate between LUSC and LUAD and applied the model to predict LUSC and 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 [67]. 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 cancer 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, filtered regions and matching trinucleotides

We divided the hg19 assembly sequence 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 [68] with alignability <1.0. In addition, we removed the regions that are unstable when converting between GRCh37 and GRCh38 [69] and the ENCODE blacklist of problematic regions of the genome [70].

Additionally, to minimize the effect of known sources of mutation rates variability at the sub-gene scale we removed CTCF binding sites (downloaded from the UCSC Table Browser), ETS binding sites (downloaded from http://funseq2.gersteinlab.org/data/2.1.0) and APOBEC mutagenized hairpins [71]. Finally, we removed all coding exon regions (+- 2nts, downloaded from the Table Browser) to minimize the effect of selection on mutation rates.

To minimize the variability in mutational spectra confounding the analyses, we adjusted 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 with less than 500,000 usable bp remaining. The final number of analyzed windows is 2,540.

Calculating the regional mutation density (RMD) of each window

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

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

To calculate the RMD applied to NMF analysis, we first subsampled mutations from the few hypermutator tumors, to prevent undue influence on overall analysis. We allow a maximum of 20 muts/Mb, that is 39,174 mutations. 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 chromosome arm]. We multiplied by the average nucleotides at risk in (i) and by the mean of the whole genome in (ii) to keep the bootstrapped values in the same range as the original values in each sample.

Applying NMF to extract RMD signatures

We applied bootstrap resampling (R function UPmultinomial from package sampling v2.10) to the RMD scores that we calculated for NMF, as above. The result for each tumor sample is a vector of counts with a total mutation burden close to the original one but normalized by the nucleotides at risk by window and also for possible chromosome arm-level copy number alterations (CNA). Next, 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 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 (SI), a clustering quality score (which here measures, effectively, how reproducible are the NMF solutions across bootstrapped runs), for each clustering parameter set, 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 superimposed the simulated ground-truth RMD 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 RMD signature contribution to the total mutation burden (10%, 20% or 40%) and the number of tumor samples affected by the RMD signature (5%, 10% or 20%). We randomly assigned the ground-truth signatures to be superimposed onto each tumor sample (e.g. tumor sample A will be affected by RMD signature 1 and 3, while tumor 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 parameter and k-medoids clustering k parameter, based on the minimum cluster 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 RMD signatures when the cosine similarity is >=0.75 only for that ground-truth simulated RMD signature, and < 0.75 for the rest.

Analysis of differential mutation supply towards cancer genes

For 460 cancer genes from the MutPanning list [49] (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 groups of tumor samples having a RMD signature high versus the group with RMD signature low activity, by tissue. Note that the possibly different number of 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 to another group of tumor samples (here split by RMD signature activity). We binarized the tumor samples by activity of 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 regression 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 same analysis but randomizing the high or low tertile assignment for the three RMD signatures prior to the regression.

Association analysis of gene mutations with RMD global signatures

We assembled a set of 1543 genes of interest: cancer driver genes from the MutPanning list [49] and Cancer Gene Census list [72], and furthermore we included genes associated with chromatin and DNA damage [73]. As control, we used a subset of 1000 random genes selected as in [73].

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 impact 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 present, or deleterious mutations present) or 0 if it has not. We considered CNA deletions and amplifications in two independent analyses.
We fit a linear model to test whether the binary genetic feature considered in a particular analysis (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 adjusted 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 case of CNAs, to adjust for the linkage between neighboring 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 CNA PCs for the deletion CNA and the first 63 CNA 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 S7. For each of these features, we downloaded the narrow peaks, calculated their weighted density for each 1Mb window as the width of the peak multiplied by the peak value.

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

ChromHMM chromatin states. We downloaded the 25 ChromHMM state segmented files ("imputed12marks_segments") for the 129 cell types available from Roadmap epigenomics [74](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 [75], the constitutive and developmental RT domains [76], RT changes upon overexpression of the oncogene *KDM4A* [77], RT signatures of replication stress [78], RT signatures of tissues [79], RT states [80], changes in RT upon *RIF1* knock-out [81] and RT changes due to RT QTLs [82]. In addition, we downloaded data for variability in DNA methylation [19,83], HMD and PMD regions [20], CpG density, gene density, lamina associated domains (LADs), asynchronous replication domains [84], early replicating fragile sites [85], SPIN states [42], A/B subcompartments [41], DHS signatures [86] and H3K27me3 and H3K9me profiles for *RB1* wild-type and knock-out [23]. Data described in Table S5. We calculated the density for each feature for each 1 Mb window, and correlated this with the RMDglobal1 signature 1 Mb window 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) [76] in multiple human cell types (n = 158 samples). In addition, we predicted RT using the Replicon software [38] from two types of 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 (https://www.encodeproject.org/); and (ii) in human tumors (n = 410 samples, most of them with technical replicates) using ATAC-seq data of TCGA tumors downloaded from [37]. We used the Replicon tool with the default settings.

Gene expression data and analyses

For the genomes from the HMF study, we downloaded gene expression data (as adjusted TPM values) from Hartwig [61], which were available for a subset of samples for which we derived the RMD signatures. In total, we had gene expression data for 1534 samples with RMD and 18889 protein coding genes therein. For the genomes from the TCGA study, 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 tumor samples and 20092 genes therein.

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

Gene expression association with the RMDglobal1 and chromatin signatures. For the activity profile of each RMD/epigenomic signature across samples (RMDglobal1_exposures, H3K9me3_PC3, etc) we predicted the signature from the gene expression of one gene;the coefficient from this regression indicates the gene effect (upregulated or downregulated) with respect to the signature and the p-value. We performed this analysis for every gene individually.

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

PCA and 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 estimate.

RMD signature exposures for clonal vs subclonal mutations

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

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

Software and packages

The analyses were performed using R (version 3.6). Relevant R packages are liftOver v1.18.0, GenomicRanges v1.46.1, sampling v2.10, NMF v0.26, glmnet v4.1-6, tclust v1.5-4, dplyr v1.1.0 and tidyr v1.3.0.

Statistics and Reproducibility

No statistical method was used to predetermine sample size; the maximum number of samples available was used. Data exclusion criteria were as described in the Methods section; principal exclusion is that of tumor genomes with low mutation burden, thus focussing on genomes with less noisy mutation rate estimates. The statistical methods applied largely do not have assumptions regarding data distributions. In this observational study there were no experiments performed to collect data, therefore randomization to conditions/groups does not apply; we note statistical tests based on randomization were used to determine statistical significance via generating permuted control data. The investigators were not blinded to allocation during analyses and assessment.

Data availability

In this study, published datasets were reanalyzed. WGS somatic mutation calls for the PCAWG study were downloaded from the International Cancer Genome Consortium (ICGC) Data portal [https://dcc.icgc.org/pcawg]. Restricted-access WGS somatic mutation calls for the HMF project were accessed via request number DR-260; details at [https://www.hartwigmedicalfoundation.nl/en/]. WGS somatic mutation calls for the POG project were downloaded from BC Cancer [https://www.bcgsc.ca/downloads/POG570/]. We downloaded restricted-access bam files for the TCGA (dbGaP accession phs000178.v11.p8), CPTAC (phs001287.v17.p6) and MMRF COMMPASS (phs000748.v7.p4) projects from the Genomic Data Commons (GDC) data portal [https://portal.gdc.cancer.gov/].

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 (described in Table S7). We downloaded experimental RT data from the Replication Domain database [https://www2.replicationdomain.com/index.php]. We downloaded ATAC-seq data of TCGA tumors [https://pubmed.ncbi.nlm.nih.gov/30361341/]. We downloaded the chromatin domain hierarchies and compartment scores generated by the Calder method from Hi-C data from 114 cell lines [https://pubmed.ncbi.nlm.nih.gov/33972523/]. We downloaded the 25 ChromHMM states segmented files ("imputed12marks_segments") for the 129 cell types available from Roadmap epigenomics [http://compbio.mit.edu/ChromHMM/].

Additionally, we downloaded other epigenomic data from various studies. The replication timing heterogeneity calculated as Twidth and Trep from high-resolution (16 phases) Repli-Seq data was from Ref. 73. The RT changes under overexpression of the oncogene KDM4A was from Ref. 75. Five RT signatures of replication stress were from Ref. 76. Ten RT cell type-specific signatures during development were from Ref. 77. Fifteen RT states were from Ref. 78. The changes (late to early or retain late) in RT upon RIF1 knock-out were from Ref. 79. The RT changes due to RT QTLs were from Ref. 80. The differences in RT between an hypomethylated cell line versus a control cell line were from Ref. 19. The regions with variability in methylation across individuals were from Ref. 81. The partially methylated domains (PMDs) and highly methylated domains (HMDs) were from Ref. 20. The CpG density, gene density and lamina associated domains

(LADs) were from the table browser [https://genome.ucsc.edu/cgi-bin/hgTables] (assembly Feb. 2009 GRCh37/hg19). The asynchronous replication domains were from Ref. 82. The early-replicating fragile sites were from Ref. 83. The SPIN states were from Ref. 40. The A/B subcompartments were from Ref. 39. Sixteen signatures generated from applying NMF to DHS peaks were from Ref. 84. The H3K27me3 and H3K9me profiles for RB1 wild-type and knock-out were from Ref. 23. The constitutive early, constitutive late and developmental domains were from http://www.replicationdomain.org.

In a FigShare repository [https://doi.org/10.6084/m9.figshare.c.6911140.v1], we provide data generated in this study: the RMD values across 2450 one-megabase windows for the 4221 tumor genomes analyzed (rmd_counts.zip), and the final RMD signatures extracted from this RMD matrix using NMF (RMDsignatures_exposures_k=13_nFact13_n=4221.csv and RMDsignatures_window_weights_hg19_k=13_nFact=13.csv). In addition, we provide the RT and chromatin remodeling PC-signatures (PCA_chrom_RT.zip). Finally, we provide the predicted DNA replication timing data at 1 Mb resolution using the Replicon tool for TCGA tumors (predRT-TCGA_1Mb.zip) and for ENCODE samples (predRT-ENCODE_1Mb.zip). Other data can be made available from the authors upon request.

Code availability

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

Acknowledgements

Work was supported by funding from an FPU fellowship of the Spanish government, Ministry of Universities to M.S., an ERC StG “HYPER-INSIGHT” (757700) to F.S., Horizon2020 project “DECIDER” (965193) to F.S., Spanish government project “REPAIRSCAPE” (PID2020-118795GB-I00) to F.S., CaixaResearch project “POTENT-IMMUNO” (HR22-00402) to F.S., an ICREA professorship to F.S., the SGR funding of the Catalan government (SGR 00616) to F.S., and the Severo Ochoa centers of excellence award of the Spanish government to the hosting institution IRB Barcelona.
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/.
We are grateful to Daniel Naro for retrieving and processing (alignment, variant calling) the WGS data from the CPTAC-3 and COMMPASS studies.

Author contributions statement

M.S. has performed data curation, wrote code, performed all analyses and visualized results. F.S. and M.S. have jointly devised analyses, interpreted results and drafted the manuscript. F.S. has conceived and supervised the study.

Competing interests statement

The authors declare no competing interests.

Figure Legends

Figure 1. Identifying RMD signatures by an NMF-based method to megabase-scale mutation density profiles. a) A principal component (PC) analysis of RMD profiles in 2540 non-overlapping 1 Mb windows for breast, head-and-neck and lung tumor WGS (n=1408). **b)** Mean RMD profiles on chromosome 1q for breast cancers in the PCA cluster 6 (n = 76) and cluster 9 (n= 211), shows enrichment of the triple-negative subtype in the former. Stars mark windows that are significantly different at FDR<25% by Mann-Whitney test (two-tailed). **c)** Mean RMD profiles on chromosome 1p for head-and-neck squamous cell cancers in the PCA cluster 11 (n = 81) and PCA cluster 13 (n= 41), where the former cluster with skin cancers and the latter with lung cancers, see Extended Data Fig. 1e. **d)** Example RMD signature from a simulation study (see Methods), comparing 1 Mb window weights for an extracted NMF signature and its matching simulated ground-truth signature along chromosome 1p. See Extended Data Fig. 2 and Tables S1-2 for additional simulation data. **e)** Clustering quality scores for various parameters for NMF run on 4221 tumor genomes. The minimum silhouette index (SI) across clusters (RMD signatures) for different numbers of NMF factors and clusters from k-medoids. The selected case (nFact=13, k=13) is marked with a cross. **f)** Overview of the distributions of the 13 extracted RMD signatures (rows) across different cancer types (columns). The circle size corresponds to the fraction of tumors in a cancer type exhibiting a high activity of a specific signature (defined as exposure >= 0.12, corresponding to the 1st percentile of the exposure of the RMDflat signature in microsatellite-instable [MSI] cancers). Total number of samples per cancer type written beneath table. Gini index quantifies the cancer type specificity, where lower Gini means signature is shared across different cancer types (i.e. the “global” RMD signatures). **g)** Relative contribution of each RMD signature to the total mutation burden across the 4221 tumor genomes analyzed. The RMDsig-tissue category represents all 10 tissue-specific RMD signatures, pooled together. The MSI tumors from various cancer types are all shown together.

Figure 2. RMDglobal1 mutation risk signature associates with domain-scale variability in heterochromatin. a) Modeling the RMDglobal1 spectrum (2540 window weights) using genome-wide profiles of various epigenomic features (x axis) using either the whole dataset, or selecting the best-predicting individual sample, or using the average across the samples. Dataset described in Table S7. **b)** Correlation between RMDglobal1 spectrum, and the differences between each pair of H3K9me3 profiles from ENCODE (blue) and randomized profiles (grey). Right: difference in H3K9me3 density for 3 example pairs of samples. Vertical lines mark the top 5% and bottom 5% windows of the RMDglobal1 spectrum (i.e. where changes in mutation risk are highest and lowest, respectively). **c)** ,Correlations between chromatin remodeling PCs (one PCA performed

per epigenomic feature) and the global RMD signatures. Black square denotes the cluster of RMDglobal1-associated PCs. **d)** Window weights on chromosome 1p for the four RMDglobal1-associated chromatin PCs and RMDglobal1 mutagenesis itself. Vertical bars as in panel **b**. **e)** Gene expression associated with the 3 relevant chromatin PCs. Gene set enrichment analysis (GSEA) scores for the MSigDB hallmark gene sets for an ordered list of genes associated with chromatin PC levels (Methods). **f)** Differences in activities of the H3K9me3_PC3 and DHS_PC3 chromatin signatures between the biosample types in ENCODE, **** is p<=0.0001 by two-sided Mann-Whitney test. H3K9me3 p-values: $p = 6.2 \cdot 10^{-7}$ (cancer CL vs tissue), $p = 3.0 \cdot 10^{-11}$ (other CL vs tissue) and $p = 1.6 \cdot 10^{-12}$ (primary cell vs tissue); with H3K9me3 n = 74 cancer CL; 40 other CL, 35 primary cell; 107 tissue. DHS p-values: $p < 2.2 \cdot 10^{-16}$ (cancer CL vs tissue), $p = 2.3 \cdot 10^{-7}$ (other CL vs tissue) and $p < 2.2 \cdot 10^{-16}$ (primary cell vs tissue); with DHS n = 153 cancer CL; 151 other CL, 229 primary cell; 143 tissue. **g)** Gene expression (square root TPM) for example genes from the proliferation-associated categories, comparing ENCODE samples with high versus low chromatin remodeling signatures. DHS_PC3 was inverted, denoted by "(-)", as an interpretation aid. H3K9me3 n = 12 PC3-high-10%; n = 13 PC3-low-10%. DHS n = 10 PC3-high-10%; n = 11 PC3-low-10%. f-g) Boxplots: the center line is the median, the box bounds the 25th and 75th percentiles and the whiskers the largest/smallest value within 1.5 times the interquartile range (IQR).

Figure 3. RMDglobal1 mutation risk redistribution is linked with RT program remodeling in cancers. a) Modelling the RMDglobal1 spectrum (2540 window weights) using genome-wide profiles of various RT features as described in Fig. 2a. Dataset described in Table S7. **b)** Correlations between the RT PCs, the 3 RMDglobal signatures, and the 4 relevant chromatin PCs from Fig 2 (in blue). Square denotes the cluster of RMDglobal1-associated RT/chromatin PCs. **c)** Median RT profiles across tumor samples with high versus low predRT-TCGA_PC5 remodelling signature. Windows with top 5% and bottom 5% RMDglobal1 mutagenesis marked with vertical lines. **d)** Gene expression associated with RT PCs (in TCGA and in ENCODE), as well as with RMDglobal1 (in HMF). Gene set enrichment analysis (GSEA) as in Fig. 2. **e)** Replicated associations of gene expression in ENCODE and the RT signature predRT-ENCODE_PC3 (y-axis) and gene expression in TCGA tumors and RT signature predRT-TCGA_PC5 (x-axis). The distribution for the genes in two significant sets are shown in orange. **f)** Associations of gene expression with a RT remodeling signature in TCGA (x axis, same as panel e), but here replicated in association of gene expression with the RMDglobal1 mutation risk redistribution in the Hartwig Medical Foundation WGS (y-axis) **g)** RMDglobal1 spectrum (window weights for n=2540 windows) across Hi-C nuclear subcompartments. **** denotes p<=0.0001 by two-sided Mann-Whitney test. p-values: $p < 2.2 \cdot 10^{-16}$ (B1 vs A1), $p < 2.2 \cdot 10^{-16}$ (B1 vs A2), $p = 1.6 \cdot 10^{-9}$ (B1 vs B2) and $p < 2.2 \cdot 10^{-16}$ (B1 vs B3). **h)** RMDglobal1 spectrum across SPIN nuclear positioning states. **** denotes p<=0.0001 by two-sided Mann-Whitney test. p-values: $p < 2.2 \cdot 10^{-16}$ (Interior_Repr2 vs Interior_Act3), $p < 2.2 \cdot 10^{-16}$ (Interior_Repr2 vs Lamina) and $p = 2.9 \cdot 10^{-14}$ (Interior_Repr2 vs Near_Lm1). g-h) Boxplots: the center line is the median, the box bounds the 25th and 75th percentiles and the whiskers the largest/smallest value within 1.5*IQR. **i)** RMDglobal1 and RMDflat signature window weights, stratified by distance to telomeres. Mean value across all chromosomes shown, separately for p and q arms. **j)** Correlation of RMDglobal1 spectrum with the CORES score describing Hi-C alterations that a domain undergoes during a whole genome

doubling [43]. The line represents the linear regression and the gray shadow the 95% intervals of the linear regression.

Figure 4. Genetic alterations associated with the activity of RMDglobal1 (RMDg1) mutagenesis. a) Schematic of the causal gene analysis in panels b-d (Methods)**. b)** Associations between somatic CNA deletions and a higher RMDglobal1 exposures in a pan-cancer analysis (n = 2875) (Methods). n=1543 cancer genes and chromatin-related genes (dots), control set of n=1000 randomly chosen genes (crosses). P-values from two-sided Z-test on regression coefficient (see Methods). **c)** Differences in RMDglobal1 exposures between *RB1* deletion (-1 or -2 CNA state) and the *RB1* wild-type tumors; separated by cancer type in Extended Data Fig. 7e-g. n=1662 tumors. P-values from Mann-Whitney test, two-tailed. **d)** Mean local CNV profile in groups of tumors, binned by RMDglobal1-high (n=81) and low exposure (n=93), in the segment of chromosome 13 containing the gene *RB1*. Each dot represents one gene. **e-f)** Overlap between the RMDglobal1 redistribution-affected domains, and the domains affected by heterochromatin remodeling (**e**) or DNA damage redistribution (**f**) in an isogenic pair of *RB1* k.o. cell lines [23]. **g**) RMDglobal1 weights near telomere, separately by chromosome arms with a *RB1* k.o. > WT change versus a WT > *RB1* k.o. change by DNA damage. **h)** Associations between deleterious SNV and indel mutations in genes as in panel **b**, and the RMDglobal1 activity of tumor samples (n = 2785, pan-cancer; Methods). P-values from two-sided Z-test on regression coefficient (see Methods). **i)** Expression level associations the MSigDB hallmark gene sets with *RB1* deletions (in Hartwig Medical Foundation and in TCGA studies), and with the RMDglobal1 mutation risk redistribution itself. **j)** Schematic of the analysis in panels k-m (Methods). **k)** Distribution for the difference in intronic mutation density for 460 cancer genes, comparing between RMDglobal1-high and low tumors. Shown separately using the actual values of RMDglobal1 and a randomized baseline**.** Vertical lines show 5th and 95th percentile of the random distribution, used as cutoffs for significance. **l)** Intronic mutation density for RMDglobal1-high versus low tumor samples (top tertile versus bottom tertile) at 5 example genes (common cancer drivers with the highest effect size). Points are cancer types: n = 4 RMDg1_high and 4 RMDg1_low. **m)** RMD profile in a region on chromosome 3p, showing the mean RMD across the RMDglobal1-high versus low tumor groups (here, top and bottom decile). Vertical lines mark the position for the *BAP1* tumor suppressor gene. c,l) Boxplots: the center line is the median, the box bounds the 25th and 75th percentiles and the whiskers the largest/smallest value within 1.5*IQR.

Figure 5. TP53 loss-of-function underlies the RMDglobal2 mutation risk redistribution signature. a) A quadratic and linear association of RMDglobal2 spectrum (window weights, n=2540) with RT. The blue lines represent the regression and the gray shading the 95% confidence intervals. **b)** Mean mutation density profiles on chromosome 4q for the RMDglobal2-high (n=7) versus low (n=129) tumors in esophagus cancer. Latest RT windows are marked with black dots. **c)** Relative RMD profile across 10 RT bins, showing the mean RMDglobal2-high (exposure > 0.17, n = 30) versus RMDglobal2-low tumor (exposures < 0.01, n=78). Cancer types with high RMDglobal2 considered: breast, lower-GI, upper-GI and prostate. P-values (*** denotes

$p<10^{-6}$) from two-sided Z-test on the regression coefficients (mean_RMD ~ RT_groups + RT_groups^2). **d)** Associations between RMDglobal2 in 2785 tumors, and deleterious mutations in cancer genes and chromatin genes and control genes (gray dots); p-values from two-sided Z-test on regression coefficient. **e)** RMDglobal2 exposures of 2785 tumors stratified by: *TP53* wild-type (wt), with 1 mutation (TP53_mut), with 1 deletion (TP53_del), TP53-loss phenocopy via CNA gain in *MDM2*, *MDM4* or *PPM1D* (TP53_pheno), or *TP53* with any two hits (TP53_2hit). P-values from two-sided Mann-Whitney test. n = 297 wt; 919 TP53_mut, 124 TP53_del, 416 TP53-loss phenocopy; 973 TP53_2hit. **f)** Relative intronic mutation density for 460 cancer genes, comparing between RMDglobal2 high and low tumors. Histograms are shown using the actual values of mutation supply difference , and using a randomized baseline. Genes known to have mutually exclusive mutations with *TP53* mutations are marked with crosses. **g)** Log2 relative intronic mutation density (normalized to flanking DNA, see Fig 4j), estimating mutation supply, for RMDglobal2-high versus RMDglobal2-low tumors for two example TP53 mutually exclusive genes. The dots are cancer types. For ARID1A n = 12 RMDg2_high and 12 RMDg2_low. For GATA3 n = 13 RMDg2_high and 11 RMDg2_low. **h)** Percentage of genes with significantly lowered mutation supply (below 5th percentile of random) upon RMDglobal2 redistribution. **i)** Mean RMD profile across the RMDglobal2-high versus low tumors in a region of chromosome 1p harboring *ARID1A*. e,g) Boxplots: the center line is the median, the box bounds the 25th and 75th percentiles and the whiskers the largest/smallest value within 1.5*IQR.

References

1. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
2. Zheng, C. L. *et al.* Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. *Cell Rep.* **9**, 1228–1234 (2014).
3. Pope, B. D. *et al.* Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405 (2014).
4. Akdemir, K. C. *et al.* Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nat. Genet.* **52**, 1178–1188 (2020).
5. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci.* **107**, 139–144 (2010).
6. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).

7. Makova, K. D. & Hardison, R. C. The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **16**, 213–223 (2015).
8. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
9. Salvadores, M., Mas-Ponte, D. & Supek, F. Passenger mutations accurately classify human tumors. *PLOS Comput. Biol.* **15**, e1006953 (2019).
10. Jiao, W. *et al.* A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* **11**, 1–12 (2020).
11. Kübler, K. *et al.* Tumor mutational landscape is a record of the pre-malignant state. 517565 Preprint at https://doi.org/10.1101/517565 (2019).
12. Koren, A. *et al.* Genetic Variation in Human DNA Replication Timing. *Cell* **159**, 1015–1026 (2014).
13. McRae, A. F. *et al.* Identification of 55,000 Replicated DNA Methylation QTL. *Sci. Rep.* **8**, 17605 (2018).
14. Oliva, M. *et al.* DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat. Genet.* **55**, 112–122 (2023).
15. Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).
16. Kinker, G. S. *et al.* Pan-cancer single cell RNA-seq uncovers recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
17. Barkley, D. *et al.* Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.* **54**, 1192–1201 (2022).
18. Du, Q. *et al.* Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat. Commun.* **10**, 416 (2019).
19. Du, Q. *et al.* DNA methylation is required to maintain both DNA replication timing precision

and 3D genome organization integrity. *Cell Rep.* **36**, 109722 (2021).

20. Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).
21. Brinkman, A. B. *et al.* Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation. *Nat. Commun.* **10**, 1749 (2019).
22. Gurrion, C., Uriostegui, M. & Zurita, M. Heterochromatin Reduction Correlates with the Increase of the KDM4B and KDM6A Demethylases and the Expression of Pericentromeric DNA during the Acquisition of a Transformed Phenotype. *J. Cancer* **8**, 2866–2875 (2017).
23. Wong, K. M., King, D. A., Schwartz, E. K., Herrera, R. E. & Morrison, A. J. Retinoblastoma protein regulates carcinogen susceptibility at heterochromatic cancer driver loci. *Life Sci. Alliance* **5**, e202101134 (2022).
24. Huang, Y., Gu, L. & Li, G.-M. H3K36me3-mediated mismatch repair preferentially protects actively transcribed genes from mutation. *J. Biol. Chem.* **293**, 7811–7823 (2018).
25. Poetsch, A. R., Boulton, S. J. & Luscombe, N. M. Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. *Genome Biol.* **19**, 215 (2018).
26. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
27. Supek, F. & Lehner, B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell* **170**, 534-547.e23 (2017).
28. Degasperi, A. *et al.* Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. *Science* **376**, abl9283 (2022).
29. 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).
30. 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.

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. Yaacov, A. *et al.* Cancer Mutational Processes Vary in Their Association with Replication Timing and Chromatin Accessibility. *Cancer Res.* **81**, 6106–6116 (2021).
33. Supek, F. & Lehner, B. Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair* **81**, 102647 (2019).
34. Liu, Y. *et al.* Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. *Nat. Commun.* **12**, 2439 (2021).
35. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417 (2015).
36. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
37. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
38. Gindin, Y., Meltzer, P. S. & Bilke, S. Replicon: a software to accurately predict DNA replication timing in metazoan cells. *Front. Genet.* **5**, (2014).
39. Pratto, F. *et al.* Meiotic recombination mirrors patterns of germline replication in mice and humans. *Cell* **184**, 4251-4267.e20 (2021).
40. Gnan, S. *et al.* Kronos scRT: a uniform framework for single-cell replication timing analysis. *Nat. Commun.* **13**, 2329 (2022).
41. 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).
42. SPIN reveals genome-wide landscape of nuclear compartmentalization | Genome Biology | Full Text. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02253-3.
43. Lambuta, R. A. *et al.* Whole-genome doubling drives oncogenic loss of chromatin

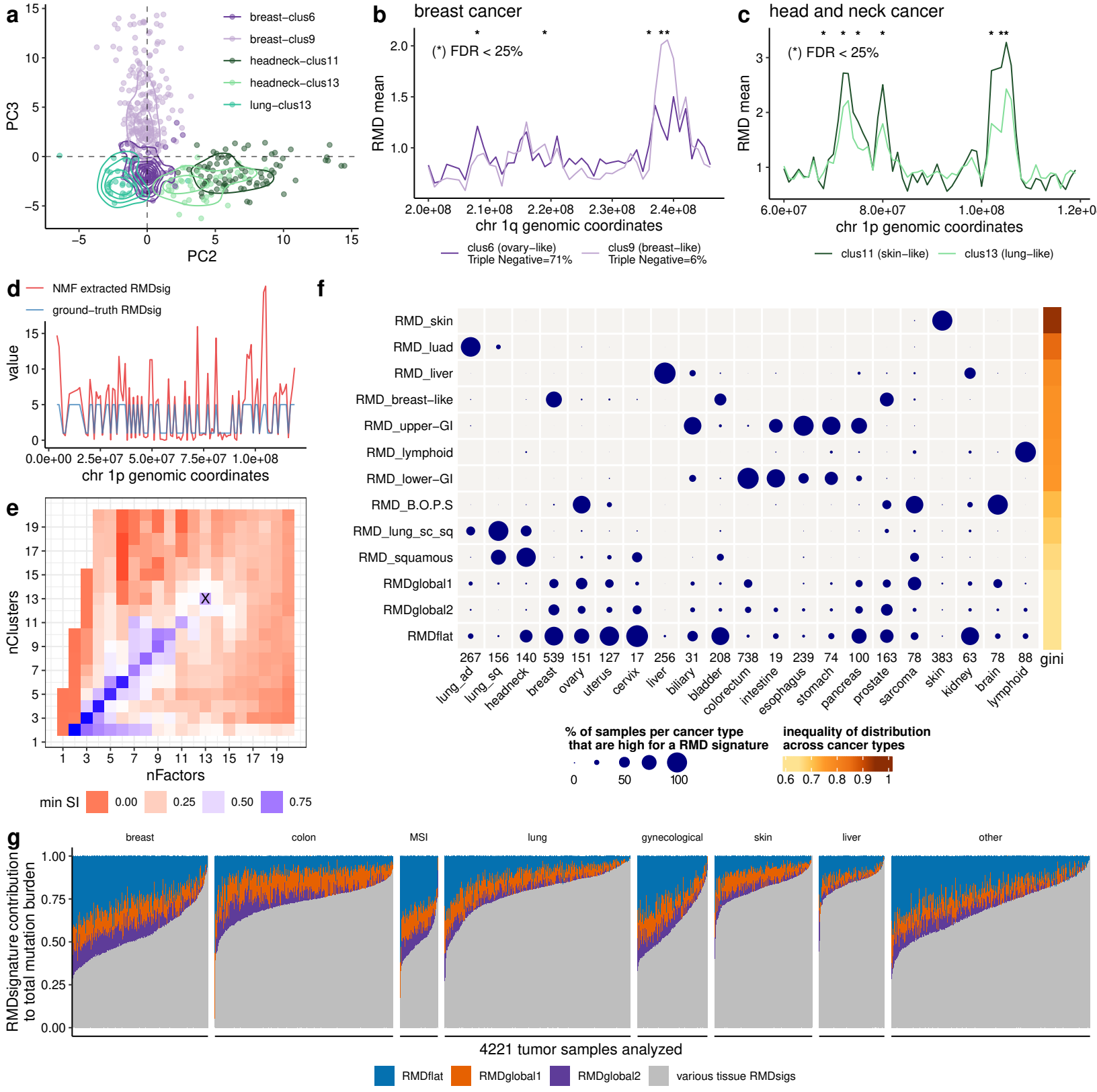
segregation. *Nature* **615**, 925–933 (2023).

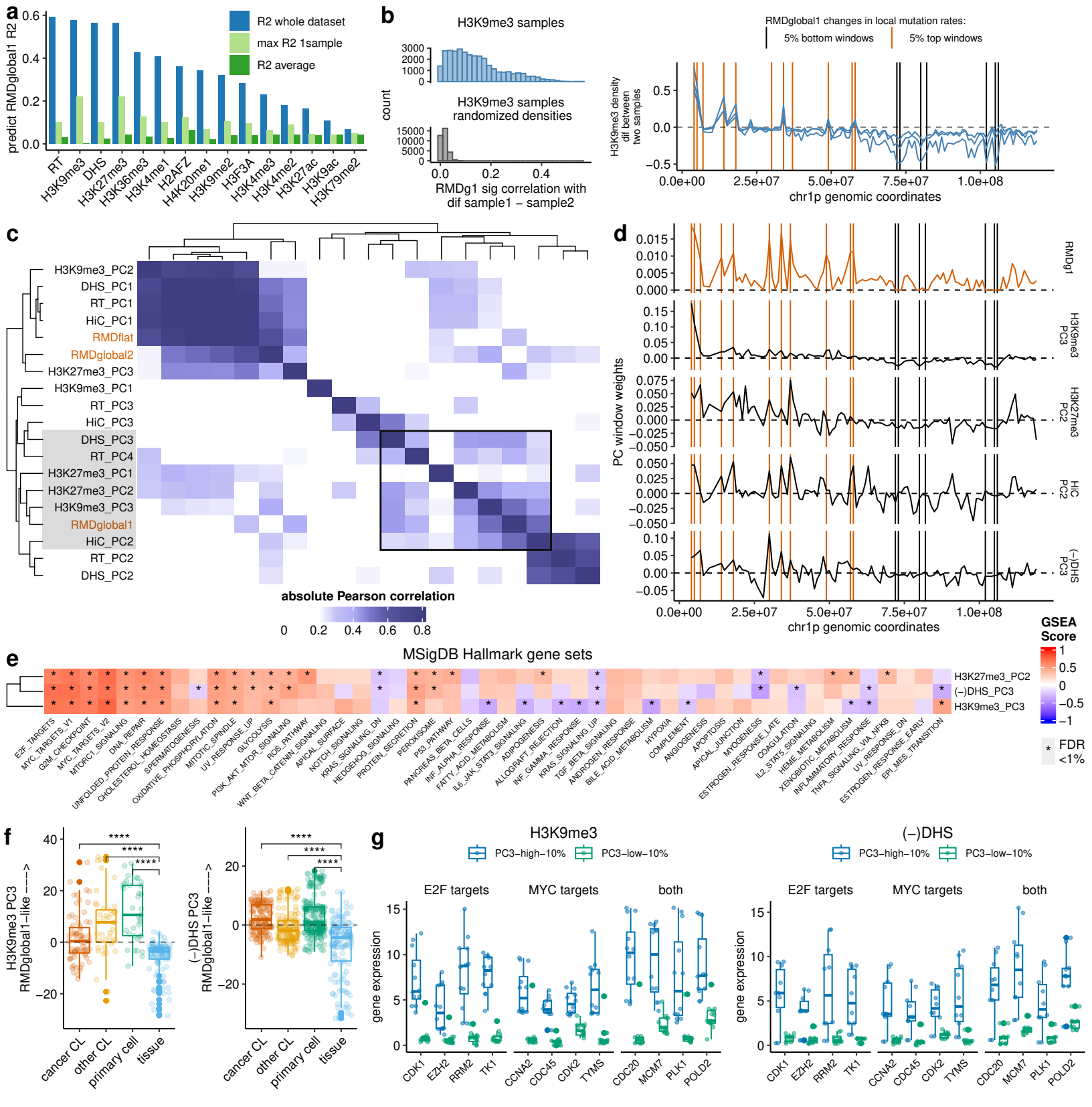
44. Gonzalo, S. *et al.* Role of the RB1 family in stabilizing histone methylation at constitutive heterochromatin. *Nat. Cell Biol.* **7**, 420–428 (2005).
45. Krishnan, B. *et al.* Active RB causes visible changes in nuclear organization. *J. Cell Biol.* **221**, e202102144 (2022).
46. Dick, F. A., Goodrich, D. W., Sage, J. & Dyson, N. J. Non-canonical functions of the RB protein in cancer. *Nat. Rev. Cancer* **18**, 442–451 (2018).
47. 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).
48. 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).
49. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).
50. Yaeger, R. *et al.* Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer. *Cancer Cell* **33**, 125-136.e3 (2018).
51. 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).
52. 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).
53. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
54. Hodgkinson, A., Chen, Y. & Eyre-Walker, A. The large-scale distribution of somatic mutations in cancer genomes. *Hum. Mutat.* **33**, 136–143 (2012).
55. Woo, Y. H. & Li, W.-H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat. Commun.* **3**, 1004 (2012).

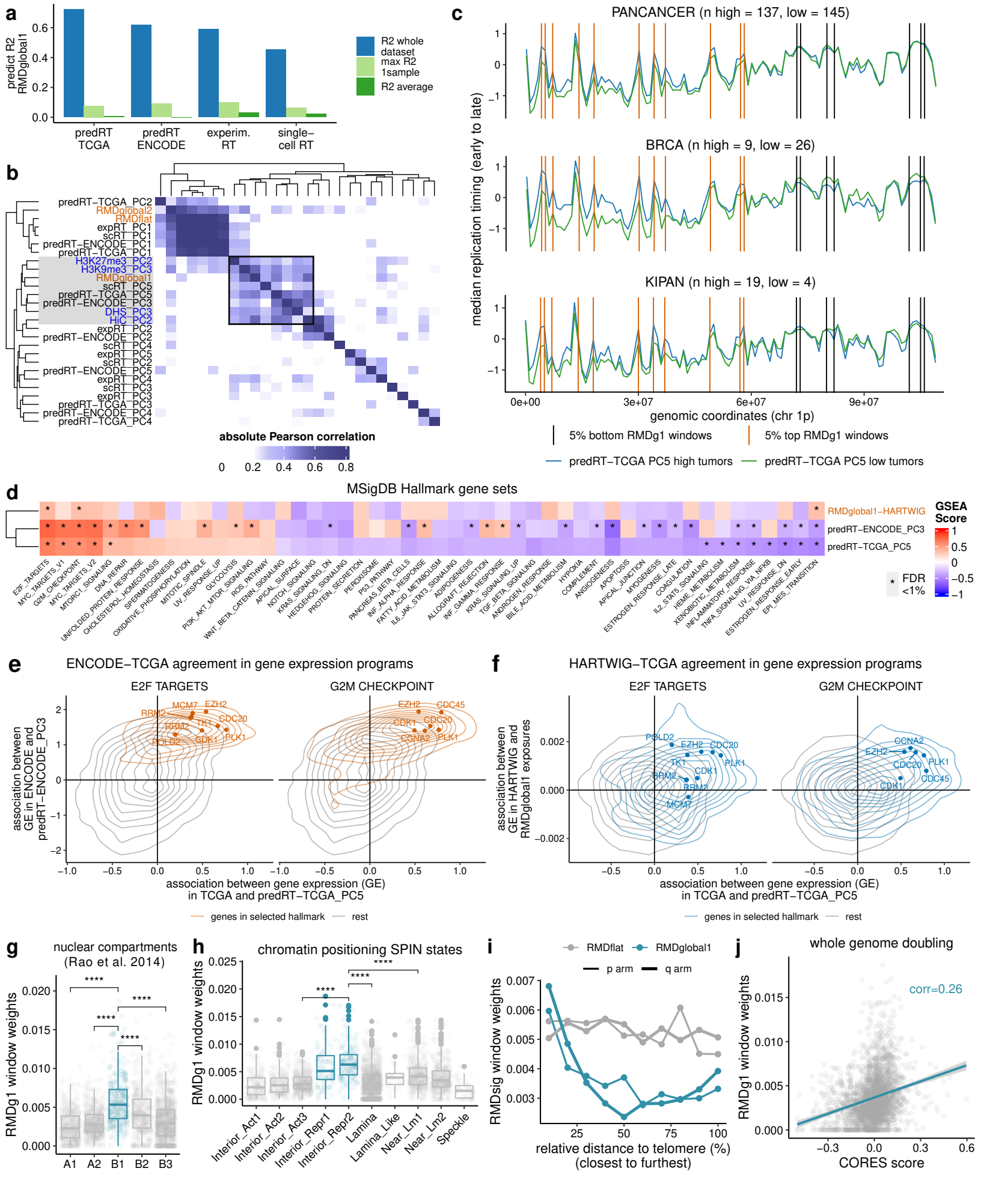
56. Liu, L., De, S. & Michor, F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.* **4**, 1502 (2013).
57. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32**, 71–75 (2014).
58. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).
59. García-Nieto, P. E. *et al.* Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. *EMBO J.* **36**, 2829–2843 (2017).
60. Hudson (Chairperson), T. J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
61. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
62. Pleasance, E. *et al.* Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* **1**, 452–468 (2020).
63. 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).
64. Edwards, N. J. *et al.* The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J. Proteome Res.* **14**, 2707–2713 (2015).
65. Walker, B. A. *et al.* A high-risk, Double-Hit, group of newly diagnosed myeloma identified by genomic analysis. *Leukemia* **33**, 159–170 (2019).
66. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
67. 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).

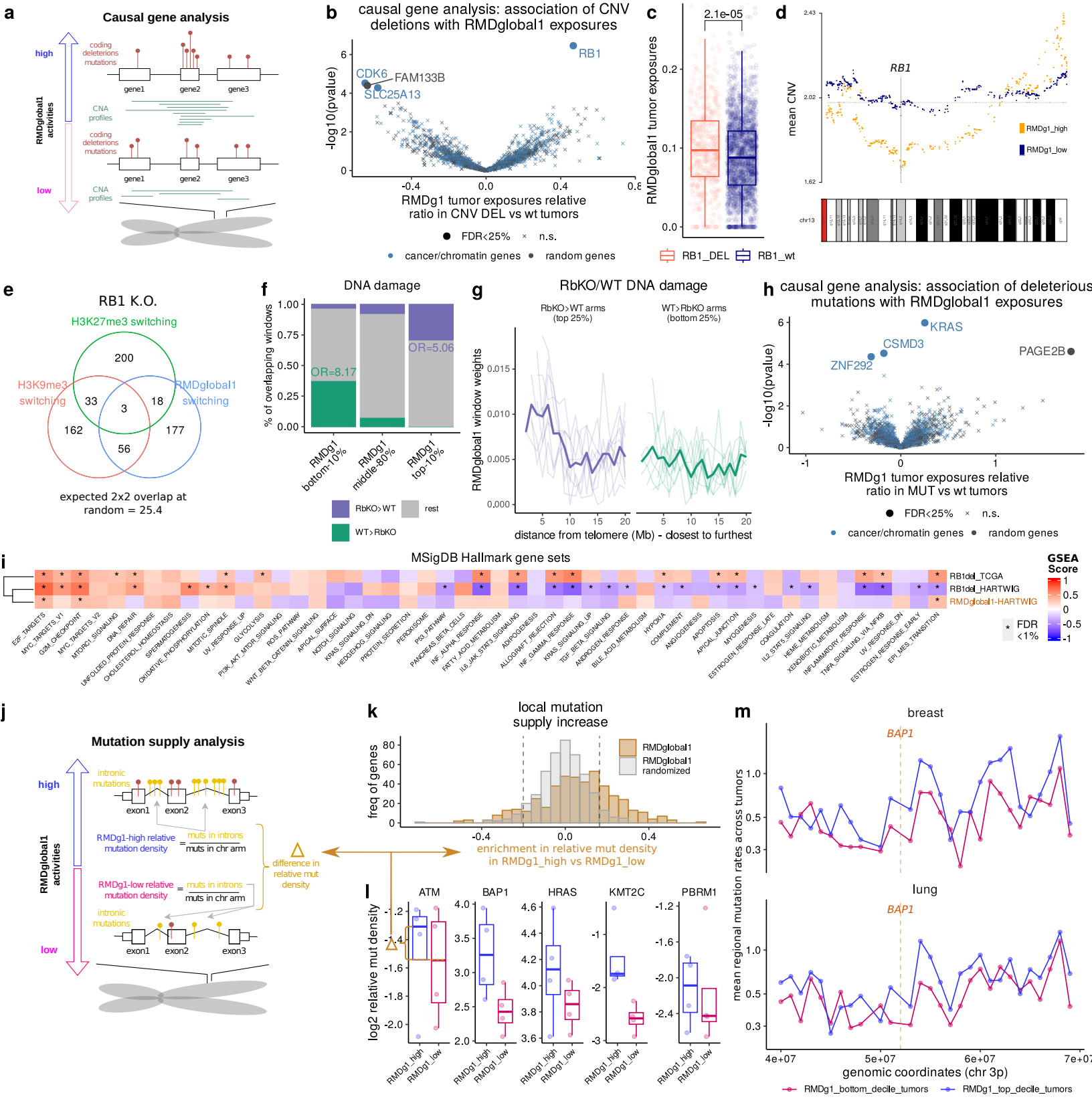
68. Derrien, T. *et al.* Fast Computation and Applications of Genome Mappability. *PLOS ONE* **7**, e30377 (2012).
69. Ormond, C., Ryan, N. M., Corvin, A. & Heron, E. A. Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Brief. Bioinform.* **22**, bbab069 (2021).
70. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).
71. Buisson, R. *et al.* Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, (2019).
72. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
73. Vali-Pour, M., Lehner, B. & Supek, F. The impact of rare germline variants on human somatic mutation processes. *Nat. Commun.* **13**, 3724 (2022).
74. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
75. 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).
76. Sima, J. *et al.* Identifying cis Elements for Spatiotemporal Control of Mammalian DNA Replication. *Cell* **176**, 816-830.e18 (2019).
77. Van Rechem, C. *et al.* Collective regulation of chromatin modifications predicts replication timing during cell cycle. *Cell Rep.* **37**, 109799 (2021).
78. 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).
79. 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).

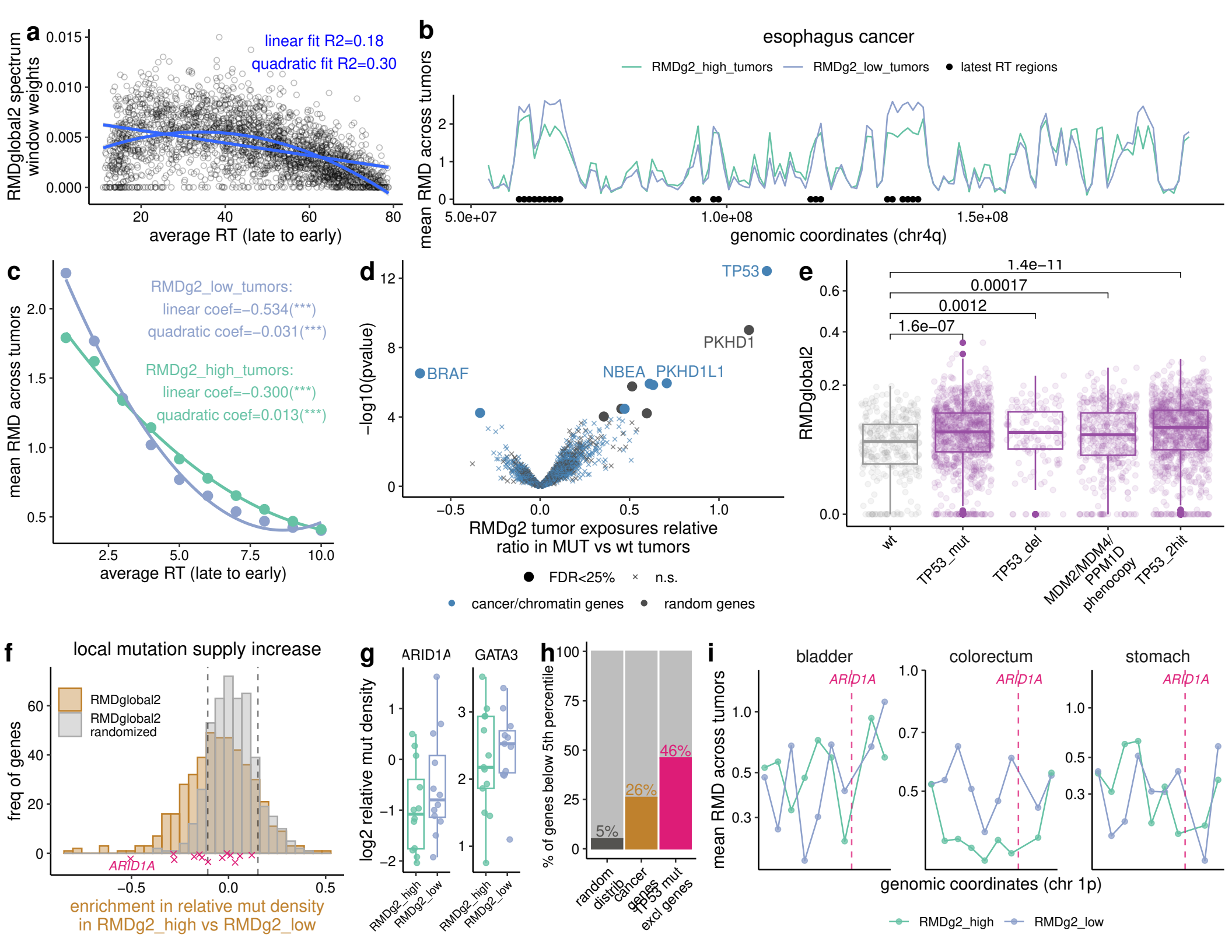
80. Poulet, A. *et al.* RT States: systematic annotation of the human genome using cell type-specific replication timing programs. *Bioinformatics* **35**, 2167–2176 (2019).
81. Klein, K. N. *et al.* Replication timing maintains the global epigenetic state in human cells. *Science* **372**, 371–378 (2021).
82. Ding, Q. *et al.* The genetic architecture of DNA replication timing in human pluripotent stem cells. *Nat. Commun.* **12**, 6746 (2021).
83. Gunasekara, C. J. *et al.* A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol.* **20**, 105–105 (2019).
84. Mukhopadhyay, R. *et al.* Allele-Specific Genome-wide Profiling in Human Primary Erythroblasts Reveal Replication Program Organization. *PLoS Genet.* **10**, e1004319 (2014).
85. Barlow, J. H. *et al.* Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. *Cell* **152**, 620–632 (2013).
86. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
87. Jönsson, J.-M. *et al.* Molecular Subtyping of Serous Ovarian Tumors Reveals Multiple Connections to Intrinsic Breast Cancer Subtypes. *PLOS ONE* **9**, e107643 (2014).
88. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
89. 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).
90. Chen, J., Miller, B. F. & Furano, A. V. Repair of naturally occurring mismatches can induce mutations in flanking DNA. *eLife* **3**, e02001 (2014).
91. Chen, D. *et al.* BRCA1 deficiency specific base substitution mutagenesis is dependent on translesion synthesis and regulated by 53BP1. *Nat. Commun.* **13**, 226 (2022).
92. 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).


a
breast−clus6
breast−clus9
headneck−clus11
headneck−clus13
lung−clus13
PC3
PC2
b
breast cancer
(*) FDR < 25%
RMD mean
chr 1q genomic coordinates
clus6 (ovary−like)
Triple Negative=71%
clus9 (breast−like)
Triple Negative=6%
c
head and neck cancer
(*) FDR < 25%
RMD mean
chr 1p genomic coordinates
clus11 (skin−like)
clus13 (lung−like)
d
NMF extracted RMDsig
ground−truth RMDsig
value
chr 1p genomic coordinates
e
nClusters
nFactors
min SI
0.00
0.25
0.50
0.75
f
RMD_skin
RMD_luad
RMD_liver
RMD_breast−like
RMD_upper−GI
RMD_lymphoid
RMD_lower−GI
RMD_B.O.P.S
RMD_lung_sc_sq
RMD_squamous
RMDglobal1
RMDglobal2
RMDflat
267 156 140 539 151 127 17 256 31 208 738 19 239 74 100 163 78 383 63 78 88
gini
lung_ad
lung_sq
headneck
breast
ovary
uterus
cervix
liver
biliary
bladder
colorectum
intestine
esophagus
stomach
pancreas
prostate
sarcoma
skin
kidney
brain
lymphoid
% of samples per cancer type that are high for a RMD signature
inequality of distribution across cancer types
g
breast
colon
MSI
lung
gynecological
skin
liver
other
RMDsignature contribution to total mutation burden
4221 tumor samples analyzed
RMDflat
RMDglobal1
RMDglobal2
various tissue RMDsigs


a
predict RMDglobal1 R2
R2 whole dataset
max R2 1sample
R2 average
RT
H3K9me3
DHS
H3K27me3
H3K36me3
H3K4me1
H2AFZ
H4K20me1
H3K9me2
H3F3A
H3K4me3
H3K4me2
H3K27ac
H3K9ac
H3K79me2
b
H3K9me3 samples
count
H3K9me3 samples
randomized densities
RMDg1 sig correlation with
dif sample1 – sample2
RMDglobal1 changes in local mutation rates:
5% bottom windows
5% top windows
H3K9me3 density
dif between
two samples
chr1p genomic coordinates
c
H3K9me3_PC2
DHS_PC1
RT_PC1
HiC_PC1
RMDflat
RMDglobal2
H3K27me3_PC3
H3K9me3_PC1
RT_PC3
HiC_PC3
DHS_PC3
RT_PC4
H3K27me3_PC1
H3K27me3_PC2
H3K9me3_PC3
RMDglobal1
HiC_PC2
RT_PC2
DHS_PC2
absolute Pearson correlation
d
PC window weights
RMDg1
H3K9me3 PC3
H3K27me3 PC2
HiC PC2
(–)DHS PC3
chr1p genomic coordinates
e
MSigDB Hallmark gene sets
H3K27me3_PC2
(–)DHS_PC3
H3K9me3_PC3
GSEA Score
FDR <1%
E2F_TARGETS
MYC_TARGETS_V1
G2M_CHECKPOINT
MYC_TARGETS_V2
MTORC1_SIGNALING
DNA_REPAIR
UNFOLDED_PROTEIN_RESPONSE
CHOLESTEROL_HOMEOSTASIS
SPERMATOGENESIS
OXIDATIVE_PHOSPHORYLATION
MITOTIC_SPINDLE
UV_RESPONSE_UP
GLYCOLYSIS
PI3K_AKT_MTOR_SIGNALING
ROS_PATHWAY
WNT_BETA_CATENIN_SIGNALING
APICAL_SURFACE
NOTCH_SIGNALING_DN
KRAS_SIGNALING_DN
HEDGEHOG_SIGNALING
PROTEIN_SECRETION
PEROXISOME
P53_PATHWAY
PANCREAS_BETA_CELLS
INF_ALPHA_RESPONSE
FATTY_ACID_METABOLISM
IL6_JAK_STAT3_SIGNALING
ADIPOGENESIS
ALLOGRAFT_REJECTION
INF_GAMMA_RESPONSE
KRAS_SIGNALING_UP
TGF_BETA_SIGNALING
ANDROGEN_RESPONSE
BILE_ACID_METABOLISM
HYPOXIA
COMPLEMENT
ANGIOGENESIS
APOPTOSIS
APICAL_JUNCTION
MYOGENESIS
ESTROGEN_RESPONSE_LATE
COAGULATION
IL2_STAT5_SIGNALING
HEME_METABOLISM
XENOBIOTIC_METABOLISM
INFLAMMATORY_RESPONSE
TNFA_SIGNALING_VIA_NFKB
UV_RESPONSE_DN
ESTROGEN_RESPONSE_EARLY
EPI_MES_TRANSITION
f
H3K9me3 PC3
RMDglobal1–like ——>
(–)DHS PC3
RMDglobal1–like ——>
cancer CL
other CL
primary cell
tissue
g
H3K9me3
(–)DHS
PC3–high–10%
PC3–low–10%
E2F targets
MYC targets
both
gene expression
CDK1
EZH2
RRM2
TK1
CCNA2
CDC45
CDK2
TYMS
CDC20
MCM7
PLK1
POLD2


a
predict R2 RMDglobal1
R2 whole dataset
max R2 1sample
R2 average
predRT TCGA
predRT ENCODE
experim. RT
single–cell RT
b
absolute Pearson correlation
c
PANCANCER (n high = 137, low = 145)
BRCA (n high = 9, low = 26)
KIPAN (n high = 19, low = 4)
median replication timing (early to late)
genomic coordinates (chr 1p)
5% bottom RMDg1 windows
5% top RMDg1 windows
predRT–TCGA PC5 high tumors
predRT–TCGA PC5 low tumors
d
MSigDB Hallmark gene sets
RMDglobal1–HARTWIG
predRT–ENCODE_PC3
predRT–TCGA_PC5
GSEA Score
FDR <1%
e
ENCODE–TCGA agreement in gene expression programs
E2F TARGETS
G2M CHECKPOINT
association between GE in ENCODE and predRT–ENCODE_PC3
association between gene expression (GE) in TCGA and predRT–TCGA_PC5
genes in selected hallmark
rest
f
HARTWIG–TCGA agreement in gene expression programs
association between GE in HARTWIG and RMDglobal1 exposures
g
nuclear compartments (Rao et al. 2014)
RMDg1 window weights
h
chromatin positioning SPIN states
i
RMDflat
RMDglobal1
p arm
q arm
RMDsig window weights
relative distance to telomere (%) (closest to furthest)
j
whole genome doubling
corr=0.26
CORES score


a
Causal gene analysis
high
low
RMDglobal1 activities
coding deleterious mutations
CNA profiles
gene1
gene2
gene3
b
causal gene analysis: association of CNV deletions with RMDglobal1 exposures
RB1
CDK6
FAM133B
SLC25A13
-log10(pvalue)
RMDg1 tumor exposures relative ratio in CNV DEL vs wt tumors
FDR<25%
n.s.
cancer/chromatin genes
random genes
c
2.1e-05
RMDglobal1 tumor exposures
RB1_DEL
RB1_wt
d
RB1
mean CNV
RMDg1_high
RMDg1_low
chr13
e
RB1 K.O.
H3K27me3 switching
H3K9me3 switching
RMDglobal1 switching
200
33
3
18
162
177
56
expected 2x2 overlap at random = 25.4
f
DNA damage
% of overlapping windows
OR=8.17
OR=5.06
RMDg1 bottom-10%
RMDg1 middle-80%
RMDg1 top-10%
RbKO>WT
WT>RbKO
rest
g
RbKO/WT DNA damage
RbKO>WT arms (top 25%)
WT>RbKO arms (bottom 25%)
RMDglobal1 window weights
distance from telomere (Mb) - closest to furthest
h
causal gene analysis: association of deleterious mutations with RMDglobal1 exposures
KRAS
CSMD3
ZNF292
PAGE2B
-log10(pvalue)
RMDg1 tumor exposures relative ratio in MUT vs wt tumors
FDR<25%
n.s.
cancer/chromatin genes
random genes
i
MSigDB Hallmark gene sets
RB1del_TCGA
RB1del_HARTWIG
RMDglobal1-HARTWIG
GSEA Score
FDR <1%
E2F_TARGETS
MYC_TARGETS_V1
G2M_CHECKPOINT
MYC_TARGETS_V2
MTORC1_SIGNALING
DNA_REPAIR
UNFOLDED_PROTEIN_RESPONSE
CHOLESTEROL_HOMEOSTASIS
SPERMATOGENESIS
OXIDATIVE_PHOSPHORYLATION
MITOTIC_SPINDLE
UV_RESPONSE_UP
GLYCOLYSIS
PI3K_AKT_MTOR_SIGNALING
ROS_PATHWAY
WNT_BETA_CATENIN_SIGNALING
APICAL_SURFACE
NOTCH_SIGNALING
KRAS_SIGNALING_DN
HEDGEHOG_SIGNALING
PROTEIN_SECRETION
PEROXISOME
P53_PATHWAY
PANCREAS_BETA_CELLS
INF_ALPHA_RESPONSE
FATTY_ACID_METABOLISM
IL6_JAK_STAT3_SIGNALING
ADIPOGENESIS
ALLOGRAFT_REJECTION
INF_GAMMA_RESPONSE
KRAS_SIGNALING_UP
TGF_BETA_SIGNALING
ANDROGEN_RESPONSE
BILE_ACID_METABOLISM
HYPOXIA
COMPLEMENT
ANGIOGENESIS
APOPTOSIS
APICAL_JUNCTION
MYOGENESIS
ESTROGEN_RESPONSE_LATE
COAGULATION
IL2_STAT5_SIGNALING
HEME_METABOLISM
XENOBIOTIC_METABOLISM
INFLAMMATORY_RESPONSE
TNFA_SIGNALING_VIA_NFKB
UV_RESPONSE_DN
ESTROGEN_RESPONSE_EARLY
EPI_MES_TRANSITION
j
Mutation supply analysis
high
low
RMDglobal1 activities
intronic mutations
exon1
exon2
exon3
RMDg1-high relative mutation density = muts in introns / muts in chr arm
RMDg1-low relative mutation density = muts in introns / muts in chr arm
difference in relative mut density
k
local mutation supply increase
freq of genes
RMDglobal1
RMDglobal1 randomized
enrichment in relative mut density in RMDg1_high vs RMDg1_low
l
ATM
BAP1
HRAS
KMT2C
PBRM1
log2 relative mut density
RMDg1_high
RMDg1_low
m
breast
lung
BAP1
mean regional mutation rates across tumors
genomic coordinates (chr 3p)
RMDg1_bottom_decile_tumors
RMDg1_top_decile_tumors


a
RMDglobal2 spectrum window weights
linear fit R2=0.18
quadratic fit R2=0.30
average RT (late to early)
b
esophagus cancer
RMDg2_high_tumors
RMDg2_low_tumors
latest RT regions
mean RMD across tumors
genomic coordinates (chr4q)
c
mean RMD across tumors
RMDg2_low_tumors:
linear coef=−0.534(***)
quadratic coef=−0.031(***)
RMDg2_high_tumors:
linear coef=−0.300(***)
quadratic coef=0.013(***)
average RT (late to early)
d
−log10(pvalue)
TP53
PKHD1
BRAF
NBEA
PKHD1L1
RMDg2 tumor exposures relative ratio in MUT vs wt tumors
FDR<25%
n.s.
cancer/chromatin genes
random genes
e
RMDglobal2
1.4e−11
0.00017
0.0012
1.6e−07
wt
TP53_mut
TP53_del
MDM2/MDM4/ PPM1D phenocopy
TP53_2hit
f
local mutation supply increase
RMDglobal2
RMDglobal2 randomized
freq of genes
ARID1A
enrichment in relative mut density in RMDg2_high vs RMDg2_low
g
ARID1A
GATA3
log2 relative mut density
RMDg2_high
RMDg2_low
h
% of genes below 5th percentile
5%
26%
46%
random distrib
cancer genes
TP53 mut excl genes
i
bladder
colorectum
stomach
ARID1A
mean RMD across tumors
genomic coordinates (chr 1p)
RMDg2_high
RMDg2_low