**Methods**

**Simulating synthetic cancer datasets**

Synthetic cancer datasets were simulated using SigProfilerSimulator (<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03772-3>).

**Annotating somatic indels based on transcribed versus non-transcribed strand**

We followed the method in (<https://doi.org/10.1016/j.celrep.2023.112930>). Briefly somatic indels were called with respect to the + strand of the reference genome and further annotated in regard to the pyrimidine base(s) of the insertion/deletion. Thus, indels with only C or T bases were annotated as + strand mutations; indels with only A or G bases were annotated as – strand mutations. The remaining indels were not included in the analysis. Next, + strand indels in protein coding genes were further subclassified as transcribed (template) if the gene’s sense strand was on the + strand of the genome, or else un-transcribed (sense). The logic was inverted for – strand indels. 3wsIndels in bidirectionally transcribed regions were ignored. (non transcribed?)

**Annotating somatic indels based on leading versus lagging replication strand**

Replication strand was determined by wavelet-smoothed replication-timing signal data that indicated both “valleys” (replication termination zones) and “peaks” (replication initiation zones) (<https://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/>). Valleys and peaks were sorted by the genomic coordinate in ascending order. In regard to + strand of the reference genome, replication timing signal were examined for consecutive stretches of the genome (from valley to peak or form peak to valley), with positive slope corresponded to leading strand regions and negative slope corresponded to lagging strand regions. Then for the - strand of the reference genome, leading regions (- slope) and lagging regions (+slope) were automatically acquired. Similar to the annotation for transcription, indels were first annotated as + or – strand mutations based on the pyrimidine bases. Next, indels were counted as being on leading strand or lagging strand based on their occupancy in a leading or lagging region.

**Detecting strand asymmetries across cancer types**

Strand asymmetry analyses were based on the assignment of signature probabilities to each individual indel mutation (expand if needed). Only indels with the probability greater than or equal to 0.50 to a certain ID signature were retained. For each ID signature and for all cancer types having this mutational signature, we retrieved the number of indels on each strand/region. In strand asymmetries analyses, only cancer types with at least 1,000 somatic mutations unambiguously attributed to an individual mutational signature were included.

For each strand asymmetry analyses (genic and intergenic region asymmetry, transcription strand asymmetry, replication strand asymmetry), indel mutations were split into two types and counted (genic vs intergenic mutation, leading strand vs lagging strand mutation, un-transcribed strand vs transcribed strand mutation). The two types were denoted as +/- strand mutations strand mutations in all three cases. The ratio of real somatic indels and the ratio of simulated somatic indels was calculated separately:

Odds ratio between the ratio of real somatic indels and the ratio of simulated somatic indels was calculated:

p values were calculated for the odds ratio using Fisher’s exact test. Only strand asymmetries with p value > 0.05 and odds ratios above 1.10 were considered showing strand asymmetries (Do we want to use this criteria?).

**Analyses of replication timing across cancer types**

Replication timing data were obtained from XXX(per\_base\_territories\_20kb (2).mat – where does this table come from?) <was this the source of the peaks and valleys above?>. The replication time signals were sorted in a descending order and subsequently divided into deciles. Somatic indels were counted within the corresponding deciles based on their overlap with the replication domains in the examined deciles. As with other analyses, for each individual ID signature, the reported replication timing analyses included only cancer types with at least 1,000 somatic mutations unambiguously attributed to it.

Replication timing mutation counts were generated for both real and simulated somatic indels.

Normalized mutation density vectors across replication timing deciles were calculated as following:

To classify whether the replication timing mutation density was increasing, flat, or decreasing, a linear regression model was fitted to the values of the normalized mutation densities. An ID signature was considered to be generally unaffected by replication timing if the slope m was not statistically significant from a flat line. Otherwise, with the slope m statistically significant from a flat line, an ID signature was considered to be increasing from early to late replicating regions if the slope m > 0, and was considered to be decreasing from early to late replicating regions if the slope m < 0.

**Results**

**Transcriptional strand bias: Transcribed (red) vs. Un-transcribed (blue)**

A diagram of a crossword puzzle

Description automatically generated

**Replicational strand bias: Lagging (red) vs. Leading (blue)**

A diagram of a crossword puzzle

Description automatically generated

**Genic (Red) vs. intergenic regions (Blue)**

A diagram of dna and genetic code

Description automatically generated with medium confidence

**Replication timing (from early to late)**

**A chart of green and red bars

Description automatically generated**

**Replication timing (from early to late)**

with real mutation normalized by simulated mutation

**A chart of green and white bars

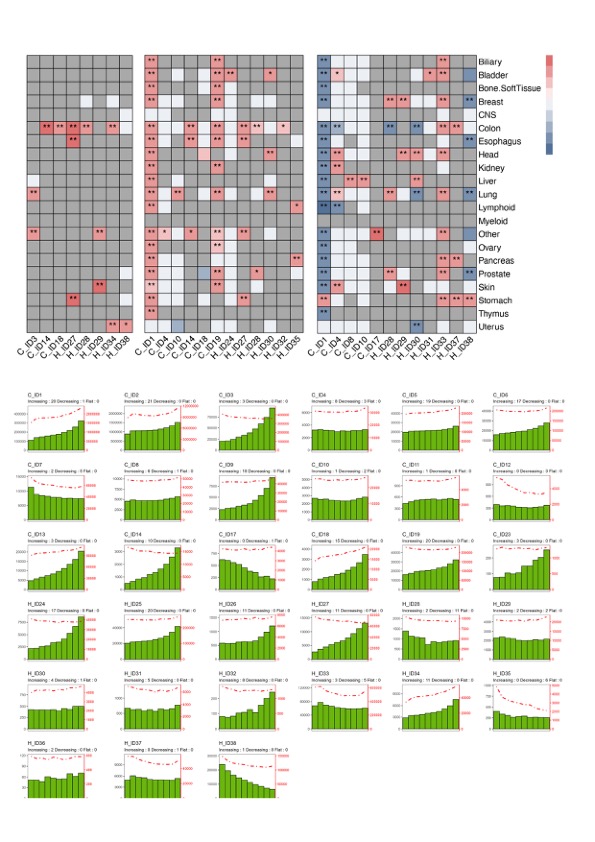
Description automatically generated with medium confidence**

**Replication timing ID89 (from early to late)**

with only real mutation data

****

**Possible topography analyses figure:**

****