Supplemental Materials and Methods

Somatic variants from cancer data

Data were obtained from whole genome sequenced cancers for breast cancer (n=560) from (Nik-Zainal et al. 2016) and from 9 cancer types publicly available in ICGC (The International Cancer Genome Consortium 2010). The ICGC project codes for the cancer types were: PACA-CA (n=148) and PACA-AU (n=94) for pancreatic cancer (Waddell et al. 2015), (Notta et al. 2016), OV-AU (n=72) for ovarian cancer (Patch et al. 2015), LIRI-JP (n=264) for liver cancer (Fujimoto et al. 2016), PRAD-CA (n=120) for prostate cancer (Fraser et al. 2017), ESAD-UK (n=98) for esophageal adenocarcinoma (The Cancer Genome Atlas Research Network 2017), GACA-CN (n=40) for gastric cancer (The International Cancer Genome Consortium 2010), RECA-EU (n=74) for renal cell cancer (The International Cancer Genome Consortium 2010), PBCA-DE (n=239) for pediatric brain cancer (The International Cancer Genome Consortium 2010) and MALY-DE (n=100) for malignant lymphoma (The International Cancer Genome Consortium 2010). In total, 1809 whole genome sequenced cancers were analysed. Sequencing coverage exceeded 25X for all tumours and matched normal samples.

Short insert paired-end reads were aligned to the reference human genome (GRCh37) using Burrows-Wheeler Aligner, BWA (v0.5.9).

High quality curated somatic variant calls (substitutions, insertions/deletions and structural variations) were derived from the Wellcome Trust Sanger Institute's Cancer Genome Project whole genome sequencing pipeline as previously described (Nik-Zainal

et al. 2016). This is constituted by a bespoke, Expectation-Maximisation- based substitution-calling algorithm (CaVEMan), (Jones et al. 2016), (Nik-Zainal et al. 2012) a modified version of an insertion/deletion detection algorithm, Pindel (Ye et al. 2009) and a bespoke structural variant algorithm which uses de Bruijn graphing for discovery of somatic rearrangements and local reassembly for mapping breakpoints to base pair level.

A subset of all somatic variants for breast cancer samples had been previously validated using alternative sequencing platforms to ensure high specificity of data (Nik-Zainal et al. 2016). In short, 70 samples were used for validation across a mix of histopathological subtypes and were sourced from different collaborating centres.

- On average 3% (range 0.6-20%) of the total burden of substitutions per sample were used for validation (total 11,581 mutations). The positive predictive value was ~95.5% (average) for substitutions.
- On average of 40% (range 8%-68%) of the total number of indels were validated per sample (total 7,192). The positive predictive value was 85% for indels.
- Rearrangements were discovered using Brass I and an additional *in silico* method was used (*de novo* breakpoint assembly) to validate the finding. Only breakpoints that were *de novo* assembled with high confidence (80% and above only) were included in order to reduce the likelihood of false positive calls. PCR-based Sanger sequencing validation confirmed the presence of 803 randomly sampled breakpoints from this conservative dataset.

Supplementary Table1: Number of substitutions, indels and rearrangement breakpoints per tumour type.

Cancer Name	Samples	Substitutions	Indels	Rearrangement breakpoints
BRCA	560	3,479,651	371,993	131,068
LIRI	264	3,575,056	852,361	51,034
OVCA	72	732,189	141,296	39,078
ESAD	98	2,890,654	347,680	48,394
GACA	40	525,850	185,213	12,268
PBCA	239	299,241	231,874	13,120
PACA	242	1,881,336	625,803	48,404
RECA	74	584,144	123,180	1,972
MALY	100	1,242,356	203,051	10,752
PRAD	120	602,729	799,583	24,104

Furthermore, these datasets have all been published and therefore been through peer-review previously.

Simulations were performed for 10% randomly selected substitutions for each tumour type. The controls generated controlled for trinucleotide content. For each generated simulated mutation, the site of the substitution was excluded from the search space, and a site of the same trinucleotide content was randomly selected within a window of 50kb, therefore also controlling for genomic location.

The reference genome hg19 was used throughout the manuscript. The results would not differ, if the analysis was performed in GRCh38. This is because GRCh38 differs from hg19 in its annotation of low mappability and centromeric sites, at which we do not call mutations.

Reference non-B DNA annotations

Non-B DNA sequence motif annotations were derived from (Cer et al. 2013). We have focused on the following categories in this analysis: Mirror repeats, H-DNA, short tandem repeats, Z-DNA, inverted repeats, direct repeats and G-quadruplexes.

- A mirror repeat is a section of sequence that is repeated with a center of symmetry on the same strand, length of at least 20nt and arm size of at least 10nt. A subset of mirror repeats are termed Hinged DNA (H-DNA), because they are predisposed to forming a triple helical structure connected through alternative chemical bonds called Hoogsteen bonds. H-DNA have a high (>90%) AG content, arm lengths of >=10nt and spacer size of less than 8nt.
- Z-DNA is a left-handed double helical structure that is formed by alternating purine-pyrimidine tracts of at least 12nt (excluding AT repeats).

- Direct repeats are defined as repeated sequences with arm length of >=10nt, with maximum size 300nt.
- Short tandem repeats are (also called microsatellite repeats) defined as motifs of 1-9nt, repeated at least 3 times with a minimum length of 9nt and without any interruptions. Short tandem repeats are prone to misalignment and formation of looped or slipped structures.
- Inverted repeats are palindromic sequences with minimum arm length of 6nt, spacer size up to 100nt and have a tendency to form hairpin or cruciform structures.
- G-quadruplexes are defined as 4 or more runs of at least 3 guanines, separated by spacers of 1-7nt of other nucleotides. For G-quadruplex motifs we referred to G-runs as the guanine runs that can form Hoogsteen bonds and the loops as the spacer between the G-runs.

BEDTools utilities v2.21.0 were used to manipulate genomic files and intervals (Quinlan and Hall 2010).

To count the number of nucleotides shared between different non-B DNA motifs, “bedtools intersect” and “bedtools coverage” functions were used. The command “bedtools jaccard” was used to calculate the Jaccard index for each pair of motifs (Fig. S1b).

Epigenomic Data

DNase and histone modification narrowpeak files were downloaded from (Roadmap Epigenomics Consortium 2015) (http://www.roadmapepigenomics.org/data/) and BAM files were derived from (The ENCODE Project Consortium 2012) (http://genome.ucsc.edu/ENCODE/downloads.html). HMEC cell line epigenetic narrowpeak data were used to model breast cancer, whereas PANC1, HepG2 and GM12878 cell line narrowpeak epigenomic data were used to model pancreatic cancer, liver cancer and malignant lymphoma. Ovary, esophagus, fetal female brain, stomach mucosa, fetal kidney primary tissue narrowpeak epigenomic data were used to model ovarian cancer, esophageal carcinoma, pediatric brain cancer, gastric cancer and renal cell cancer respectively. BAM files for the same epigenetic modifications were derived from (The ENCODE Project Consortium 2012) to validate the findings derived from narrowpeak files, for MCF-7 cell line, which is used to model breast cancer.

Chromatin States

Chromatin states represent partitions of the genome derived using chromatin modification patterns in commonly used cell lines. Here chromatin states were defined with Segway as described in (Hoffman et al. 2012), (Hoffman et al. 2013) using chromatin modifications from (The ENCODE Project Consortium 2012) for 6 human cell lines (GM12878, H1-Hesc, HepG2, HUVEC, K562, HelaS3) and downloaded from http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgSegmentation/. Segway regulatory segment tracks display labeled, non-overlapping partitions of the genome of regulatory sites with biological functions and are produced by an unsupervised pattern discovery algorithm. The labeled partitions were grouped in: CTCF, DNase, transcription associated, candidate strong enhancer, candidate weak enhancer, low

activity proximal to active states, promoter flanking, inactive promoter, heterochromatin-repetitive-copy number variation, Polycomb repressed and active promoter.

The background or "expected" density of each non-B DNA motif (DN-background) was calculated as the total number of occurrences of the motif (TO) over all mappable nucleotides across the Segway states (TN). The density of each non-B DNA motif at a particular state (DN-specific) was calculated as the fraction of the number of occurrences of the non-B DNA motif at that state (SO) over the number of mappable nucleotides covered in that state (SN). The enrichment of a non-B DNA motif at a given state was the fraction of the density at the state over the background density of the motif.

Background density of a non-B DNA motif: DN-background: TO /TN
Density of a non-B DNA motif at the state = DN-specific: SO /SN
Enrichment: DN-specific / DN-background

The mean enrichment across 6 human cell lines (GM12878, H1-Hesc, HepG2, HUVEC, K562, HelaS3) was calculated. Hierarchical clustering of chromatin states and plotting was performed with the python package "seaborn" using default parameters (Fig. 1i).

Repli-Seq Data

Reference coordinates for replication landmarks were inferred from Repli-Seq data of 14 cell-lines, which were NHEK, IMR90, HUVEC, HeLa-S3, GM12813, GM12812, GM12801, GM06990, BJ, BG02ES, MCF-7, GM12878, HepG2 and K562. Repli-Seq data were obtained from (The ENCODE Project Consortium 2012)

(https://www.encodeproject.org/) and processed as described in (Morganella et al. 2016). Replication timing was measured at each genomic interval using "bedtools map" utility function. Repli-Seq data for MCF-7 were used for breast cancer, HepG2 Repli-seq data were used for liver cancer, GM12878 for malignant lymphoma and MCF-7 for all other cancer types. A positive correlation between mutations and replication time indicates positive correlation for early replication time domains and mutations, while a negative correlation denotes a positive correlation for late replication time domains (Fig. 2a, Fig. S4). Pearson correlation between any two cell lines with Repli-Seq data exceeded 0.69 in all cases, using 500kb genomic windows (Fig. S2).

Modelling the relationship between mutations and genomic features (epigenetics, replication time domains and non-B DNA motifs)

The human genome (hg19) was partitioned in equal-sized regions of 500kb segments. Centromeric sites, simple repeats and regions of excessive sequencing depth (UCSC Top 0.01 Hi Seq Depth) were downloaded from the UCSC genome browser and used to identify bins with low mappability. The command "bedtools coverage" was used to calculate the coverage of centromeric sites and low complexity sequences at each genomic interval. We excluded the first and last bin from each chromosome as well as any bin where <50% of the bases were mappable or where replication time data is missing, and the sex chromosomes. This resulted in 5,581 non-overlapping bins. All quantities except for the replication times were transformed as $x' = \log2(1 + x)$ for the downstream analysis.

To map reads from histone modification BAM files from (The ENCODE Project Consortium 2012) at each genomic interval, “bedtools multicov” utility function was used. In case of multiple replicates per file, the mean number of reads per segment was calculated across replicates.

To calculate the number of non-B DNA motifs, mutations, genomic features and narrowpeak files from (Roadmap Epigenomics Consortium 2015) at each genomic segment, BEDtools “intersect” utility was used (bedtools intersect -a segments.file -b mutation.file with flags -u, -v, -c). The statistics of non-B DNA motifs at 500kb windows across the humans are provided in Supplementary Table 2.

The command “bedtools nuc” was used to calculate the GC content at each interval as well as the number of As, Gs, Cs, Ts and Ns at each interval for the hg19 reference genome.

Partial correlation is a measure of association between two variables, controlling for the effect of covariates. Partial correlations were applied to measure the relationship between mutations and non-B DNA motifs, controlling for the effect of epigenetic markers and replication timing. Partial correlations were calculated in R with the package ‘ppcor’ (Kim 2015). Results are noted in Fig. S5.

Linear and random forest regression

To model the relationship between the number of mutations and a plethora of explanatory variables we applied two predictive models; linear regression and random forest regression. In the former, additive relationships are modeled using linear predictor

functions, whereas a random forest model is an ensemble-learning model in which multiple regression trees are constructed and evaluated. In both models, the relative importance of each predictor variable can be measured. The two models were applied independently to each cancer type. Prostate cancer, for which epigenetic data from a relevant cell of origin were not available, was excluded.

Both models were evaluated using 10-fold cross-validation, whereby the model was trained using 90% of the data and tested using the held out 10%. The same bins and transformations that were used for the correlation calculations were used for the regression. For the linear model we used the command “lm” in R and for the random forest regression, we used the R-package “randomForest” with default parameters. For the random forest regression model feature importance was measured using the predictive measure of the original and the permuted dataset. In particular, the variable importance for Fig. 2 panels d and e was evaluated using the R-package “pRF”, which uses a permutation test. The parameters for the pRF function were “n.perms = 200” and “mtry = 4”. The biplot in Fig. 2f represents the first two loadings obtained using the princomp command in R.

Supplementary Table 2: Statistics of non-B DNA motifs at genomic bins of 500kb across the human genome.

Non-B DNA	Median Occurrences	No Occurrences in Bin	Mutability Enrichment of Motif at Subs	Mutability Enrichment of Motif at Indels	Mutability Enrichment of Motif Rearrangements
IR	1029	184	1.098	1.626	1.210
STR	330	185	1.562	5.767	1.025
DR	88	186	1.134	2.378	0.833
MR	162	184	1.083	2.494	0.983
G4	38	185	1.188	1.449	0.901
Z-DNA	33	223	1.742	10.668	0.981
H-DNA	11	206	1.677	5.965	0.887

Enrichment of mutagenesis within non-B DNA motifs

For each bin of size *B* and non B-DNA motif, we calculated the number of bps covered, *b*, as well as the number of mutations that overlapped the motif type, *m*, and the number of mutations not overlapping the motif, *n*. The fraction of mutations overlapping non B-DNA motifs is *m/b,* and the fraction of mutations not overlapping motifs is *n/(B-b)*. The enrichment of mutations overlapping non B-DNA motifs is given by $r = m(B-b)/nb$. When calculating *r* we exclude the bins where $b = 0$. When calculating ratios in Fig. 3a, Fig. S9, the expected values and the variances are adjusted to account for correlations as:

$$E[X/Y] = E[X]/E[Y] - Cov[X, Y]/E[Y]^2 + Var[Y]E[X]/E[Y]^3$$

and

$$Var[X/Y] = (E[X]/E[Y])(Var[X]/E[X]^2 - 2Cov[X,Y]/E[X]E[Y] + Var[Y]/E[Y]^2).$$

For panels in Fig. S11c, Fig. S12 this correction was not applied since it results in negative values for some of the spacer or arm lengths. For all cases, the correlation between the mutation densities of spacer and arms is positively correlated and for all but a handful, the adjusted ratios are an order of magnitude higher than the unadjusted, suggesting that the latter is an underestimate.

When discussing mirror repeats, direct repeats and inverted repeats “spacer” is used to denote the part of the motif that is not repeated, whereas “arm” is used to denote the repeating parts (Fig. 1d-f). The number of mutations overlapping spacers and arms was recorded separately.

For each direct repeat, inverted repeat and mirror repeat motif we calculated the length of the spacer and the arms. The mutation density was calculated as the number of mutations overlapping each motif part divided by the length of the spacer or arm respectively, averaged across all instances of the motif-type. Figure 3c is a summary figure across the ten tumour types of Fig. S11c and Fig. S12, measuring the average mutational density in the spacer and arm for spacer sizes 1-10nt. For Figure S11b, the mutational enrichment at spacers over arms was corrected for that expected based on the trinucleotide content frequencies of substitutions at each cancer type.

We measured the mutational density in each sub-component of each G-quadruplex motif (G-run and loop) and calculated the enrichment as the fraction of the mutational densities of the two sub-components averaged across instances. Furthermore, we separated G-quadruplexes into two groups based on the average size of the loops (less or equal than 3nt or longer than 3nt) and compared the mutational density of each group. In all cases, error bars displayed standard error measured using bootstrapping with replacement (n=10,000). For Figure S11a, the mutational enrichment at loops over G-runs was corrected for that expected based on the trinucleotide content frequencies of substitutions at each cancer type.

To investigate the relationship between mutagenesis and the distribution of non-B DNA motifs we generated a window of 2kb centered at mutations and measured the distribution of non-B DNA motifs within each position at that window. Next, we calculated the median number of each non-B DNA motif across the window and from that we defined the enrichment as the number of occurrences at a position over the median number of occurrences across the window (Figure S10a-c). For Figure 2b the average enrichment across the ten tumour types is shown for each non-B DNA motif. Micrococcal nuclease sequencing (MNase-seq) data for K562 cell line were obtained from (The ENCODE Project Consortium 2012) (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhNsome/wgEncodeSydhNsomeK562Sig.bigWig) to assess the relationship between nucleosome occupancy and G-quadruplex motifs, given the mutational patterns observed (Figure S10a-c). The signal profile and heatmap plot for nucleosome occupancy at a window of 2kb around G4s (Figure S10d) was generated using deepTools (Ramírez et al. 2014).

Analysis of recurrent mutations

The number of substitutions and indels at each genomic site was calculated per cancer type across patients using a python script, which has been uploaded as Supplemental Material. The overlap between recurrently mutated sites for each mutation type and each non-B DNA motif was subsequently calculated using “bedtools intersect” utility. A truncated Poisson model was applied as the null model.

The truncated Poisson model was estimated using the “mle” function from the “stats4” R-package. Mann-Whitney *U* test for recurrent and non-recurrent indels and substitutions overlapping each non-B DNA motif was calculated to measure significance.