Differential DNA mismatch repair underlies mutation rate variation across the human genome

Fran Supek1,2,3 & Ben Lehner1,2,4

Cancer genome sequencing has revealed considerable variation in somatic mutation rates across the human genome, with mutation rates elevated in heterochromatic late replicating regions and reduced in early replicating euchromatin1–4. Multiple mechanisms have been suggested to underlie this5–11, but the actual cause is unknown. Here we identify variable DNA mismatch repair (MMR) as the basis of this variation. Analysing ~17 million single-nucleotide variants from the genomes of 652 tumours, we show that regional autosomal mutation rates at megabase resolution are largely stable across cancer types, with differences related to changes in replication timing and gene expression. However, mutations arising after the inactivation of MMR are no longer enriched in late replicating heterochromatin relative to early replicating euchromatin. Thus, differential DNA repair and not differential mutation supply is the primary cause of the large-scale regional mutation rate variation across the human genome.

We examined 1 megabase (Mb) mutation densities along 652 fully sequenced human cancer genomes with >3,000 single-nucleotide variants (SNVs) per genome, originating from 16 tissues. This threshold enables more robust estimates of regional SNV densities in the examined samples, but it excludes cancer types with a very low mutation burden (Methods). Despite vastly different mutational loads between tissues of origin and between individual tumours11, the relative regional densities were, overall, consistent between samples. In a principal components analysis (PCA), the first principal component (PC1) corresponds closely to the average densities over all samples ($R^2 = 0.99$) and captures 86.2% of the non-random variability between the 1 Mb windows (Fig. 1a–c). This estimate of baseline variability per PC (Methods) encompasses the non-biological sources of randomness in the data (for example, low mutation counts per bin in some cancer types) but it may also include genuine biological variability, if it is particular to individual tumour genomes. The second most prominent PCA trend (PC2, 5.9% variability; Fig. 1a, b) precisely captures the known hypermutation of the X chromosome in a subset of tumours12. Across the 652 tumours, we estimate that a further 7.9% of non-random variability exists that is not explained by the general pattern of regional rates or by the hypermutation of X (in PC3–8, with 4.4% in PC3 only; Fig. 1a, b).

This signal can, in part, be ascribed to tissue-specific mutation rates: cancer samples from 7/16 tissues were significantly shifted in the distribution of their PC3 loadings (Mann–Whitney test, false discovery rate (FDR) < 1%). Liver, colorectal and B-lymphocyte tumours were differentiated by lower PC3 loading scores, whereas melanoma, breast, ovarian and lung tumours had higher PC3 loadings (tissues highlighted in Fig. 1e). To examine if these significant regional rate changes are associated to DNA replication timing changes, we used Repli-Seq data from ENCODE cell lines as a proxy for the changes occurring in the cancers of the corresponding tissues. In all cases, the correlation of the cell-line-specific Repli-Seq signal to the cancer-type-specific 1 Mb mutation rates was most prominent in the cancer type matching the cell line, with a significant difference from the non-matching cancer types (FDR < 10%, Fig. 1f). Similarly, the changes in average gene expression levels in 1 Mb windows in tumour samples paralleled the changes in mutation rates in the same samples, with significantly stronger correlations for the matching cancer type (FDR < 10%, Fig. 1g; example in Extended Data Fig. 1a, b).

Visualizing the cancer samples in a PC plot revealed a group of outliers in a low-density area of the plot with extreme PC3 and PC4 loadings (box in Fig. 1e). These samples derive almost exclusively from colorectal, stomach or uterine cancers, even though most samples from these three cancer types clustered elsewhere on the PC plot. One feature particular to tumours from these three tissues is that they frequently display inactivation of the MMR pathway through mutation of MMR genes or hypermethylation of the MLH1 gene promoter12–13. Inactivation of MMR results in a high incidence of small insertions and deletions (indels) at simple sequence repeats—referred to as microsatellite instability (MSI) and used to identify MMR deficiency phenotypically—but also in an increased SNV load (Extended Data Fig. 1c).

The outlier colon, gastric and uterine cancer samples on the PC plot (Fig. 1e and Extended Data Fig. 1d) were almost always MSI samples (phenotypically scored as MSI-high (MSI-H)), suggesting MMR deficiency as the cause of their unusual regional mutation rate variation.

To understand how the regional mutation rate differs in these samples, we plotted their mutation densities along the genome. The regional rate variation was substantially decreased in these tumours compared to other microsatellite-stable (MSS) tumours from the same tissues (Fig. 2a–c). The change towards more homogeneous mutation rates in MSI cancers was significant across many chromosomal regions, with ~72% of the genome being affected (1,977/2,748 1 Mb windows, Mann–Whitney test on pooled tissues, 10% FDR; examples in Fig. 2d and Extended Data Fig. 1e, f). Consistently, the regional rates in MSI cancers from all three tissues correlated poorly with replication timing (Fig. 2e–g), gene expression levels and heterochromatin (Extended Data Fig. 2a–f). Moreover, the slopes of the regression lines between binned replication timing and SNV rates revealed consistent changes in MSI cancer samples when calculated separately for intergenic and for genic (intronic) DNA in the whole genome sequences (Fig. 2h, i). Very similar trends were observed in a broader set of 950 exome sequences of colorectal, uterine and gastric cancers (Extended Data Fig. 2g–j).

High mutation rates in uterine and colorectal cancers can also be caused by inactivation of the proofreading domain of DNA polymerase ɛ13–14 (PolE). Proofreading is a result of a 3′–5′ exonuclease activity that enhances the accuracy of PolE by excising incorrectly placed nucleotides during synthesis. MSS PolE tumours exhibited a significantly larger spread of the regional SNV density distribution than MSI tumours (Fig. 2a, b), even though their mutational load is typically higher (Extended Data Fig. 1c). Similar conclusions are reached with stomach cancer15 hypermutators of unknown aetiology (Fig. 2c). Thus, increased mutation supply does not explain the loss of regional mutation rate variability in MSI cancers.

The relative frequencies of 5′ and 3′ contexts of different SNVs—the mutation spectrum—are informative of the mutational processes operative in a particular cancer type16. We observed the previously reported17 signatures of MMR deficiency in MSI cancers: C > T transitions in an NpCpG sequence context and C > A transversions at CpCpG (all
Figure 1 | Changes in megabase-scale regional mutation rate variation between tumour samples. a–e, PCA of the 1 Mb regional rates of 652 whole-genome sequences. a, Amount of variance conveyed by the prominent PCs. Baseline estimated by ‘broken stick’ method (Methods). b, Same as in a but expressed as percentage above baseline (putatively non-noise) variance. c, PC1 reflects average rates. d, PC2 captures the variability in chromosome X mutation rates. a.u., arbitrary units; Avg, average; Chr, chromosome. e, Tumour sample loadings on PC3/4, highlighting cancer types significantly shifted by PC3 (Mann–Whitney test, FDR < 1%).

Figure 2 | Reduced regional mutation rate variability in genomes of MSI cancer samples. a–c, Decreased variance between mutation rates of 1 Mb windows in MSI samples, when compared to MSS samples (including MSI-L) or to ultramutated PolE/MSS samples. MSI/PolE samples are in the MSI group. In the case of STAD, comparison is to PolE wild-type hypermutators (hyper). Data points in distributions are medians of relative mutation frequencies of each 1 Mb window across all cancer samples in group. **P ≤ 0.01 by F-test for decrease in variance. Number of tumour samples given in brackets. e, Standard deviation. d, Relative SNV frequencies across 1 Mb windows of chromosome 1p in colorectal cancer (CRAD). Unbroken and dotted lines are the median across tumour samples and its 95% confidence interval, respectively. For each tumour sample, relative mutation frequencies are obtained by dividing by the mean of all 1 Mb windows. *FDR ≤ 10% for rates significantly closer to unity in MSI-H samples (Mann–Whitney test). Striped bars are low- and high-significance regions (Methods). e–g, Reduced correlation of regional mutation rates to replication timing in MSI cancer samples. Genomic 1 Mb windows were pooled into five equal-frequency bins by the median Repli-Seq signal over 11 cell lines. For each bin, median and interquartile range of relative mutation rates across 1 Mb windows is shown. R² values are on original (not binned) data. *P < 0.01 for a difference of R, after Fisher Z-transform. Prior to binning, cancer samples in a group were combined by taking the median of the relative mutation frequencies in each 1 Mb window (as shown in d). h, i, Same as e–g but examined separately for genic (intronic) and intergenic regions.
Figure 3 | Association of mutational signatures to MSI and to replication timing. a, b, Relative mutation rates of example MSI-associated (a) or non-MSI-associated (b) contexts across genome-wide replication timing bins. Dotted lines are linear fits to the bins with slope $a$, a measure of association to replication timing. c, Association between MSI propensity of a mutational context (log$_2$ ratio of its frequency in MSI versus MSS tumours) to its replication timing slope in MSI tumours. Ts, transition; Tv, transversion. d, e, Association of percentage MSI-specific signatures (CCN $>$ CAN + GCN $>$ GTN + [C/T]AN $>$ [C/T]GN) in an MSI tumour sample and the binned replication timing slopes for all contexts (d), or for various non-MSI transversions (e) in the same tumour sample. Repli-Seq slopes are averaged over those mutational contexts that are displayed in the respective panel. In all panels, mutation rates were normalized to the number of nucleotides at risk in a 1 Mb window before determining the slopes. See also Extended Data Fig. 3.

Flat the overall regional rates are with respect to replication timing in each sample (Fig. 3d, $R^2 = 0.37, P = 0.0017$). Importantly, this is also true for mutations in contexts that increase little or decrease in relative frequency in MSI genomes; all transversions (excluding C $>$ A in CCN) and the transitions A $>$ G in AAN and C $>$ T in TCN (Fig. 3e and Extended Data Fig. 3e, f, $R^2 = 0.38$–0.45, $P < 0.0015$). Thus, the greater the proportion of its history a tumour has spent in an MMR-deficient state, the flatter its distribution of mutations across the genome, and this flattening is observed for mutations in both MSI-associated and non-MSI-associated contexts.

To quantify more precisely the contribution of MMR to the observed regional mutation variation across the genome, we next estimated the time spent in the MMR-proficient and MMR-deficient states in each individual tumour. We used a simple model that assumes that genome-wide mutation rates for each mutational context are equal across samples and the binned replication timing slopes for all contexts (Fig. 3d). In addition, the 99% confidence interval (CI) of the fitted line crosses zero at $<100$% mutations post-MSI, indicating ($P < 0.01$) that the mutation rate landscape in MMR-deficient cells does not show replication-timing-associated regional variability. Further testing was performed with a Mann–Whitney test, two-tailed.

Figure 4 | Inferring the time of MMR failure by a deconvolution of the mutational signatures. a, b, Examples illustrating the parameters estimated: relative mutation rates in the MSS (a) or MSI states (b), different for mutational contexts but constant across samples, and the relative time spent in MSS (c) or MSI (1 − z), which can vary across samples. TCGA-AA-3516 is a tumour sample identifier. c, Estimated rates for mutational contexts. Significance given for increase of b over a in the eight MSI contexts (Wilcoxon test). Contexts with different 3’ flanking nucleotides were pooled, except the C $>$ T in NCG contexts; other C $>$ T changes are labelled NCH. Ts, transitions; Tv, transversions. d, The estimated proportion of MSS mutations for the MSI versus a set of MSS samples. P value by Mann–Whitney test, two-tailed. e, Estimated number of mutations arising after MMR failure correlates to the loss of variability in regional mutation rates (slope of the relative rates across replication timing bins, see Figs 2 and 3) across the MSI samples. The 99% confidence interval (CI) of the fitted line crosses zero at $<100$% mutations post-MSI, indicating ($P < 0.01$) that the mutation rate landscape in MMR-deficient cells does not show replication-timing-associated regional variability. f, Relative mutation rates of chromosome 13 for the median of five samples with the largest per cent of post-MSI failure mutations (e, rightmost points) versus five MSI samples with the least per cent post-MSI failure mutations (e, leftmost points). Both groups consist of two CRAD, two UCEC and one STAD.

In addition, the 99% confidence interval of a linear fit to the points in Fig. 4e crosses the horizontal zero line—corresponding to a fully flat regional landscape—before the point at which all mutations are predicted to have arrived in the MMR-deficient state. This indicates that the mutations that arose after the inactivation of MMR in these tumours are not distributed with the characteristic regional variation across the genome. In the absence of MMR, mutation rates are not reduced in early replicating euchromatic regions compared to late replicating heterochromatin.

Through an analysis of human tumours we have shown that MMR is more effective in euchromatic early replicating regions of the genome, suppressing the accumulation of mutations in these regions. MMR is known to be coupled to DNA replication, with elevated repair efficiency during S phase. Differences in DNA accessibility to the repair machinery, the coupling of this machinery to the replication fork, or the time available for repair might contribute to the increased efficiency of repair in early replicating euchromatin. Across cell types, most active genes performing essential functions are euchromatic and replicated early. It is thus sensible to envisage that enhanced MMR in euchromatin is a beneficial trait and one that has been selected for during evolution.
1. Hodgkinson, A., Chen, Y. & Eyre-Walker, A. The large-scale distribution of somatic mutations in cancer genomes. Hum. Mutat. 33, 136–143 (2012).
2. Schuster-Boeckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 488, 504–507 (2012).
3. Woo, Y. H. & Li, W.-H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nature Commun. 3, 1004 (2012).
4. Pleasance, E. D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463, 191–196 (2010).
5. Liu, L., De, S. & Michor, F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. Nature Commun. 4, 1502 (2013).
6. Stamatoyannopoulos, J. A. et al. Human mutation rate associated with DNA replication timing. Nature Genet. 41, 393–395 (2009).
7. Waters, L. S. & Walker, G. C. The critical mutagenic translesion DNA polymerase Rev1 is highly expressed during G2/M phase rather than S phase. Proc. Natl Acad. Sci. USA 103, 8971–8976 (2006).
8. Hsu, T. C. A possible function of constitutive heterochromatin: the bodyguard hypothesis. Genetics 79 (suppl.), 137–150 (1975).
9. Sima, J. & Gilbert, D. M. Complex correlations: replication timing and mutational landscapes during cancer and genome evolution. Curr. Opin. Genet. Dev. 25, 93–100 (2014).
10. Chen, C.-L. et al. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res. 20, 447–457 (2010).
11. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214–218 (2013).
12. Jäger, N. et al. Hypermutation of the inactive X chromosome is a frequent event in cancer. Cell 155, 567–581 (2013).
13. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337 (2012).
14. The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. Nature 497, 67–73 (2013).
15. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513, 202–209 (2014).
16. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
17. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. Nature Rev. Genet. 15, 585–598 (2014).
18. Hombauer, H., Srivatsan, A., Putnam, C. D. & Kolodner, R. D. Mismatch repair, but not heteroduplex rejection, is temporally coupled to DNA replication. Science 334, 1713–1716 (2011).
19. Edelbrock, M. A., Kalysaperumal, S. & Williams, K. J. DNA mismatch repair efficiency and fidelity are elevated during DNA synthesis in human cells. Mutat. Res. 662, 59–66 (2009).
20. Amouroux, R., Campalans, A., Epe, B. & Radicella, J. P. Oxidative stress triggers the preferential assembly of base excision repair complexes on open chromatin regions. Nucleic Acids Res. 38, 2878–2890 (2010).
21. Chaudhuri, S., Wyrick, J. J. & Smardon, M. J. Histone H3 Lys79 methylation is required for efficient nucleotide excision repair in a silenced locus of Saccharomyces cerevisiae. Nucleic Acids Res. 37, 1690–1700 (2009).
22. Murga, M. et al. Global chromatin compaction limits the strength of the DNA damage response. J. Cell Biol. 178, 1101–1108 (2007).
23. Hiratani, I. et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryosogenesis. Genome Res. 20, 155–169 (2010).
24. Lubelsky, Y. et al. DNA replication and transcription programs respond to the same chromatin cues. Genome Res. 24, 1102–1114 (2014).
25. Hiratani, I. et al. Global reorganization of replication domains during embryonic stem cell differentiation. PloS Biol. 6, e245 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by grants from the Spanish Ministry of Economy and Competitiveness (BFU2011-26206 and ‘Centro de Excelencia Severo Ochoa 2013-2017’ SEV-2012-0208), a European Research Council Consolidator grant IR-DC (616434), Agència de Gestió d’Ajuts Universitaris i de Recerca (AGAUR), the EMBO Young Investigator Program, the EMBL-CRG Systems Biology Program, FP7 project 4DCellFate (277899), FP7 project MAESTRA (ICT-2013-612944) and by Marie Curie Actions.

Author Contributions F.S. performed all analyses. F.S. and B.L. designed analyses, interpreted the data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.L. (ben.lehner@crg.eu).
METHODS

TCGA genomes and calling somatic mutations. We downloaded aligned short reads (to hg19/GRCh37) for the available whole-genome sequences of tumours (n = 630) and the matched normal tissue from the The Cancer Genome Atlas (TCGA) repository at CGHub. We then called somatic SNVs in each tumour–normal pair using Illumina’s Strelka 1.0.6 workflow32, Strelka is a highly accurate caller, with a low false positive rate of SNVs at the default settings27,28. We further increased the stringency of Strelka’s post-caller filtering to prevent spurious mutation calls. By default, Strelka requires that the overall confidence score (QSS_NT) is ≥ 15, that the fraction of filtered basecalls at the site (BCNoise) is ≤ 40%, and that also the fraction of reads crossing site with spanning deletions (SpanDeD) is ≤ 75%. Here, we allow very few filtered or gapped reads at the site: BCNoise and SpanDeD must both be ≤ 3% for the tumour sample and < 10% for the normal tissue. Exceptionally, for extremely high-confidence calls (QSS_NT ≥ 45), BCNoise for the tumour sample may be ≤ 6%. For TCGA leukaemia (LAML), we downloaded the called somatic SNVs from the corresponding publication27 (n = 50 samples).

Other whole-genome sequences. We downloaded the previously called somatic SNVs for whole-genome sequences (n = 507) from the online supplementary material of Alexandrov et al.27; these samples did not overlap the TCGA data set. Next, we downloaded the somatic SNVs from the whole-genome sequences in the International Cancer Genome Consortium (ICGC) v.15.1 database, in case the same genomic information was not already available in Alexandrov et al.27 or TCGA; this encompasses the ICGC projects RECA-EU, MALY-DE and EOPC-DE (n = 150 genomes). Finally, we removed all samples of B-cell lymphoma from the Alexandrov et al.27 set due to a suspected partial overlap with a broader set in ICGC MALY-DE.

Filtering and dividing the genome into 1 Mb windows. To rule out errors due to misalignment of short reads, we masked out all regions in the genome defined in the ‘CRG Alignability 36’ track30, requiring a 36-nucleotide fragment to be unique in the genome even after allowing for two differing nucleotides. Next, we masked out the regions in the UCSC Browser blacklists (Duke and DAK), chosen for often exhibiting anomalous signals in next-generation sequencing experiments (http://genome.ucsc.edu/cgi-bin/hgTrackU?g=EncodeMapability). Finally, we discarded the exons (plus the flanking 2 nucleotides) of the UCSC gene set to avoid the signal stemming from selection on gene coding regions. Then we divided the genome into non-overlapping 1 Mb windows and discarded those with <250 kb of DNA passing the above filters; we also exclude the remaining windows on chromosome Y (n = 7) as Repli-seq data were not available for them (see later). This yielded 2,748 1 Mb windows with an average of 664 kb highly alignable, non-blacklisted, non-exonic DNA per window. When determining SNV densities, the SNV counts in each window were divided by the effective length of the window (after masking) to obtain frequencies per Mb. The choice of 1 Mb resolution is a trade-off between the inclusion of low mutation burden tumour samples/types (better with coarser resolution) and the level of detail in describing the regional variability (better with finer resolution). Furthermore, replication timing is known to be organized in megabase-scale domains23,25.

Conditions for inclusion in the final genome set. We merged the three sets of genomes (TCGA, Alexandrov et al.27 and ICGC) and discarded the genomes that had fewer than 3,000 SNVs in the alignable regions of the genome in order to obtain more reliable regional mutation density estimates. This filter will completely remove the cancer types with very low SNV loads, such as TCGA thyroid, prostate, leukaemia, or kidney chromophobe. We also removed four cancer types represented in the TCGA genomes and calling somatic mutations.

In particular, AD-A5EJ and QG-A5Z2 were putatively labelled as MSI tumours (both MSI-L) because of the overall load of somatic SNVs and introns called by Strelka (see earlier). In particular, AD-A5EJ and QG-A5Z2 were putatively labelled as MSI tumours (both MSI-L) because of the overall load of somatic SNVs and introns called by Strelka (see earlier). In particular, AD-A5EJ and QG-A5Z2 were putatively labelled as MSI tumours (both MSI-L) because of the overall load of somatic SNVs and introns called by Strelka (see earlier).

Next, we assigned MSI-H, MSI-L or MSS status to the samples used therein, of which 10 were CRAD, 10 were UCEC and 4 were STAD. Gene expression sequences. The called somatic mutations from exome sequencing (MAF files) were downloaded from the TCGA for the COAD/READ (CRAD), UCEC and STAD cancer types in October 2014. For each tumour sample (TCGA patient), we selected the newest available MAF that had mutation data for that sample, and we did not load further mutation data for that sample from other MAFs (files listed in Supplementary Table 1). Next, we assigned MSI-H, MSI-L or MSS status to the exomes from the same data sources used for the whole genomes (see earlier) and discarded samples for which the MSI status was not known. This left 950 samples, of which 195 were MSI-H, and the rest were MSS or MSI-L (pooled together). PolE status was not inferred for exomes. The genomic mask for the exome analysis was constructed differentially than for whole genomes: we similarly used the ‘CRG Alignability 36’ filter and the two UCSC blacklists. However, we excluded all DNA except the protein coding exons (but not exons of commonly mutated cancer genes, which were also excluded). Finally, we retained only the 1 Mb windows with at least 5 kb of alignable, non-backlisted exonic DNA. This reduced the initial set of 2,748 windows (for the whole-genome analysis) to 1,709 windows for the exome analysis. Density of mutations was expressed per Mb of available DNA in each window, and again normalized by dividing by the average of all windows in a sample. For the analysis where we considered each exome separately (Extended Data Fig. 2), we limited the analysis to exomes with ≥50 SNVs in the selected genomic windows.

Gene expression data. The expression levels in tumours were downloaded from the TCGA RnaSeqV2 data sets34, where they are expressed as transcript per million (TPM) values for each gene. For 15 TCGA cancer types that had RnaSeqV2 data available, we downloaded TPM levels for those tumour samples for which we had called whole-genome somatic SNVs, in total 8–92 RnaSeqV2 samples per cancer type (average = 29). The TPM levels of genes overlapping (including partially) each 1 Mb window were averaged for each sample, and we then averaged over all samples of each cancer type to get the final expression level for that 1 Mb window in that cancer type; if lower than 0.01 TPM, it was adjusted to 0.01 TPM. The overall, cross-tissue 1 Mb expression levels (in Extended Data Fig. 2a–c) are then the medians across 15 cancer types.

The tissue specificity (TS) of gene expression of 1 Mb windows for a particular cancer type (in Fig. 1g) is the log2 ratio of TPM in that cancer type and the average TPM across all cancer types. In comparing the TS of gene expression to the TS of mutation rates, we limited the analyses to the 2,442/2,748 1 Mb windows that were at least somewhat expressed (>0.01 TPM) in at least one examined cancer type. Moreover, we excluded chromosome Y for consistent treatment with the replication timing TS analysis (see later). Finally, we considered only the eight cancer types (seven tissues) with significant shifts in the PC3 of regional mutation rates.

Replication timing data. We downloaded the Repli-seq measurements29 (as wavelet-based replication timing (RT) signal of ENCODE cell lines from the UCSC Genome Browser (also available in the Gene Expression Omnibus under accession number GSE43399). To avoid biasing the sample, we excluded multiple lymphoblastoid cell lines and retained Gm12878 as a representative. We computed the average Repli-seq signal within 1 Mb genome windows of the remaining 11 cell lines, except chromosome Y, which was unavailable in the original data. The resulting values ranged from 0–100, where higher values indicate earlier replication. The overall, cross-tissue replication timing signal (used for genome binning in Figs 2 and 3 and Extended Data Figs 2 and 3) is the median value across the 11 cell lines.

The TS of replication timing in a cell line (in Fig. 1f) is the difference between the Repli-seq signal of that cell line and the average signal across all cell lines. For the TS analysis, we considered those cell lines that (1) could be matched to a cancer type, based on their tissue of origin, and that (2) corresponded to one of the eight cancer types found to have a significant shift in PC3 of the regional mutation rates (Fig. 1e). In particular, replication timing TS in MCF7 cells served as a proxy for BRCA
(breast cancer), B cells for SKCM (melanoma), IMR-90 cells for LUSC/LUAD (lung cancers), HEK G2 cells for LIHC (liver cancer) and Gm12878 cells for DLBC (lymphomas). For consistency with the gene expression TS analyses, we considered the same set of 1 Mb windows (see earlier) in the replication timing TS analysis.

**Heterochromatin data.** We downloaded the RoadMap epigenomics H3K9me3 ChIP-seq signal for a diverse set of healthy tissues or cell lines from the Gene Expression Omnibus, encompassing GSM261216 (adult kidney), GSM537710 (adult liver), GSM670028 (mesenchymal stem cells), GSM772917 (CD4 naive), GSM669939 (fetal lung, chromosome 20), GSM679903 (fetal lung, chromosome 3), GSE53701 (IMR-90 cell line). Overall, we include the H3K9me3 levels from Barski et al. that were previously shown to have a strong correlation to regional mutation rates in cancer. We calculated the mean H3K9me3 signal in 1 Mb windows for each sample and trimmed the distribution at the 99.9th percentile. Then we log-transformed the values and found the median over eight tissues/cell lines to get the overall H3K9me3 levels used for binning (in Extended Data Fig. 2d–f).

**Slopes over genomic bins.** We used the overall, cross-tissue Repli-Seq signal to create five equal-frequency (same number of 1 Mb windows) genome bins for further analyses. Additional sets of bins were created also for the gene expression and for the heterochromatin signal (Extended Data Fig. 2a–f). The slope of a regression line fitted through the average 1 Mb relative mutation rates of each bin is a measure of the association of mutation rates to replication timing (Figs 2–4). This measure has the desirable property of being robust to differences in the overall mutation load between smaller groups of tumour samples or individual genomes, or between mutational signatures.

For the analysis in which intergenic and genic regions were examined separately in MSI and MSS cancers, we used the UCSC gene set to define these regions. Since the gene exons had already been excluded in a genome pre-processing step, the genic regions effectively consist only of introns. Only for the purposes of this intergenic versus genic analysis, we relaxed the requirement of ≥250 kb (of total alignable DNA) per 1 Mb window to >100 kb (either intergenic or genic) alignable DNA per 1 Mb. When calculating relative mutation rates, regardless whether we were analysing intergenic or intronic mutations, each sample was always normalized by dividing by the mean of the aggregate (intergenic plus intronic DNA) SNVs rates across all its windows. The five Repli-Seq bins were the same as in the whole-genome analysis.

**Statistical analysis.** PCA was performed in R 3.1.1 (R Core Team) and in XLSTAT 2014.2 (Addinsoft) on the relative 1 Mb mutation rates of each tumour sample, where tumour samples were features (columns in data table), and 1 Mb windows were examples (rows). To find the relative rates, first the SNV densities per Mb of alignable, non-exonic, non-blacklisted DNA (see earlier) were determined for each 1 Mb window in every tumour sample. Then, these densities in each sample were normalized by dividing by the mean SNV density of all windows in that sample. Thus, the relative mutation frequencies >1 correspond to above-average SNV densities in a tumour sample, and <1 to below-average densities. The approximate 95% confidence intervals of the median (across tumour samples: Fig. 2d and Extended Data Fig. 1e,f) for the 1 Mb windows were estimated using the formula $\pm 1.58*\text{IQR}/\text{sqrt(n}_{\text{samp}})$, as defined in the R function boxplot.stats and references therein. In the PCA, given that the tumour samples were features (columns) and the 1 Mb windows were examples (rows), the PC’s will be linear combinations of tumour samples, and the loadings of the samples on PCs 3 and 4 are shown in Fig. 1e. The expected (baseline) per cent variance in each PC stemming from noise in data was estimated using the ‘broken stick’ method, found to outperform related approaches.

The cancer types were tested for shifts in PC3 loadings using a Mann–Whitney test (two-tailed), where the loadings of samples in one cancer type were contrasted to the loadings of samples in all other types.

**Mutational signatures.** As in previous work, the mutational signatures are defined as relative frequencies of SNVs at different nucleotides in all possible 9 possible changes exist: $\text{A, C, G, T}$ correspond to above-average SNV densities. The out-
36. Hansen, R. S. et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA* **107**, 139–144 (2010).

37. Thurman, R. E., Day, N., Noble, W. S. & Stamatoyannopoulos, J. A. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.* **17**, 917–927 (2007).

38. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).

39. Jackson, D. A. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* **74**, 2204–2214 (1993).

40. Mebane, W. R. & Sekhon, J. S. Genetic optimization using derivatives: the rgenoud package for R. *J. Stat. Softw.* **42**, 473–487 (2010).
Extended Data Figure 1 | Overall mutational burden and megabase-scale regional rate variability in tumour samples of MSI-prone cancer types. 

a, b, Correlations of tissue specificity (TS; see Methods) in regional mutation rates of diffuse large B-cell lymphoma (DLBC) with TS of gene expression in DLBC (a), or with TS of replication timing in the Gm12878 lymphoblastoid cell line (b). 

c, Overall mutational load, as SNVs per Mb of alignable genomic DNA (Methods) for MSI-H, MSS (includes MSI-L), PolE mutant tumours, or otherwise hypermutated tumour samples. 

d, PC plot with PCs 3 and 4, as in Fig. 1e, but showing only tumour samples for colorectal (CRAD), uterine (UCEC) and stomach (STAD) cancers for visual emphasis. 

e, f, Relative SNV frequencies across 1 Mb windows of chromosome 1p in UCEC and STAD. Unbroken and dotted lines are the median across tumour samples and its 95% confidence interval, respectively. For each tumour sample, relative mutation frequencies are always obtained by dividing by the mean of all 1 Mb windows. 

MSS/PolE samples are in the MSI-H group; hyper/ultramutators are not in the MSS group. *FDR ≤ 10% for rates significantly closer to unity in MSI-H samples (Mann–Whitney test; not applicable to STAD because of too few MSI-H samples).
Extended Data Figure 2 | Reduced correlation of regional mutation rates to gene expression, heterochromatin and replication timing in genomes and exomes of MSI tumours. a–c, The 1 Mb windows in the genome were pooled into five equal-frequency bins by the average gene expression levels (log2 transcripts per million (TPM)) in each window. The median and interquartile range of relative mutation rates across 1 Mb windows is shown for each bin. $R^2$ values were always determined on original (not binned) data. *P, 0.01 for difference of $R$ after Fisher $Z$-transform. Gene expression levels are medians over TPM across 15 cancer types. Relative SNV frequencies of each tumour sample were obtained by normalizing by the average SNV density of all genomic 1 Mb windows of that sample. Prior to binning the windows, cancer samples in a group were combined by taking the median of the relative mutation frequencies for each 1 Mb window, as illustrated for CRAD in Fig. 2d. PolE/MSI samples are in the MSI group; ultramutators are not in the MSS group. MSI-L samples are pooled with MSS. d–f, Same as in a–c but for five heterochromatin bins (median H3K9me3 signal over eight tissues and cell lines). g–i, Regional mutation rates in exome sequences of a broader set of 195 MSI-H tumour samples. The 1,709 genomic 1 Mb windows with at least 5 kb alignable protein-coding DNA each were grouped into five equal-frequency bins by the median Repli-Seq signal over 11 cell lines (Methods). Mutations were pooled across all samples in one cancer type with a known MSI-H or MSS status (Methods). $a$ is the slope of the regression line fit to binned data. j, Slopes $a$ determined for individual cancer exomes with a sufficient number of mutations ($\geq 50$ SNVs). Number of samples n shown below each group. For all cancer types, MSI-H samples have significantly less negative slopes than MSS ($P < 0.01$, Mann–Whitney test, one tailed). MSI-H also includes the MSI-H/PolE mutant samples, and MSS includes the MSI-L samples. In the exome analyses, ultramutators were not considered separately.
Extended Data Figure 3 | Association of mutational signatures to microsatellite instability and to replication timing. a, Relative frequencies of the 96 mutation contexts (strand symmetric) in MSI versus MSS cancers; the MSS group includes MSI-L samples but not MSS/PolE ultramutators. Mutations were pooled across samples of MSI-prone tissues (CRAD, UCEC and STAD). b, c, Similar to Fig. 3a, b, showing two additional examples of mutational contexts with different MSI propensities and their relative mutation rates across five genomic replication timing bins. d, Lack of correlation between the MSI propensity of a mutational context with its replication timing slope in MSS tumour samples (compare to Fig. 3c, which shows slopes in MSI samples). Ts, transition; Tv, transversion. e, f, Association of per cent MSI-specific signatures ([CCN > CAN] + [GCN > GTN] + [CTN > CTN]) across cancer samples and the binned replication timing slopes for two non-MSI transition signatures in the same samples. Slopes averaged over contexts are displayed in each plot. In all panels except a, mutation rates were normalized to number of nucleotides at risk in a 1 Mb window before determining the replication timing slopes.
Extended Data Figure 4 | The deconvolution of MSI mutational spectra robustly converges onto two equivalent solutions. a, Agreement of the observed relative frequencies of mutational contexts in each tumour sample with the predictions of model 1 (having median \(a\), \(b\) and \(z\) coefficients across all solutions in cluster 1). b, Sets of best-fit solutions determined in a hundred optimization runs initialized with different starting conditions. The solutions cluster into two homogeneous clusters (Pearson \(R > 0.9\) between >90% of the solutions within a cluster, in UPGMA hierarchical clustering).

c, d, Solutions within both clusters have similar fit to observed data (c) and make extremely similar predictions for mutation spectra in tumour samples (d). e–h, Similar to Fig. 4a, b. Example mutation accumulation diagrams for two mutation contexts typical of MSI tumours, shown for an example MSI tumour TCGA-BR-4280 (e, g) and for an MSS tumour TCGA-CD-8529 (f, h).

i, j, Values of the parameters in two solution clusters, with medians and interquartile ranges (shown as whiskers). Each solution encompasses 104 parameters: relative mutation rates \(a\) and \(b\) for each of 28 mutational contexts (i), and the relative pre-MMR failure time \(z\) for each tumour sample of the 24 MSI and 24 MSS samples (j).