Supplementary Information for

Mutational heterogeneity in cancer and the search for new cancer genes

Lawrence et al. (2013) *Nature*

Contents

Figure S1: Simplified illustration of the problem of mutational heterogeneity

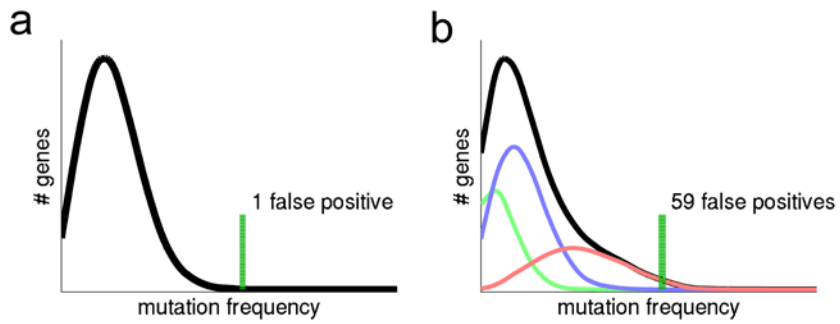

Figure S1. Effect of heterogeneity in background mutation frequency on detection of significantly mutated genes. Here, a simplified exome consists of 20,000 genes, each with the same coding length of 1500 nucleotides, and a dataset comprises 200 patients. The average background mutation frequency[1] of a gene is 10/Mb, and all mutations are assumed to be due to the background mutation processes; there are no true driver genes. Two variants of this scenario are compared. **(a)** Uniform model. All genes have a background mutation frequency (BMF) of 10/Mb. The plot shows a histogram of genes by their *observed* mutation frequency, which follows a distribution based on stochastic binomial sampling. The green line indicates a significance threshold that allows a single false positive gene (that is, P<1/20,000), which corresponds to ~40/Mb. **(b)** Heterogeneous model. One quarter of the genes (light green) have a BMF equal to 4/Mb; another 50% of the genes (light blue) have a BMF of 8/Mb; and the final one quarter of the genes (light red) have a BMF equal to 20/Mb. Applying the same threshold for significance (green light corresponding to ~40 mutations/Mb), 59 genes will be called significantly mutated. The heterogeneity can also decrease sensitivity. In the uniform case, a driver gene with a mutation rate 4-fold higher than the background rate will be detected. In the heterogeneous case, a true cancer gene among the low-BMF genes (green distribution) will not be detected unless it has a mutation rate that is 10-fold higher than its background rate.

Figure S2: Effect of biological factors on mutation frequencies

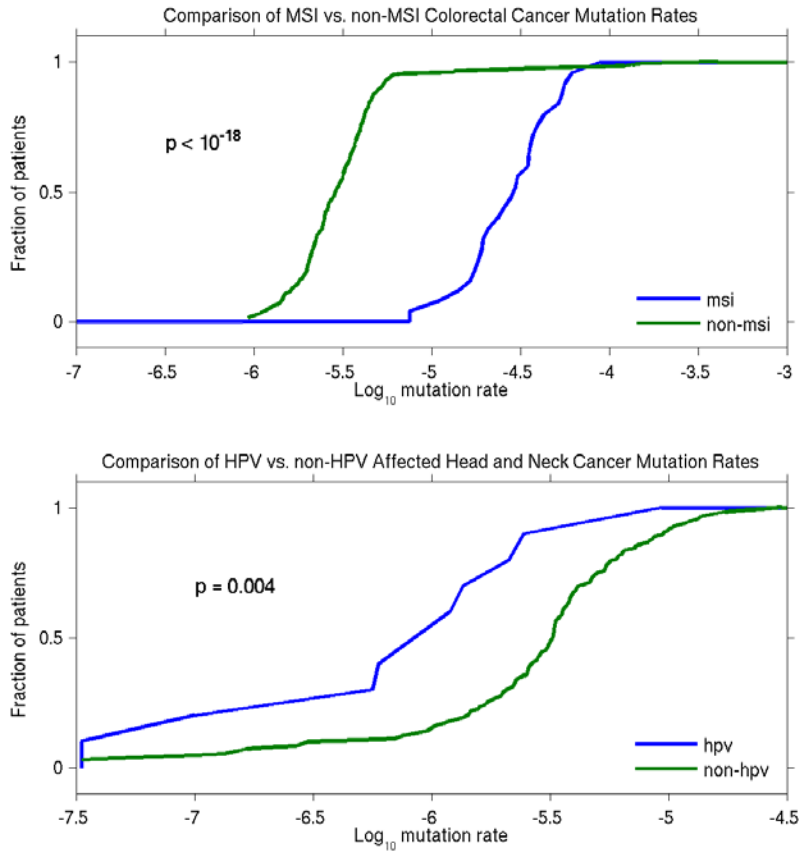

Figure S2. In some cases, heterogeneity in mutation rates can be ascribed to key biological factors. **(a)** Effect of microsatellite instability (MSI) status on mutation frequency in colorectal tumors[2]. (**b**) Effect of human papillomavirus (HPV) status on mutation frequency in head-and-neck tumors[3].

Figure S3: Concordance of genomic measures across tissue types.

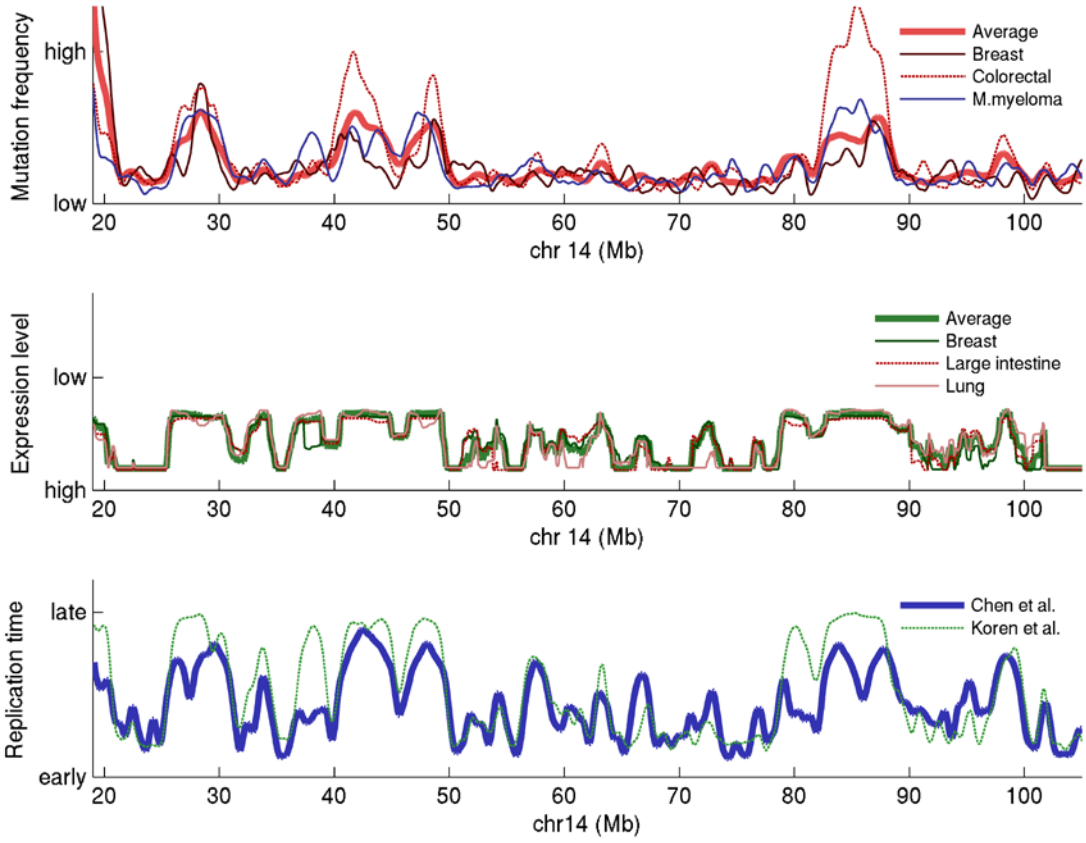

Figure S3. Concordance of genomic measures across tissue types, illustrated on chromosome 14 as in **Figure 3**. **(a)** Correlation of mutation rates in non-coding regions, compared across breast cancer[4], colorectal cancer[5], and multiple myeloma[6], compared to the average of 126 WGS samples shown in **Figure 3**. **(b)** Correlation of expression levels across breast cancer, large intestine cancer, and lung squamous carcinoma (data from Cancer Cell Line Encyclopedia[7]), compared to the average of 91 CCLE cell lines shown in **Figure 3**. **(c)** Correlation of DNA replication timing data in HeLa cells[8] (shown in **Figure 3**), and human blood cell lines[9].

Figure S4: Six mutational spectra detected by NMF

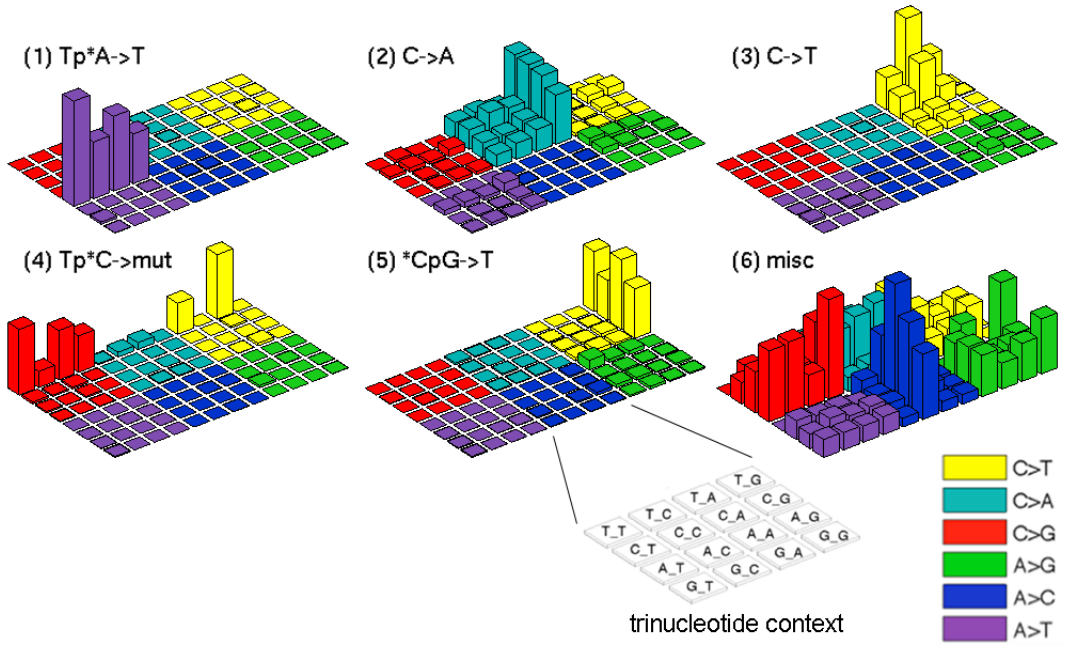

Figure S4. "Lego" plots for each of the six mutational patterns determined by NMF. Each plot organizes the 96 possible mutation types into six large blocks, color-coded to reflect the base substitution type. Each large block is further subdivided into the 16 possible pairs of 5' and 3' neighbors, as listed in the "trinucleotide context" legend. The height of each block corresponds to the mutation frequency for that kind of mutation. The patterns are named according to their dominant type of mutation.

Figure S5: Relationship between tumor types and the six mutational processes.

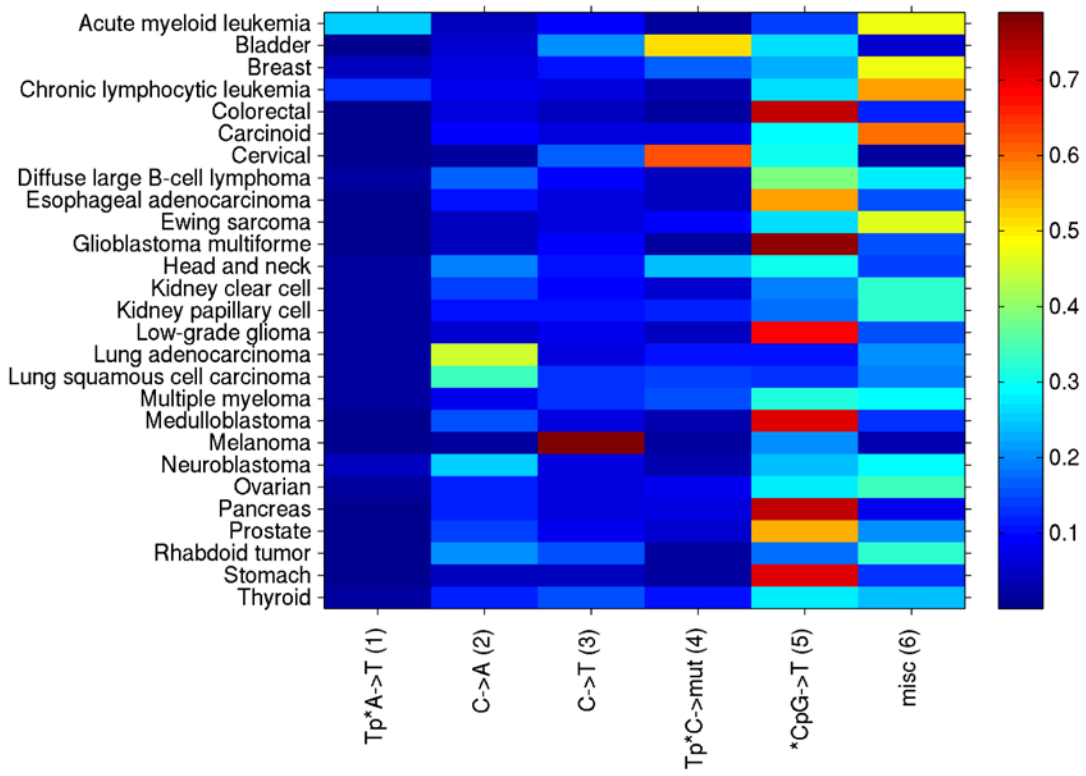

Figure S5. Heatmap showing the contribution to each tumor type from each of the six mutational processes (factors) discovered by NMF. For example, melanoma has a high weight corresponding to C→T mutations (factor 3). Cervical and bladder cancer (as well as head-and-neck cancer to a lesser degree) are dominated by the Tp*C mutations that represent the APOBEC signature (factor 4). Gastrointestinal tumor types such as colorectal, esophageal, and stomach—as well as central nervous system cancers such as glioblastomas, low-grade gliomas, and medulloblastomas—and pancreatic cancer are distinguished by their elevated frequency of transitions at CpG dinucleotides (factor 5).

Figure S6: Samples ordered according to clustering of weights of mutational spectra

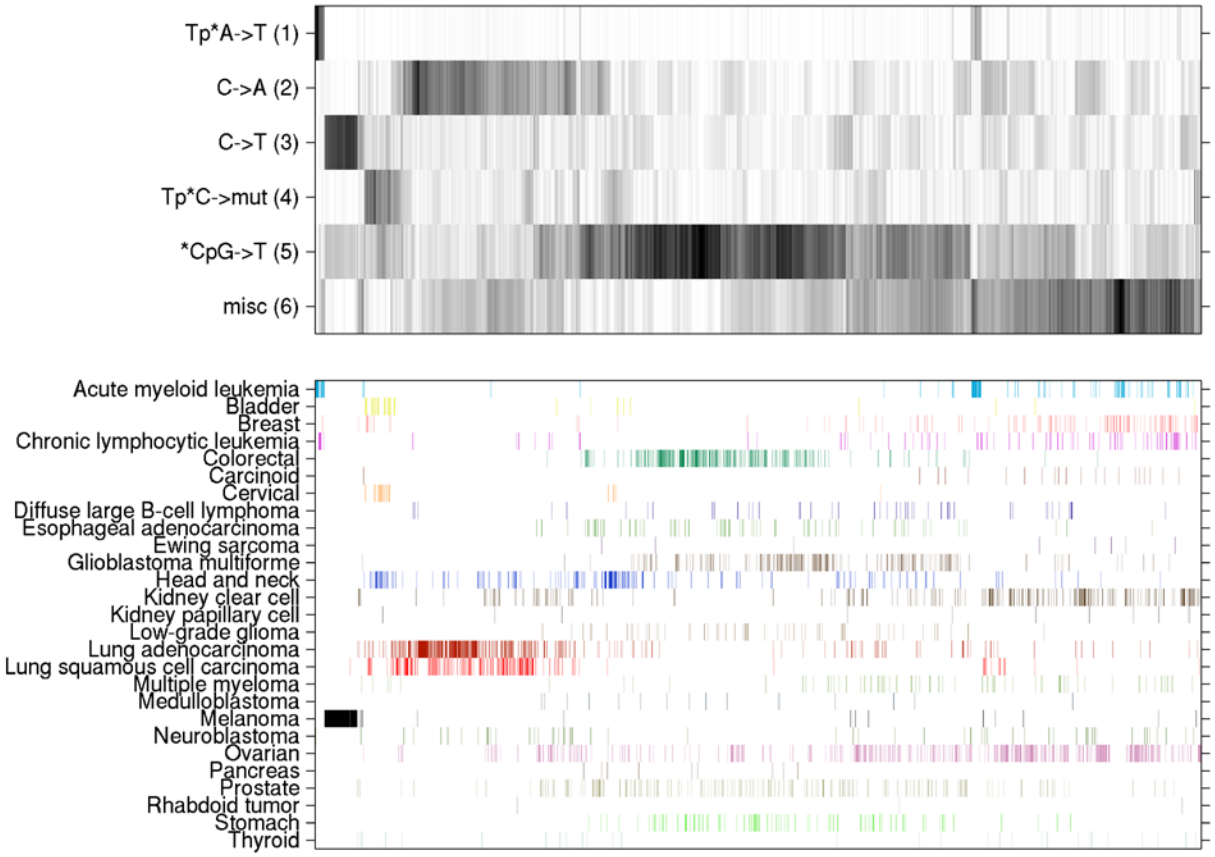

Figure S6. Samples were clustered based on their weights of the six mutational spectra found by NMF (W matrix of **Method S1**), then displayed horizontally in the order of their clustering dendrogram. The top matrix shows the six factors and their contribution to each sample. The lower matrix shows tumor-type membership for each sample. For example, the tight black cluster at left (melanoma) corresponds to dominant activation of the C→T signature (factor 3). Lung tumors (red) align with the C→A signature (factor 2). A subset of leukemias at the extreme left of the figure (AML, cyan; and CLL, magenta) are distinguished by a high frequency of Tp*A→T mutations (factor 1).

Figure S7: Radial plot based on clustering of spectra weights

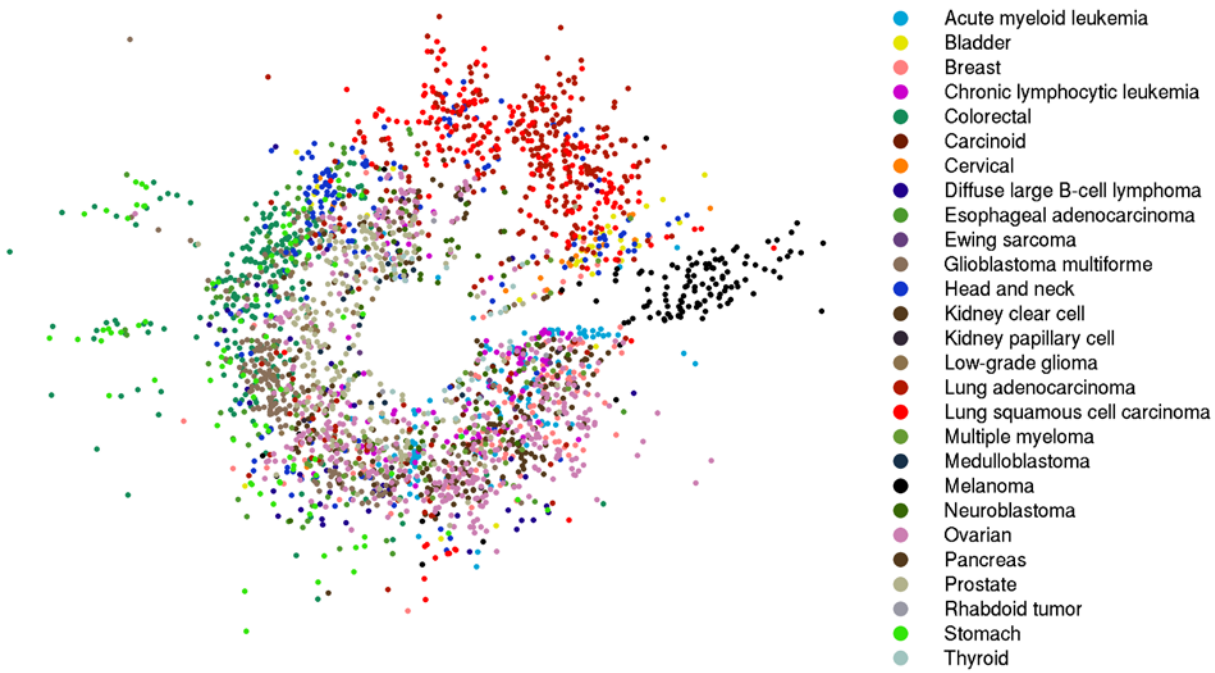

Figure S7. Alternate method of distributing the samples around the circle. Instead of assigning angular positions based on rank order of the NMF factors (as in Figure 2), this figure assigns angular positions based on clustering of factor weights. The distance between any pair of samples was defined based on the correlation of their weights (W matrix of **Method S1**). The samples were then clustered using the average-linkage clustering method[10]. The order of samples was then based on the dendrogram from that clustering. As before, tumor samples with more than 10 mutations are shown. Again, the samples tended to cluster according to tumor types or subtypes with similar active processes, e.g. lung cancer at 12–1 o'clock, melanoma at ~3 o'clock, and samples with potential APOBEC activity[11] between them at ~2 o'clock. Clustering results were similar when clustering was performed using the entire set of 96 categories (not shown); however reducing the dimensions and decoupling the contributions of different factors enabled illumination of clearer associations with distinct mutational processes (**Supplementary Figures S5,6**).

Figure S8: Mutational spectra of 27 tumor types.

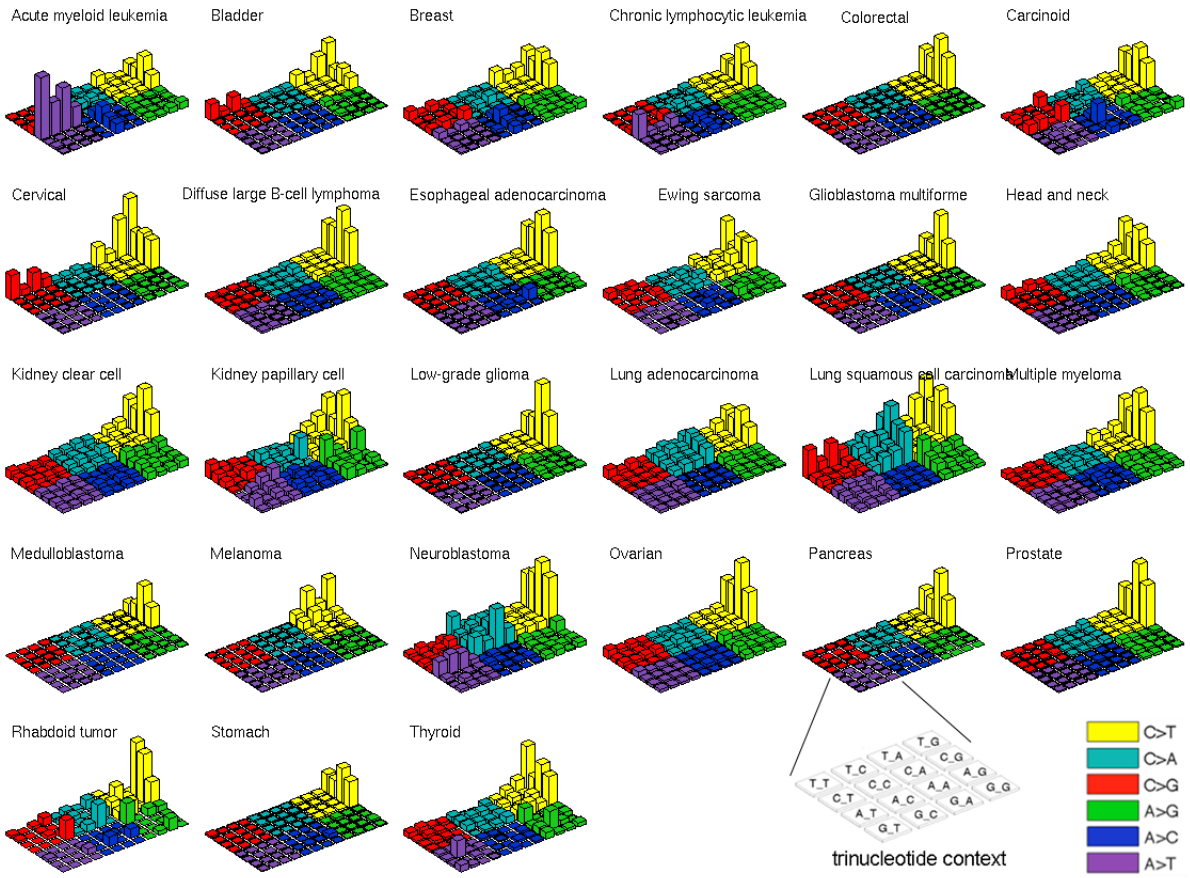

Figure S8. "Lego" plots for each of the 27 tumor types investigated in the study. Representation is same as that used in Figure S4. Of particular interest are the unique spectra of AML (tall purple bars corresponding to factor 1), bladder and cervical (elevated back row corresponding to factor 4), and lung cancer and neuroblastoma (raised cyan block, corresponding to factor 2).

Figure S9: Scatter plots of somatic mutation frequency against gene characteristics.

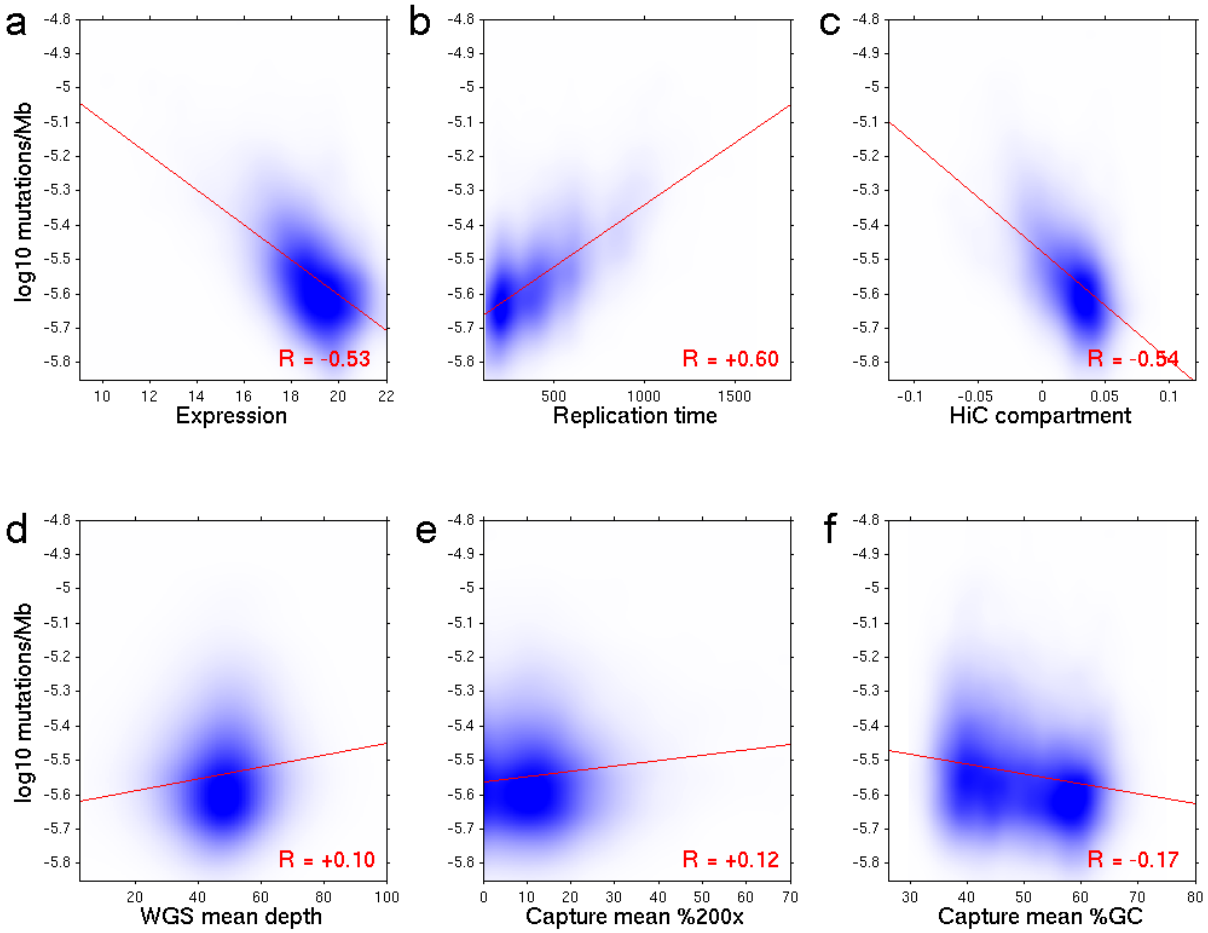

Figure S9. Six gene characteristics **(Table S5)** were evaluated as potential predictors of somatic noncoding mutation frequency. In each case, the y-axis shows log10 of the somatic noncoding mutation frequency (mutations/Mb) measured from the panel of 126 cancer samples that were subjected to whole-genome sequencing. The x-axis shows **(a)** average expression level across 91 cell lines in the CCLE[7]; (b) DNA replication time[8], expressed on a scale of 100 (early) to 1500 (late); (c) a HiC[12]-derived metric indicating which chromosomal compartment the gene is in (negative values = closed compartment "B", positive values = open compartment "A"); (d) the mean sequencing depth (fold coverage) of the gene, measured in the panel of 126 WGS samples; (e) the mean percentage of bases that were covered to at least 200 fold, measured in the panel of ~3000 exome capture samples; and (f) the mean percent GC content of the reads that covered the gene, also measured from the exome samples. Linear regression fits and coefficients ("R") are indicated on the plots: see also **Table S6**. Overall, the genomic characteristics **(a–c)** were much more effective predictors than the technical metrics **(d–f)**.

Figure S10: Kataegis events detected in WGS data.

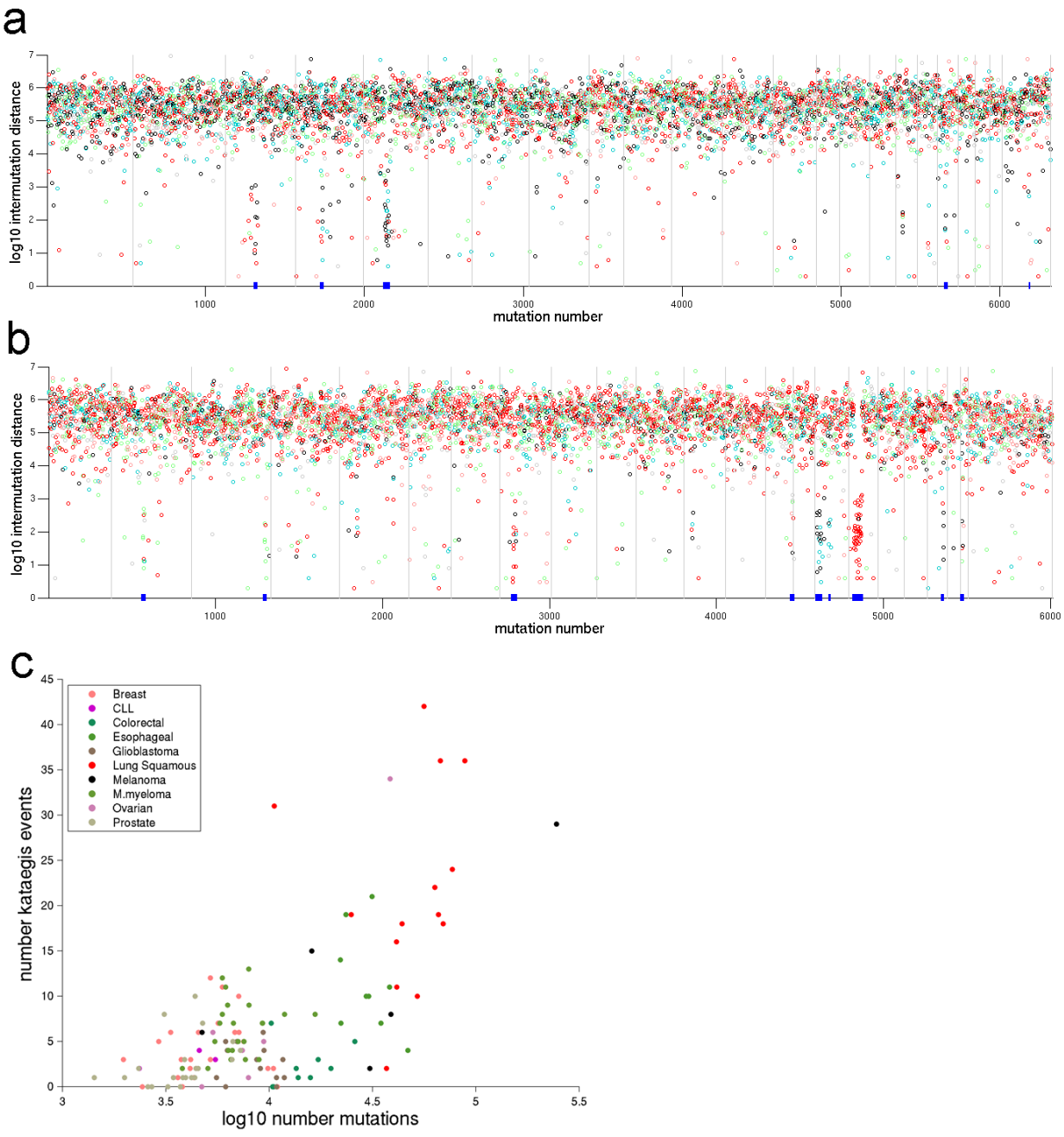

Figure S10. Kataegis[13] (or clustered mutations[11]) events detected in samples that were subjected to whole-genome sequencing. "Rainfall plots" were constructed as described[13] for two example samples: **(a)** breast tumor BR-0004 and **(b)** multiple myeloma sample MM-0344. In each of these plots, the log10 inter-mutation distances between mutations (ordered by chromosomal position) are plotted on the y-axis. X-axis indicates mutation number. Chromosome boundaries are shown as grey vertical divisions. The color of each point tells what type of basepair change it is: C→A (cyan) , C→G (black), C→T (red), A→C (pink), A→G (light green), or A→T (grey). Kataegis events, detected by identifying stretches of at least six mutations having inter-mutation distances at least two standard deviations smaller than the sample median, are indicated with blue boxes at the bottom of the plots. In some cases it can be observed that the individual mutations of a kataegis event are all of the same category. **(c)** Summary of number of kataegis events observed per WGS patient genome. In general the more highly mutated samples are seen to have larger numbers of kataegis events.

Figure S11: Transcription-coupled repair across pan-cancer dataset.

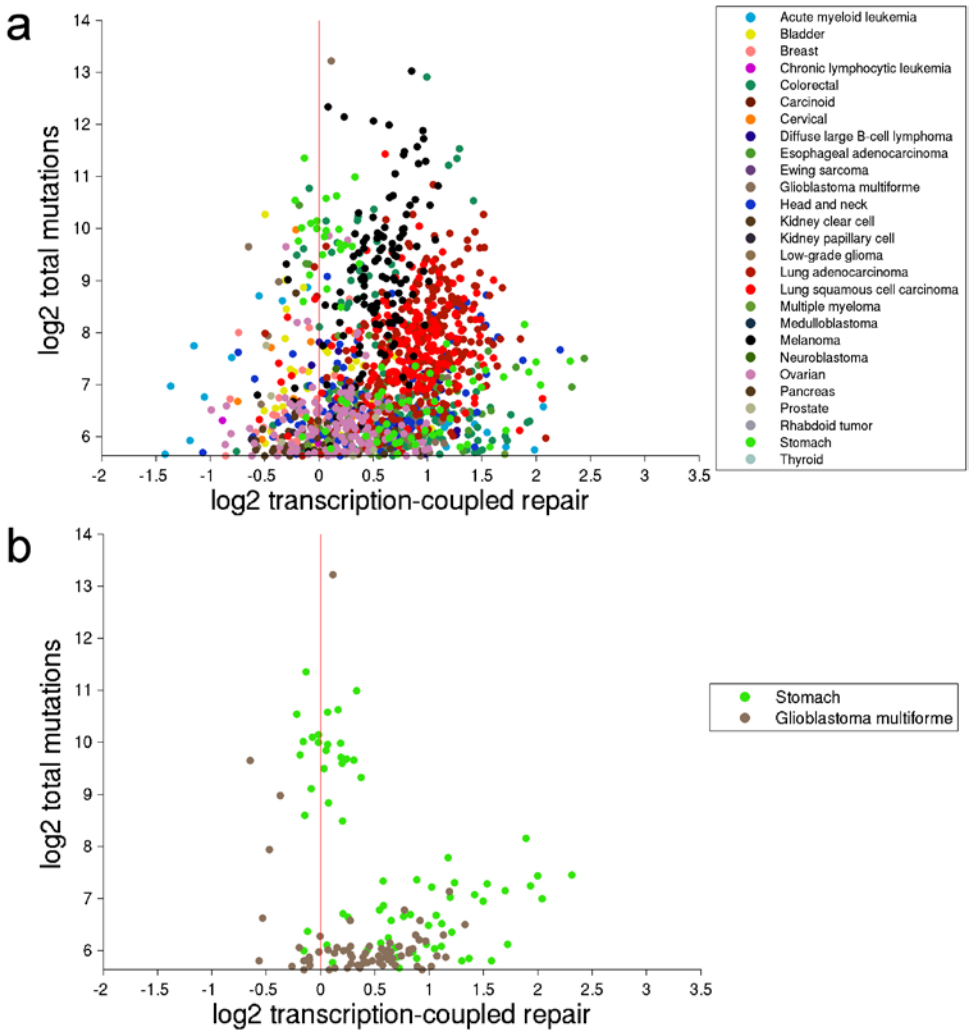

Figure S11. Survey of transcription-coupled repair (TCR) across the ~3000 exomes of the pan-cancer dataset. In order to quantitate TCR in exomes (where only transcribed regions are interrogated), the ~18,000 genes were partitioned into expression tertiles (highest, middle, lowest) according to their average general expression level across 91 cell lines in the Cancer Cell Line Encyclopedia (CCLE)[7]. Then, the size of the TCR effect was defined as the log2 ratio of mutation rate in the lowest-expression tertile divided by mutation rate in the highest-expression tertile. The scatterplot x-axis shows this TCR metric, and the y-axis shows the log2 total number of mutations supporting each measurement, with a minimum of 50 mutations required for inclusion in the plot. Each point is a sample, color-coded as in **Figure 2**. Panels show **(a)** all samples, and **(b)** stomach and GBM tumors, highlighting a subclass of samples that are hypermutated and TCR-inactive (or TCR-overwhelmed[14]).

Methods S0: Sample collection and DNA sequencing analysis

Sample Collection

All samples were obtained under institutional IRB approval and with documented informed consent. A total of 3,083 tumors were analyzed, with 2,842 (92%) having been sequenced at the Broad Institute and the remaining 241 having been published as part of The Cancer Genome Atlas (TCGA) projects[2,15,16]. A total of 1647 were from The Cancer Genome Atlas (TCGA) projects (394 ovarian cancer[16], 225 kidney cancer, 219 glioblastoma[15], 222 colorectal cancer[2], 134 AML, 88 stomach cancer, 94 head and neck, 81 prostate, 57 lower grade glioma, 52 thyroid cancer, 35 bladder cancer, 20 cervical cancer, 13 pancreas, 13 lung cancer). An additional 956 tumors were from NHGRI Center Initiated Projects (501 lung cancer[17], 141 prostate[18], 121 melanoma[19], 91 CLL[20], 76 esophageal[21], 26 medulloblastoma[22]). An additional 299 samples (121 breast[4], 87 head and neck[3], 49 diffuse large B-cell lymphoma[23], 22 rhabdoid tumors[24], 20 Ewing sarcoma) were from the Slim Initiative for Genomic Medicine collaboration, affiliated with the International Cancer Genome Consortium (ICGC). An additional 81 neuroblastoma[25] samples were from the NCI TARGET project (http://target.cancer.gov). Additional samples were analyzed in collaboration with research foundations: 63 multiple myeloma[6] with the Multiple Myeloma Research Foundation and 23 carcinoid samples with the Care for Carcinoid Foundation. Several whole-genome datasets were from published papers (9 colorectal[5]; 5 prostate[26]; 5 melanoma[27]). A complete list of samples is given in **Supplementary Table S2**.

Whole exome sequencing

Whole-exome capture libraries were constructed from 100ng of tumor and normal DNA following shearing, end repair, phosphorylation and ligation to barcoded sequencing adapters[28] [29]. Ligated DNA was size-selected for lengths between 200-350bp and subjected to exonic hybrid capture using SureSelect v2 Exome bait (Agilent). Samples were multiplexed and sequenced on multiple Illumina HiSeq flowcells to average target exome coverage of 118x.

Whole genome sequencing

Whole-genome sequencing library construction was done with 1-3 micrograms of native DNA from primary tumor and germline samples for each patient. The DNA was sheared to a range of 101-700 bp using the Covaris E210 Instrument, and then phosphorylated and adenylated according to the Illumina protocol. Adapter ligated purification was done by preparatory gel electrophoresis (4% agarose, 85 volts, 3 hours), and size was selected by excision of two bands (500-520 bp and 520-540 bp respectively) yielding two libraries per sample with average of 380 bp and 400 bp respectively[6,5,26]. Qiagen Min-Elute column based clean ups were performed after

each step. For a subset of samples, gel electrophoresis and extraction was performed using the automated Pippin Prep system (Sage Science, Beverly MA). Libraries were then sequenced with the Illumina GA-II or Illumina HiSeq sequencer with 76 or 101 bp reads, achieving an average of ~30X coverage depth.

Sequence data processing and quality control

Exome and whole-genome sequence data processing and analysis were performed as follows. Illumina reads were aligned to the reference human genome build hg19 using an implementation of the Burrows-Wheeler Aligner[30], and a BAM file was produced for each tumor and normal sample by the Picard pipeline[6]. The Firehose pipeline (www.broadinstitute.org/cancer/cga) was used to manage input and output files and submit analyses for execution in GenePattern[31].

Quality control modules in Firehose were used to compare genotypes derived from Affymetrix arrays and sequencing data to ensure concordance. Genotypes from SNP arrays were also used to monitor for low levels of cross-contamination between samples from different individuals in sequencing data using the ContEst algorithm[32].

Mutation calling

The MuTect algorithm[33] was used to identify SSNVs in targeted exons and whole-genome data. MuTect identifies candidate SSNVs by Bayesian statistical analysis of bases and their qualities in the tumor and normal BAMs at a given genomic locus. We required a minimum of 14 reads covering a site in the tumor and 8 in the normal for declaring a site is adequately covered for mutation calling. We used a minimal allelic fraction cutoff of 0.1. The MuTect publication[33] describes the specificity and sensitivity of the MuTect calling algorithm, performance parameters that are of crucial importance to downstream analyses such as the ones reported here. Small somatic insertions and deletions were detected using the Indelocator algorithm (Sivachenko, A. et al., manuscript in preparation) after local realignment of tumor and normal sequences[34]. All point mutations and short indels were subjected to filtering against a large panel of normal samples, in order to remove common alignment artifacts that escaped the original calling algorithms.

Methods S1: Dimensionality reduction and decomposition to mutational processes

We calculated the matrix $R_{96\times3083}$, which represents the relative mutation frequency of each of the 96 mutation types in each of the each of the 3,083 samples (**Supplementary Table S3)**. (Specifically, for each sample s and each mutation type c, we determined n_{cs}, the total number of mutations observed and N_{cs}, the total number of bases with adequate coverage to observe such mutations ($\geq$ 14 reads in tumor and $\geq$8 in normal). We defined a sample specific overall mutation frequency as $\mu_s = \Sigma_c\, n_{cs} / \Sigma_c\, N_{cs}$. We then calculated the relative rate of each category as $R_{cs} = (n_{cs} / N_{cs}) / \mu_s$. We then performed dimensionality reduction and detection of distinct mutational processes on $R_{96\times3088}$. We used non-negative matrix factorization (NMF)[35,36] to approximate R by a product of two non-negative matrices with lower ranks, $R \approx H_{96\times K} \times W_{K\times3083}$ where K represents the number of spectra (different mutational patterns) used to summarize the data (**Supplementary Figures S4, S5)**. K was chosen to be six because it was found to ensure that known mutational processes are captured as distinct spectra (such as mutations in CpG dinucleotides, C>A mutations in lung cancer and C>T mutations in melanoma and mutations at C's in TpC dinucleotides associated with HPV in cervical and head-and-neck cancers) **(Supplementary Figure S8)**; larger values of K did not significantly change the results. Columns of $H_{96\times6}$ represent the six distinct spectra and the columns of $W_{6\times3083}$ represent the relative weight of each of the six spectra (level of activation of each process) in each sample. The angle assigned to each sample in **Figure 2** reflects the rank order of the six weights in each sample. For sample s, we define i_{sr} to be the index (from 1 to K) of the spectrum with the r-th highest weight and the angle to be $\alpha_s = 2\pi\Sigma^{K}_{r=1}\, i_r(1/K)^r$. In this representation, the largest weight thus assigns the sample to one of the six sectors around the circle, the next largest assigns it to one of six sub-sectors, and so on. We also compared an alternative approach in which the weights in W were clustered and then the angle was determined according to the clustering dendrogram (**Supplementary Figures S6, S7)**.

Methods S2: Standard significance analysis method (MutSig1.0)

The standard method for detecting significantly mutated genes, which was used in many publications in the past recent years[1-3,5,6,15,16,20,23,26,27,37-49], is based on using a single average genome-wide background mutation rate μ estimated for the tumor type, and a number of category-specific relative rates (r_c, with $\mu_c=r_c\mu$) for mutations of a handful of different categories. The set of mutation categories used for analysis of the lung squamous cell carcinoma dataset[50] was as follows: (i) transitions in C's or G's in CpG dinucleotides; (ii) transversions in C's or G's in CpG dinucleotides; (iii) transitions in other C's or G's; (iv) transversions in other C's or G's; (v) transitions at A's or T's; (vi) transversions in A's or T's; and (vii) small insertions/deletions, nonsense and splice site mutations. This simple set of categories was chosen to be consistent with the output of the NMF procedure when run on the lung data alone. Note that a more complex category set may be required for analyzing other datasets.

In order to calculate gene-specific p-values, first a score is calculated for each gene based on the observed number of mutations (n) and number of covered bases (N) in each category c:

$s_g = \sum_c [-\log10 \text{ binomial}(n_c, N_c, \mu_c)]$

Next, a p-value is calculated for the gene by convoluting the background distributions of all the mutation types, and determining the probability of meeting or exceeding that score by background mutation alone. Finally, in order to control the false discovery rate, the Benjamini-Hochberg FDR procedure[51] is applied obtaining q-values. Genes with $q \leq 0.1$ are typically regarded as significantly mutated. When we applied this method ("MutSig1.0") to the lung squamous cell carcinoma dataset[50], it identified 450 genes as significantly mutated (**Supplementary Table S1**).

Methods S3: Significance analysis based on covariates (MutSigCV)

MutSigCV (Mutation Significance with *covariates*) extends the previously described MutSig algorithm[1,6,16,4]. Briefly, MutSig1.5[4] scores every mutation against the corresponding patient-specific background rate μ_p in which it is observed. The null distribution for the gene's score is calculated by convoluting across patients the patient-specific null distribution based on μ_p. A p-value for the gene is then calculated by comparing the observed score to this null distribution (as described in MutSig1.0[1]). Additionally, to prioritize genes that are mutated in many different samples, in preference to those having several mutations in the same sample, a scoring technique called *Projection* was introduced in MutSig1.5. First, the events in each sample are summarized by projecting to a space of *degrees* corresponding to the different categories of mutations it could have (or no mutations) – the lowest degree is associated with no mutations and the degrees increase with rarity of the event. The degree associated with each sample represents the rarest event observed in the sample. The probability for each sample to be of each degree is computed based on μ_p, and the score associated with that degree is given by the $-\log$(probability of the degree under the null hypothesis). As described above, the null distribution is then calculated by convoluting the sample-specific nulls (which also depend on μ_p).

The crucial advance behind MutSigCV is its accounting for gene-specific differences in background mutation rate. It approximates the mutation frequency in different genes, categories, and patients, $\mu_{g,c,p}$, (where g represents the gene, c the category, and p the patient), by using genomic covariates (such as expression level and DNA replication time). For very long genes, we can directly estimate the local background mutation rate (BMR) from (a) synonymous mutations in the gene's coding sequence, and (b) noncoding mutations in the flanking UTR and intronic sequences, safely beyond functional splice site mutations. However, for shorter genes, where there is not enough data to confidently estimate the local BMR, we extended the binning approach developed previously[6], where genes were binned by estimated expression level, and an average mutation rate was calculated for each bin, with the observation that mutation rate generally decreased with increasing expression. In MutSigCV, expression data, averaged across many tissue types in the Cancer Cell Line Encyclopedia[7], is augmented with other gene

characteristics observed empirically to co-vary with mutation rate, such as local DNA replication time[8], open vs. closed chromatin status measured by HiC mapping[12], local GC content, and local gene density (Supplementary Table S5). Note that gene expression levels and local replication time are highly correlated across tissue types (Supplementary Figure S3).

We have developed a general framework to encompass an arbitrary collection of such covariates. Briefly, each gene is placed in a high-dimensional covariate space, and the gene's nearest neighbors are identified. A set of nearest neighbors surrounding the gene of interest (which we term a *bagel* of genes, to reflect the fact that the gene itself is excluded and thus the set has a hole at its center) is built up around the original gene, and the local BMR is re-evaluated pooling the data across the genes in the bagel, gradually decreasing the uncertainty of the estimate as the total amount of genomic territory reflecting the genes in the bagel increases. A stopping criterion is imposed to balance the increased precision with the decreased accuracy (*i.e.* increased bias) that results from expanding outward to increasingly distant neighbors. Finally, a gene-specific contribution to the BMR is estimated using the frequency of synonymous and noncoding mutations in the gene plus its surrounding bagel. This gene-specific factor is combined with patient- and category- specific factors to yield the final estimated distribution for the expected value of $\mu_{g,c,p}$, calculated for each gene g, category c, and patient p combination. These $\mu_{g,c,p}$ values are then fed into the *Projection* method described above, here extended to take into account up to *two* mutations (instead of just one) in each patient, thus allowing an extra scoring opportunity for genes that have both alleles mutated in one or more patients (*e.g.* classic two-hit tumor suppressors like APC). A further advance of MutSigCV is its propagation of measurement error in the estimate of $\mu_{g,c,p}$, by preserving the mutation and coverage counts separately as $x_{g,c,p}$ and $X_{g,c,p}$ respectively, instead of merging them in the ratio $\mu = x/X$ and thereby losing the uncertainty in μ (i.e. error bars).

1 Input data

The input data to MutSigCV consists of three files. Each of these is a tab-delimited text file with a header row.

1.1 Mutation table

This file contains information about the mutations detected in the sequencing project. It lists one mutation per row, and the columns (named in the header row) report several pieces of information for each mutation. The five columns required by MutSigCV are

- `Hugo_Symbol` = name of the gene that the mutation was in
- `Tumor_Sample_Barcode` = name of the patient that the mutation was in
- `categ` = number of the *category* that the mutation was in (categories must match those in the coverage table)
- `is_coding` = 1 (the mutation in a coding region or splice-site) or 0 (the mutation is in a noncoding flanking region)
- `is_silent` = 1 (the mutation is a synonymous change) or 0 (the mutation is a coding change or is noncoding)

For the specific data file used in the present manuscript, the category numbers in `categ` are

1. transition mutations at CpG dinucleotides
2. transversion mutations at CpG dinucleotides
3. transition mutations at C:G basepairs not in CpG dinucleotides
4. transversion mutations at C:G basepairs not in CpG dinucleotides
5. transition mutations at A:T basepairs
6. transversion mutations at A:T basepairs
7. *null+indel* mutations, including nonsense, splice-site, and indel mutations

Categories 1-6 correspond to the mutation categories discovered in the NMF mutation spectrum analysis described in the main text.

1.2 Coverage table

This file contains information about the sequencing coverage achieved for each gene and patient. Within each gene-patient bin, the coverage is broken down further according to the *category* (*e.g.* A:T basepairs, C:G basepairs), and also according to the *zone* (silent/nonsilent/noncoding). The columns of the file are

- `gene` = name of the gene that this line reports coverage for
- `zone` = either `silent`, `nonsilent`, or `noncoding`
- `categ` = number of the *category* that this line reports coverage for (must match the categories in the mutation table)
- PATIENT1_NAME = number of covered bases for PATIENT1 in this gene, zone, and category
- PATIENT2_NAME = number of covered bases for PATIENT2 in this gene, zone, and category
- ...
- PATIENTn_p_NAME = number of covered bases for PATIENTn_p in this gene, zone, and category

Note, covered bases will typically contribute fractionally to more than one *zone* depending on the consequences of mutating to each of three different possible alternate bases. For example, a particular covered C base may count $\frac{2}{3}$ toward the `nonsilent` zone and $\frac{1}{3}$ toward the `silent` zone, if mutation to A or G causes an amino acid change whereas mutation to T is silent (synonymous).

1.3 Covariates table

This file contains the genomic covariate data for each gene, for example expression levels and DNA replication times, that will be used in MutSigCV to judge which genes are near to each other in covariate space. In general, the columns of this file are

- `gene` = name of the gene that this line reports coverage for

- COVARIATE1_NAME = value of COVARIATE1 for this gene
- COVARIATE2_NAME = value of COVARIATE2 for this gene
- ...
- COVARIATEn_v_NAME = value of COVARIATEn_v for this gene

For the specific data file used in the present manuscript, the columns are

- `gene` = name of the gene that this line reports coverage for
- `expr` = expression level of this gene, averaged across many cell lines in the Cancer Cell Line Encylcopedia
- `reptime` = DNA replication time of this gene, ranging approximately from 100 (very early) to 1000 (very late)
- `hic` = chromatin compartment of this gene, measured from HiC experment, ranging approximately from -50 (very closed) to +50 (very open)

Note, gene and patient names must agree across these three tables. Similarly, the `categ` category numbers must agree between the mutation table and the coverage table.

2 Algorithmic procedure

2.1 Representation of data matrices

In the first step of the algorithm, the input data files are loaded from disk and converted in memory to the following matrix forms. The matrix indices g, c, p, v range from 1 to n_g, n_c, n_p, n_v, representing the total number of genes, categories, patients, and covariates respectively. The special case $c = n_c + 1$ is used to represent the *total* counts. For mutation counts n, this is simply the sum across 1 to n_c. However, for coverage counts N, the total may be different than the sum across 1 to c, due to categories with overlapping territories, *e.g.* the territory of A:T mutations (which can happen at any A:T basepair) is included within the territory of indel mutations (which can happen at any basepair). In practice, the total coverage N will be equal to the coverage of the *null+indel* category.

2.1.1 Mutation counts

The mutation table is converted to the following three matrices

$$n_{g,c,p}^{silent}$$

$$n_{g,c,p}^{nonsilent}$$

$$n_{g,c,p}^{noncoding}$$

Each of these n matrices represents the number of mutations for a given gene g, category c, and patient p.

2.1.2 Coverage counts

The coverage table is converted to the following three matrices

$$N_{g,c,p}^{silent}$$

$$N_{g,c,p}^{nonsilent}$$

$$N_{g,c,p}^{noncoding}$$

Each of these N matrices represents the number of covered sequenced bases for a given gene g, category c, and patient p.

2.1.3 Covariate values

The covariate table is converted to the following matrix

$$V_{v,g}$$

It represents the value of covariate v for gene g.

2.2 Embedding of genes in covariate space

In the next step of the algorithm, each covariate is converted to a Z-score, *i.e.* centered and normalized, by subtracting the mean and dividing by the standard deviation across genes.

$$Z_{v,g} = \frac{V_{v,g} - \frac{1}{n_g}\sum_{i=1}^{n_g} V_{v,i}}{\sqrt{\frac{1}{n_g - 1}\sum_{j=1}^{n_g}\left(V_{v,j} - \frac{1}{n_g}\sum_{i=1}^{n_g} V_{v,i}\right)^2}}$$

Each gene is now represented as a point in $\mathbb{R}^{n_v}$, such that the coordinate v of gene g is equal to $Z_{v,g}$. Pairwise distances between genes are calculated in Euclidean fashion, such that the distance between genes i and j is

$$D_{i,j} = \sqrt{\sum_{v=1}^{n_v}(Z_{v,i} - Z_{v,j})^2}$$

2.3 Local regression using *bagels*

In this step of the computation, the local BMR (background mutation rate) of each gene is estimated from the silent and noncoding mutations of the gene itself, plus (if necessary) those of its neighbor genes in the covariate space. First, silent and noncoding mutations are pooled together across patients and categories to yield the following background (*bkgd*) counts

$$n_g^{bkgd} = \sum_{p=1}^{n_p}(n_{g,c+1,p}^{silent} + n_{g,c+1,p}^{noncoding})$$

$$N_g^{bkgd} = \sum_{p=1}^{n_p}(N_{g,c+1,p}^{silent} + N_{g,c+1,p}^{noncoding})$$

Note, as mentioned above, here $c + 1$ indicates the *total* counts across categories.

For each gene, a *bagel* of the closest neighboring genes in the covariate space is chosen, such that all of the genes in the bagel do not disagree with the BMR (background mutation rate) estimated for the gene itself. The neighbor genes in the bagel of gene g are represented as the largest set B_g that meets these criteria

$$\forall(i \in B_g, j \notin B_g)(D_{g,i} \leq D_{g,j})$$

and

$$\forall(i \in B_g)(Q_{i,g} \geq Q_{min})$$

and

$$|B_g| \leq n_B^{max}$$

where n_B^{max}, the *maximum neighbors*, is defined to be 50, and Q_{min}, the *minimum quality*, is defined to be 0.05. $Q_{i,g}$ is the two-sided p-value for comparing the BMRs of gene i and the center gene g given their observed mutation and coverage counts.

$$Q_{i,g} = 2\min(Q_{i,g}^{\text{left}}, 1 - Q_{i,g}^{\text{left}})$$

$$Q_{i,g}^{\text{left}} = H_C(n_i^{bkgd}, N_i^{bkgd}, n_g^{bkgd}, N_g^{bkgd})$$

H_C is the cumulative form of the beta-binomial distribution H.

$$H_C(n_1, N_1, n_2, N_2) = \sum_{n=0}^{n_1} H(n, N_1, n_2, N_2)$$

H is the beta-binomial probability mass function

$$\begin{aligned} H(n_1, N_1, n_2, N_2) &= \binom{N_1}{n_1} \frac{B(n_1 + \alpha, N_1 - n_1 + \beta)}{B(\alpha, \beta)} = \\ &= \frac{\Gamma(N_1 + 1)\Gamma(N_2 + 2)\Gamma(n_1 + n_2 + 1)\Gamma(N_1 + N_2 - n_1 - n_2 + 1)}{\Gamma(n_1 + 1)\Gamma(n_2 + 1)\Gamma(N_1 - n_1 + 1)\Gamma(N_2 - n_2 + 1)\Gamma(N_1 + N_2 + 2)} \end{aligned}$$

where $\alpha = n_2 + 1$, $\beta = N_2 - n_2 + 1$ and Γ is the gamma function. Note that H is normalized, *i.e.* $\sum_{n1=0}^{N_1} H(n_1, N_1, n_2, N_2) = 1$.

Finally, the total background counts x_g and X_g for the gene are calculated, given the background counts in the gene itself plus its bagel (note, it is possible for a gene to have no genes in its bagel).

$$x_g = n_g^{bkgd} + \sum_{i \in B_g} n_i^{bkgd}$$

$$X_g = N_g^{bkgd} + \sum_{i \in B_g} N_i^{bkgd}$$

2.4 Incorporation of category- and patient-specific rates

In this section, category- and patient-specific background mutation rates are calculated and combined with the per-gene x_g and X_g background counts from the previous section.

First, mutations and coverage are summed across the three *zones* to yield total counts

$$n_{g,c,p}^{total} = n_{g,c,p}^{silent} + n_{g,c,p}^{nonsilent} + n_{g,c,p}^{noncoding}$$
$$N_{g,c,p}^{total} = N_{g,c,p}^{silent} + N_{g,c,p}^{nonsilent} + N_{g,c,p}^{noncoding}$$

Totals are calculated across genes

$$n_{c,p}^{total} = \sum_{g=1}^{n_g} n_{g,c,p}^{total}$$

$$N_{c,p}^{total} = \sum_{g=1}^{n_g} N_{g,c,p}^{total}$$

and across patients

$$n_{c}^{total} = \sum_{p=1}^{n_p} n_{c,p}^{total}$$

$$N_{c}^{total} = \sum_{p=1}^{n_p} N_{c,p}^{total}$$

to yield marginal category-specific mutation rates

$$\mu_c = \frac{n_c^{total}}{N_c^{total}}$$

and the overall total mutation rate

$$n_{overall}^{total} = n_{c+1}^{total}$$
$$N_{overall}^{total} = N_{c+1}^{total}$$
$$\mu_{overall} = \frac{n_{overall}^{total}}{N_{overall}^{total}}$$

Patient-specific marginal mutation rates are calculated

$$n_p^{total} = n_{c+1,p}^{total}$$
$$N_p^{total} = N_{c+1,p}^{total}$$
$$\mu_p = \frac{n_p^{total}}{N_p^{total}}$$

and *relative* category- and patient-specific rates f are calculated by normalizing to $\mu_{overall}$

$$f_c = \frac{\mu_c}{\mu_{overall}}$$
$$f_p = \frac{\mu_p}{\mu_{overall}}$$

Also, the relative amounts of covered territory f^N per category and patient are calculated. The category-specific territory is normalized to the total overall territory, and the patient-specific territory is normalized to the *mean* patient-specific territory.

$$f_c^N = \frac{N_c^{total}}{N_{overall}^{total}}$$
$$f_p^N = \frac{N_p^{total}}{\frac{1}{n_p} N_{overall}^{total}}$$

Finally, $x_{g,c,p}$ and $X_{g,c,p}$ are estimated by the product of marginal relative rates and x_g and X_g:

$$x_{g,c,p} = x_g f_c f_p f_c^N f_p^N$$
$$X_{g,c,p} = X_g f_c^N f_p^N$$

2.5 Calculation of gene p-values using 2-D *Projection* method

For each gene, the mutational signal from the observed nonsilent counts are compared to the mutational background estimated above. This is done by first calculating how likely it would be by chance for each sample to have a mutation in each of the categories.

$$P^{(0)}_{g,c,p} = H(0, N^{nonsilent}_{g,c,p}, x_{g,c,p}, X_{g,c,p})$$
$$P^{(1)}_{g,c,p} = H(1, N^{nonsilent}_{g,c,p}, x_{g,c,p}, X_{g,c,p})$$
$$P^{(2+)}_{g,c,p} = 1 - P^{(0)}_{g,c,p} - P^{(1)}_{g,c,p}$$

H is the same beta-binomial probability mass function defined earlier. $P^{(0)}_{g,c,p}$ is the probability that in this gene g, patient p, has *zero* mutations in category c. $P^{(1)}_{g,c,p}$ is the probability of exactly *one* mutation, and $P^{(2+)}_{g,c,p}$ is the probability of two or more.

Next, within each patient, mutation categories are sorted into an order of priorities according to $P^{(1)}$. The categories are sorted from the category most likely by chance (lowest priority), to the category least likely by chance (highest priority). Each patient is *projected* to a two-dimensional space of *degrees* $D_{g,p} = (d_1, d_2)$, taking into account up to two of its mutations, with the mutations prioritized by category as described, *i.e.* the two with the highest priorities ($d_1 \geq d_2$). For example, a sample of degree (0,0) has no mutations. A sample of degree (1,0) has one mutation, and that mutation is of the lowest-priority category. A sample of degree (n_c,0) has one mutation, and that mutation is of the highest-priority category. A sample of degree (n_c,n_c) has at least two mutations of the highest-priority category. Then, in order to compute the distribution of patient degrees expected under the estimated model of background mutation, the probability is calculated for each patient to be of each degree by chance

$$P^{(d_1,d_2)}_{g,p} = \begin{cases} \prod_{d=1}^{n_c} P^{(0)}_{g,d,p}, & \text{if } d_1 = 0, d_2 = 0 \\ P^{(1)}_{g,d_1,p} \prod_{d=1}^{d_1-1} P^{(0)}_{g,d,p} \prod_{d=d_1+1}^{n_c} P^{(0)}_{g,d,p}, & \text{if } d_1 > 0, d_2 = 0 \\ P^{(1)}_{g,d_1,p}(P^{(1)}_{g,d_2,p} + P^{(2+)}_{g,d_2,p}) \prod_{d=d_2+1}^{d_1-1} P^{(0)}_{g,d,p} \prod_{d=d_1+1}^{n_c} P^{(0)}_{g,d,p}, & \text{if } d_1 > 0, 0 < d_2 < d_1 \\ P^{(2+)}_{g,d_1,p} \prod_{d=d_1+1}^{n_c} P^{(0)}_{g,d,p}, & \text{if } d_1 > 0, d_2 = d_1 \\ 0 \text{ (impossible by definition)}, & \text{if } d_2 > d_1 \end{cases}$$

Each degree is also associated with a score S.

$$S_{g,p}^{(d_1,d_2)} = \begin{cases} 0, & \text{if } d_1 = 0, d_2 = 0 \\ S_{null} - \log_{10} P_{g,d_1,p}^{(1)}, & \text{if } d_1 > 0, d_2 = 0 \\ S_{null} - \log_{10} P_{g,d_1,p}^{(1)} - \log_{10} P_{g,d_2,p}^{(1)}, & \text{if } d_1 > 0, 0 < d_2 < d_1 \\ S_{null} - \log_{10} P_{g,d_1,p}^{(2+)}, & \text{if } d_1 > 0, d_2 = d_1 \\ 0 \text{ (impossible by definition)}, & \text{if } d_2 > d_1 \end{cases}$$

where S_{null} represents the *null score boost* added to scores associated with the presence of a null mutation, reflecting the increased value of a null mutation towards the total evidence of a gene's driver potential.

$$S_{null} = \begin{cases} 0, & \text{if } d_1 < n_c \\ +3, & \text{if } d_1 = n_c \end{cases}$$

The gene is assigned a total overall score for the observed configuration of patient degrees, by summing the scores associated with the observed degree D of each patient.

$$S_g^{obs} = \frac{\sum_{p=1}^{n_p} S_{g,p}^{D_{g,p}}}{E_{min}}$$

where E_{min} is the *minimum effect size* considered sufficient evidence for positive selection in the gene. A value of $E_{min} = 1.25$ is used, corresponding to a required $+25\%$ effect size. Smaller effect sizes are treated as falling within the noise regime of the data.

In order to determine the probability of obtaining a given score by chance, *i.e.* from background mutation alone, a null distribution of scores is calculated by convolution. First, within each individual patient p, the null distribution of scores for that patient is computed by convoluting the probabilities and scores of each possible degree

$$P_{g,p}^{(S=x)} = \bigotimes_{d_1=0}^{n_c} \bigotimes_{d_2=0}^{n_c} P_{g,p}^{(d_1,d_2)} \delta(x - S_{g,p}^{(d_1,d_2)})$$

where δ is the Dirac delta function. Then, the distributions for each patient are convoluted together to obtain the overall null distribution for the gene.

$$P_g^{(S=x)} = \bigotimes_{p=1}^{n_p} P_{g,p}^{(S=x)}$$

The p-value of the gene, *i.e.* the probability of obtaining at least the observed score by chance, is given by

$$P_g^{(S \geq S^{obs})} = \int_{S_g^{obs}}^{\infty} P_g^{(S=x)} dx$$

In practice, it is easier to compute this by calculating the probability of obtaining *less* than the observed score and subtracting from one.

$$P_g^{(S \geq S^{obs})} = 1 - \int_{0}^{S_g^{obs}} P_g^{(S=x)} dx$$

2.6 Calculation of False Discovery Rate

Each gene is assigned a q-value, *i.e.* False Discovery Rate, using the method of Benjamini and Hochberg[51]. Genes with $q \leq 0.1$ are considered to be significantly mutated.

3 Output data

The output of the algorithm is a table listing the genes with their p- and q-values, ordered by p-value.

Table S1: Lung cancer genes called significantly mutated under the naive model

T**able S1** (see Excel spreadsheet tab 1). List of 450 genes declared to be significantly mutated in lung squamous cell carcinoma (see text), when using the standard analytical approach (**Supplementary Method S2)**, which fails to account for important sources of heterogeneity in mutational processes. Genes known to be involved in lung squamous cell carcinoma are shown in green; other known cancer genes are shown in blue. Olfactory receptors are shown in red. Other particularly suspicious genes are shown in orange.

Table S2: Cancer samples analyzed

Table S2 (see Excel spreadsheet tab 2). List of the 3,083 cancer samples analyzed in this study. Each sample is listed with its name, tumor type, and whether it underwent whole-exome or whole-genome sequencing. Also listed are the number of somatic coding mutations, and the somatic coding mutation rate. These mutations are then broken down into the six categories shown at the bottom of **Figure 1**.

Table S3: Mutation spectrum of each sample

Table S3 (see Excel spreadsheet tab 3). List of the 2,892 samples having at least 10 somatic coding mutations each, as shown in **Figure 2**. The mutations of each sample are broken down into the six color-coded categories from **Figure 1**, and then further into the 16 contexts distinguished by the identity of the 5' and 3' nucleotide neighbors. These are the 96 categories of the "Lego" plots (**Supplementary Figures S4, S8**). For instance, the first category "ACAtoAAA" refers to C→A mutations in the context Ap*CpA, i.e. both the 5' and 3' neighbors of the mutating cytosine are adenosine. The top of row of the table ("coverage") lists the average number of coding basepairs in the exome that were covered to sufficient depth in DNA sequencing.

Table S4: Genomic windows and their characteristics

Table S4 (see Excel spreadsheet tab 4). The hg19 genome was broken down into nonoverlapping windows of 100Kb. These windows are listed, along with their average expression level across 91 cell lines in the CCLE[7], their DNA replication time[8], expressed on a scale of 100 (early) to 1500 (late), and their average noncoding mutation frequency (mutations per bp), measured from the panel of 126 cancer samples that were subjected to whole-genome sequencing.

Table S5: Genes and their characteristics

Table S5 (see Excel spreadsheet tab 5). Genes are listed with their chromosomal coordinates, their average expression level across 91 cell lines in the CCLE[7], their DNA replication time[8], expressed on a scale of 100 (early) to 1500 (late), and their average noncoding mutation frequency (mutations per bp), measured from the panel of 126 cancer samples that were subjected to whole-genome sequencing. Also listed for each gene are the local GC content in the genome (measured on a 100kB scale), a HiC[12]-derived metric indicating which chromosomal compartment the gene is in (negative values = closed compartment "B", positive values = open compartment "A"). Finally, a number of technical metrics for each gene are listed, relating to the efficiency and depth of the sequencing process. Correlations between mutation frequency and the various gene characteristics are summarized in **Table S6**.

Table S6: Correlations of somatic mutation frequency with gene characteristics

Table S6 (see Excel spreadsheet tab 6). The gene characteristics in Table S5 were evaluated to determine which were useful predictors of somatic noncoding mutation frequency. For single-variable regressions, scatterplots and linear regressions were evaluated as in Figure S9. Regression coefficients "R" are listed in the table. The strongest predictor of mutation frequency was DNA replication time, yielding a correlation of 0.60. Expression level and HiC chromatin compartment were just behind this in predictive power, with correlations of 0.53 and 0.54 respectively. For multivariate regressions, the MATLAB "regress" function[52] was used to find the best fit to the set of covariates being evaluated. This fit was then used to predict the mutation frequency, and the coefficient "R" was calculated for the correlation of the predicted and observed mutation frequencies. In general, the multivariate regressions yielded only a modest improvement (R_{max} = 0.66) over the single-variable fits, reflecting the fact that the genomic characteristics were highly mutually correlated. Finally, a number of technical metrics from the WGS and hybrid capture sequencing technology were evaluated as possible covariates / confounders in the analysis. For WGS, the metrics were: mean sequencing depth, read length, fraction of paired reads, and percent of bases above 20X coverage. For hybrid capture, they were: the on-target rate (defined stringently as the number of reads that overlapped one of the gene's exons, divided by the total number of reads that overlapped the gene), the mean sequencing depth, the percent of bases covered at various thresholds from 20X to 300X, and finally the mean GC content of the reads for that gene. These technical metrics were generally uncorrelated ($|R|<0.2$) to mutation frequency, and thus much less useful in the analysis.

Table S7: Performance survey across variants of MutSig algorithm

Table S7 (see Excel spreadsheet tabs 7 and 8). Several variants of the MutSig algorithm were compared, using the lung squamous cell carcinoma dataset[50] as a benchmark. Each version was run on the entire set of ~18,000 genes in the genome, and the total number of genes called significantly mutated (q<0.1) is reported in line 9 of the spreadsheet. A set of 40 "indicator" genes were selected as an additional useful readout of the performance of the algorithm, and are listed from lines 11 onward. The two tabs of the table (tabs 7 and 8) show p-values and q-values respectively. Colored cells indicate genes that were found to be significantly mutated (q<0.1) in each analysis. Colors distinguish the various subsets of indictator genes. From top to bottom, dark green genes are known from previous studies to be involved in lung cancer. Light green genes are known to be involved in other cancer types. Grey genes have unknown cancer association status. Red genes (e.g. titin, mucins, ryanodine receptors, cub and sushi domain proteins) are the dubious hits discussed in the main text, i. e. genes with very low expression and/or very late replication times, which tend to have elevated gene-specific background mutation frequencies, and represent false-positive findings in the significance analysis. Different variants of methods for significance analysis are shown in columns from left to right. **(1)** the original MutSig1.0 algorithm[1] (**Method S2**), which found the 450 significantly mutated genes listed in **Table S1,** including all the known cancer genes but also a huge number of spurious hits. **(2)** the MuSiC algorithm[38] using convolutions, which is a reimplementation of the MutSig1.0 method. This method found 789 significantly mutated genes. **(3)** the MuSiC algorithm replacing the convolutions with a likelihood ratio test (LRT), also described previously[1]. This method identified 3,318 significantly mutated genes. **(4)** the MuSiC algorithm using Fisher's method of combining *p*-values, which identified 216 significantly mutated genes and missed some known cancer genes. Note, the MuSiC algorithm reports a gene as significantly mutated if at least two out of its three submethods called the gene significant. Applying this criterion, MuSiC finds an overall list of 721 significantly mutated genes. We also ran MuSiC with sample-specific background rates; this did not improve the results (data not shown). The fundamental problem causing the inflated list of significant genes, as with MutSig1.0, is its assumption of a single genome-wide background mutation rate per category and/or per patient. **(5)** the early refinement MutSig1.5, which take into account gene-specific nonsilent-to-silent mutation ratios ENREF 22[23]. This reduced the gene count to 150 and eliminated many of the dubious hits, but also lost some of the known cancer genes. **(6)** regression version of MutSigCV ("with covariates"), where three gene characteristics (expression level, DNA replication time, and HiC chromatin compartment, **Table S5**) were used in a multivariate linear regression (**Table S6**) to predict the gene-specific mutation frequency. This reduced the gene count to 50, but was overall unsuccessful in enriching known cancer genes over spurious hits; **(7)** standard version of MutSigCV, where the multivariate linear regression method was replaced by the local nonlinear regression "bagel-finding" technique described in detail in **Method S3**. This method was judged to produce the best results out of the MutSig variants. It found a total of 11 significantly mutated genes, including all the known cancer genes, and none of the dubious genes. **(8)** The same analysis as the previous column, but with the bagel-finding

procedure supplemented by three additional covariates, techinical metrics from the sequencing process: mean WGS sequencing depth, mean percent of bases covered at >200X in capture sequencing, and mean GC content of capture reads (**Table S5**). Addition of these technical covariates had only a modest effect on the results, with the loss of NOTCH1 and HLA-A, the total number of significantly mutated genes falling from 11 to 9. **(9)** Variation of MutSig that estimates the gene-specific mutation frequency directly from the silent and flanking (intronic/UTR) mutations of each gene itself, without the assistance of "bagels". To optimize this estimate, the silent+flanking mutations of all patients were added together (pooled). This approach missed several known cancer genes: PIK3CA, NFE2L2, RB1, NOTCH1, and FBXW7. In other words, using neighbor genes in covariate space can mitigate some of the damage to a gene's significance level that is caused by that gene having by chance a modest number of silent and/or flanking mutations. **(10)** Repeat of the previous column, but with patients kept segregated from each other. In this approach, each patient is scored separately as to whether it has a higher rate of functional (nonsilent coding) mutations over its rate of nonfunctional (silent + flanking) mutations. This removes the advantage of pooling evidence across patients, and leads to the identification of only a single significantly mutated gene: TP53. **(11)** Same as previous column, but with an even more stringent criterion for each patient: it must show a higher rate of functional (nonsilent coding mutations divided by nonsilent coding territory) mutations over its *total* mutation rate (all mutations divided by all territory). This approach corresponds to the extremely stringent "InVEx" method developed previously[19] in reaction to the very high mutation rate of melanoma. It identified no significantly mutated genes in the lung dataset, not even TP53. In order for this approach to produce useful findings it requires supplementing with external data about the likely functional impact of individual mutations, e.g. from PolyPhen[19,53]; as a purely statistical approach it is underpowered.

References

1. Getz, G. *et al.* Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* **317**, 1500 (2007).
2. TCGA. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature* (2012).
3. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-60 (2011).
4. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405-9 (2012).
5. Bass, A.J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet* **43**, 964-8 (2011).
6. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-72 (2011).
7. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-7 (2012).
8. Chen, C.L. *et al.* Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* **20**, 447-57 (2010).
9. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet* **91**, 1033-40 (2012).
10. Duda, R., Hart, P. & Stork, D. *Pattern Classification*, (John Wiley & Sons, Inc., 2001).
11. Roberts, S.A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol Cell* **46**, 424-35 (2012).
12. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
13. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149**, 979-993 (2012).
14. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-6 (2010).
15. TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8 (2008).
16. TCGA. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-15 (2011).
17. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107-20 (2012).
18. Barbieri, C.E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* **44**, 685-9 (2012).
19. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251-63 (2012).
20. Wang, L. *et al.* SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med* **365**, 2497-506 (2011).

21. Dulak, A.M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet* (2013).
22. Pugh, T.J. *et al.* Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature* **488**, 106-10 (2012).
23. Lohr, J.G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc Natl Acad Sci U S A* **109**, 3879-84 (2012).
24. Lee, R.S. *et al.* A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. *J Clin Invest* **122**, 2983-8 (2012).
25. Pugh, T.J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat Genet* **45**, 279-84 (2013).
26. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-20 (2011).
27. Berger, M.F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485**, 502-6 (2012).
28. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol* **12**, R1 (2011).
29. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-9 (2009).
30. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).
31. Reich, M. *et al.* GenePattern 2.0. *Nat Genet* **38**, 500-1 (2006).
32. Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601-2 (2011).
33. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* (2013).
34. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
35. Lee, D.D. & Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-91 (1999).
36. Brunet, J.P., Tamayo, P., Golub, T.R. & Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* **101**, 4164-9 (2004).
37. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-75 (2008).
38. Dees, N.D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22**, 1589-98 (2012).
39. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-74 (2006).
40. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801-6 (2008).
41. Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* **44**, 47-52 (2012).
42. Morin, R.D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**, 298-303 (2011).

43. Parsons, D.W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807-12 (2008).
44. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-8 (2007).
45. Wood, L.D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108-13 (2007).
46. Lin, J. *et al.* A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res* **17**, 1304-18 (2007).
47. Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869-73 (2010).
48. Stephens, P.J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-10 (2009).
49. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27**, 175-81 (2011).
50. TCGA. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* (2012).
51. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statistical Society Series B* **57**, 289 (1995).
52. Chatterjee, S. & Hadi, A.S. Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science* **1**, 379-393 (1986).
53. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).