nature genetics

SUPPLEMENTARY INFORMATION

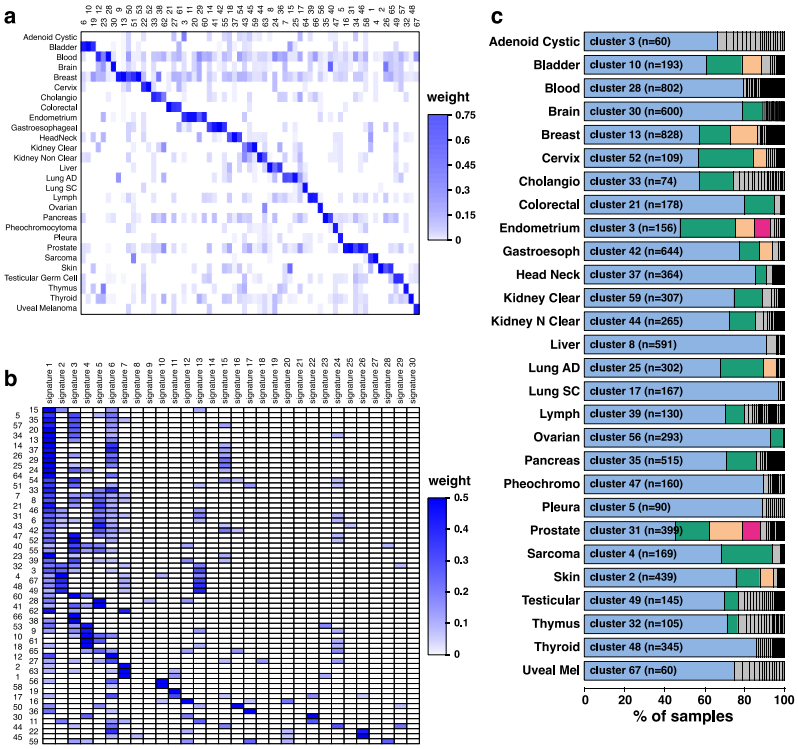
https://doi.org/10.1038/s41588-019-0572-y

In the format provided by the authors and unedited.

Identification of cancer driver genes based on nucleotide context

Felix Dietlein[1,2,7*], Donate Weghorn[3,4,5,7], Amaro Taylor-Weiner[1,2], André Richters[2,6], Brendan Reardon[1,2], David Liu[1,2], Eric S. Lander[2], Eliezer M. Van Allen[1,2,8*] and Shamil R. Sunyaev[3,4,8*]

[1]Department of Medical Oncology, Dana–Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [2]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA. [3]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [5]Centre for Genomic Regulation, Barcelona, Spain. [6]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]These authors contributed equally: Felix Dietlein, Donate Weghorn. [8]These authors jointly supervised this work: Eliezer M. Van Allen, Shamil R. Sunyaev. *e-mail: Felix_Dietlein@dfci.harvard.edu; EliezerM_VanAllen@dfci.harvard.edu; ssunyaev@rics.bwh.harvard.edu



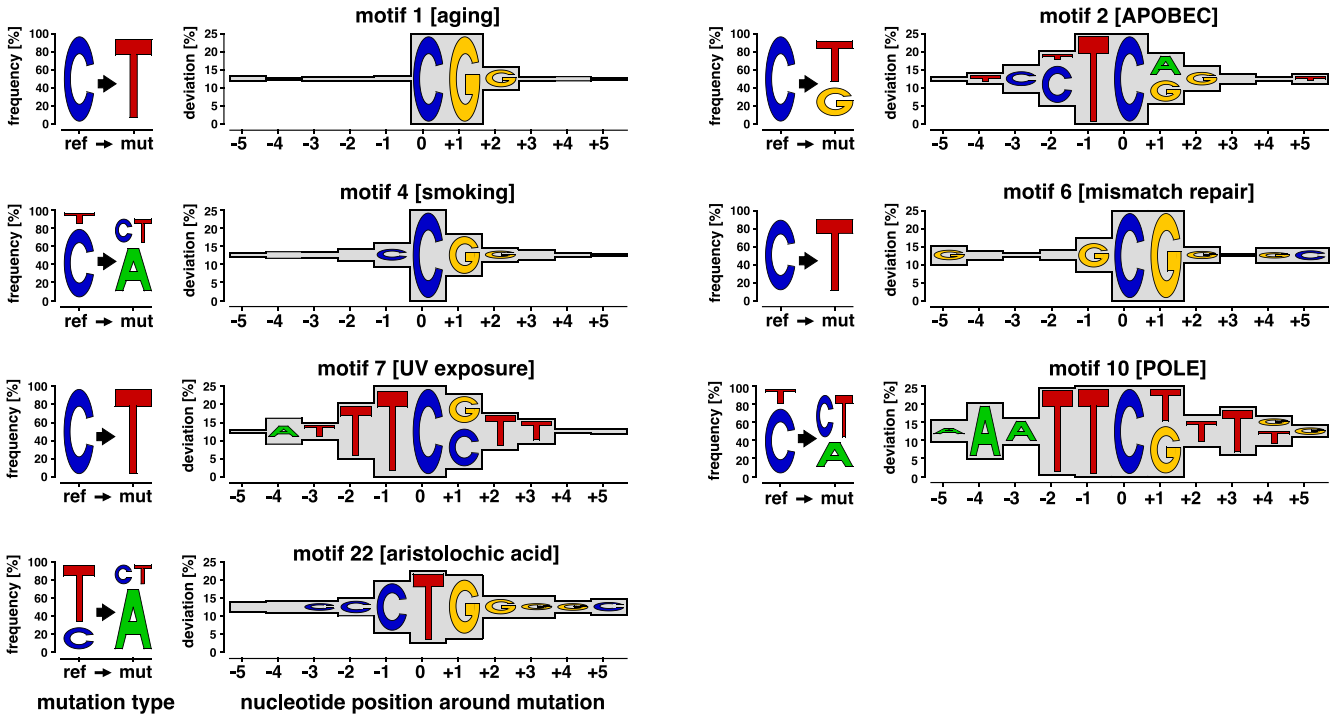
Supplementary Figure 1 | Benchmarking of the Bayesian hierarchical clustering step across cancer types and known mutational processes.

a, MutPanning first performs a Bayesian hierarchical clustering step to aggregate samples with similar underlying mutational processes and to derive composite likelihood coefficients from these clusters. We examined the results of Bayesian hierarchical clustering procedure based on our full study cohort (n = 11,873 samples, cf. Extended Data Figure 5 for the number of samples per cancer type). In brief, we counted the number of samples per cancer type (row) for each cluster (column), and normalized these counts for each cluster. The heatmap shows which clusters occur in which cancer type. These analyses revealed that the majority of the clusters predominantly contain samples of one cancer type.

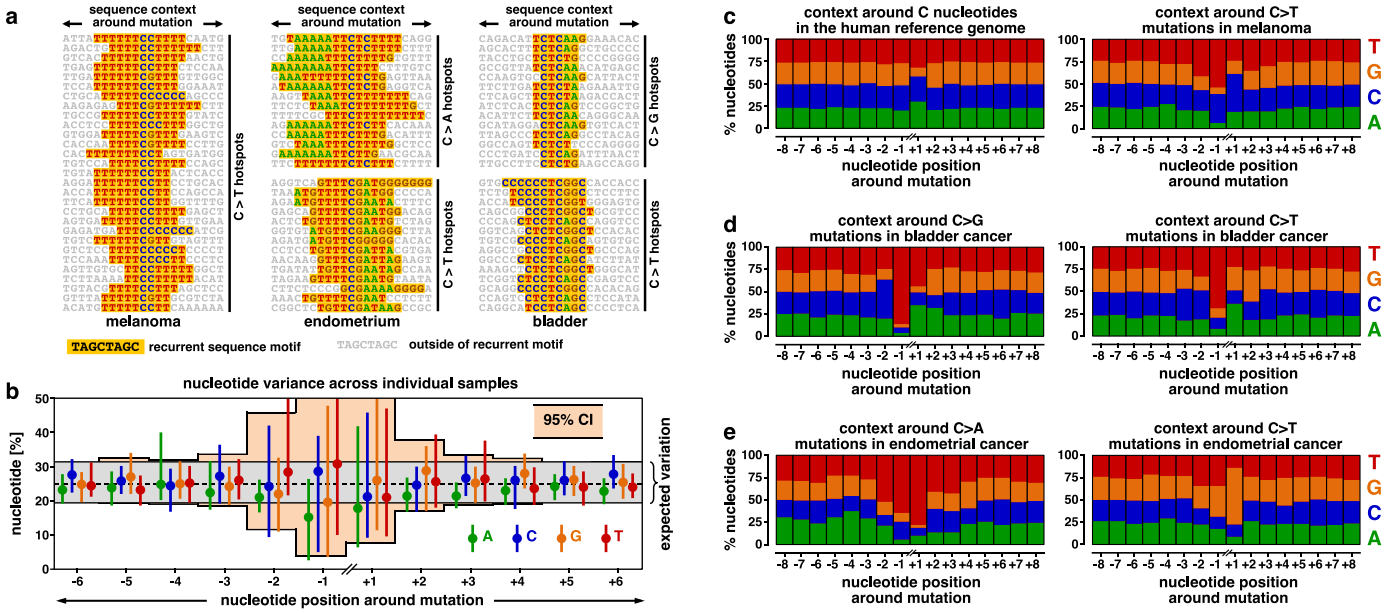
To interpret this result we note that our Bayesian hierarchical clustering procedure follows a 2-step design. In the first step, it primarily clusters samples within the same cancer type, whereas in the second step it clusters samples from different cancer types. That way, the clustering procedure maximizes the robustness of the clusters within each cancer type, i.e. clusters within each cancer type are largely independent of the other cohorts the cancer type was processed with.

b, We further compared the results from the Bayesian hierarchical clustering step against known mutational processes. For this purpose, we counted for each cluster the number of mutations in all possible 96 trinucleotide contexts, thereby deriving a 96-dimensional count vector. We then decomposed the 96-dimensional count vector of each cluster (rows) into known mutational processes (30 COSMIC signatures, columns) using the tool deconstructSigs. The heatmap visualizes the weights of this decomposition. This analysis revealed that most clusters were either dominated by a single COSMIC mutation signature, or reflected the composition of 1-3 known mutational processes. These combinations were not arbitrary, but reflected known combinations. For instance, signatures 1 and 5 occur as an admixture in several clusters. These signatures are known as mutational background signatures that occur across a wide range of different tumor types ("clock-like mutational processes"). Furthermore, signatures 2 and 13 frequently co-occur together and these signatures both originate from APOBEC deaminase activity. Similarly, signatures 6, 15, 20 and 26 result from mismatch repair deficiency and can be frequently co-occur together.

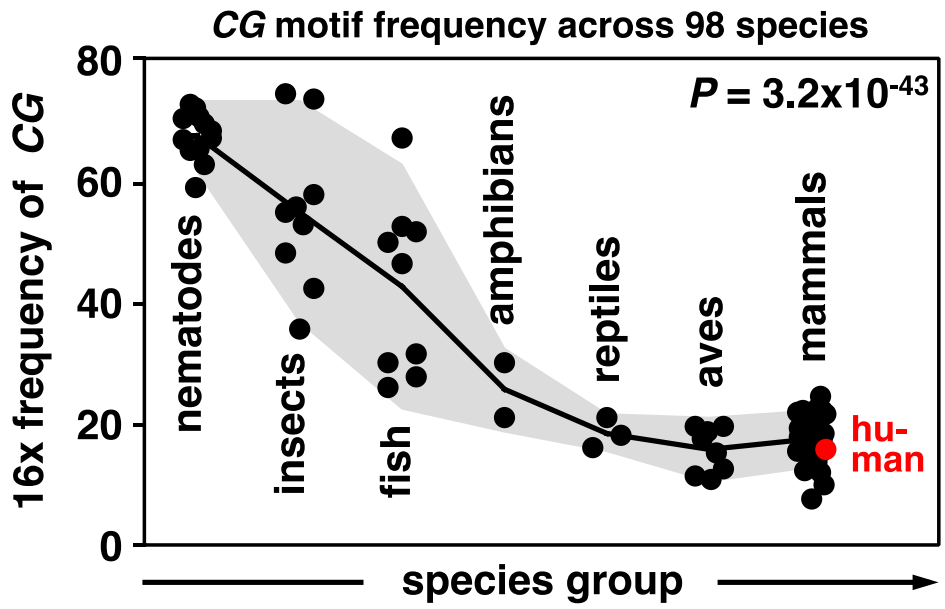
c, We further examined how many clusters were present in each cancer type. In brief, we counted the number of samples per cancer type for each cluster. We normalized these counts for each cancer type and plotted them as a stacked bar graph for each cancer type (row). The fraction of the most dominant cluster is shown in blue, the second dominant cluster in green etc. Most cancer types contain 1-3 main clusters that cover >80% of their samples, whereas the remaining samples (shown in gray) cluster into clusters with smaller fractions. This result is concordant with the goal of the Bayesian hierarchical clustering procedure to cluster most samples within each cancer type. Hence, clusters within each cancer type are largely independent of the other cohorts the cancer type was processed with.



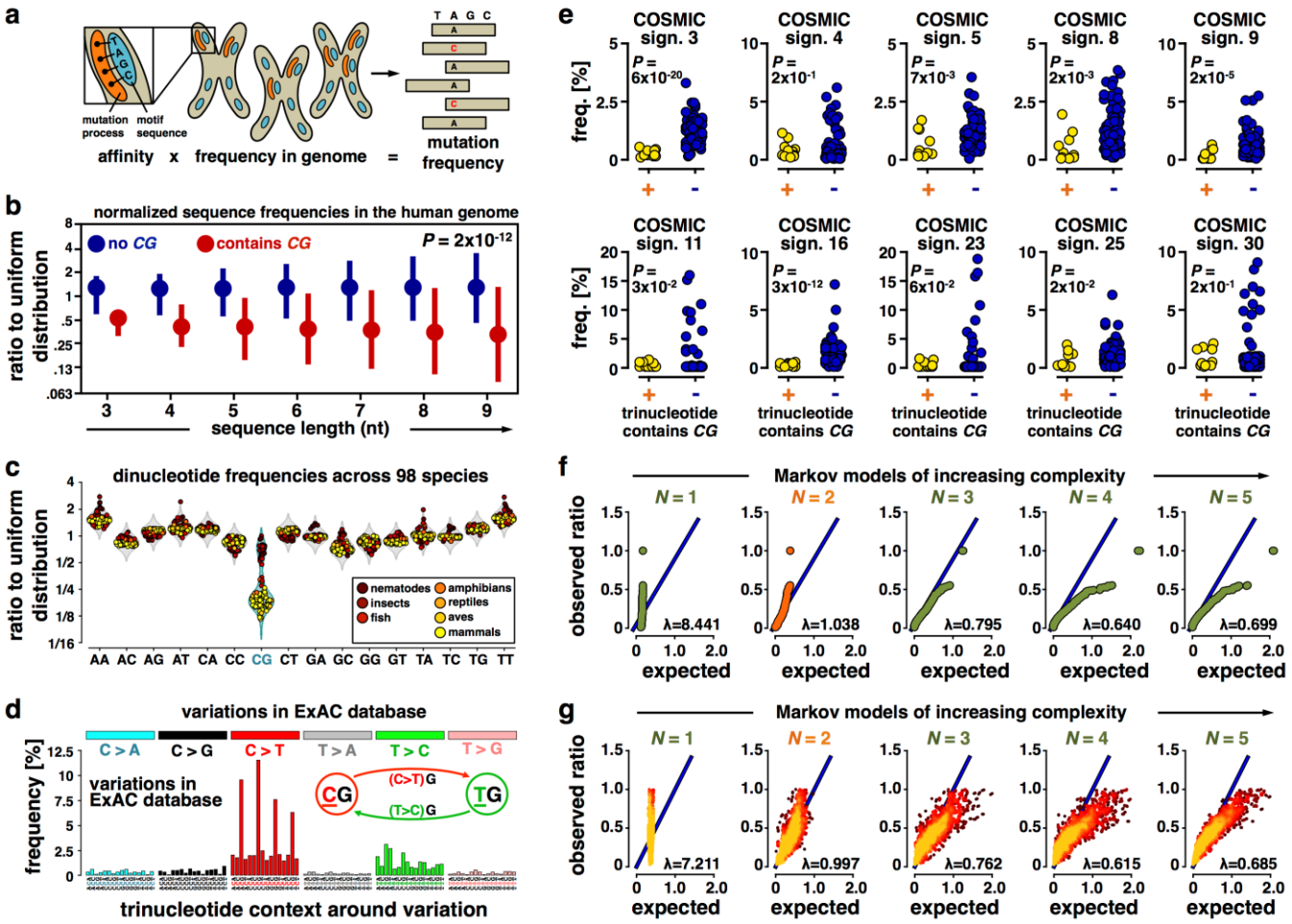
Supplementary Figure 2 | Passenger mutations are enriched in broad and characteristic nucleotide contexts.
We observed that passenger mutations were surrounded by characteristic nucleotide contexts. We visualized these characteristic contexts by sequence logo plots. Left: For each mutational process, mutation frequencies of the reference (ref) nucleotides (C vs. T) and mutation types (mut: transitions (C/T), type-I-transitions (A), type-II-transitions (G)) are visualized as logo plots. Right: Sequence logo plots represent the relative over-representation (deviation) of the nucleotides around mutations relative to their expected frequencies in the human exome. In other words, the height of each flanking nucleotide indicates its impact on the local mutation probability. Most characteristic nucleotide contexts clearly exceed the trinucleotide context (±1) around mutations. For instance, the nucleotide context associated with UV exposure (motif 7) contains almost exclusively C>T mutations; T's are overrepresented in 5' adjacency and C/G followed by T's are overrepresented in 3' adjacency. Hence, sequence logo plots provide a convenient way to visualize the dependence of local mutation probabilities on the broad nucleotide context.



Supplementary Figure 3 | Initial exploration of the nucleotide context around passenger mutations.
Context-specific mutation rates are typically modeled based on the 5' and 3' nucleotides immediately adjacent to the mutation, i.e. the trinucleotide context. We explored the nucleotide context around passenger mutations and examined whether also flanking nucleotides beyond the trinucleotide context had a substantial impact on the local mutation probability. **a,** We observed that passenger mutations were enriched in characteristic nucleotide contexts, particularly for melanoma, endometrial and bladder cancer. Exemplary reads around mutations that contained these characteristic contexts (i.e. recurrent sequence motifs, yellow) are shown. This suggested that there was a substantial biological signal in the broad nucleotide context beyond the trinucleotide context. **b,** To quantify the biological signal in the extended nucleotide context, we examined the nucleotide contexts around mutations in samples with at least 500 mutations (n = 728 samples). For each sample, we counted how often we observed which nucleotide in the context of its mutations (green: A, blue: C, orange: G, red: T). Based on these counts, we derived the fraction of each nucleotide and position per sample. The distribution of these fractions across 728 samples is plotted in this figure. Dots indicate the medians of these distributions, and vertical bars extend to the 2.5%/97.5% quantiles (95% confidence interval). Furthermore, the orange envelope represents the joint distribution (nucleotide fractions of four nucleotides together) and similarly extends to the 2.5%/97.5% quantiles (95% confidence interval). We then examined whether this variation exceeded our expectation, assuming there was no biological signal in the extended nucleotide context. The gray envelope indicates the 0.1%/99.9% quantiles of a beta distribution with α=0.25x500 and β=(1-0.25)x500. The distribution median is indicated by a dashed line. This distribution reflects the expected variation of nucleotide fractions in samples with 500 mutations, assuming the four nucleotides occur with the same frequency at each position (25%). Hence, fractions outside of the gray envelope correspond to a one-sided p-value of $<10^{-3}$ based on the beta distribution. Our distributions indicate that a substantial number of samples deviated from the expected distribution within a ±4-nucleotide window. This suggests that there is a relevant biological signal in the broad nucleotide context that is not captured by the trinucleotide context. **c-e,** These plots supplement Figure 1b. The nucleotide context around passenger mutations is visualized for three cancer types with high average background mutation rates for three cancer types (bladder, n = 317; endometrium, n = 327; skin, n = 582). In brief, we counted how often we observed each nucleotide around non-recurrent mutations (±8 nucleotides). Based on these counts, we then determined the relative nucleotide frequencies in the nucleotide context around passenger mutations (y-axis). These plots suggest that the trend that we observed in Figure 1b also continued for non-recurrent mutations.



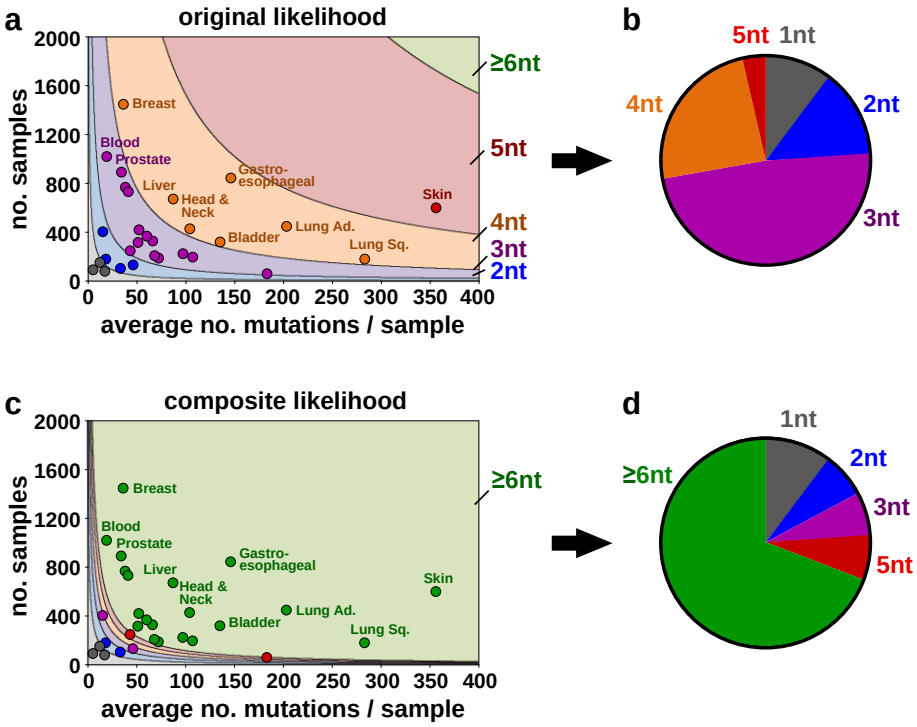
Supplementary Figure 4 | *CG* suppression in the human exome.
In concordance with previous studies, we noticed that *CG* dinucleotides were markedly underrepresented in the human exome, compared with the other 15 possible dinucleotides (red). This phenomenon is commonly referred to as CG suppression in the literature. We set out to examine the relative frequency of the *CG* dinucleotide in the reference genomes of different species (n = 98 species, 7 species groups, x-axis). Each black dot represents the relative frequency of the *CG* motif in an individual species, multiplied by 16 to normalize against the uniform distribution (y-axis). The black line connects the species group averages, and the gray envelope depicts the standard deviation for each group. The p-value was derived based on a one-way ANOVA F-test statistic (degrees of freedom: df1=6, df2=91, F-value=136.1). These analyses suggested that the marked *CG* suppression of the human exome had gradually evolved over the evolution. This *CG* suppression needs to be considered in our composite likelihood model.



Supplementary Figure 5 | Characterization and modeling of the nucleotide sequence composition of the human exome. A major prerequisite for the development of the composite likelihood model was an accurate characterization of the nucleotide sequence composition of the human exome. In other words, we aimed to develop a model that determined how often each possible nucleotide sequence occurred in the human exome. In **a-e** we demonstrate that the sequence composition of the human exome is non-trivial, i.e. that different sequences occur at different frequencies. For instance, the human exome is characterized by a marked underrepresentation of *CG* dinucleotides, commonly referred to as *CG* suppression in the literature. In **f-g** we develop a Markov model to model the sequence composition of the human exome.

a, Schematic depicting the relationship between mutational likelihood and mutational probability. For each possible nucleotide sequence, the composite likelihood model describes whether it is over- or underrepresented around passenger mutations compared with its occurrence in the human exome. In other words, for each possible nucleotide context, the composite likelihood delivers a frequency ratio, where a ratio > 1 indicates that the nucleotide context has a higher frequency around passenger mutations compared with its frequency in the human exome (the "affinity" of the mutational process to a nucleotide sequence). **b,** *CG* dinucleotide motifs are underrepresented in the human exome. We determined the relative frequency of each possible nucleotide sequence (x-axis: 3 to 9 nucleotides, nt) in the human exome (n = 8.4x10^7 genomic positions examined). We normalized the frequencies to the total number of sequences of the same length (y-axis: x-fold change relative to a uniform distribution). For each sequence length l, we plotted the distribution of these normalized ratios (n = 4^l motifs for length l) : medians are represented by dots, and vertical lines extend to the 5%/95% quantiles. Ratios fluctuated around 1 for motifs without *CG* (blue), irrespective of the sequence length. Nucleotide sequences that contained at least one *CG* dinucleotide (red) were significantly and substantially underrepresented, a phenomenon reported in previous studies. Hence, nucleotide sequences are not equally represented in the human exome and the *CG* underrepresentation had to be considered in the composite likelihood model. The p-value in the figure reflects the maximum p-value across comparisons between motifs with and without *CG* of lengths 3 to 9 nucleotides based on a two-tailed Welch's t-test. In detail, comparison between motifs with (n_+ motifs) and without *CG* (n_- motifs) yielded the following results (p-value, p; t-value, t; degrees of freedom, df) based on a two-tailed Welch's t-test. Length 3: n_+=56, n_-=8, p=1.8x10^{-12}, t=13.4, df=23.5; Length 4: n_+=209,

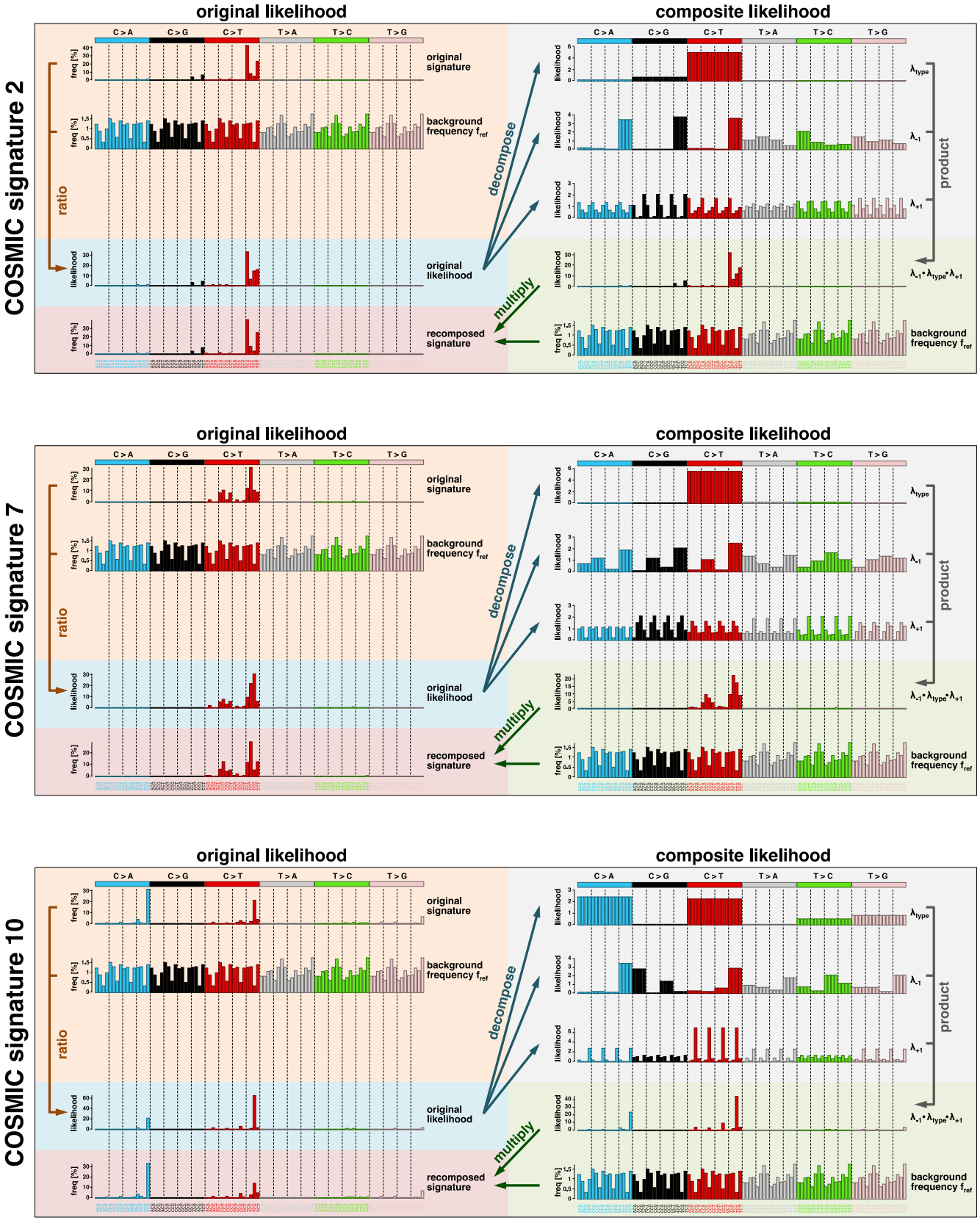
n_{-}=47, p=3.8x10^{-49}, t=21.3, df=164.6; Length 5: n_{+}=780, n_{-}=244, p=2.1x10^{-180}, t=36.6, df=897.8; Length 6: n_{+}=2911, n_{-}=1185, p=4.5x10^{-628}, t=64.7, df=4067.7; Length 7: n_{+}=10864, n_{-}=5520, p=2.0x10^{-2023}, t=112.3, df=16071; Length 8: n_{+}=40545, n_{-}=24991, p=7.3x10^{-6283}, t=193.2, df=59071; Length 9: n_{+}=151316, n_{-}=110828, p=8.9x10^{-18243}, t=321.6, df=208710. **c,** We determined the relative representation of the 16 dinucleotide motifs (y-axis) in the genomes of 98 species from 7 different species groups (x-axis). Frequencies were normalized to the total number of dinucleotide motifs, so that a frequency ratio of 1 reflected the uniform distribution. The frequency ratio of each species is represented by a dot. Envelopes (light gray and turquoise) were derived based on a Gaussian kernel (n = 98 data points for each dinucleotide motif). While there was a bimodal distribution pattern for the *CG* dinucleotide motif across species, the frequency ratios closely varied around 1 for the other dinucleotide motifs. The bimodal distribution for *CG* was clearly separated between different species groups, suggesting that this motif was lost during evolution. **d,** Based on the Exome Aggregation Consortium (ExAC) project, which contains whole-exome sequencing data from 60,706 unrelated individuals, we characterized the nucleotide context around common single nucleotide variations from the reference genome. We determined the trinucleotide context and substitution type for each of these common variants. Their genomic distribution clearly resembled the distribution of mutations in the ageing mutation signature (C>T mutations in a lagging G context) and its complementary signature (T>C mutations in a lagging G context). This suggests that the loss of *CG* dinucleotides during evolution may be due to a similar mechanism to the ageing signature (spontaneous cytosine deamination in hypermethylated CpG islands). **e,** We compared the mutation frequency of trinucleotide contexts that contained a *CG* motif (yellow dots, n = 12 contexts) against those that did not contain a *CG* motif (blue dots, n = 84 contexts) for COSMIC mutation signatures. Significance values were derived from a two-tailed Welch's t-test. In detail, the test yielded the following results (p-value, p; t-value, t; degrees of freedom, df). Signature 3: p=6.1x10^{-20}, t=12.6, df=72.4; signature 4: p=2.1x10^{-1}, t=1.3, df=26.0; signature 5: p=7.4x10^{-3}, t=3.1, df=15.9; signature 8: p=2.5x10^{-3}, t=3.5, df=19.7; signature 9: p=1.8x10^{-5}, t=5.0, df=33.7; signature 11: p=3.3x10^{-2}, t=2.2, df=93.7; signature 16: p=2.6x10^{-12}, t=8.0, df=93.7; signature 23: p=5.8x10^{-2}, t=1.9, df=92.8; signature 25: p=1.6x10^{-2}, t=2.7, df=18.1; signature 30: p=2.4x10^{-1}, t=1.2, df=38.7. **f-g,** We developed a Markov chain model to characterize the sequence composition of the human exome (Supplementary Note). Markov models assume that there is a constant n for which genomic positions with a distance larger than n do not interfere with each other's distributions (Markov order n). The major parameter of our Markov model is its maximal order, reflecting its context dependency. Data in (**f**) and (**g**) are based on the frequencies of all 4096 possible nucleotide sequences of length 6 in the human exome. We tested the accuracy of Markov models with increasing orders (left to right) to predict the frequency of 6-nucleotide sequences in the reference genome. **f,** Q-Q-plots evaluate the calibration of our model for the sequence composition of the reference genome. The quantiles of the observed distribution (y-axis) are plotted against the quantiles of the expected distribution (x-axis) based on the Markov model (n = 4096). A close fit of the Q-Q plot to the diagonal (inflation factor λ close to 1) suggests that the model is accurately calibrated. **g,** In parallel to the Q-Q-plots, we compared the observed and modeled frequency ratios (frequencies normalized against a uniform distribution) of nucleotide sequences in the human exome directly. Each dot represents an individual 6-nucleotide sequence (n = 4096). Expected frequency ratios (x-axis) are plotted against observed ratios (y-axis). Dot density is represented by dot colors, ranging from dark red (low density) to yellow (high density). In both analyses in (**f**) and (**g**), a Markov chain order of 2 nucleotides (16 Markov states) delivered the most precise model (orange). This finding suggests that underrepresentation of the *CG* motif was the main characteristic of the human exome, which had to be incorporated into the Markov model to capture the sequence composition of the human exome.



Supplementary Figure 6 | A new mathematical model is needed to characterize the broad nucleotide context around passenger mutations.

The analyses shown in this figure are based on 11,873 samples. The number of samples per cancer type can be found in Extended Data Figure 5. **a-b,** Trinucleotide contexts (i.e., the 5' and 3' nucleotides immediately adjacent around a mutation) are commonly used to model context-dependent mutation probabilities. For this purpose, the mutation probabilities of all possible trinucleotide contexts are determined independently. However, as the number of flanking nucleotides increases, the number of possible nucleotide contexts grows exponentially. For instance, there are 96 possible trinucleotide contexts, but 24,576 possible 7-nucleotide contexts. As the number of possible contexts soon exceeds the number of mutations per tumor, the traditional approach (characterizing each possible context independently, original likelihood) is intrinsically limited to the trinucleotide contexts for most cancer types. **c-d,** The composite likelihood model integrates the effect each flanking nucleotide in the context around passenger mutations as a multiplicative factor. Hence, the degrees of freedom increase linearly with the number of flanking nucleotides included into the composite likelihood model. This allowed us to characterize a context of up to 20 nucleotides around passenger mutations. To determine the number of samples needed (y-axis) depending on the no. mutations per sample (x-axis), we assumed that a fixed number of observations (arbitrarily set to 100 mutations per degree of freedom in this figure) are required to characterize each parameter sufficiently.

a, c, For each cancer type, the average number mutations per sample (x-axis) is plotted against the number of samples included in our study cohort (y-axis). Dot colors indicate the number of nucleotides, which can be integrated using the original likelihood (**a**) or the composite likelihood (**c**), respectively. **b**, **d,** Pie charts show the number of cancer types grouped according to the length of the mutation context in nucleotides, using the original likelihood (**b**) or the composite likelihood (**d**), respectively.



(figure legend on next page)

Supplementary Figure 7 | Exemplary calculation of the mutability scores returned by the composite likelihood model. This figure explains the calculation of the composite likelihood score based on the trinucleotide context-specific mutation frequencies in three COSMIC mutation signatures. A more detailed explanation of the exact formulas underlying the composite likelihood model can be found in the Supplementary Note. The steps illustrated in this figures are as follows:

1.) We started from a COSMIC trinucleotide signature, which consists of 96 mutation probabilities, each representing specific trinucleotide context. Each mutation probability reflects the fraction of mutations that are surrounded by the specific nucleotide context. We then normalized these context-specific probabilities by their baseline occurrence in the human exome. In other words, for each nucleotide context we computed the ratio of (i) the fraction of mutations surrounded by the nucleotide context to (ii) the fraction of all positions in the human exome surrounded by the same nucleotide context. That way, we obtained 96 mutational likelihoods. A likelihood of >1 indicates that the trinucleotide context is enriched around mutations, compared to its prevalence in the human exome. Our statistical model works with these likelihoods since they reflect the mutability of an individual genomic position; mutation probabilities reflect the mutability of a nucleotide context in aggregate.

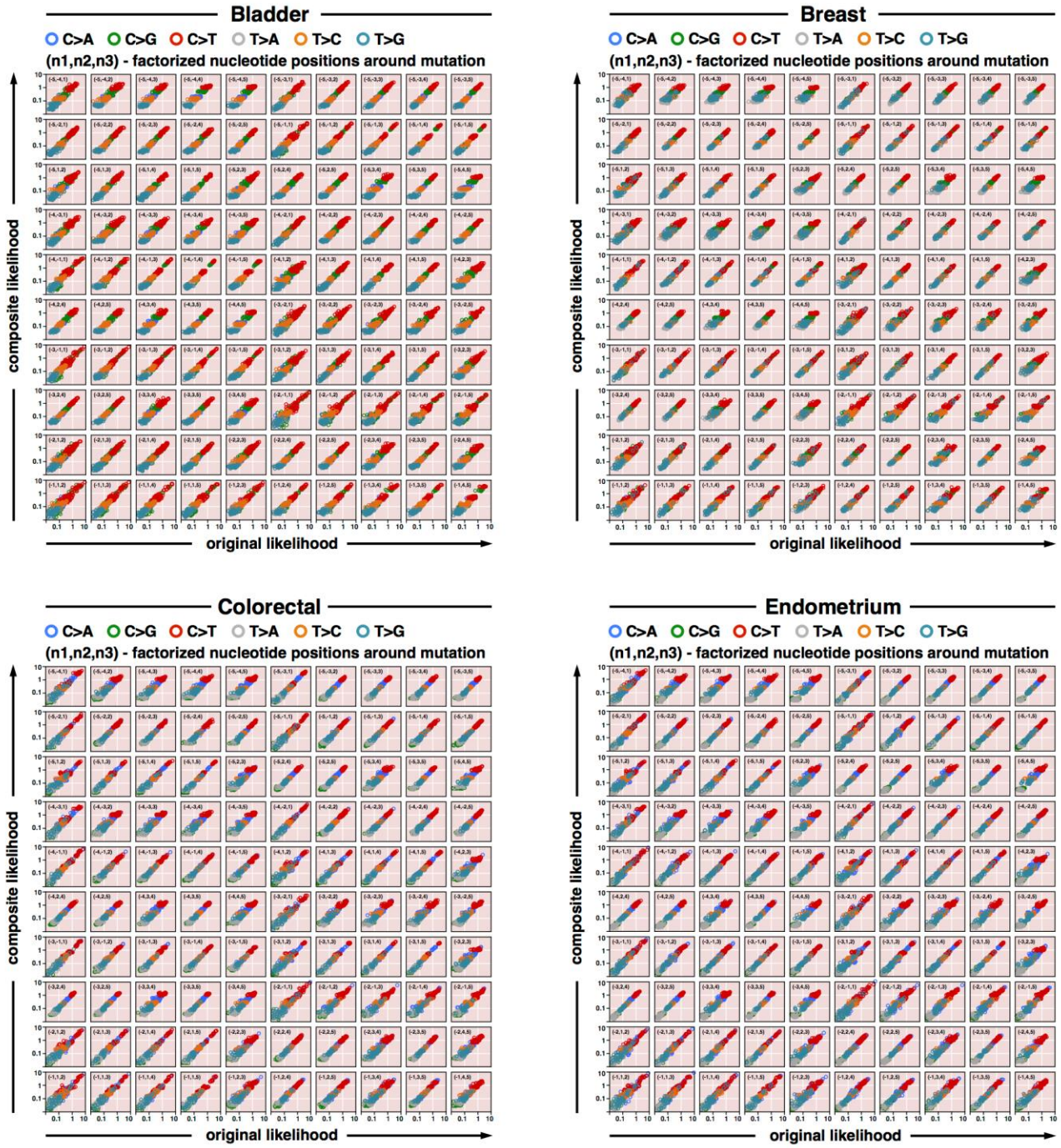
2.) In a second step, we decomposed the mutational likelihood of each trinucleotide context into multiplicative factors, representing the effects of the base substitution type (λ_{type}), the flanking 5' nucleotide (λ_{-1}) and the flanking 3' nucleotide (λ_{+1}), respectively. For broader nucleotide contexts, we obtained similar decompositions in which each factor represented a flanking nucleotide. Details on this decomposition are described in the Methods.

3.) To determine the mutability of an individual position in the human exome, we computed the product of the likelihood factors associated with its surrounding nucleotide context. That way, we obtained the likelihood score returned by the composite likelihood model for each genomic position, or - in this figure - possible trinucleotide context.

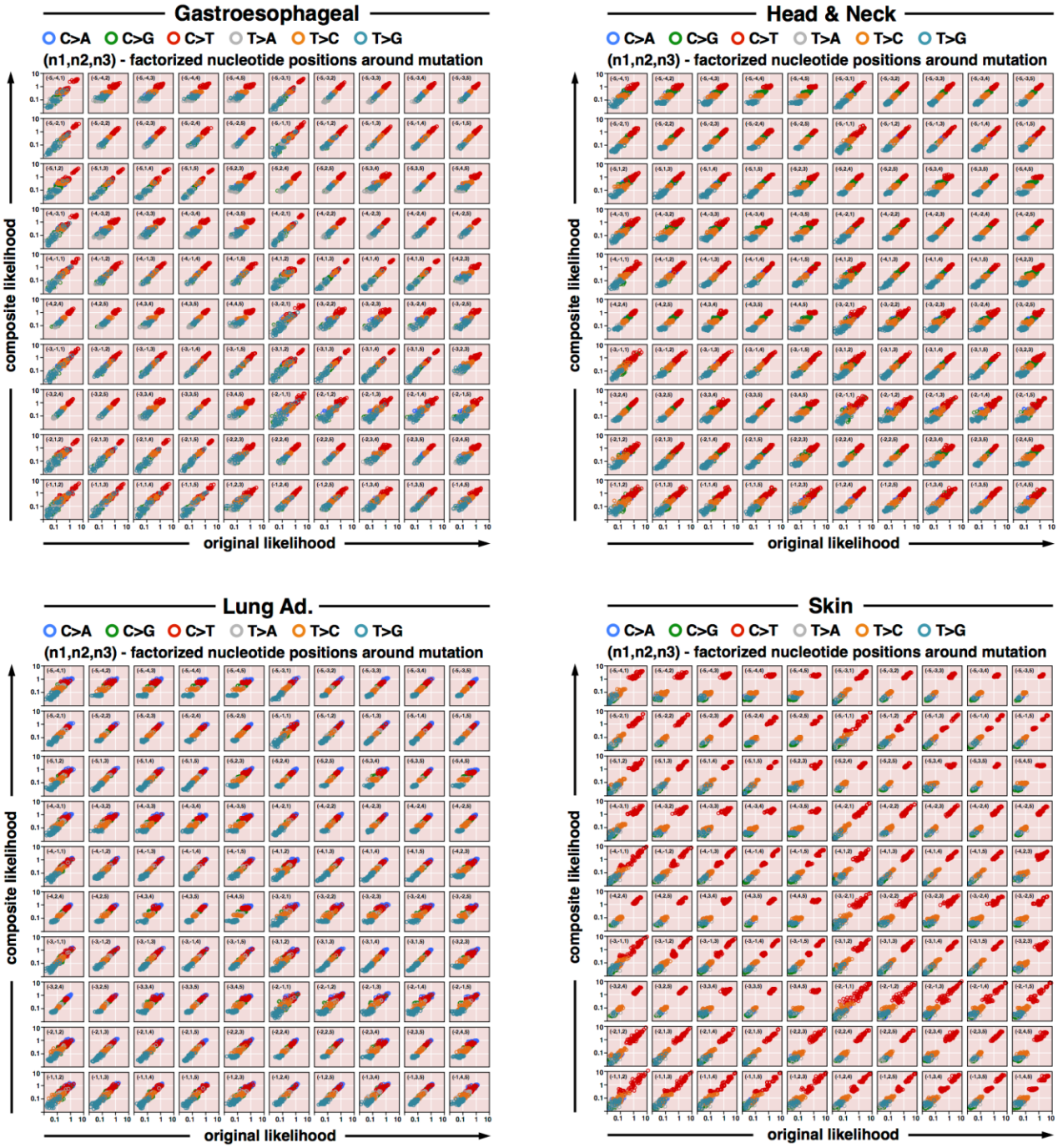
4.) To compare composite likelihoods with mutation probabilities, we multiplied each likelihood with its associated baseline frequency in the human exome. We note that this step is not part of our model and only serves for the comparison with the original COSMIC mutation signature in Figure 1d.

By employing a composite likelihood model, we followed two goals: reduction of parameters and avoiding of overfitting: The number of likelihood factors increases linearly with the flanking nucleotides, whereas the number of possible nucleotide contexts increases exponentially. Hence, modeling of all possible nucleotide contexts individually would result into sparsity of mutations per nucleotide context. Due to its lower number of parameters, the composite likelihood model enabled us to consider broader nucleotide contexts. Further, determining the mutational likelihood of each nucleotide context individually would likely result into overfitting of the background signal. For instance, a highly recurrent hotspot would strongly increase the mutational probability of a single and highly specific nucleotide context. Overfitting of this nucleotide context would annihilate the signal, mediated by unusual nucleotide contexts, for the hotspot. The composite model decomposes the likelihood of each nucleotide context into multiplicative factors, rather than computing the mutation probability of each possible nucleotide context individually. Hence, recurrent mutations in a hotspot do not increase the composite likelihood score of a single nucleotide context, thus preventing our model from signal annihilation due to overfitting of the background signal.

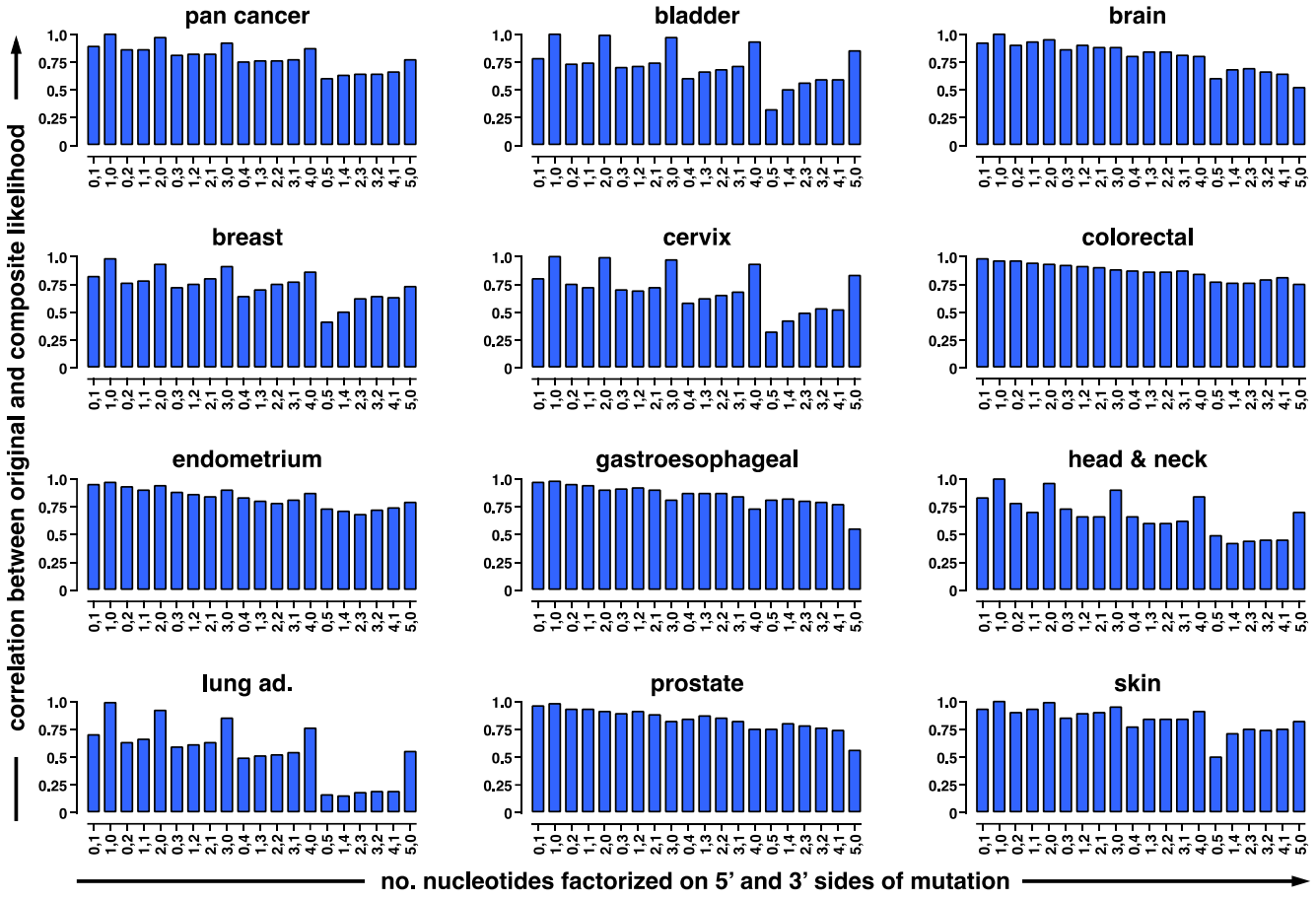
The basic assumption of the composite likelihood model is that each flanking nucleotide either increases or decreases the mutation probability of a position. Hence, the composite likelihood model does not necessarily assume symmetry of the underlying mutation signature. A straightforward example is given for C>T mutations in COSMIC signature 10 (POLE signature). All 5' nucleotides except for T and all 3' nucleotides except for G will strongly decrease the mutational likelihood. As a result, only C>T mutations in a T_G context have a substantial mutational likelihood. An example for a symmetric signature is given by C>T mutations in COSMIC signature 7 (UV signature). Here, both C and T in 5' positions increase the mutational likelihood and the mutational likelihood is similarly high for all possible 3' nucleotides, thus leading to the highly symmetrical shape of this signature. That way, a large variety of different signatures can be captured using the composite likelihood model.



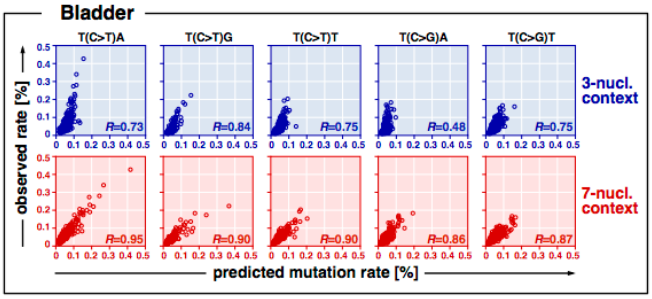
Supplementary Figure 8 | Modeling of context-specific mutation probabilities based on broad nucleotide contexts.
To test the independence assumption of the composite likelihood, we examined the interaction between any three positions in the 11-nucleotide context for four cancer types (bladder, *n* = 317 samples; breast, *n* = 1443; colorectal, *n* = 223; endometrium, *n* = 327). For each triplet (100 possible combinations, at least 1 position on 5' and 3' side), there are 384 possible nucleotide contexts (64 nucleotide combinations, 6 substitution types), and we plotted the observed mutation count of each nucleotide context (x-axis) against the prediction of the composite likelihood model (y-axis). The results of each position pair are visualized in a separate correlation plot, and the positions of the pair are annotated on the bottom right of the correlation plot. Dot colors indicate the base substitution types. For any three nucleotide positions, the Pearson correlation coefficient between observed and predicted data around the diagonal served as a measure of their independence. The same analysis for additional cancer types can be found in Supplementary Figure 9.

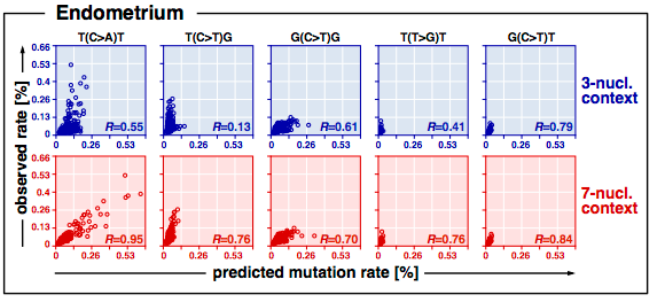


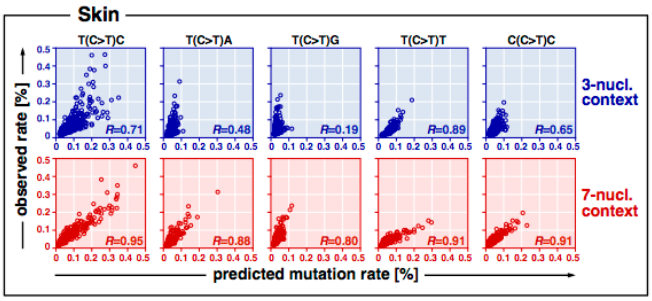
Supplementary Figure 9 | Modeling of context-specific mutation probabilities based on broad nucleotide contexts. To test the independence assumption of the composite likelihood, we examined the interaction between any three positions in the 11-nucleotide context for four cancer types (gastroesophageal, n = 833 samples; head and neck, n = 425; lung adeno, n = 446; skin, n = 582). For each triplet (100 possible combinations, at least 1 position on 5' and 3' side), there are 384 possible nucleotide contexts (64 nucleotide combinations, 6 substitution types), and we plotted the observed mutation count of each nucleotide context (x-axis) against the prediction of the composite likelihood model (y-axis). The results of each position pair are visualized in a separate correlation plot, and the positions of the pair are annotated on the bottom right of the correlation plot. Dot colors indicate the base substitution types. For any three nucleotide positions, the Pearson correlation coefficient between observed and predicted data around the diagonal served as a measure of their independence. The same analysis for additional cancer types can be found in Supplementary Figure 8.

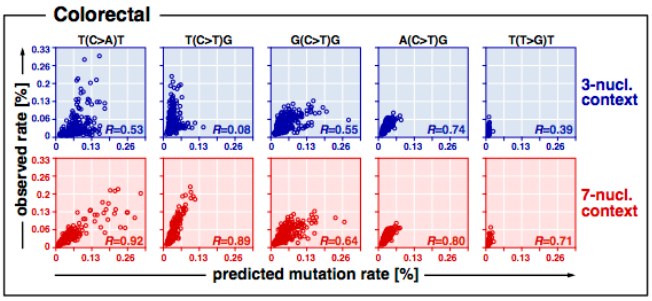


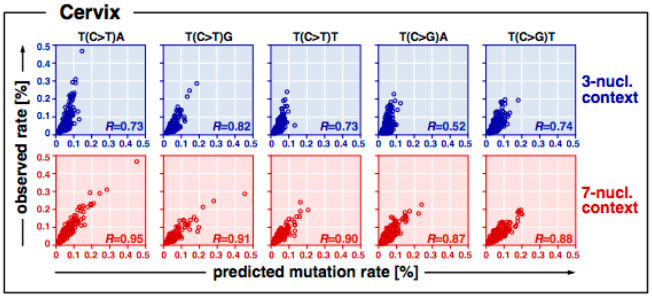
Supplementary Figure 10 | Correlation analysis of the composite likelihood model on long nucleotide contexts.
To examine whether the composite likelihood model generalized to long nucleotide contexts, we tested it on nucleotide contexts of 20 different lengths. These nucleotide lengths are denoted as length pairs (n_1, n_2) on the x-axis, where n_1 and n_2 reflect the number of nucleotides on the 5' and 3' sides of the mutation that were included in the model. For instance, (1,1) denotes the traditional trinucleotide context; (3,2) denotes a mutation context with 3 nucleotides on the 5' side, and 2 nucleotides on the 3' side.
For each context length pair (x-axis), we determined the Pearson correlation coefficient (y-axis) between observed mutation counts and the mutation probabilities returned by the composite likelihood model (y-axis) based on sequencing data of 11 cancer types (bladder, n = 317 samples; brain, n = 760; breast, n = 1443; cervix, n = 192; colorectal, n = 223; endometrium, n = 327; gastroesophageal, n = 833; head & neck, n = 425, lung adeno., n = 446; prostate, n = 880; skin, n = 582) as well as the full study cohort (pan cancer, n = 11873). The number of possible nucleotide contexts increases exponentially with the context length ($6 \cdot 4^{l-1}$ possible nucleotide contexts of context length l). Hence, the number of mutations per possible nucleotide context is lower for longer nucleotide contexts. This leads to data noise, which lowers the Pearson correlation coefficient between observed and predicted data for longer nucleotide contexts. However, the Pearson correlation coefficient was positive for long nucleotide contexts, despite data sparsity. In combination with the reduction in residual variance (cf. Extended Data Figure 4), this supports that the composite likelihood model robustly generalizes to longer nucleotide contexts.

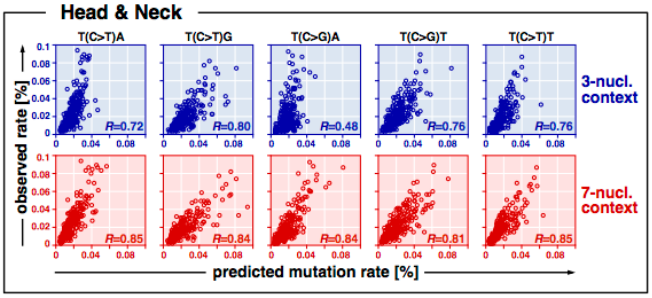


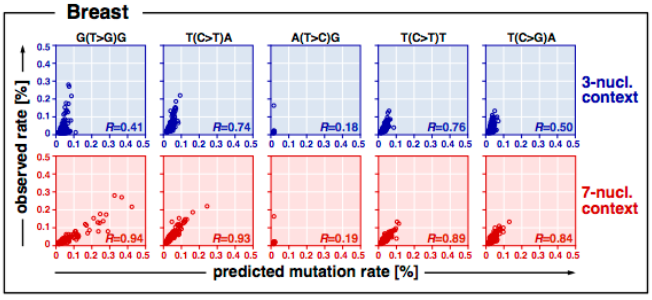


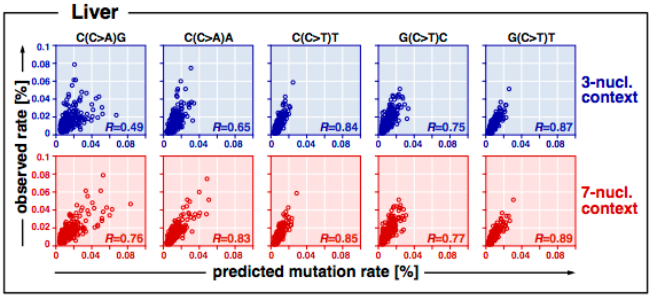


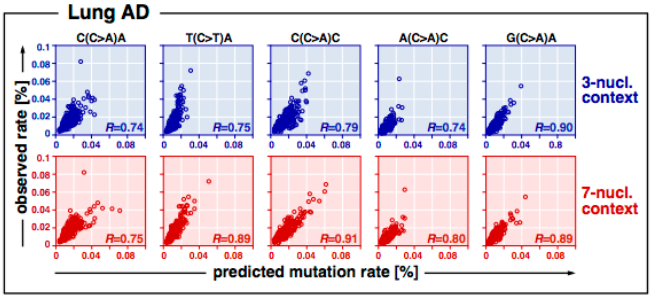


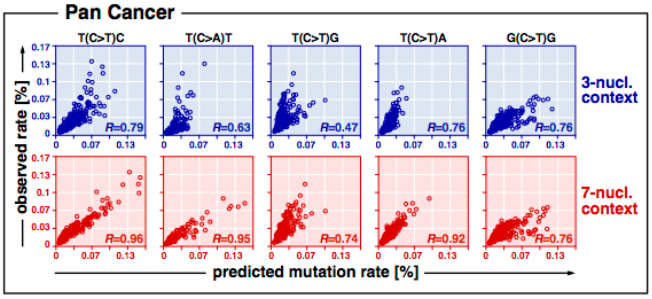










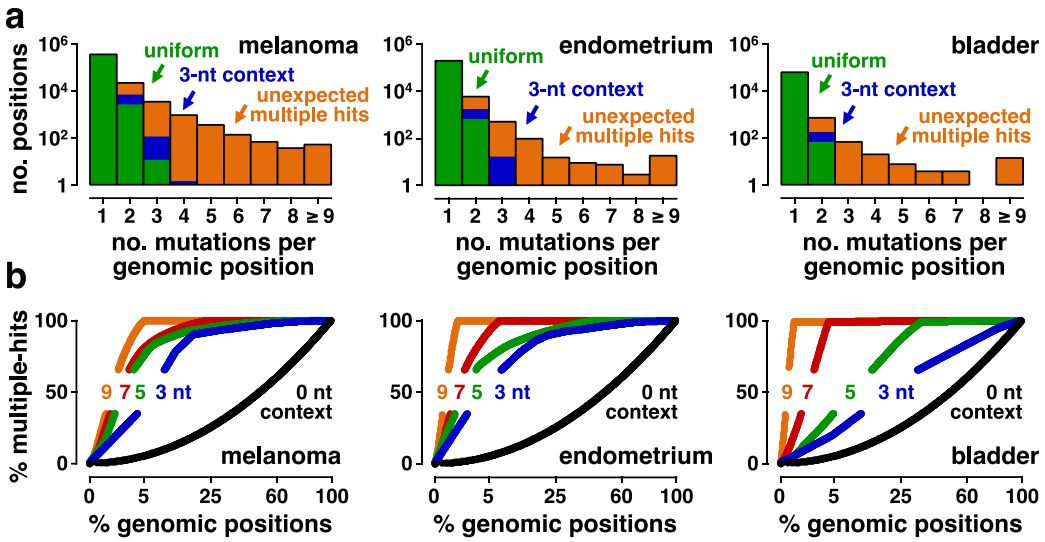


(figure legend on next page)

Supplementary Figure 11 | Quantification of the impact of flanking nucleotides outside the trinucleotide context on the likelihood score returned by the composite likelihood model.

Mutation probabilities of individual positions in the human exome are commonly modeled based on their trinucleotide contexts (i.e., the flanking 5' and 3' nucleotides). We used a composite likelihood model to quantify the mutation probability of each genomic position based on its broad nucleotide context. We examined whether integrating the effect of flanking nucleotides outside of the trinucleotide context refined the accuracy of the scores returned by the composite likelihood model. To this end, we counted for each possible 7-nucleotide context the number of mutations that fell into this specific 7-nucleotide context for 9 cancer types (bladder, n = 317 samples; breast, n = 1443; cervix, n = 192; colorectal, n = 223; endometrium, n = 327; head & neck, n = 425; liver, n = 650; lung adeno., n = 446; skin, n = 582) as well as the pan-cancer cohort (n = 11873). We then normalized these counts against the total number of mutations, resulting in the observed mutation rate of each possible 7-nucleotide context (y-axis). We further derived the expected mutation rate of each possible 7-nucleotide from the composite likelihood model using either the full 7-nucleotide information (red, bottom, x-axis) or using the information of the inner trinucleotide context only (blue, top, x-axis). We plotted the predicted mutation rates (x-axis) against the observed mutation rates (y-axis) for 7-nucleotide contexts with the same inner trinucleotide context (annotated on the top). Each dot in these plots represents an individual 7-nucleotide context. The Pearson correlation coefficient (R) of each plot is annotated on the bottom right. In contrast to the plots shown in Extended Data Figure 1c, we compared the raw mutation counts, i.e. we did not normalize the raw mutation counts to their representation in the human reference exome.

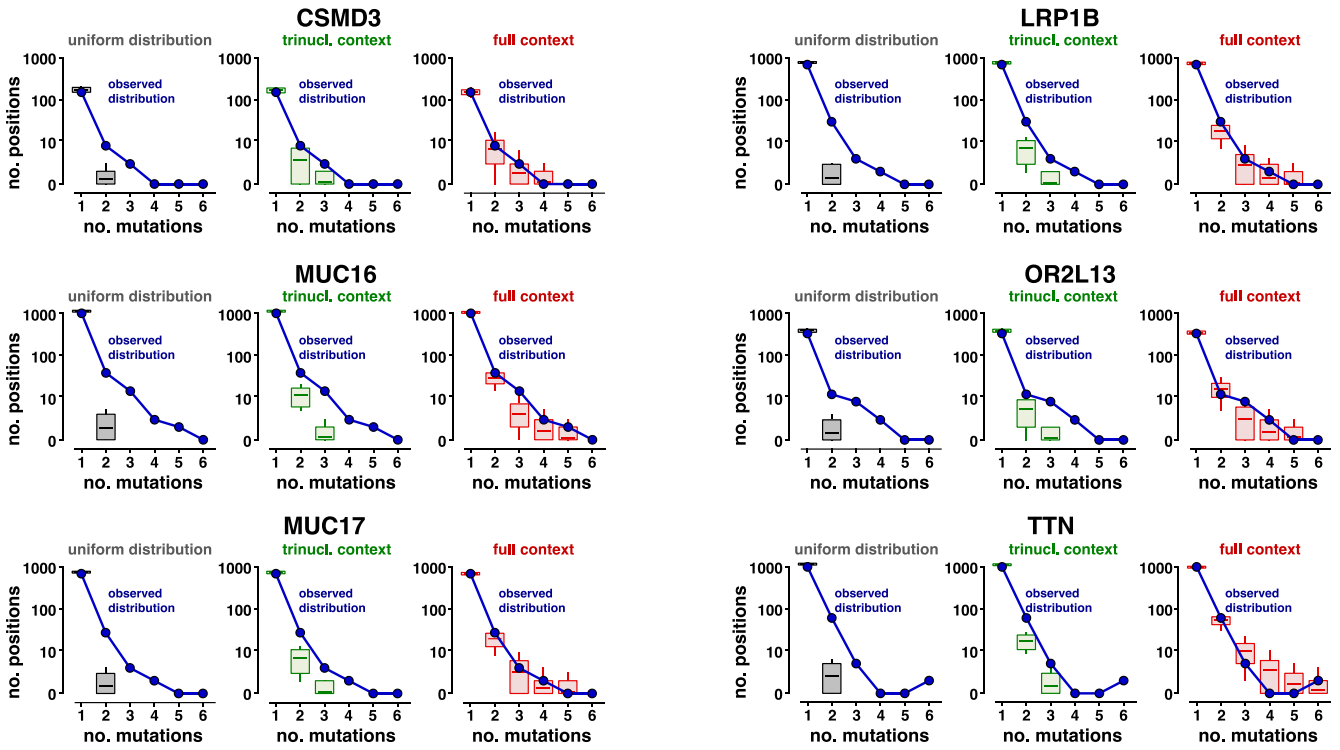
We noticed that several of the trinucleotide correlation plots (blue) displayed more than one correlated groups of dots. For instance, plots in the upper rows of skin, endometrium and colorectal cancer display this phenomenon. Incorporating the nucleotides outside of the trinucleotide sequence context into the model accounts for this phenomenon (red plots), resulting a higher correlation between observed and predicted mutation rates. This observation reflects the impact of nucleotides outside of the trinucleotide sequence context on the raw mutation counts of individual genomic positions (blue plots). We thus concluded that incorporating the flanking nucleotides outside of the trinucleotide context into the composite likelihood model refined our approximation of local mutation probabilities.



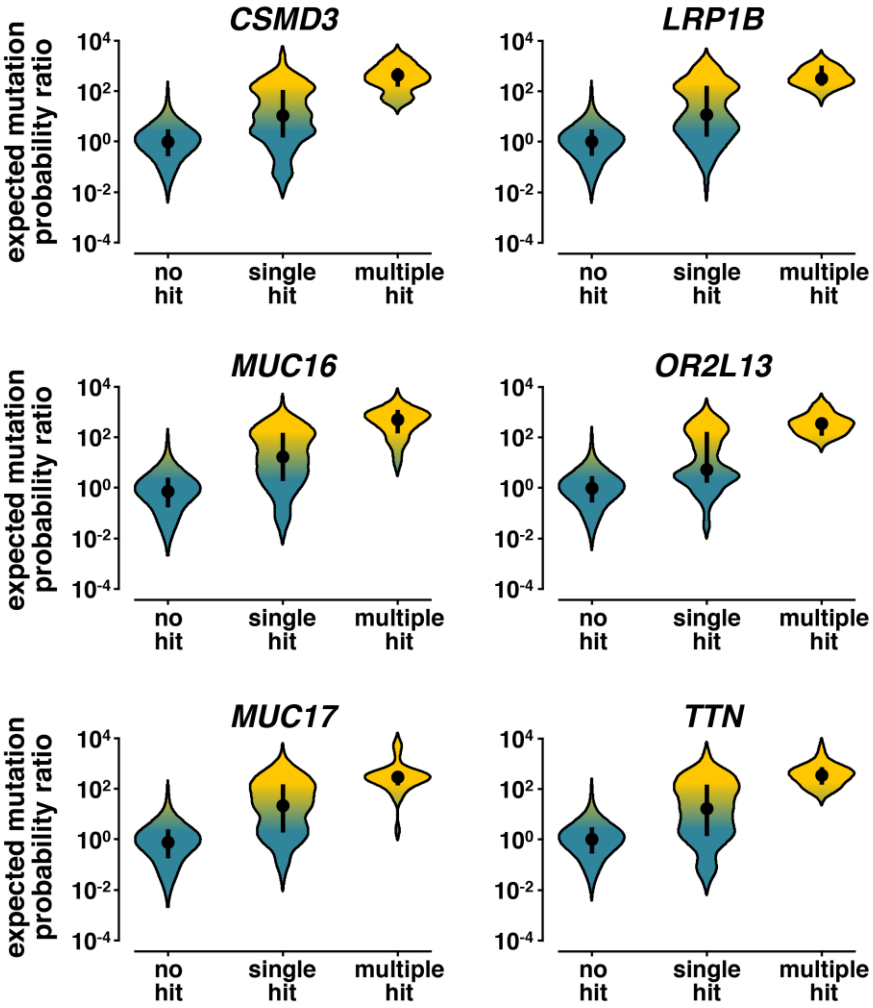
Supplementary Figure 12 | Human cancer exomes contain a large number of recurrent mutations that can be insufficiently explained based on their surrounding trinucleotide contexts.

Data in (**a**) and (**b**) relate to the genomic distribution of recurrent mutations in 582 melanomas (left), 327 endometrial carcinomas (middle) and 317 bladder cancers (right). **a,** We counted the number of genomic positions with recurrent mutations. These positions contain either passenger mutations in highly mutable nucleotide contexts or driver mutations in mutational hotspot positions under positive selection. The histograms display the number of genomic positions in the exome, containing 1, 2, … , 8 or ≥9 mutations, respectively (orange: observed histogram; green: histogram expected based on a uniform distribution; blue: histogram, obtained by redistributing mutations conditional on their trinucleotide context). The histograms suggest that trinucleotide contexts alone do not fully recapitulate the large number of positions with recurrent mutations in human cancer exomes (orange). Since the number of unexplained recurrent mutations exceeds the number of known mutational cancer hotspots, trinucleotide contexts are unable to explain the context-dependent distribution of passenger mutations completely.

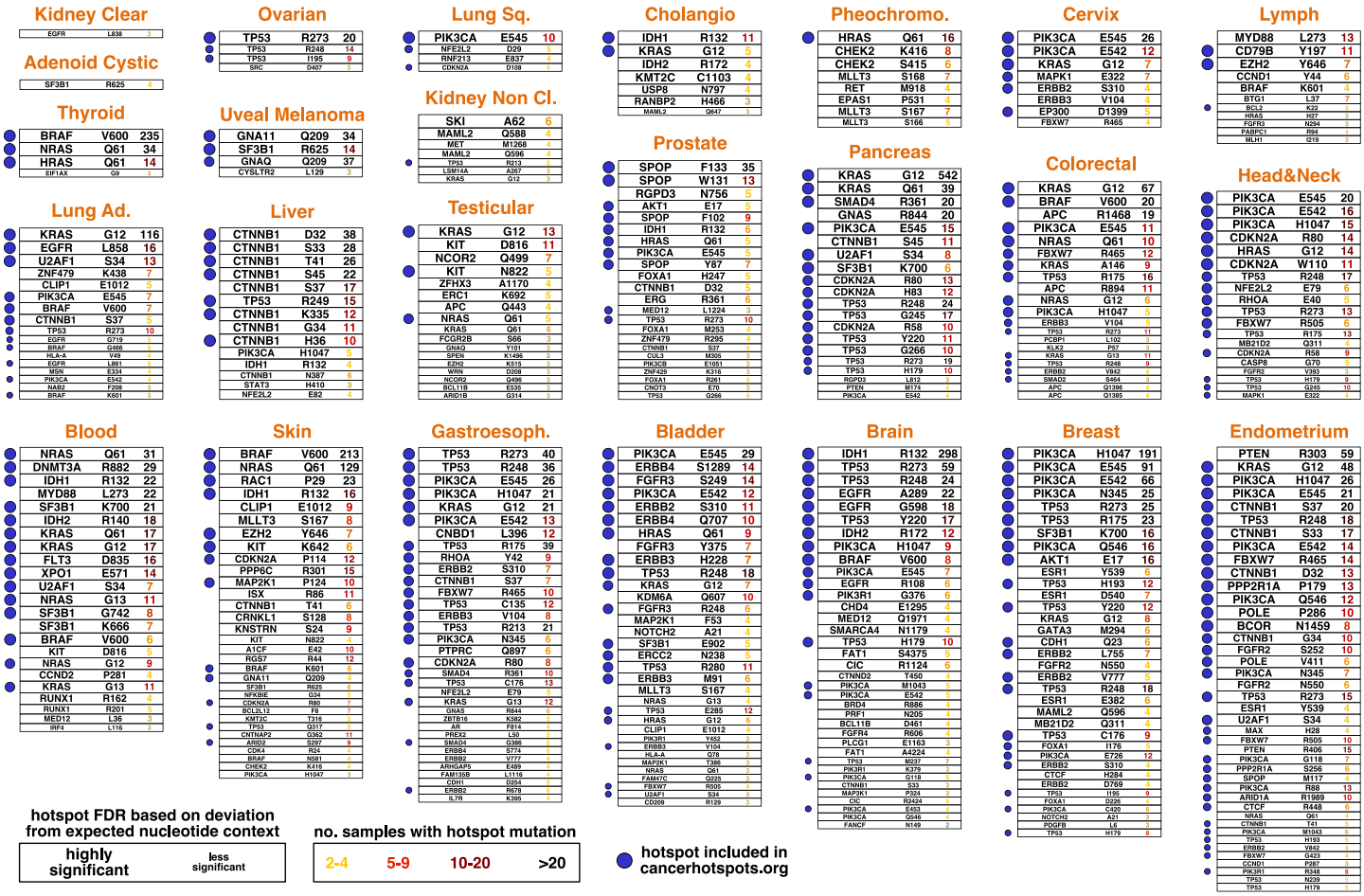
b, We asked whether integration of additional flanking nucleotides outside of the trinucleotide context might help understand the context-dependent distribution of passenger mutations across the human exome. We determined the probability of each genomic position in the exome of being hit by recurrent mutations, based on the number of nucleotides (nt) incorporated into the surrounding sequence context. We then sorted the genomic positions in descending order of those probabilities (x-axis, logarithmic). The genomic positions are plotted against the cumulative fraction of multiple hit positions (y-axis, positions which contain more than one mutation). A uniform distribution of mutations served as a negative control (black). These data suggest that considering additional nucleotides in the surrounding context improves the model of the context-dependent distribution of passenger mutations, which is an important prerequisite to accurately distinguish between accumulations of passenger mutations in highly mutable nucleotide contexts from mutational hotspots under positive selection.



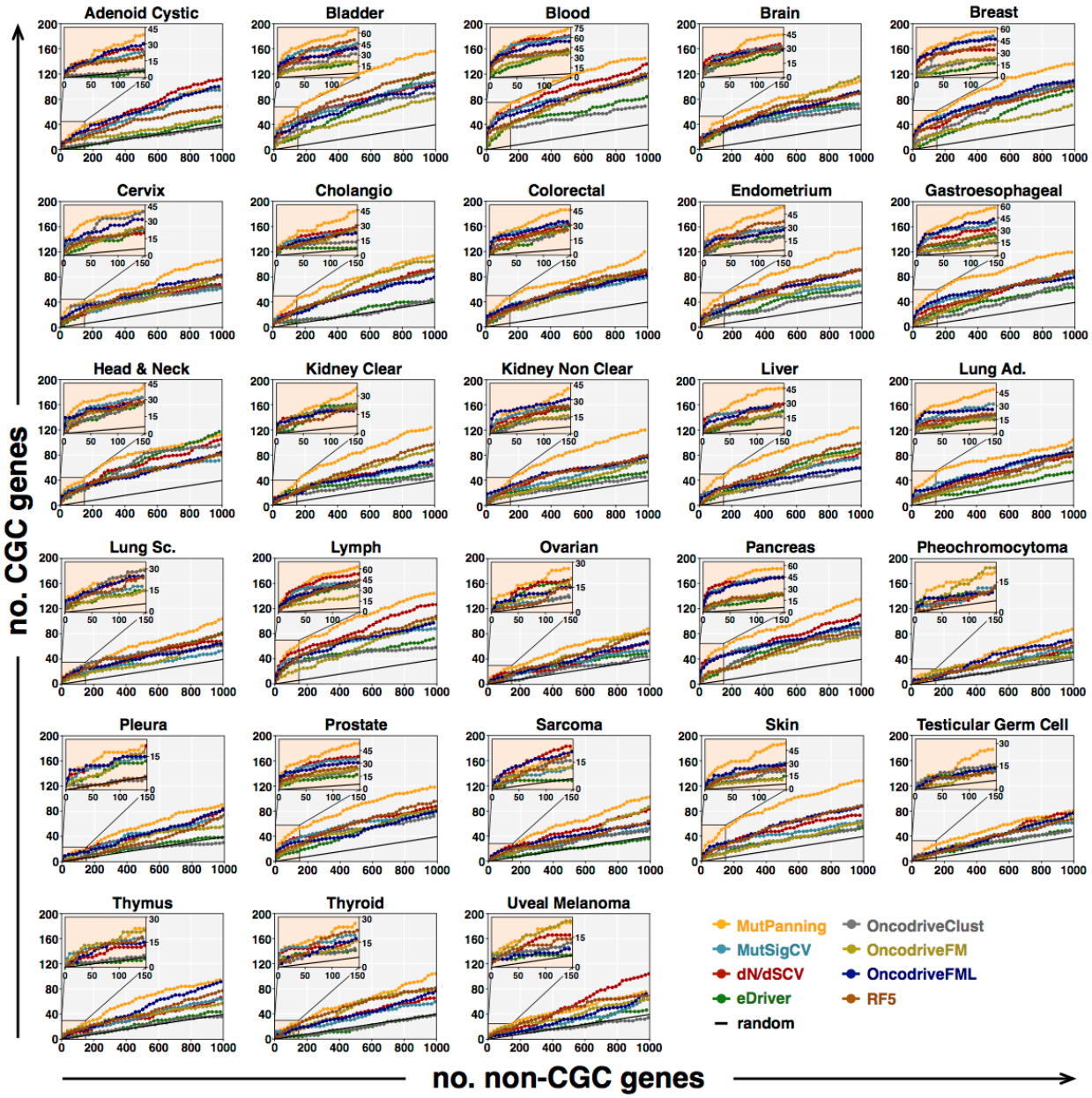
Supplementary Figure 13 | Modeling the distribution of passenger mutations based on nucleotide context.
A major prerequisite for the discovery of novel cancer genes was our model sufficiently calibrated to the background distribution of passenger mutations across the human exome. We examined whether our model was accurately calibrated to the distribution of passenger mutations in non-cancer-related genes that had been reported as false-positive findings in previous studies (*CSMD3*, *MUC16*, *MUC17*, *LRP1B*, *OR2L13*, *TTN*). To this end, we counted the number of positions that were hit by recurrent mutations in 582 melanoma samples (x-axis, blue). We then randomly re-distributed mutations in these genes (10,000 iterations), assuming a uniform distribution (gray), trinucleotide-specific mutation probabilities (green), or the likelihood scores returned by the composite likelihood model (red). In each iteration, we counted the number of positions containing 1, … , 5, and ≥6 mutations, respectively. The distributions of these simulated counts are represented as box plots. Boxes mark the interquartile range; vertical lines denote the 5%-95% percentile range; horizontal lines mark the group medians. These experiments revealed that the uniform model (gray) and the trinucleotide context model (green) systematically underestimated the number of positions that were hit by recurrent mutations, thereby potentially leading to false-positive findings. Considering the broad nucleotide context through the composite likelihood model (red) provided a closer approximation of the distribution of recurrent mutations in these non-cancer-associated genes.



Supplementary Figure 14 | Accounting for the broad nucleotide context identifies genomic positions in human cancer that are prone to accumulate recurrent passenger mutations.
We asked whether considering the broad nucleotide context informed the discovery of mutational hotspots in cancer. Critically, accumulations of passenger mutations in positions surrounded by a highly mutable nucleotide context need to be distinguished from mutational hotspots under positive selection. For this purpose, we examined the distribution of the mutation probability ratios derived from our composite likelihood model in six non-cancer-related genes with high background mutation rates (*CSMD3*, *LRP1B*, *MUC16*, *OR2L13*, *MUC17*, *TTN*) in 582 melanoma samples. Violin plots display the distribution of the mutational likelihood in individual positions, depending on whether they contained no mutations (no hit), a single mutation (single hit) or recurrent mutations (multiple hit). The distribution median is indicated by a dot and vertical lines extend to the 25%/75% quantiles. Envelopes are based on a Gaussian kernel. A mutational likelihood >1 indicates that the position is expected to contain more mutations than based on a uniform distribution. Probability ratios of the non-mutant positions stably varied around 1, and there was ~100× increase in the mutation probability ratios for multiple-hit positions. Positions with a single mutation displayed a bimodal distribution in between. This finding suggests that our context-dependent composite likelihood model accurately identifies positions, which have a high probability of accumulating recurrent passenger mutations. Thus, considering the broad nucleotide context informs discovery of mutational cancer hotspots and cancer driver mutations.



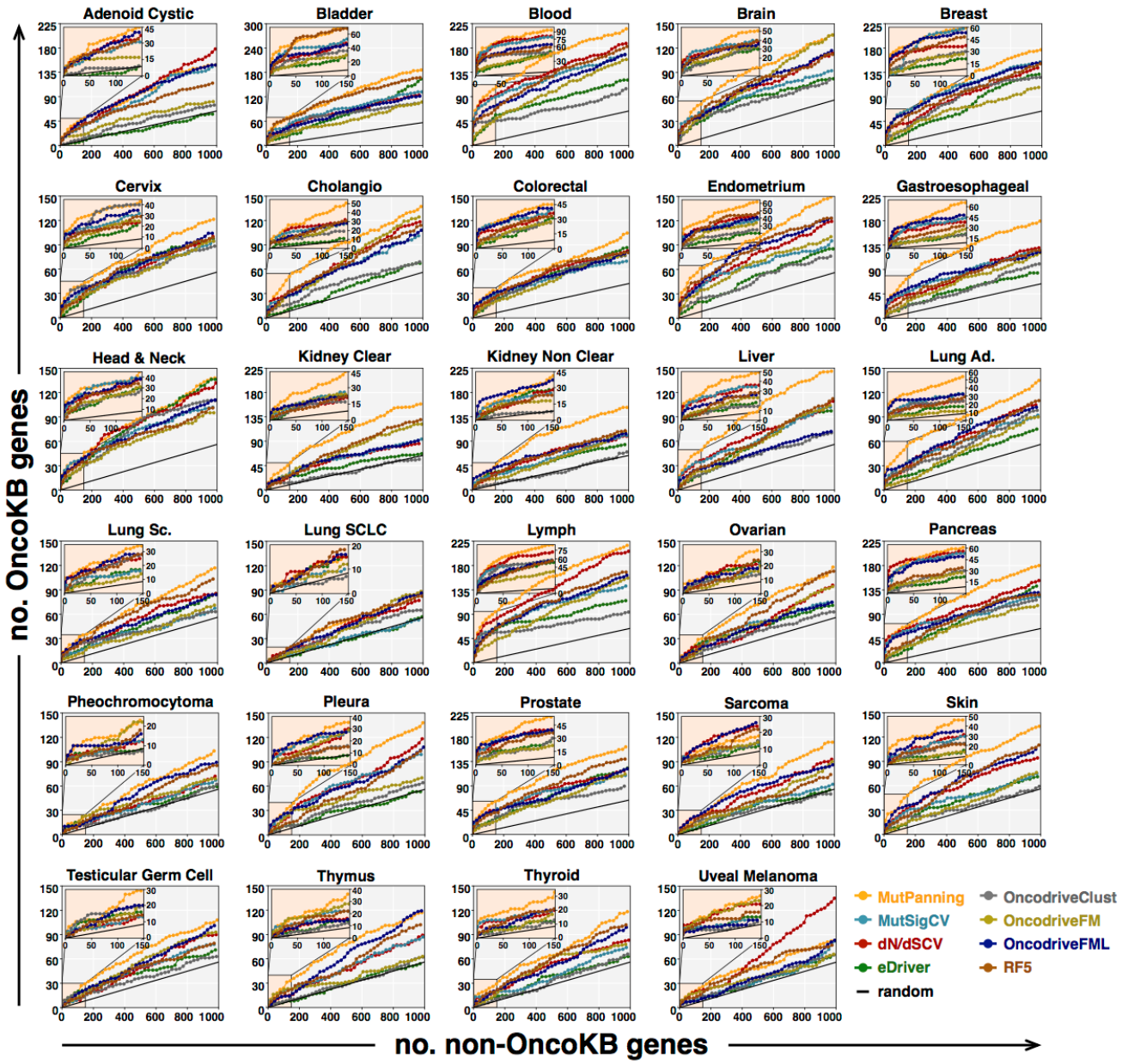
Supplementary Figure 15 | Identification of mutational hotspots in cancer based on nucleotide contexts.
We examined the feasibility of nucleotide contexts to distinguish recurrent passenger mutations in highly mutable nucleotide contexts from recurrent driver mutations in mutational hotspots in Cancer Gene Census genes. In total, these analyses are based on sequencing data of 11,873 samples. The exact number of samples included in this analysis per cancer type can be found in Extended Data Figure 5. In brief, we determined for each position with recurrent mutations whether the mutation count exceeded our expectation based on its surrounding nucleotide context. To this end, we determined the composite likelihood for each genomic position and summed up the likelihood scores across positions for each codon. We normalized these likelihood sums to the total sum of likelihood scores in the same gene. That way, we derived a mutation probability p for each codon in a gene. Given a gene with N mutations, we determined for each codon with probability p and n mutations a p-value based on a one-sided binomial test (Binom (N, p), which reflects the probability of observing n or more mutations based on a binomial distribution). Similar to MutPanning, these p-values were then calibrated to a uniform distribution using the Brown method and then adjusted for multiple testing using the Benjamini-Hochberg procedure (false-discovery rate, FDR). Similar to MutPanning we used an FDR threshold of 0.25. Positions are listed in descending order according to their number of mutations. Further, we annotated for each significant position the gene name as well as the amino acid residue. Font sizes indicate hotspot significance values; font colors reflect mutation counts. Blue dots indicate whether the hotspot could be confirmed in the same cancer type based the data on cancerhotspots.org (inclusion criteria: $q < 0.1$, $n \geq 5$).



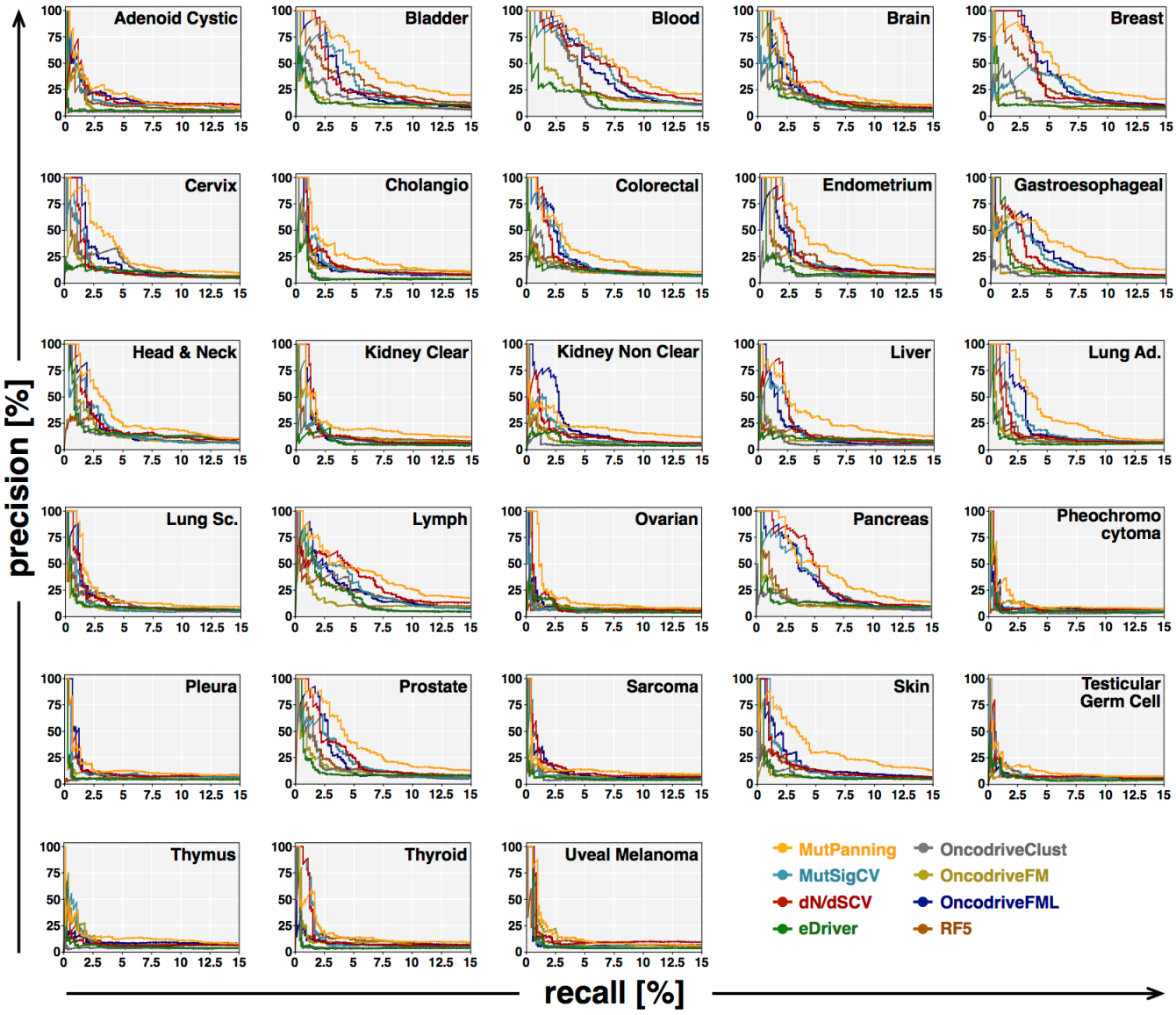
Supplementary Figure 16 | Performance evaluation of different methods for cancer gene identification across 28 cancer types.

We benchmarked the specificity of our method against 7 previously published methods for cancer gene identification. These methods captured the major biological signals commonly used for cancer gene detection. We tested each method separately on 28 different cancer types. In total, these analyses are based on sequencing data of 11,873 samples. The exact number of samples included in this analysis per cancer type can be found in Extended Data Figure 5.

Each plot shows a comparison of the specificity between our method (orange) and 7 previously published methods. To compare the performance between methods, we plotted an ROC curve for each method. As a conservative approximation of the true-positive rate we used canonical cancer genes in the Cancer Gene Census (CGC). In brief, we went through the list of genes returned by each method in the order of their significance (i.e., the significance values computed by MutPanning and the 7 other methods and adjusted for multiple testing); we plotted the number of significant genes in the CGC (y-axis) against the number of significant genes not in CGC (x-axis). Deviation of the curve to the upper left reflects enrichment for CGC genes as a measure of performance.



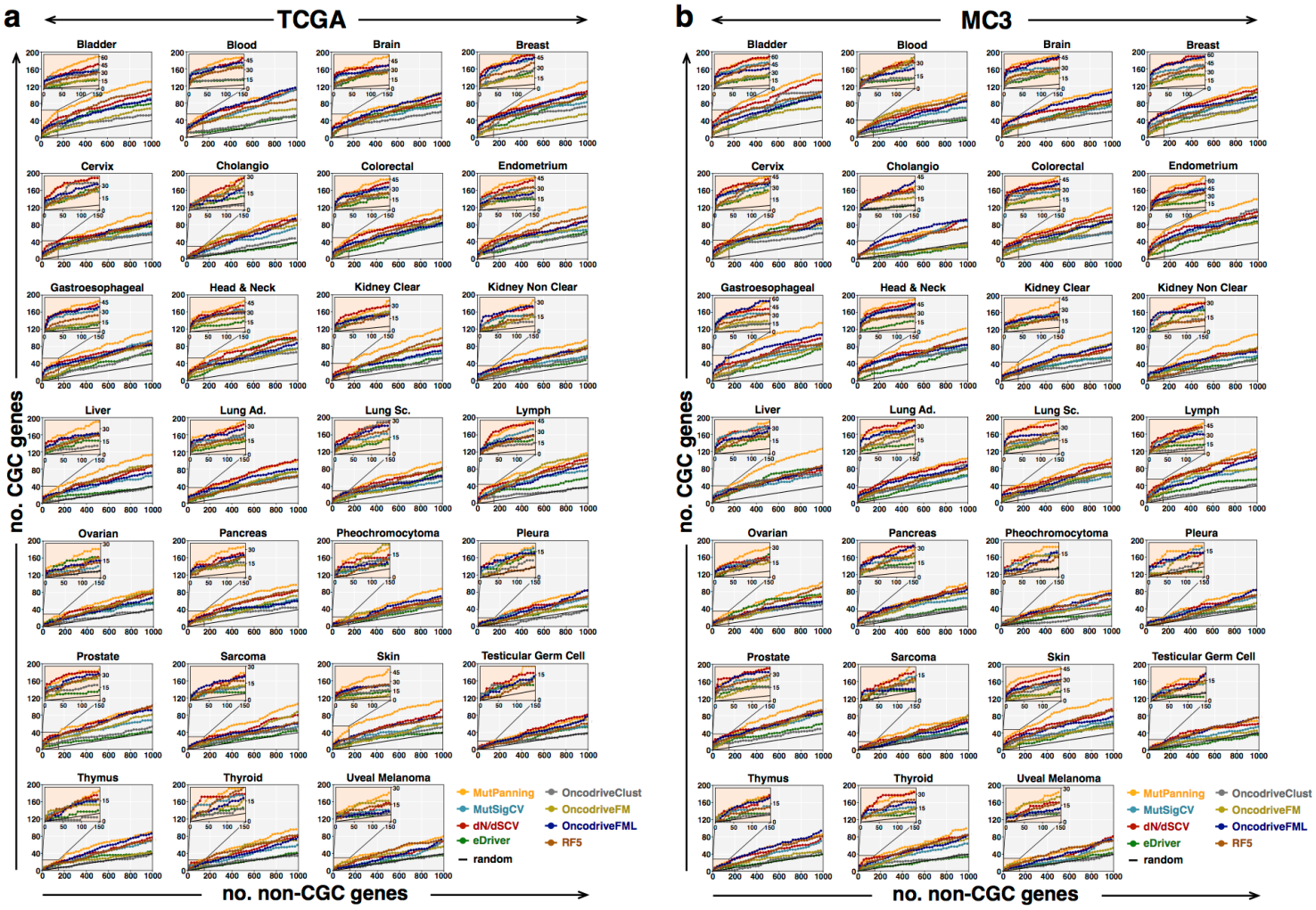
Supplementary Figure 17 | Benchmarking of the specificity of different methods based on OncoKB.
We benchmarked the specificity of our method against 7 previously published methods for cancer gene identification. These methods captured the major biological signals commonly used for cancer gene detection. We tested each method separately on 28 different cancer types. In total, these analyses are based on sequencing data of 11,873 samples. The exact number of samples included in this analysis per cancer type can be found in Extended Data Figure 5.
The plots shown in this figure parallel the plots shown in Supplementary Figure 16. To compare the performance between methods, we plotted an ROC curve for each method (similar to Supplementary Figure 16). Instead of the CGC, we used the OncoKB dataset as a conservative approximation of the true-positive rate.



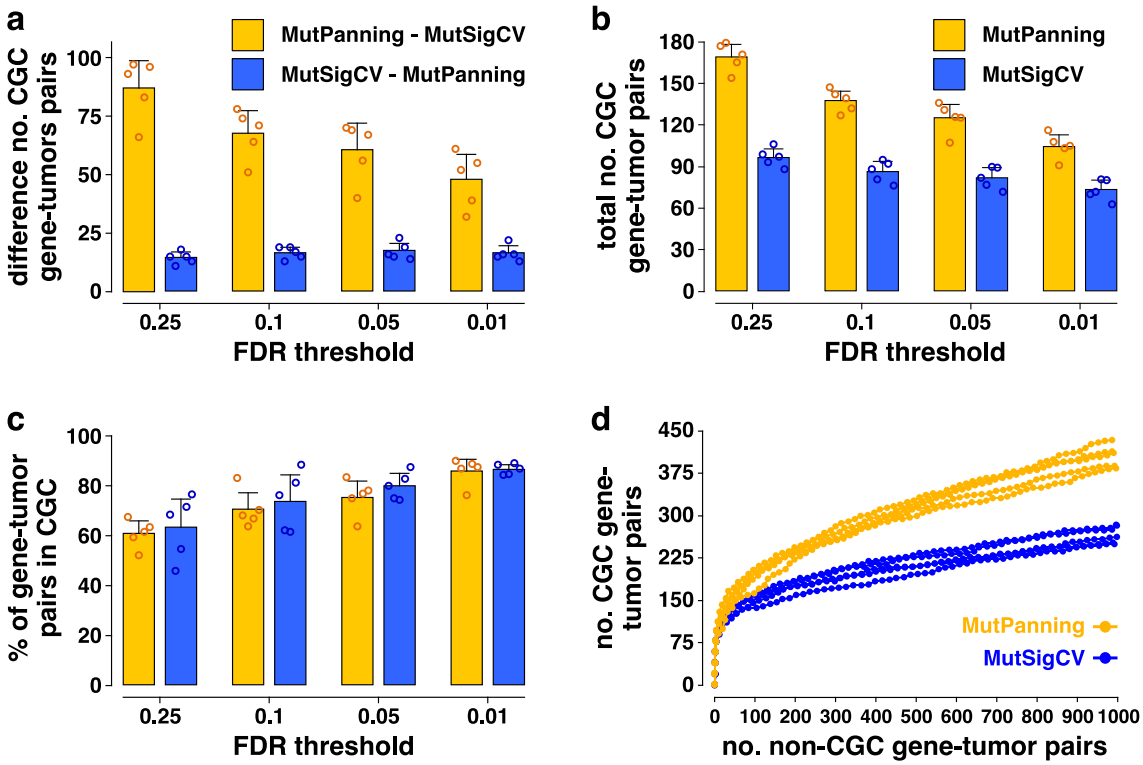
Supplementary Figure 18 | Precision-recall curves provide an additional way to compare the specificity between different methods for driver gene identification.

We benchmarked the specificity of our method against 7 previously published methods for cancer gene identification. These methods captured the major biological signals commonly used for cancer gene detection. We tested each method separately on 28 different cancer types. In total, these analyses are based on sequencing data of 11,873 samples. The exact number of samples included in this analysis per cancer type can be found in Extended Data Figure 5.

As an alternative way to capture the specificity of each method, we used precision-recall curves. In parallel to our previous analyses, we counted the number of significant genes in CGC; we then compared these counts with the total number of CGC genes (recall, x-axis) and the total number of significant genes (precision, y-axis). From left to right, the number of significant genes increases (the significance values computed by MutPanning and the 7 other methods and adjusted for multiple testing), so that the fluctuation of the precision-recall curves decreases. To compare the specificity between different methods, precisions (y-axis) are compared for the same recall (x-axis) in these curves. From these curves, we derived the precision at 5% recall as a measure to compare the performance between different methods. Hence, precision-recall curves provide an additional way to compare the specificity between different methods.



Supplementary Figure 19 | Benchmarking of method performance in two homogeneously processed datasets.
We set out to confirm the superior performance of MutPanning on two independently processed datasets. The TCGA subcohort of our study had been homogeneously processed with the same variant calling pipeline. Similarly, samples in the MC3 dataset were homogeneously processed, and this dataset was the basis of the Bailey et al. study. We applied MutPanning, as well as the other 7 methods used for comparison (MutSigCV, dN/dNCV, eDriver, OncodriveClust, OncodriveFM, OncodriveFML, RF5), to the TCGA (n = 7,060 samples, **a**) and MC3 (n = 9,079, **b**) datasets. We compared the performance of these methods on the TCGA and MC3 datasets. For a conservative approximation of the true-positive rate, we used genes in the Cancer Gene Census (CGC). We went through the top 1,000 genes returned by each method in their order of significance (significance values returned by MutPanning and the 7 other methods used for comparison, adjusted for multiple testing), and we plotted the number of non-CGC genes (x-axis) against the number of CGC genes (y-axis). The AUC of these ROC curves served as a measure of performance. Cohort sizes for TCGA/MC3 datasets: bladder: 130/386; blood: 197/139; brain: 576/821; breast: 975/779; cervix: 192/274; cholangio: 35/34; colorectal: 223/316; endometrium: 305/451; gastroesophageal: 467/529; head&neck: 279/502; kidney clear: 417/368; kidney non-clear: 227/340; liver: 194/354; lung adenocarcinoma: 230/431; lung squamous: 173/464; lymph: 48/37; ovarian: 316/408; pancreas: 149/155; pheochromocytoma: 179/179; pleura: 82/81; prostate: 323/477; sarcoma: 247/204; skin: 342/422; testicular: 149/145; thymus: 123/121; thyroid: 402/492; uveal melanoma: 80/80

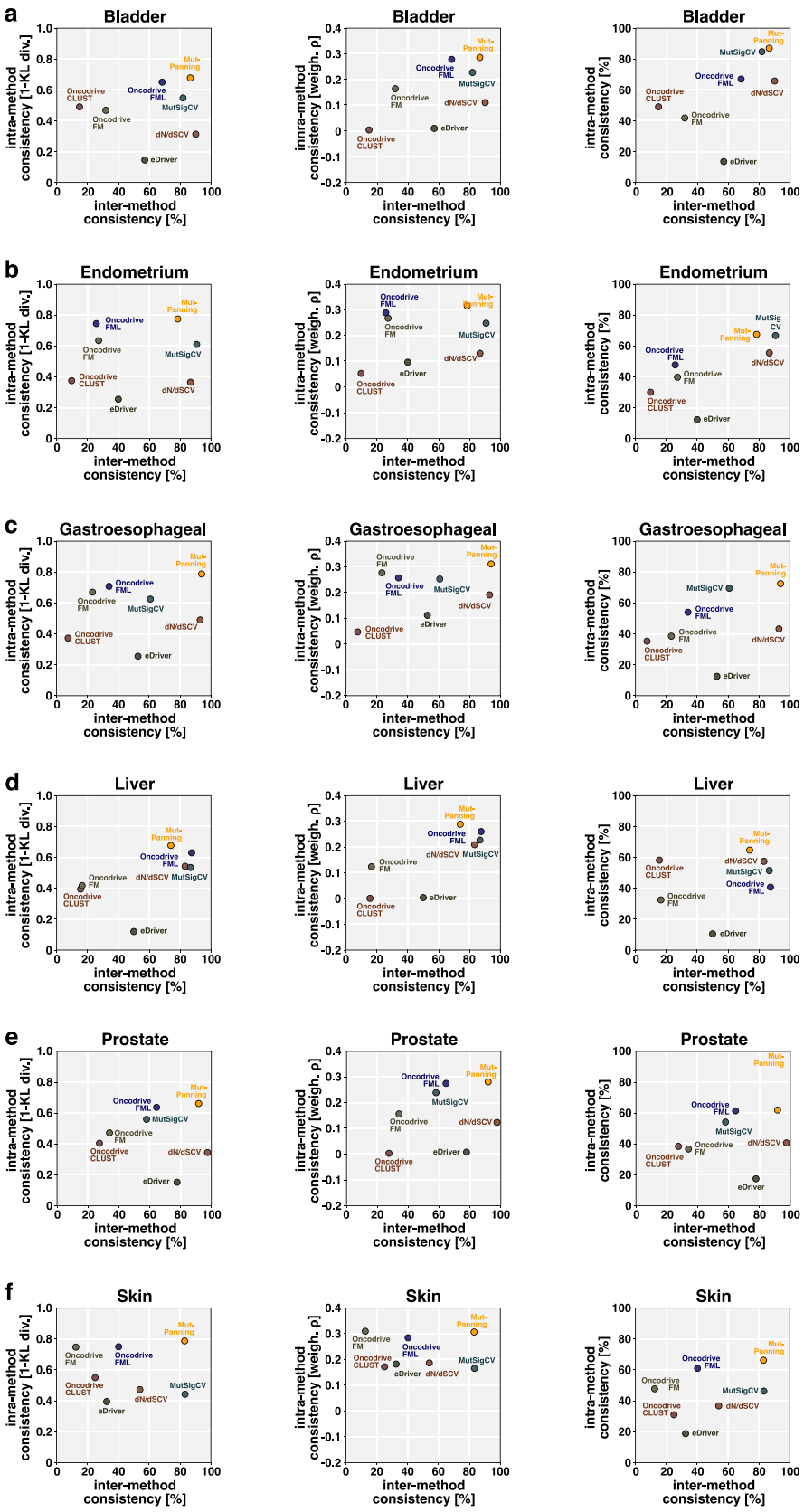


Supplementary Figure 20 | Comparison of the performance between MutPanning and MutSigCV in a downsampling analysis.

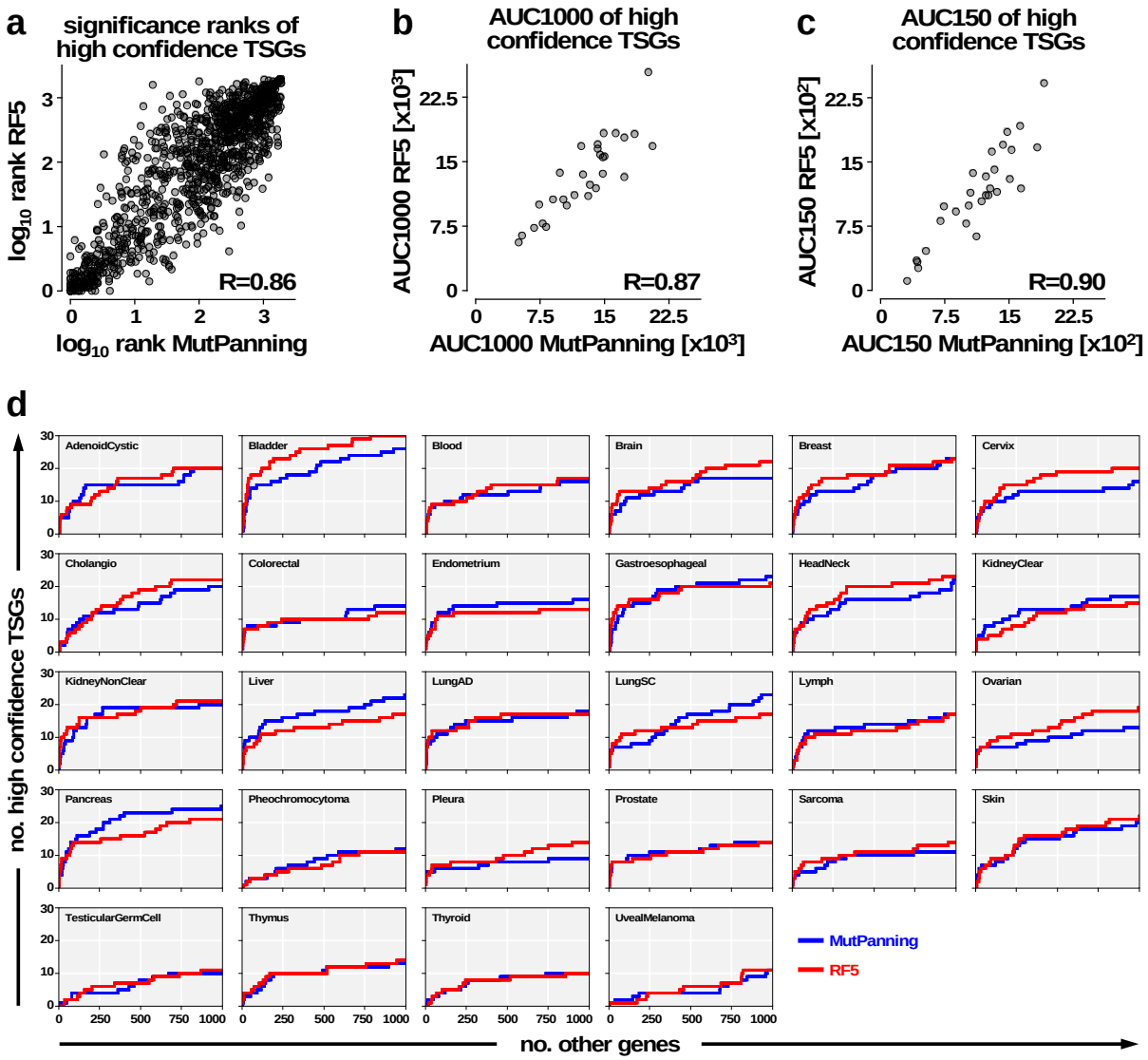
We used a downsampling analysis to explore differences in performance between MutPanning and MutSigCV. MutSigCV has been used widely in previous studies to identify driver genes. In brief, we randomly selected 20% of the samples from our study cohort (n = 11,873 samples in total, 5 downsampling iterations). The exact number of samples included in this analysis per cancer type can be found in Extended Data Figure 5. On each of these small random subcohorts, we ran MutPanning and MutSigCV. Significance values were computed using MutPanning and MutSigCV and adjusted for multiple testing (false discovery rate, FDR).

a, The bar graphs shows the average number of CGC gene-tumor pairs identified by MutPanning, but not by MutSigCV (yellow) as well as the number of CGC gene-tumor pairs identified by MutSigCV, but not by MutPanning (blue). Error bars indicate the standard deviation between different downsampling runs. Dots indicate the results from each of the five downsampling runs. The same FDR thresholds (x-axis) were applied to MutSigCV and MutPanning for the comparison of their results. **b,** The bar graphs compares the average number of CGC gene-tumor pairs identified by MutPanning (yellow) with MutSigCV (blue). Error bars indicate the standard deviation between different downsampling runs. Dots indicate the results from each of the five downsampling runs. The same FDR thresholds (x-axis) were applied to MutSigCV and MutPanning for the comparison of their results. **c,** The bar graph shows the average fraction of canonical cancer genes in the Cancer Gene Census (CGC), in the set of genes identified by MutPanning, but not by MutSigCV (yellow) as well as the set identified by MutSigCV, but not by MutPanning (blue). Error bars indicate the standard deviation between different downsampling runs. Dots indicate the results from each of the five downsampling runs. The same FDR thresholds (x-axis) were applied to MutSigCV and MutPanning for the comparison of their results. This analysis demonstrates that the increased number of genes identified by MutPanning does not result from a lower specificity rate. In other words, the additional results of MutPanning do not result from an increased rate of false-positives. **d,** We further plotted the ROC curve for each downsampling analysis run (n = 11,873 samples in total, 5 downsampling iterations, results from all cancer types combined). As a conservative approximation of the true-positive rate, we used the Cancer Gene Census (CGC). These analyses corroborated the increased performance of MutPanning in the downsampling runs without defining an FDR threshold. In particular, these results show that our observations in (**a**)-(**c**) do not result from differences in calibration of the significance values between MutPanning and MutSigCV.

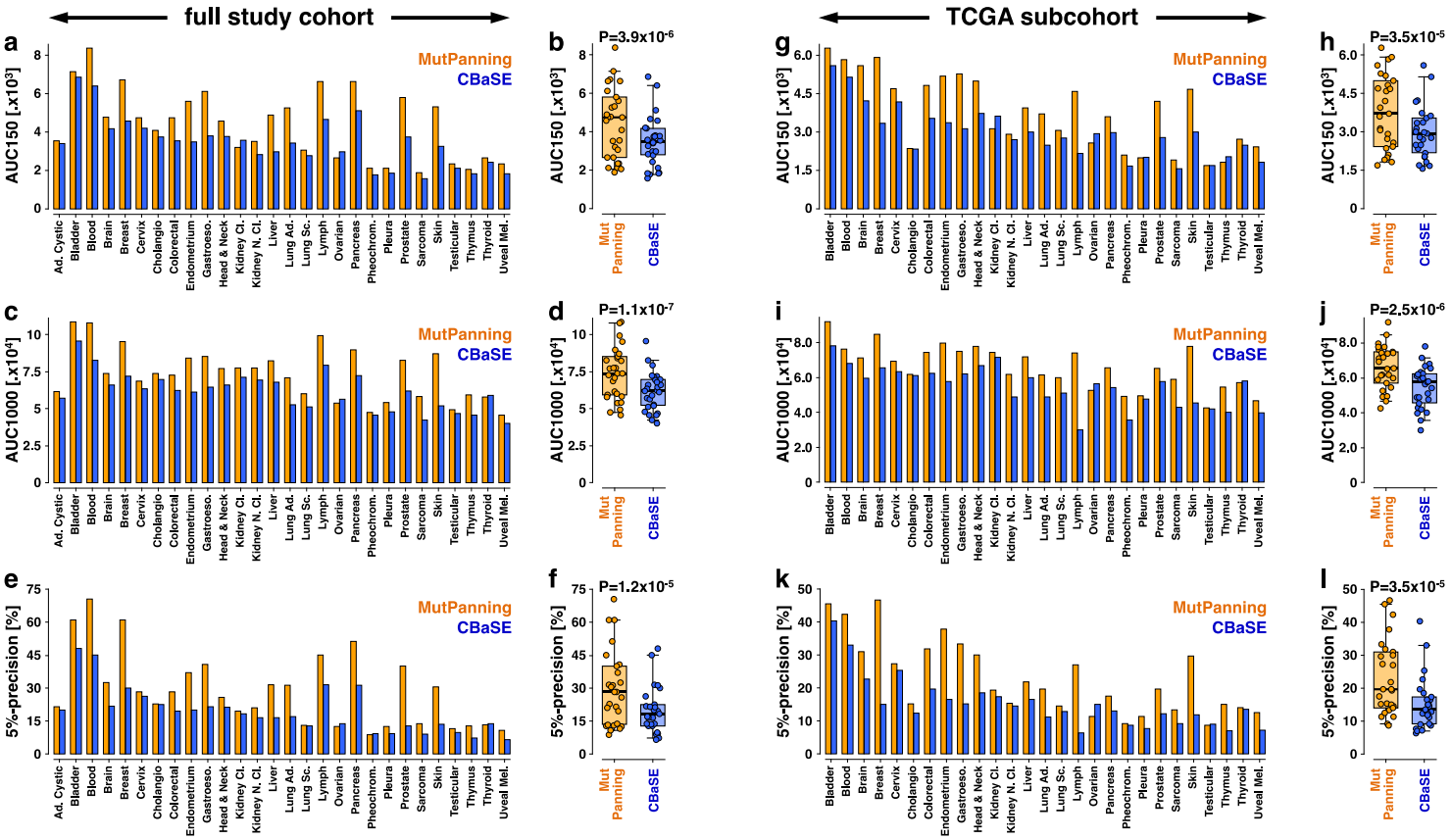
Taken together, the results in (**a**)-(**d**) demonstrate that MutPanning identified several additional canonical cancer genes in random small subcohorts (20%) that were not identified by MutSigCV in the same subcohorts. The analyses shown in (**c**) and (**d**), demonstrate that these additional findings do not result from differences in specificity or p-value calibration.



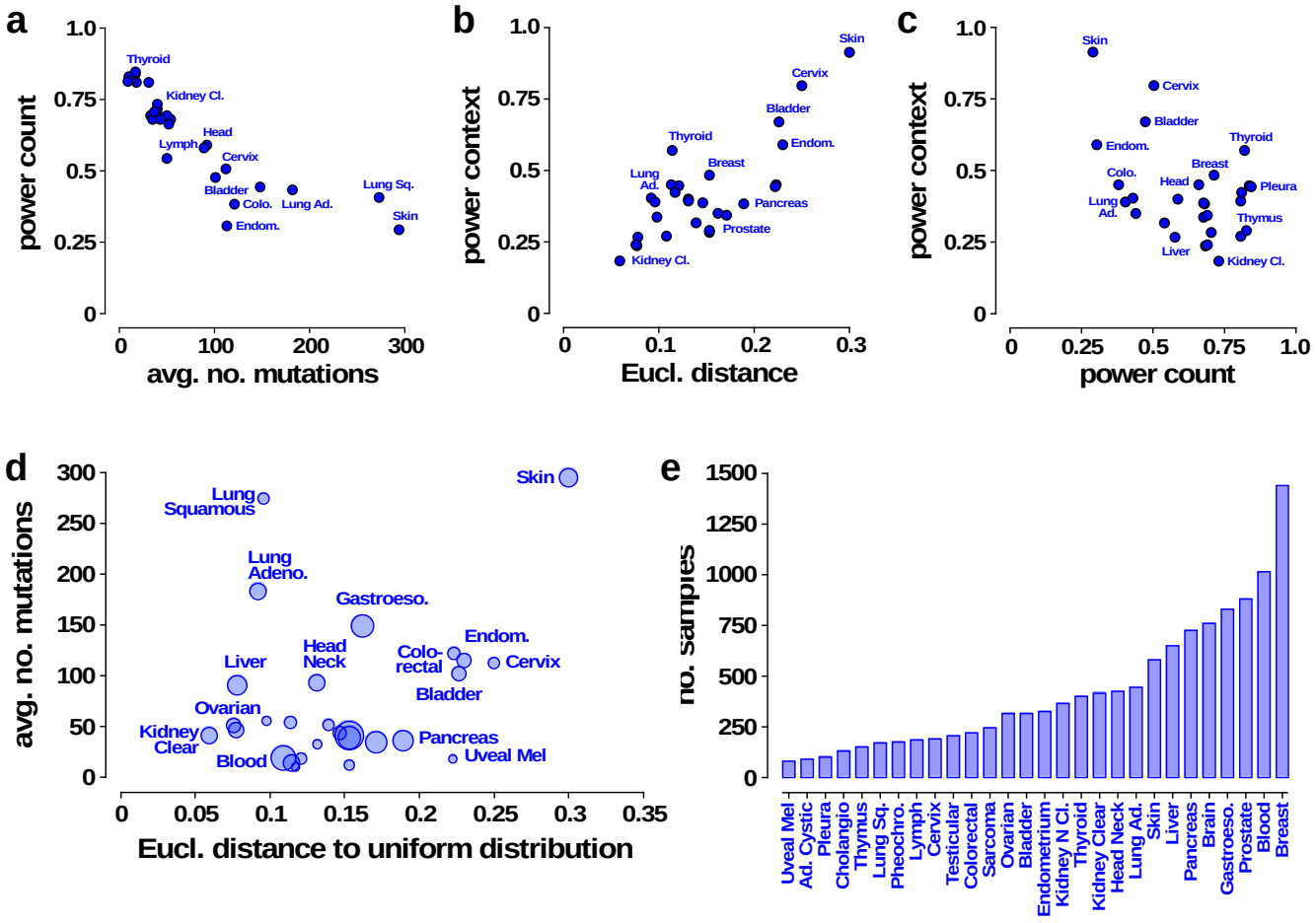
Supplementary Figure 21 | Analysis of the consistency of results as an additional measure to compare the performance between methods.
For six cancer types (bladder, n = 317 samples; endometrium, n = 327; gastroesophageal, n = 833; liver, n = 650; prostate, n = 880; skin, n = 582) we randomly split the cohort into halves and ran MutPanning and other methods on both halves of the cohort (five random splits for each method and cancer type). We then determined the similarity between the results obtained from each half of the cohort (y-axis) based on three similarity measures: the KL-divergence between p-values (left), the weighted correlation between p-values (middle), and the fraction of significant genes from the first cohort half that passed the significance threshold for the second cohort half (right). We further determined the inter-method consistency, i.e. the fraction of significant genes (significance values based on MutPanning and the other methods used for benchmarking and corrected for multiple testing) that emerged as significant by at least one independent method (x-axis). We plotted the inter-method consistency against the intra-method consistency to compare the performance between methods.



Supplementary Figure 22 | Performance evaluation of MutPanning on known tumor suppressor genes.
We examined the performance of MutPanning on a set of 48 known tumor suppressor genes. We compared MutPanning with the performance of the RF5 method, which implements a separate test to identify tumor suppressor genes as part of its statistical model. In total, the analyses in (**a**)-(**d**) are based on sequencing data of 11,873 samples. The exact number of samples included in this analysis per cancer type can be found in Extended Data Figure 5.
a, We plotted the significance ranks of the 48 high confidence tumor suppressor genes based on the results of MutPanning (x-axis) and RF5 (y-axis) across 28 cancer types. Each dot reflects a pair of a high confidence tumor suppressor gene and a cancer type. We compared MutPanning and RF5 ranks based on a Pearson correlation coefficient (*R*, bottom right). **b-d,** For each cancer type, we further plotted an ROC curve that used the set of 48 known tumor suppressor genes as a reference. We determined the AUC of these ROC curves for the top 1000 (**b**) and 150 (**c**) genes returned by MutPanning (x-axis) and RF5 (y-axis). These AUC values reflect the overall performance on tumor suppressor genes and were highly correlated (Pearson correlation coefficient, *R*, bottom right) between MutPanning and RF5. The ROC curves, which these AUC values are based on, are shown in (**d**).

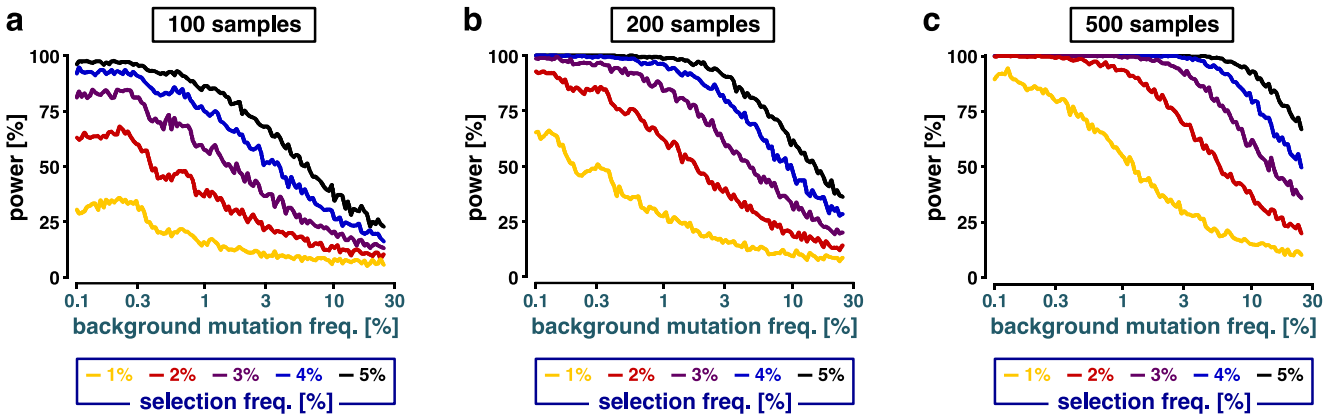


Supplementary Figure 23 | Comparison of the performance between MutPanning and CBaSE.
We compared the performance of MutPanning and CBaSE. MutPanning and CBaSE share the same statistical model for the count component with differences in the implementation. MutPanning amends this count component by a statistical component that accounts for mutations in "unusual" nucleotide contexts. Hence, in contrast to the other 7 methods in our benchmarking panel, this analysis does not reflect a comparison against a fully independent model. Instead, this comparison reflects the impact of the nucleotide context component on the performance of MutPanning. However, due to differences between MutPanning and CBaSE in the implementation of the count component, additional factors can impact this comparison. An analysis of the contribution of nucleotide contexts on the performance of MutPanning can be found in Supplementary Figure 24. The analyses shown in figures (**a**)-(**f**) are based on sequencing data of the full study cohort of 11,873 samples spanning 28 different cancer types. The exact number of samples included in this analysis per cancer type can be found in Extended Data Figure 5. The analyses shown in figures (**g**)-(**l**) are based on sequencing data of the TCGA subcohort of 7,060 samples spanning 27 different cancer types. The cohort sizes per cancer type of the TCGA subcohort are as follows: bladder: 130; blood: 197; brain: 576; breast: 975; cervix: 192; cholangio: 35; colorectal: 223; endometrium: 305; gastroesophageal: 467; head&neck: 279; kidney clear: 417; kidney non-clear: 227; liver: 194; lung adenocarcinoma: 230; lung squamous: 173; lymph: 48; ovarian: 316; pancreas: 149; pheochromocytoma: 179; pleura: 82; prostate: 323; sarcoma: 247; skin: 342; testicular: 149; thymus: 123; thyroid: 402; uveal melanoma: 80.
To compare the level of performance, we tested whether the most significant genes (significance values computed by MutPanning and CBaSE and adjusted for multiple testing) returned by each method were enriched for canonical cancer genes in the Cancer Gene Census. We quantified this measure as the area under the curve (AUC) that plotted CGC vs. non-CGC genes for the top 150 (**a, b, g, h**), and top 1,000 (**c, d, i, j**) significant genes returned by each method. Further, we computed the precision at 5% recall (**e, f, k, l**). **a, c, e, g, i, k** These performance measures are shown for MutPanning (orange) and CBaSE (blue) as a bar graph for individual cancer types. **b, d, f, h, j, l** We plotted the overall distribution of these measures as a box plot . Each cancer type is represented by a dot (28 cancer types in **b, d, f**; 27 cancer types in **h, j, l**). Boxes indicate the 25%/75% interquartile range, and whiskers extend to the 5%/95% percentiles. Medians are indicated by vertical lines. P-values are based on a two-sided paired sample t-test (28 paired data points, 27 degrees of freedom in **b**, **d** and **f**; 27 paired data points, 26 degrees of freedom in **h**, **j** and **l**). T-values of the test are as follows: 5.8 (**b**), 7.2 (**d**), 5.4 (**f**), 5.0 (**h**), 6.0 (**j**) and 5.0 (**l**).



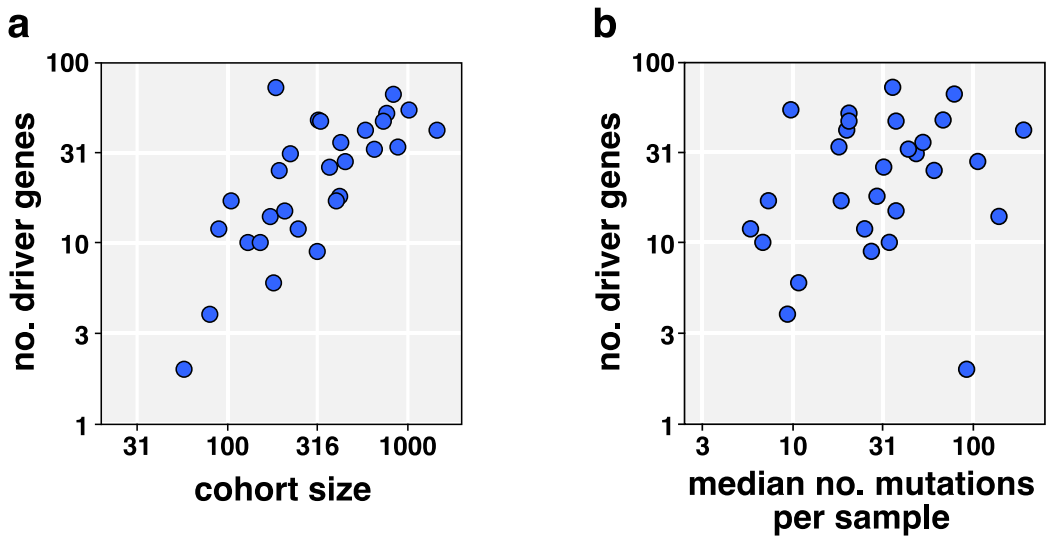
Supplementary Figure 24 | Power analyses of the differential performance of MutPanning across cancer types.
The statistical framework of MutPanning relies on two principal components - mutational excess above the regional background mutation rate and increased number of mutations in "unusual" nucleotide contexts. Our performance benchmarking analyses revealed that these two statistical components yielded differential returns across different cancer types. We aimed to further understand why the performance of MutPanning varied across cancer types. To this end, we used two power analyses to examine the behavior of these two components of MutPanning across different cancer types. In total, the analyses shown in figures (**a**)-(**e**) are based on sequencing data of 11,873 samples to evaluate the background mutation rate, composite likelihood scores, and Euclidean distance to uniform distribution per cancer type. The exact number of samples included in this analysis per cancer type is shown in Extended Data Figure 5 and is visualized as a bar graph in (**e**).
a, To evaluate the power of the count component, we simulated the number of passenger mutations according to a Poisson distribution in a simulated cohort of 150 samples per cancer type and a gene of 1,500bp. The rate of the Poisson distribution was dependent on the individual background mutation rate per cancer type (rate=no. samples x gene length x mutation rate). We further simulated the number of driver mutations according a Binomial distribution, assuming a positive selection frequency of 1%. For each cancer type, we performed 1,000 simulation runs. In each simulation run, we computed a p-value by determining whether the simulated count (passenger + driver mutations) was higher than expected based on background model (passenger mutations only, Poisson distribution, one-tailed test). We then determined the power as the fraction of simulation runs in which the p-value was significant ($p<0.05$). We plotted the results of this power analysis (y-axis) against the background mutation rate per cancer type (x-axis). The power to identify driver genes based on increased mutation counts varied across cancer types, depending on their background mutation rates. **b,** To evaluate the power of the nucleotide context component, we quantified the power to distinguish a context-dependent passenger mutation distribution pattern from a uniform distribution. For this purpose, we performed 1,000 simulation runs and we simulated in each run 15 passenger mutations across 1,500bp based on a Multinomial distribution with probabilities based the composite likelihood model of each cancer type. Furthermore, we simulated 5 mutations according to a uniform distribution across 1,500bp. We then determined the probability of each scenario (15+5 mutations) based on a Multinomial distribution of 20 mutations with

probabilities based the composite likelihood model. In each simulation run, we computed a p-value by determining whether the probability of the simulated scenario was lower than expected based on background model (20 mutations with probabilities based the composite likelihood model, Multinomial distribution, one-tailed test). We then determined the power as the fraction of simulation runs in which the p-value was significant ($p<0.05$). In addition, we quantified for each cancer type the dependence of passenger mutations on the surrounding nucleotide context. To this end, we counted for each cancer type the number of mutations within all 96 possible trinucleotide contexts, thereby obtaining a 96-dimensional count vector. We then normalized this count vector to 1 and determined the Euclidean distance to a normalized unit vector (y-axis). This distance reflected the "flatness" of the underlying mutation signature, i.e. mutation types with a low distance had a low dependency on nucleotide context.We plotted the results of this power analysis (y-axis) against the Euclidean distance metric per cancer type (x-axis). The statistical power to distinguish a uniform distribution from the passenger mutation distribution per cancer type correlated with the "flatness" of the underlying mutation signature. **c,** We plotted the results from the power analysis of the count component (x-axis, cf. **a**) against the results from the power analysis of the context component (y-axis, cf. **b**). Overall, the results of both power analyses were negatively correlated, which suggested that they defined complementary statistical components that could compensate each other across cancer types. Both lung cancer cohorts were an exception from this trend and had relatively small power according to both components. **d,** Given that both the background mutation rate (y-axis) and the context dependency (x-axis) affected the ability of MutPanning to identify driver genes, we compared these two measures across cancer types using Pearson correlation coefficients (R). Intriguingly, background mutation rates were correlated with the context dependencies for cancer types ($R = 0.66$). The two lung cancer types (lung adenocarcinoma, $n = 446$; squamous-cell lung cancer, $n = 173$) were an exception from this correlation pattern, since they had both a high mutation rate and a low context dependency ($R = 0.77$ without lung cancer). **e,** The bar graph plots the number of samples per tumor type in our study cohort. These three factors (cohort size, mutation rate, context dependency) should be considered more generally when evaluating the performance of MutPanning for driver gene identification.

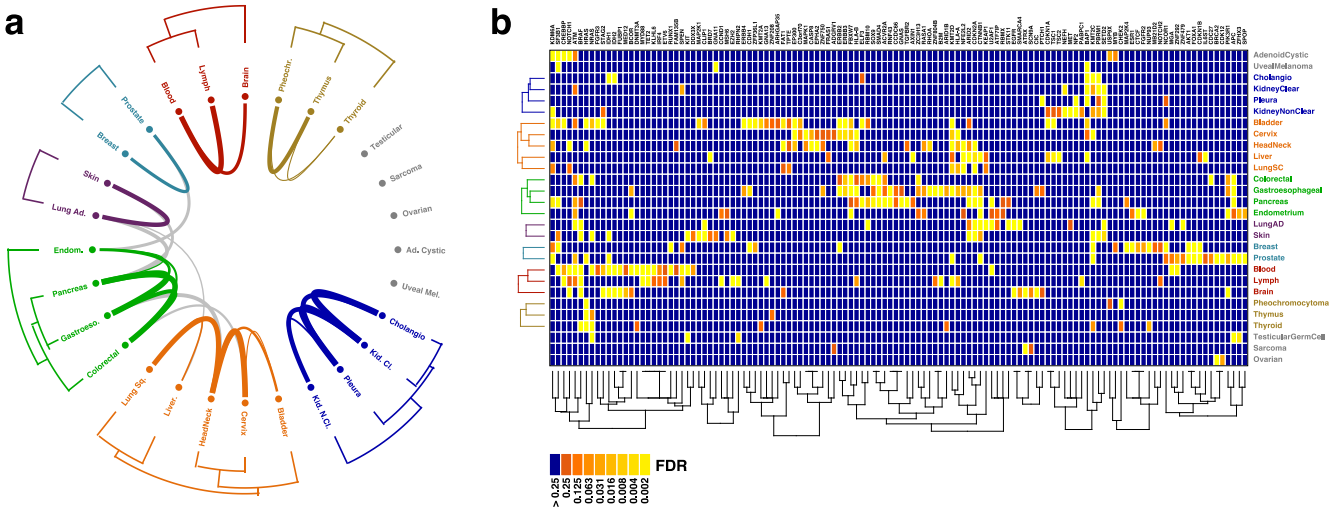


Supplementary Figure 25 | Passenger mutations limit the applicability of mutational recurrence for driver gene identification.

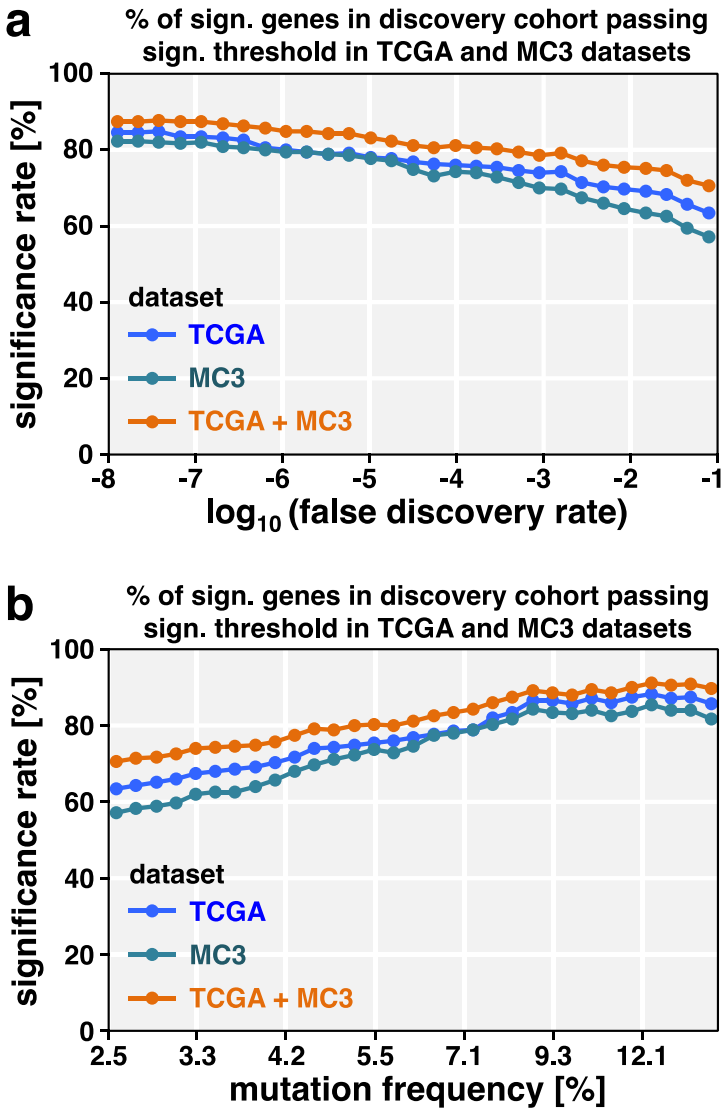
a-c, These analyses support Figure 3 and demonstrate the limited applicability of mutational recurrence in tumor types with high background mutation rates. To understand the effect of passenger mutations on the statistical power of mutational recurrence to detect driver genes, we used the following toy model. Given N samples and a background mutation frequency f_{bkgd}, we modeled the total number of passenger mutations in a non-cancer-related gene using a Poisson distribution, $\text{Pois}(N \cdot f_{\text{bkgd}})$. Further, given a positive selection frequency f_{sel} (i.e., the mutation frequency of a cancer driver gene above the background mutation rate) we assumed that the number of driver mutations followed a Binomial distribution, $\text{Binom}(N, f_{\text{sel}})$. Based on these two distributions, we simulated the number of mutations in cancer-related (driver mutations + passenger mutations) and in non-cancer related genes (passenger mutations only). We then determined the statistical power to discover a cancer gene as the fraction of simulation experiments in which the simulated mutation counts in cancer- and non-cancer-related genes differed significantly. We performed these simulation experiments with various background mutation frequencies f_{bkgd} (x-axis), for different selection frequencies above background f_{sel} (colors of the curves), and for different cohort sizes N (**a,** N=100 samples; **b,** N=200 samples; **c,** N=500 samples). When the background mutation frequency was smaller than the positive selection frequency, the statistical power was relatively stable. However, as soon as the background mutation frequency exceeded the positive selection frequency, the statistical power dropped rapidly. This observation is in concordance with previous studies on this issue and might explain why the detection of rare cancer genes with mutation frequencies <5% was challenging in previous studies. Based on this observation, we concluded that simultaneous integration of multiple biological criteria into a statistical framework would increase the ability to detect rare driver genes across heterogeneous tumor types.



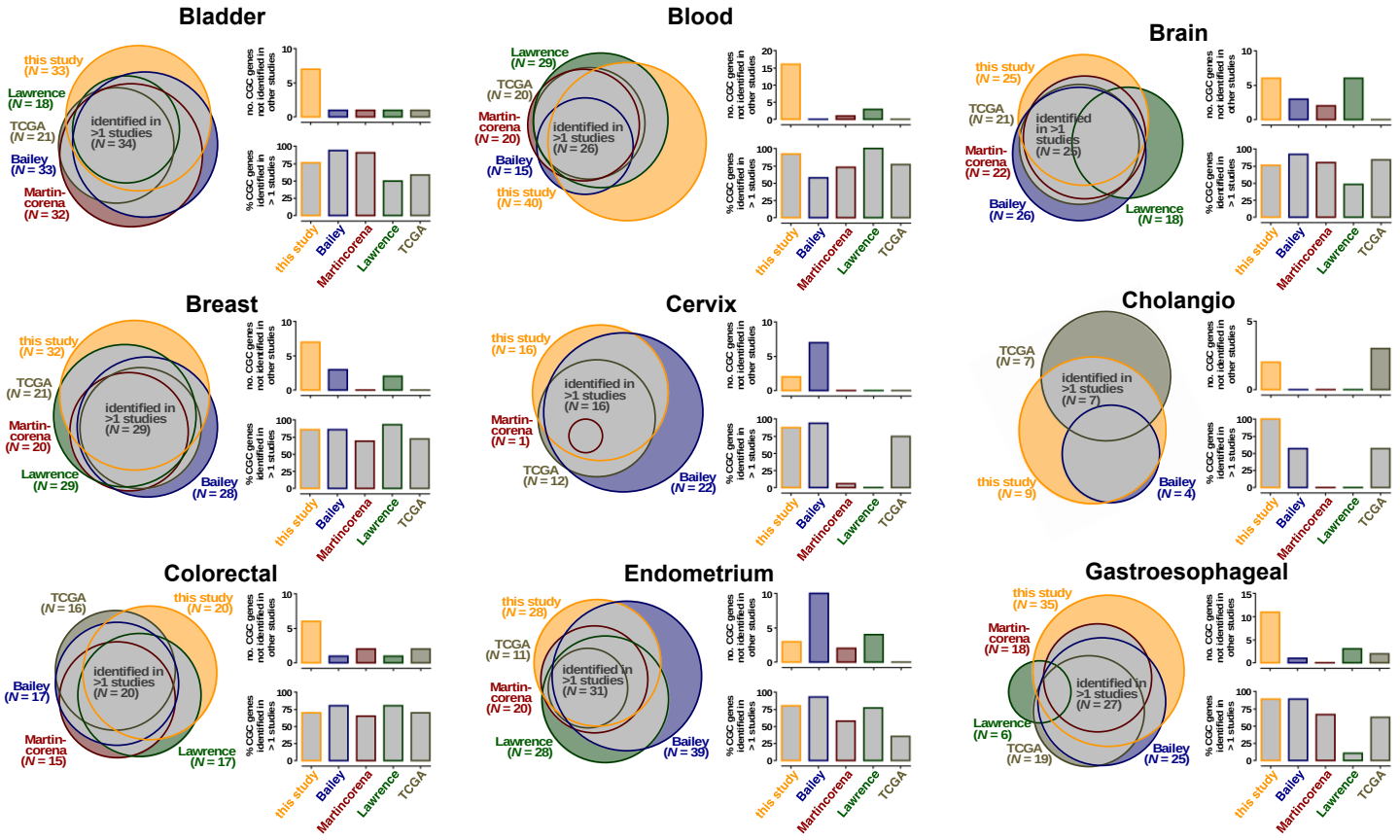
Supplementary Figure 26 | The number of driver genes depends on the cohort size and background mutation rate. For each cancer type, we plotted the number of driver genes (y-axis) against the number of samples (x-axis, **a**) and the background mutation rate (x-axis, **b**). In total, the analyses in this figure are based on sequencing data of 11,873 samples. The exact number of samples included in this analysis per cancer type can be found in Extended Data Figure 5. The number of driver genes increased with the cohort size. This observation suggested that adding more samples into our analysis would likely increase the number of rare driver genes, and that the search for rare driver genes was not saturated (**a**). Further, the number of driver genes was positively correlated with the background mutation rate (Pearson correlation), a trend that had been similarly observed in previous studies (**b**).



Supplementary Figure 27 | Clustering of mutation types by their mutual overlap in driver genes.
We quantified the pairwise overlap in driver genes between cancer types by Tanimoto coefficients, considering all driver genes that occurred in at least two cancer types and in a maximum of ten cancer types. Based on these Tanimoto coefficients, we performed hierarchical clustering (average linkage). **a,** The results of this hierarchical clustering procedure are visualized as a Circos plot. Tumor types are arranged into a circle. Hierarchical clustering between tumor types is visualized as a dendrogram. The Circos plot connects tumor types with overlaps in driver genes (Tanimoto coefficient > 0.2). Tanimoto coefficients are reflected by line strengths. Clusters are segregated by colors. **b,** The results of hierarchical clustering are visualized as a heatmap, which plots the FDR (false discovery rate, significance values derived by MutPanning and adjusted for multiple testing, cf. Supplementary Note) of each driver gene (columns) across tumor types (rows) (cf. color scale in figure). Hierarchical clustering of tumor types and driver genes is visualized as dendrograms. To enhance the cluster overview, non-tumor type specific cancer genes, which ubiquitously occurred in >10 tumor types, were not included in this analysis. Tumor type clusters are segregated by colors. In total, the analyses shown in figures (**a**) and (**b**) are based on sequencing data of 11,873 samples. The exact number of samples included in this analysis per cancer type can be found in Extended Data Figure 5.

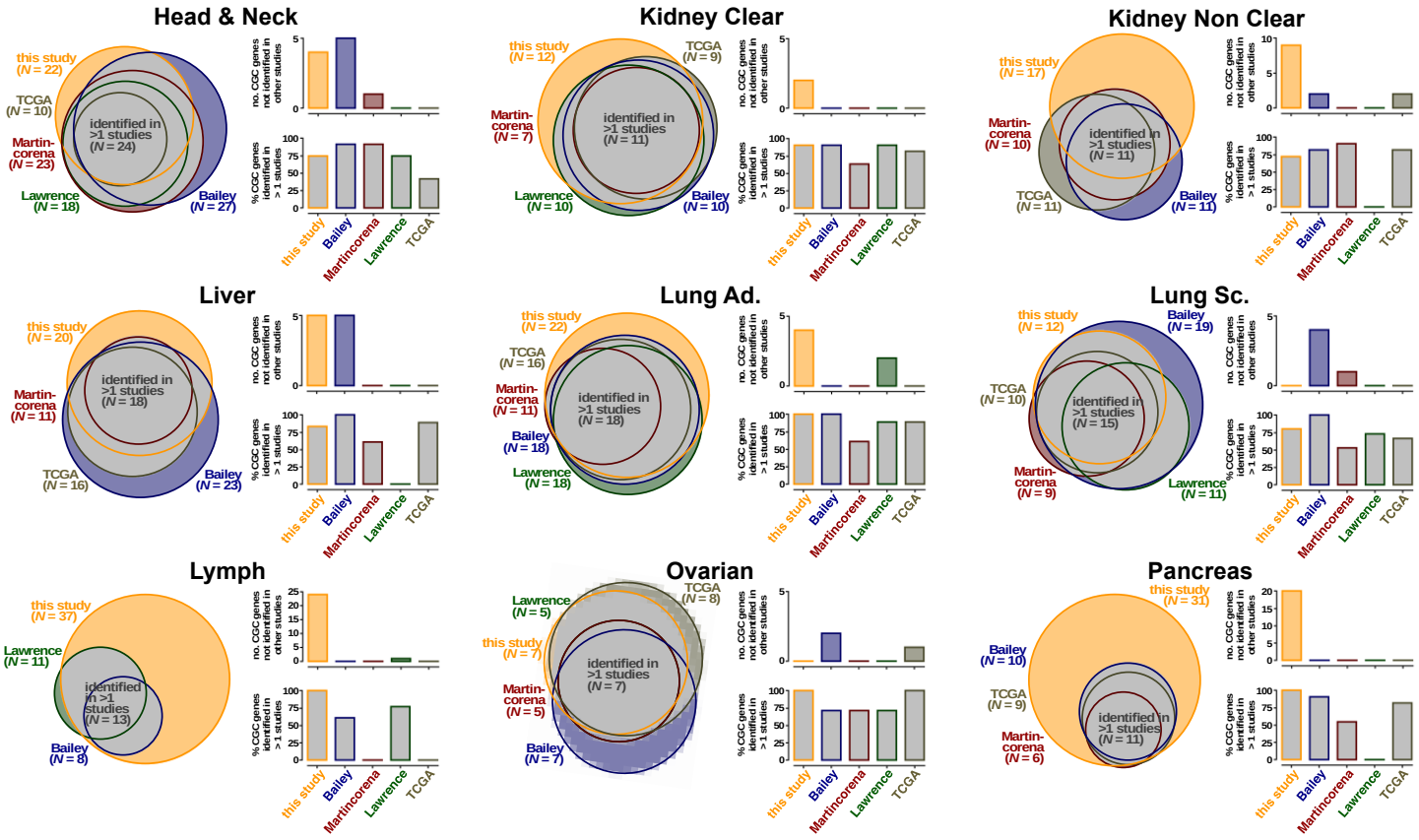


Supplementary Figure 28 | Reidentification of driver genes in two homogeneously processed datasets.
We set out to confirm the superior performance of MutPanning on two independently processed datasets. The TCGA subcohort has been homogeneously processed with the same variant calling pipeline. Similarly, the MC3 dataset has been homogeneously processed, and this dataset was the basis of the Bailey et al. study. We re-ran MutPanning with same parameters and FDR thresholds as in the original study cohort (n = 11,873 samples) on the TCGA (n = 7,060) and MC3 datasets (n = 9,079). We counted how many driver gene-tumor pairs in our original catalog (n = 827 pairs) could be re-identified in the TCGA (blue) and MC3 datasets (turquoise), or both datasets in combination (orange). Validation rates depend on the FDR-values derived by MutPanning (x-axis, false discovery rate, significance values derived by MutPanning and adjusted for multiple testing, **a**) and the mutation frequencies (x-axis, **b**) of the driver genes in the original cohort. Considering that the TCGA dataset was substantially smaller than the original dataset (7,060 vs. 11,873 samples), and that the overlap in samples between the original and the MC3 dataset was ~62%, these results corroborate the robustness of our original driver gene catalog. The validation status of each gene is annotated in Supplementary Tables 3 and 4.



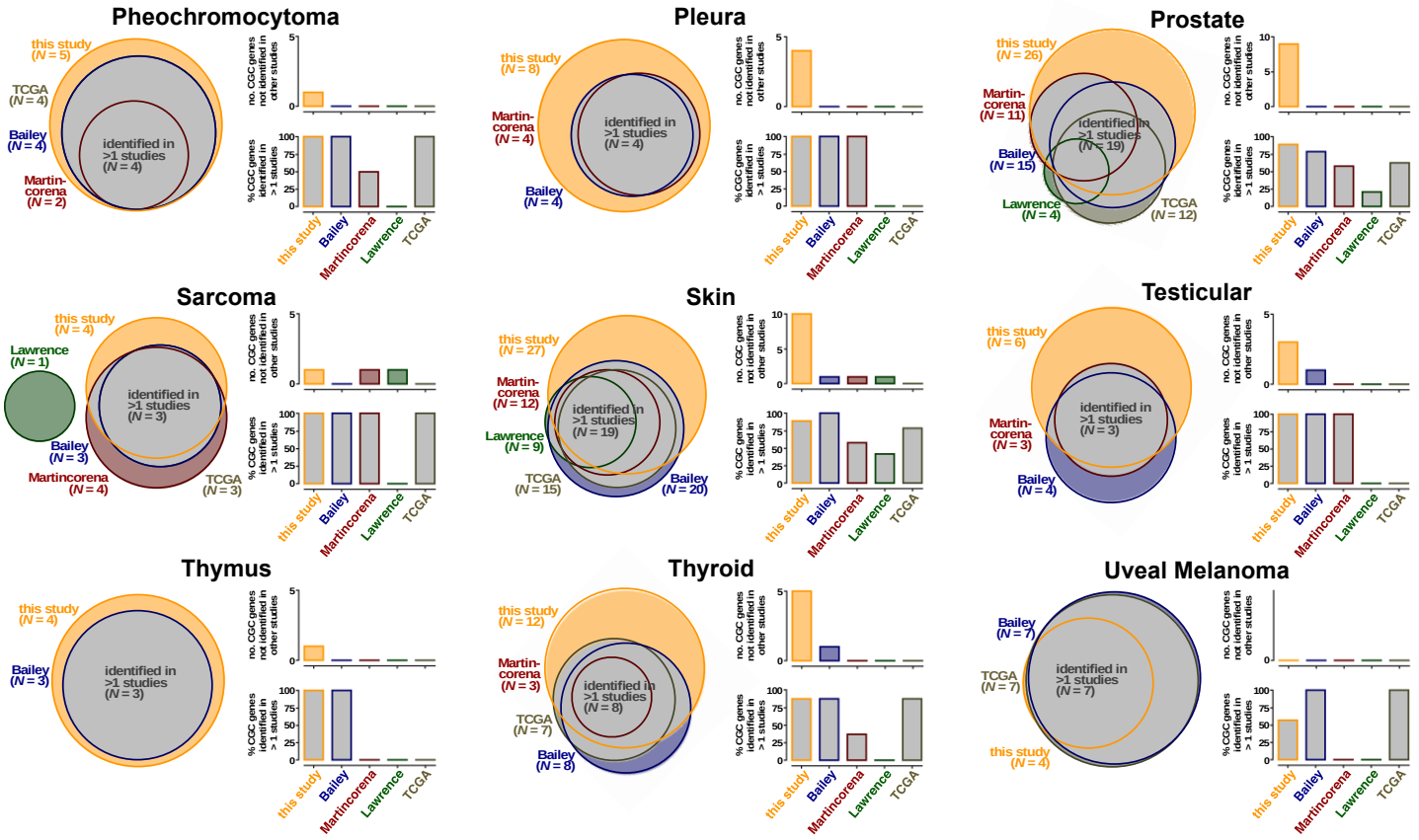
Supplementary Figure 29 | Analysis of the overlap of driver genes between our study and previous pan-cancer studies. (Part I)

This figure supports the analyses shown in Figure 5. We used area-proportional Venn diagrams to examine the overlap of CGC genes in our catalog (orange: this study) and comprehensive catalogs from previous pan-cancer studies (green: Lawrence et al, red: Martincorena et al., blue: Bailey et al., brown: TCGA papers). The gray area reflects the number of findings that were consistently reported by ≥2 studies for the same tumor type. While Figure 5 displays the overlap for all cancer types in aggregate, this figure visualizes the same data for individual cancer types. That way, our catalog can be compared with catalogs from previous pan-cancer studies for each cancer type individually, particularly in terms of its consistency (gray areas/bars) and additional findings (colored areas/bars).



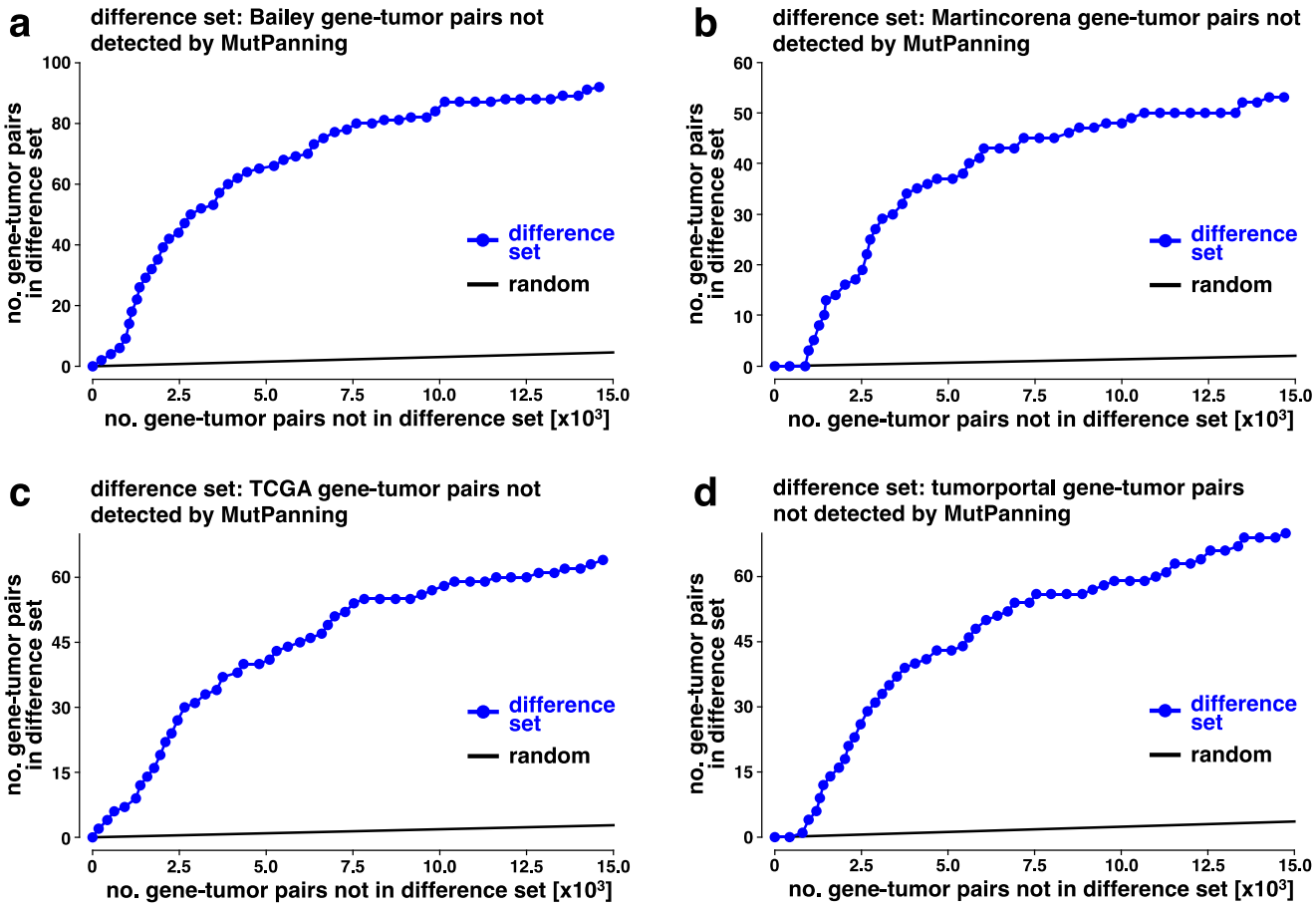
Supplementary Figure 30 | Analysis of the overlap of driver genes between our study and previous pan-cancer studies. (Part II)

This figure supports the analyses shown in Figure 5. We used area-proportional Venn diagrams to examine the overlap of CGC genes in our catalog (orange: this study) and comprehensive catalogs from previous pan-cancer studies (green: Lawrence et al, red: Martincorena et al., blue: Bailey et al., brown: TCGA papers). The gray area reflects the number of findings that were consistently reported by ≥2 studies for the same tumor type. While Figure 5 displays the overlap for all cancer types in aggregate, this figure visualizes the same data for individual cancer types. That way, our catalog can be compared with catalogs from previous pan-cancer studies for each cancer type individually, particularly in terms of its consistency (gray areas/bars) and additional findings (colored areas/bars).



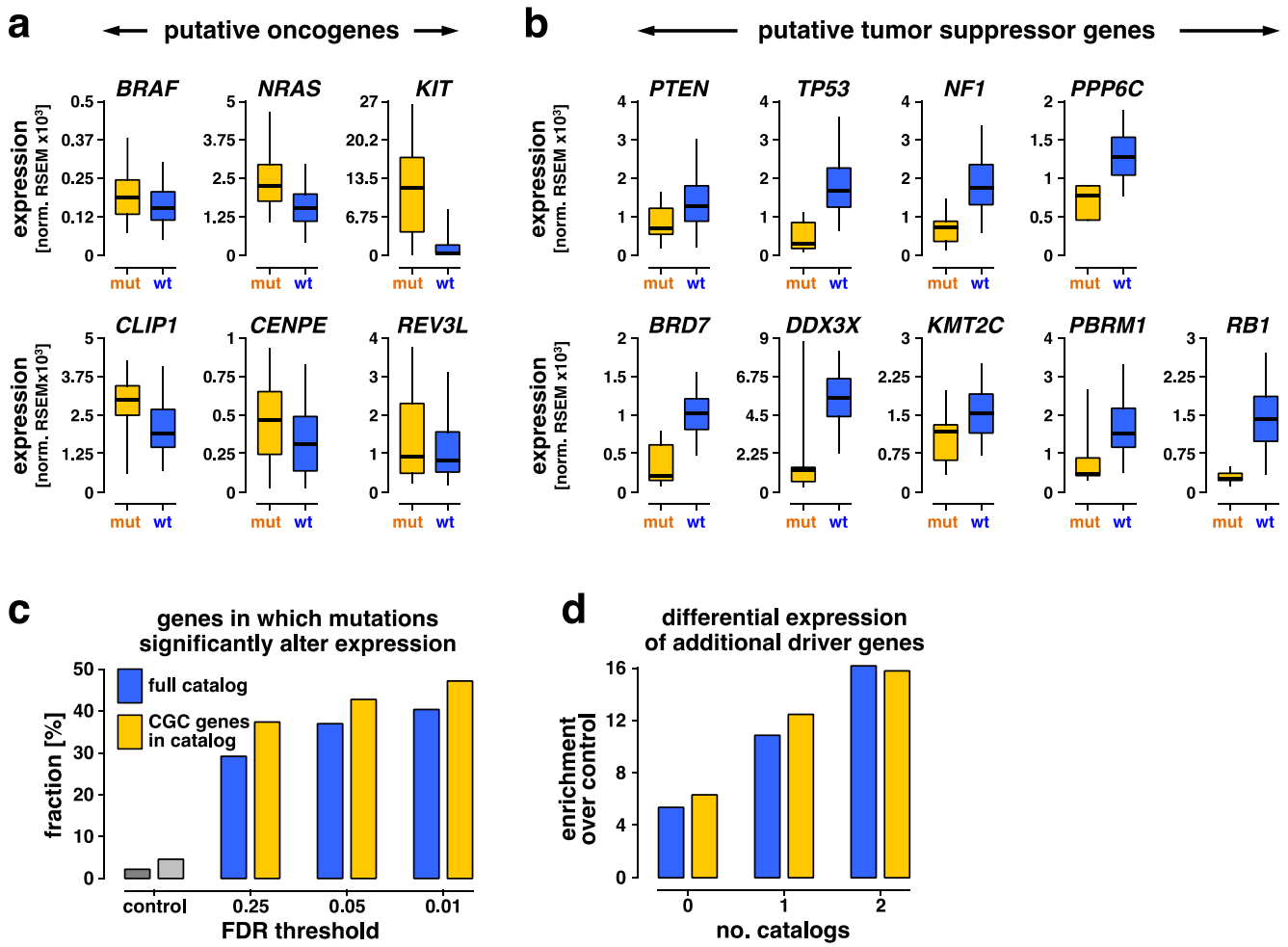
Supplementary Figure 31 | Analysis of the overlap of driver genes between our study and previous pan-cancer studies. (Part III)

This figure supports the analyses shown in Figure 5. We used area-proportional Venn diagrams to examine the overlap of CGC genes in our catalog (orange: this study) and comprehensive catalogs from previous pan-cancer studies (green: Lawrence et al, red: Martincorena et al., blue: Bailey et al., brown: TCGA papers). The gray area reflects the number of findings that were consistently reported by ≥2 studies for the same tumor type. While Figure 5 displays the overlap for all cancer types in aggregate, this figure visualizes the same data for individual cancer types. That way, our catalog can be compared with catalogs from previous pan-cancer studies for each cancer type individually, particularly in terms of its consistency (gray areas/bars) and additional findings (colored areas/bars).



Supplementary Figure 32 | Non-significant genes from other studies exhibit a trend towards significance compared with random controls.

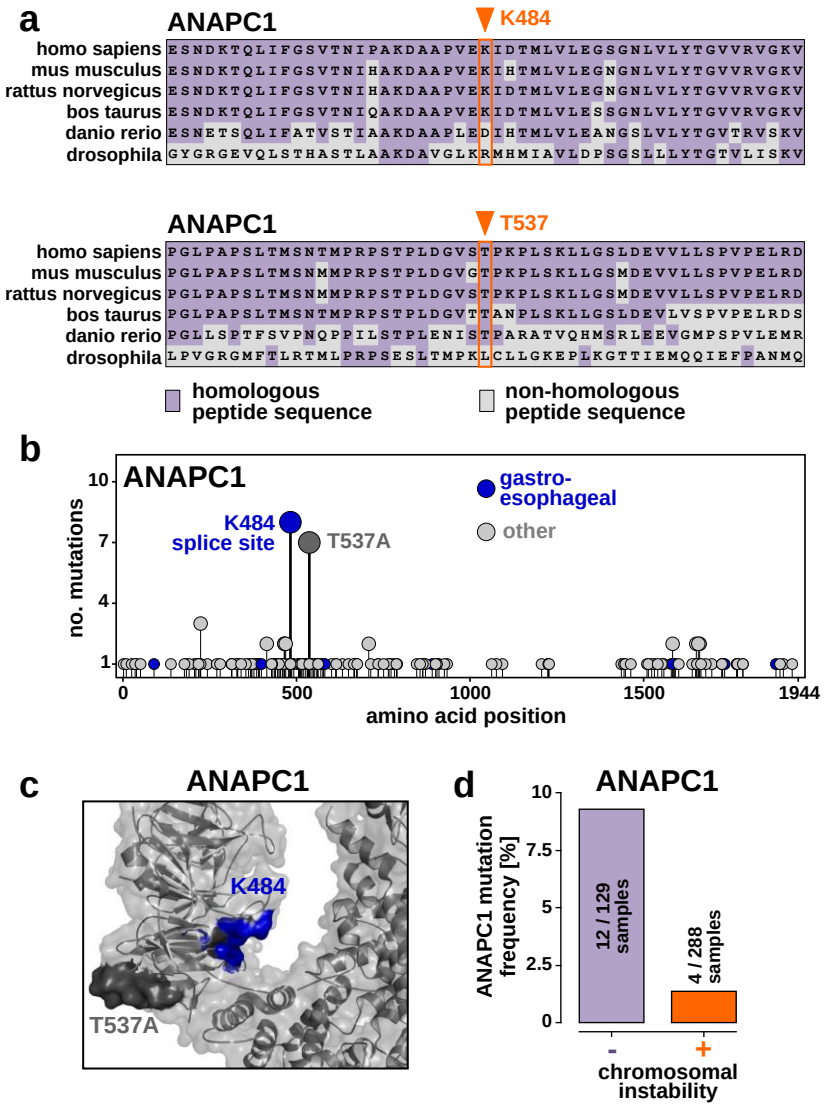
We noticed that some of the gene-tumor pairs from other studies did not emerge as significant (significance values returned by MutPanning and adjusted for multiple testing). We asked whether the mutation patterns of these genes were intrinsically different from the driver genes identified by MutPanning. For this purpose, we examined whether these genes had a residual statistical signal according to MutPanning. In other words, we asked whether their mutability scores were higher than for arbitrary genes, although they did not pass the significance threshold. We compared our results (based on n = 11,873 samples) with the results of four studies (**a,** Bailey et al., **b,** Martincorena et al., **c,** TCGA marker papers, **d,** Lawrence et al.), and we determined which driver genes were not part of our cohort ("difference set"). We then plotted an ROC curve according to this difference set (number of gene-tumor pairs in difference set vs. number of gene-tumor pairs not in difference set). These analyses revealed that genes in the difference sets showed a trend towards significance according to MutPanning. Hence, although these genes did not pass the significance threshold, most of them carried an elevated mutability score. This suggests that mutation patterns of genes in these difference sets are not necessarily different from the significant genes identified by MutPanning. The trend towards significance further suggests that MutPanning may identify these genes in a cohort with more samples.



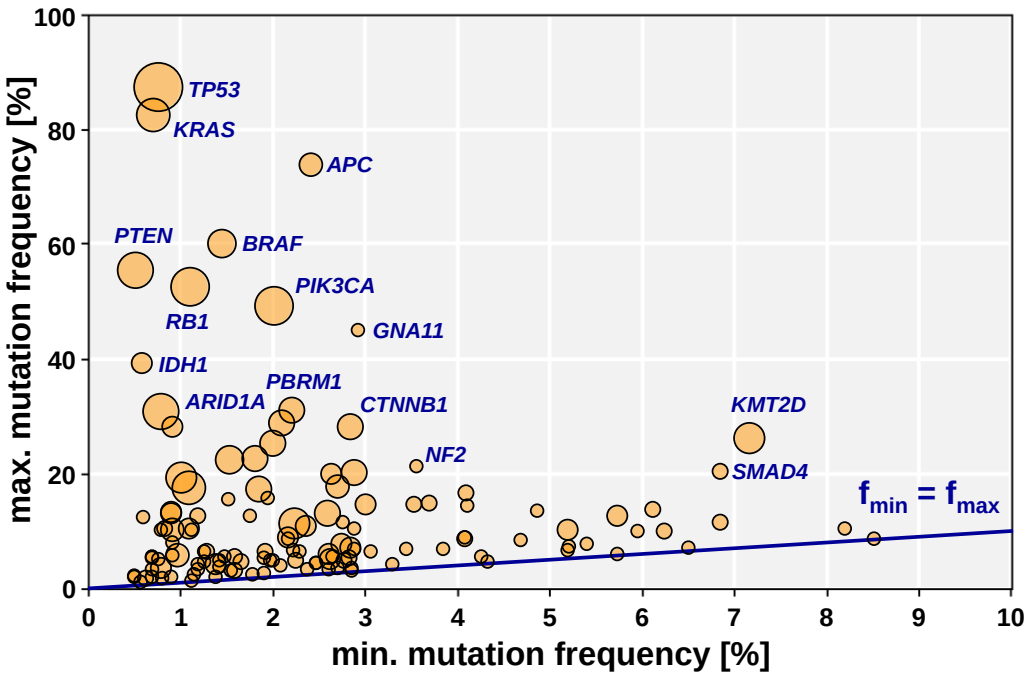
Supplementary Figure 33 | Differential expression provides an additional way to examine the driverness of mutations. The analyses shown in this figure are based on 6,648 samples for which both expression and mutation data were available. The number of samples with expression and mutation data per cancer type was as follows: bladder (*n* = 129), blood (*n* = 170), brain (*n* = 434), breast (*n* = 972), cervix (*n* = 191), cholangio (*n* = 35), colorectal (*n* = 216), endometrium (*n* = 241), gastroesophageal (*n* = 449), head & neck (*n* = 279), kidney clear (*n* = 415), kidney non-clear (*n* = 227), liver (*n* = 191), lung adeno. (*n* = 230), lung squamous (*n* = 173), lymph (*n* = 48), ovarian (*n* = 185), pancreas (*n* = 149), pheochromocytoma (*n* = 179), pleura (*n* = 82), prostate (*n* = 323), sarcoma (*n* = 245), skin (*n* = 337), testicular (*n* = 149), thymus (*n* = 119), thyroid (*n* = 400), uveal melanoma (*n* = 80).

a, b, We scanned our catalog for genes, which were significantly mutated in melanoma (significance values based on MutPanning and adjusted for multiple testing, *n* = 582 samples with mutation data) and for which mutations were associated with changes in expression. In other words, we asked whether mutated samples expressed their mutated gene (yellow) at a different level compared with wildtype samples (blue). Boxes indicate the 25%/75% interquartile range, and whiskers extend to the 5%/95%-quantiles of the expression value distribution based on 337 melanoma samples with mutation and expression data. Further, distribution medians are indicated by vertical lines. We found this differential expression pattern for several known melanoma oncogenes (**a**, top) and known melanoma tumor suppressor genes (**b**, top). Similar patterns could be observed for putative oncogenes (**a**, bottom) and putative tumor suppressor genes (**b**, bottom) in our driver gene catalog that are not traditionally associated with melanoma. This adds an additional layer of support for their role as driver genes. However, not necessarily each true driver gene is differentially expressed. **c, d,** We next studied whether gene-tumor pairs in our catalog (significance values based on MutPanning and adjusted for multiple testing, *n* = 11,873 samples with mutation data) were enriched for this differential expression pattern. For this purpose, we systematically determined for each gene and cancer type whether mutations and differential expression of the gene were correlated (*n* = 6,648 samples with mutation and expression data). In brief, we performed two tests for each gene: We determined whether the gene was differentially expressed between mutated and non-mutated samples (two-tailed Welch's t-test) and whether the expression values differed

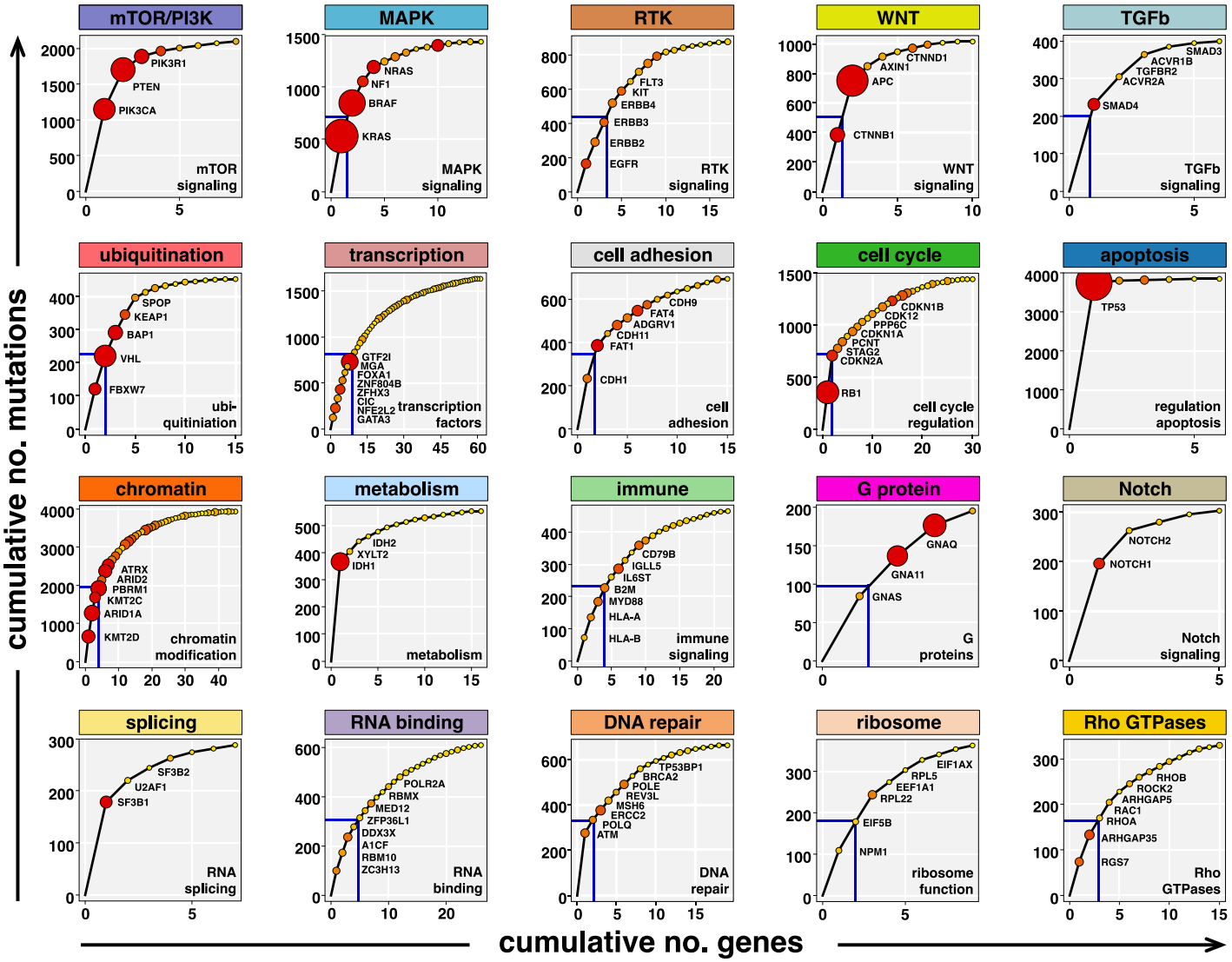
between samples with and without nonsense mutations (two-tailed Welch's t-test, particularly relevant for tumor suppressor genes). A gene was termed "differentially expressed" if one of these two tests was significant ($P < 0.05$). P-values were not adjusted for multiple testing, i.e. ~5% of the genes are expected to show differential expression by random chance. **c,** The bar graph displays the fraction of gene-tumor pairs in our catalog that showed the differential expression pattern. The genes included in these analyses were selected for three different FDR thresholds of their mutational significance (false discovery rate, significance values derived by MutPanning and adjusted for multiple testing). We computed this fraction either for all gene-tumor pairs in our catalog (blue) or for the CGC gene-tumor pairs in our catalog (yellow). As a negative control, we calculated the fraction of arbitrary gene-tumor pairs (dark gray) or arbitrary CGC gene-tumor pairs (light gray) with this pattern. This analysis revealed that driver genes in our catalog were enriched for this differential expression pattern. **d,** Similarly, we investigated whether gene-tumor pairs in our catalog that are not traditionally implicated in their respective cancer types ("additional" gene-tumor pairs) were enriched for this differential expression pattern. To this end, we computed the ratio (y-axis) between the frequency of this pattern in "additional" gene-tumor pairs (x-axis, gene-tumor pairs that were previously reported in 0, 1, or 2 studies) and the frequency of this pattern in random controls. This analysis suggests that also "additional" gene-tumor pairs are enriched for this pattern.



Supplementary Figure 34 | Characterization of the driver gene *ANAPC1*.
This figure provides exemplary evidence for the candidate cancer gene *ANAPC1*, which emerged as a significantly mutated gene (FDR =1.04x10^{-1}, significance based on MutPanning and adjusted for multiple testing) in gastroesophageal cancer (*n* = 833 samples). **a,** We aligned the peptide sequence of *ANAPC1* across six species (rows) around the positions (orange) that were recurrently altered by somatic mutations (*n* = 11873 samples). Homologous amino acids are highlighted in violet; non-homologous amino acids are colored in gray. These alignments suggest that mutational hotspots in *ANAPC1* target evolutionarily conserved protein domains. **b,** The distribution of *ANAPC1* mutations is visualized as a needle plot. For each amino acid substitution the number of samples (y-axis) is plotted against its amino acid position in the peptide sequence (x-axis). Dot colors reflect the tumor types (blue: gastroesophageal cancer, *n* = 833 samples; gray: other tumor types, *n* = 11040 samples) in which the amino acid substitution was detected. **c,** The position of the two mutational hotspots is visualized using a previously published crystal structure (PDB: 5G05). It has been reported previously that Cdk1-mediated phosphorylation of threonine 537 (T537) in ANAPC1 increases the catalytic activity of this complex by ~6-fold. Substitution of this T537 residue, which is located in the 500s loop of the WD40 repeat domain, by the nonpolar amino acid alanine may decrease the activity of the anaphase-promoting complex and thus prolong the transition from the metaphase to the anaphase. **d,** A recent study suggested that a prolonging the transition from the metaphase to the anaphase limits excessive chromosomal instability (CIN) in cancer. Using the CIN annotations from the TCGA gastroesophageal marker paper we investigated this hypothesis. In concordance with the functional results reported previously, we found that mutations in *ANAPC1* were negatively associated with chromosomal instability in our study cohort.

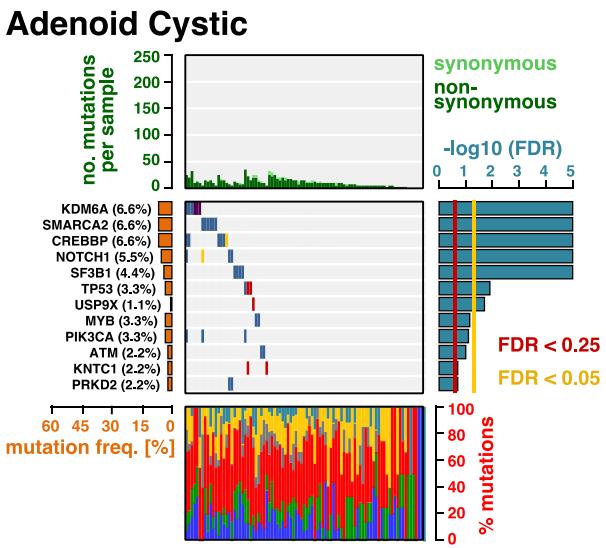


Supplementary Figure 35 | Occurrence of frequent cancer genes in rare tumor type contexts.
Besides expanding driver gene catalogs by additional genes, our study identified several known cancer genes in additional tumor type contexts. Several of these driver gene / tumor type relationships had not been captured by previous comprehensive pan-cancer catalogs. For instance, *TP53* and *PTEN* are frequently mutated in ovarian (87%, n = 316 samples) and endometrial cancers (55%, n = 327). Despite their low mutation frequencies in thyroid cancer (0.75% and 0.49%, n = 402), they both emerged as significantly mutated in thyroid cancer (FDR < 0.25, significance values based on MutPanning and adjusted for multiple testing). This suggests that these genes may have a functional role in individual thyroid cancer patients.
This plot visualizes the occurrence of several known cancer genes in rare tumor type contexts. For each gene that we detected as significantly mutant in >1 cancer type (FDR < 0.25, false discovery rate, significance values based on MutPanning and adjusted for multiple testing), we plotted its maximal (y-axis) against its minimum (x-axis) mutation frequency across cancer types. In total, these analyses are based on sequencing data of 11,873 samples. The exact number of samples included in this analysis per cancer type can be found in Extended Data Figure 5. Our catalog hence allows a systematic exploration of the occurrence of known cancer genes in rare tumor type contexts. These relationships may contribute to our understanding of the biological heterogeneity of these cancer types, such as primary or acquired therapy resistance.

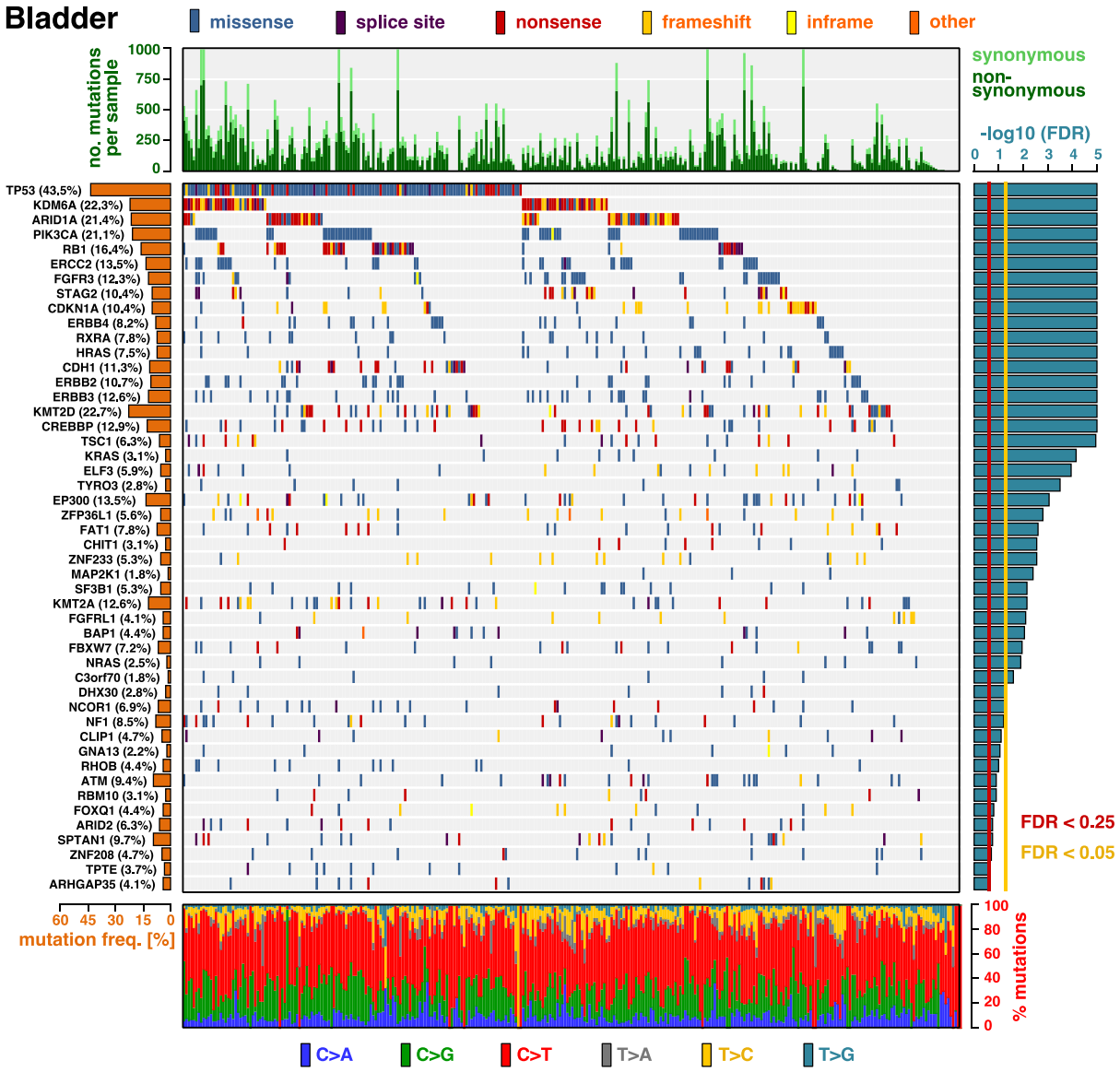


Supplementary Figure 36 | Pathways display different distributions of their mutational signal across driver genes. To dissect the contribution of different pathways to the mutational landscape of 28 cancer types, we aggregated the mutational signal across driver genes in the same pathway. We used Lorenz curves to analyze the distribution of the mutational signal across driver genes in the same pathway. In brief, we sorted the genes in decreasing order according to their mutation frequency; we then plotted the cumulative number of genes against the cumulative number of mutations they contained. The diagonal of these plots represents a uniform distribution, whereas a strong deviation from the diagonal indicates that the mutational signal is concentrated into few genes. In total, the analyses shown this figure are based on sequencing data of 11,873 samples. The exact number of samples included in this analysis per cancer type can be found in Extended Data Figure 5.

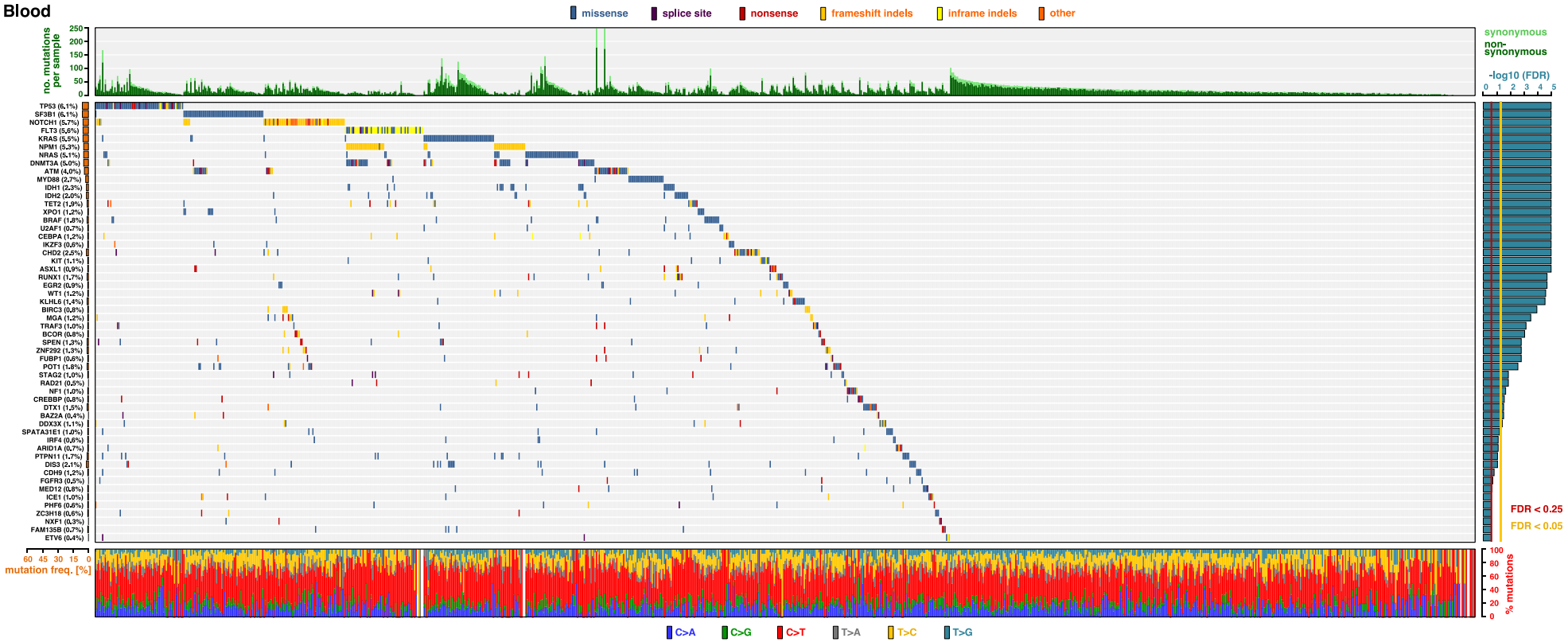
These plots revealed that a substantial fraction of the mutational signal was spread throughout a long tail of rare driver genes for most pathways. Thus, aggregating multiple rare driver genes into a comprehensive pan-cancer catalog was a major prerequisite to dissect the contribution of individual pathways to the mutational landscape of human cancer.



Supplementary Figure 37 | The landscape of driver mutations in adenoid cystic carcinomas.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 90 adenoid cystic carcinoma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.

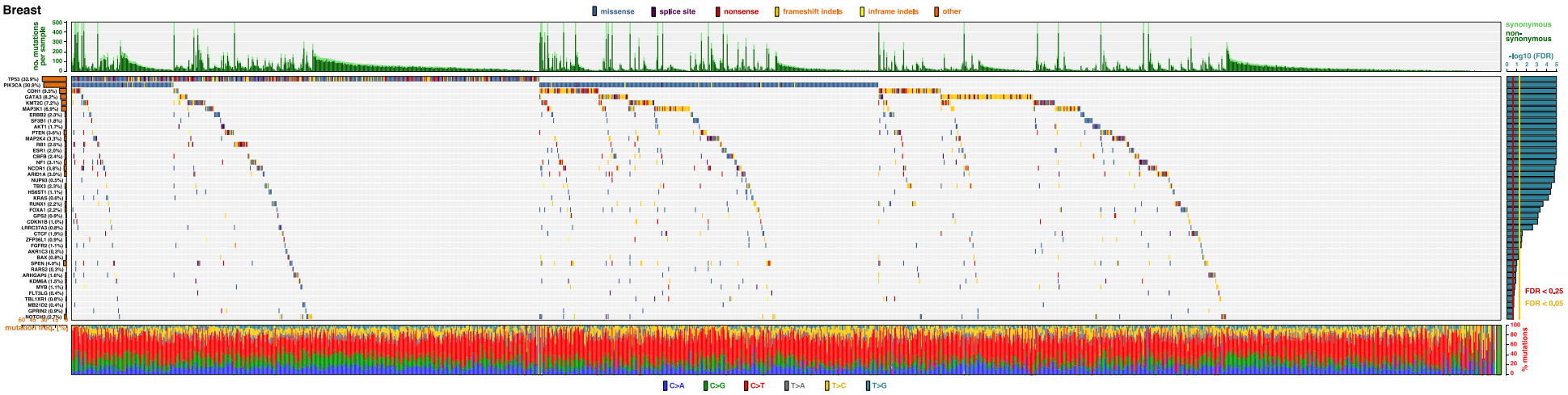


Supplementary Figure 38 | The landscape of driver mutations in bladder cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 317 bladder cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.

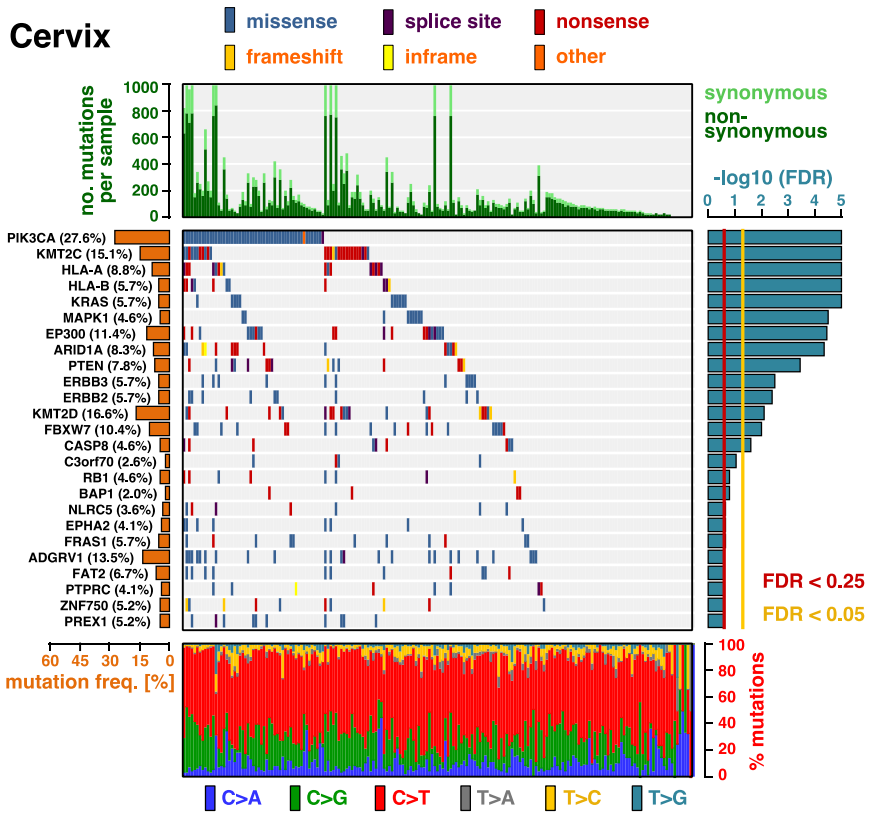


Supplementary Figure 39 | The landscape of driver mutations in hematological malignancies.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 1018 samples from hematological malignancies by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.

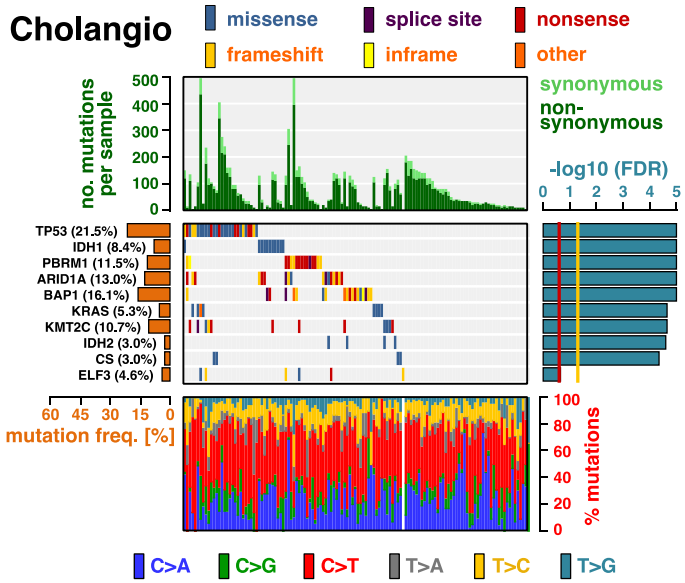
Supplementary Figure 40 | The landscape of driver mutations in brain tumors.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 760 brain tumor samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



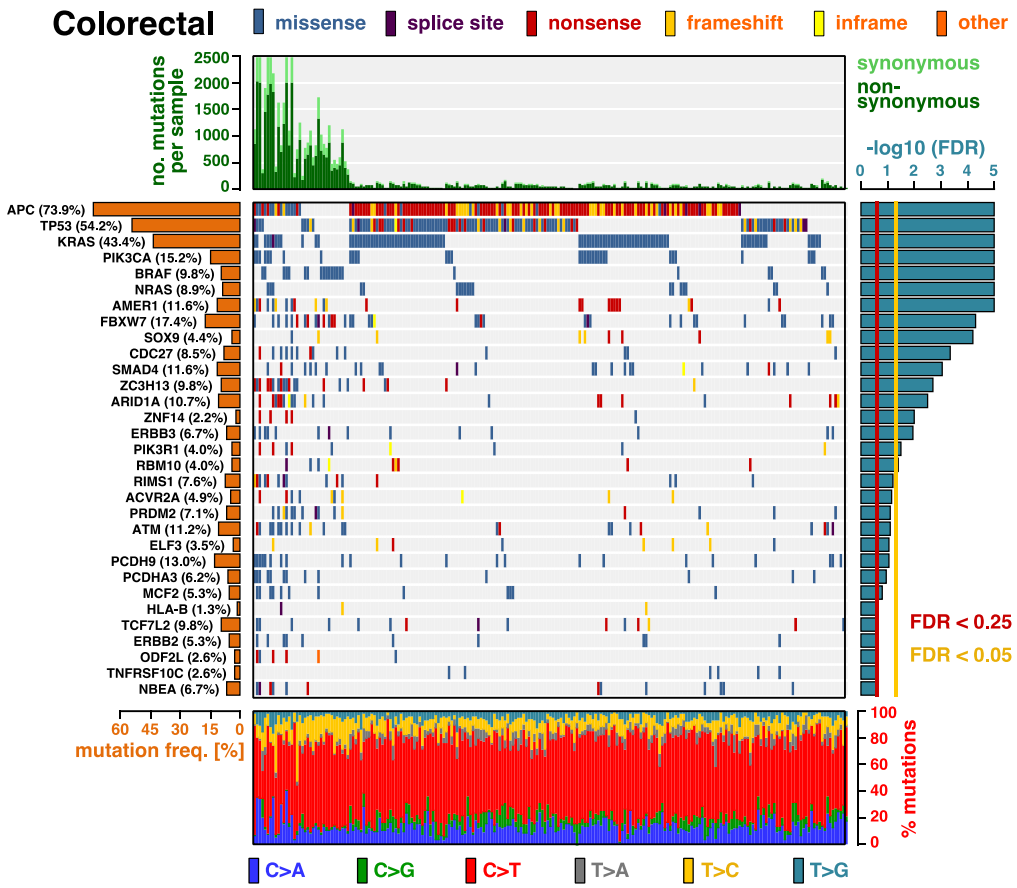
Supplementary Figure 41 | The landscape of driver mutations in breast cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 1443 breast cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



Supplementary Figure 42 | The landscape of driver mutations in cervical cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 192 cervical cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



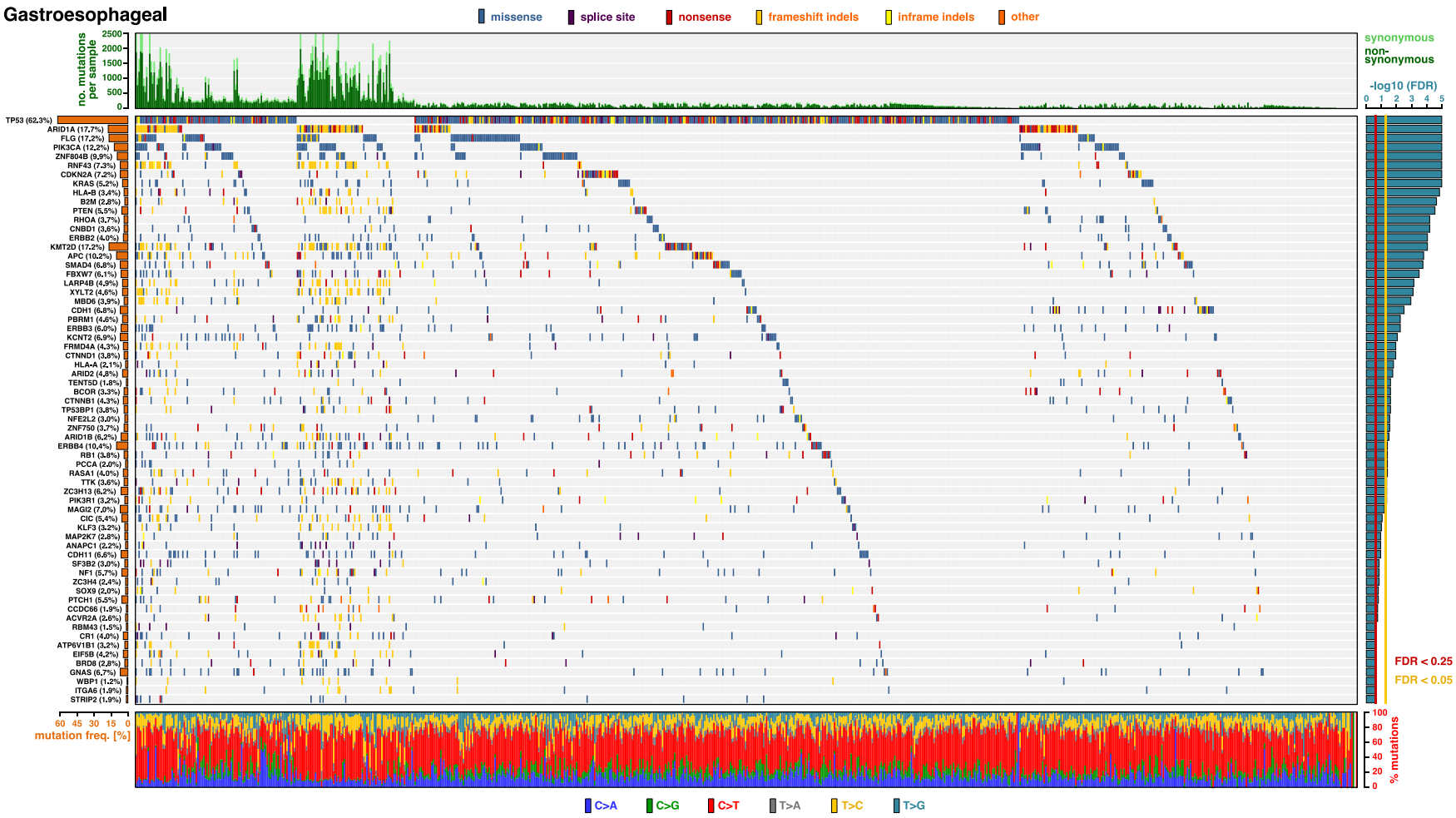
Supplementary Figure 43 | The landscape of driver mutations in cholangiocarcinoma.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 130 cholangiocarcinoma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



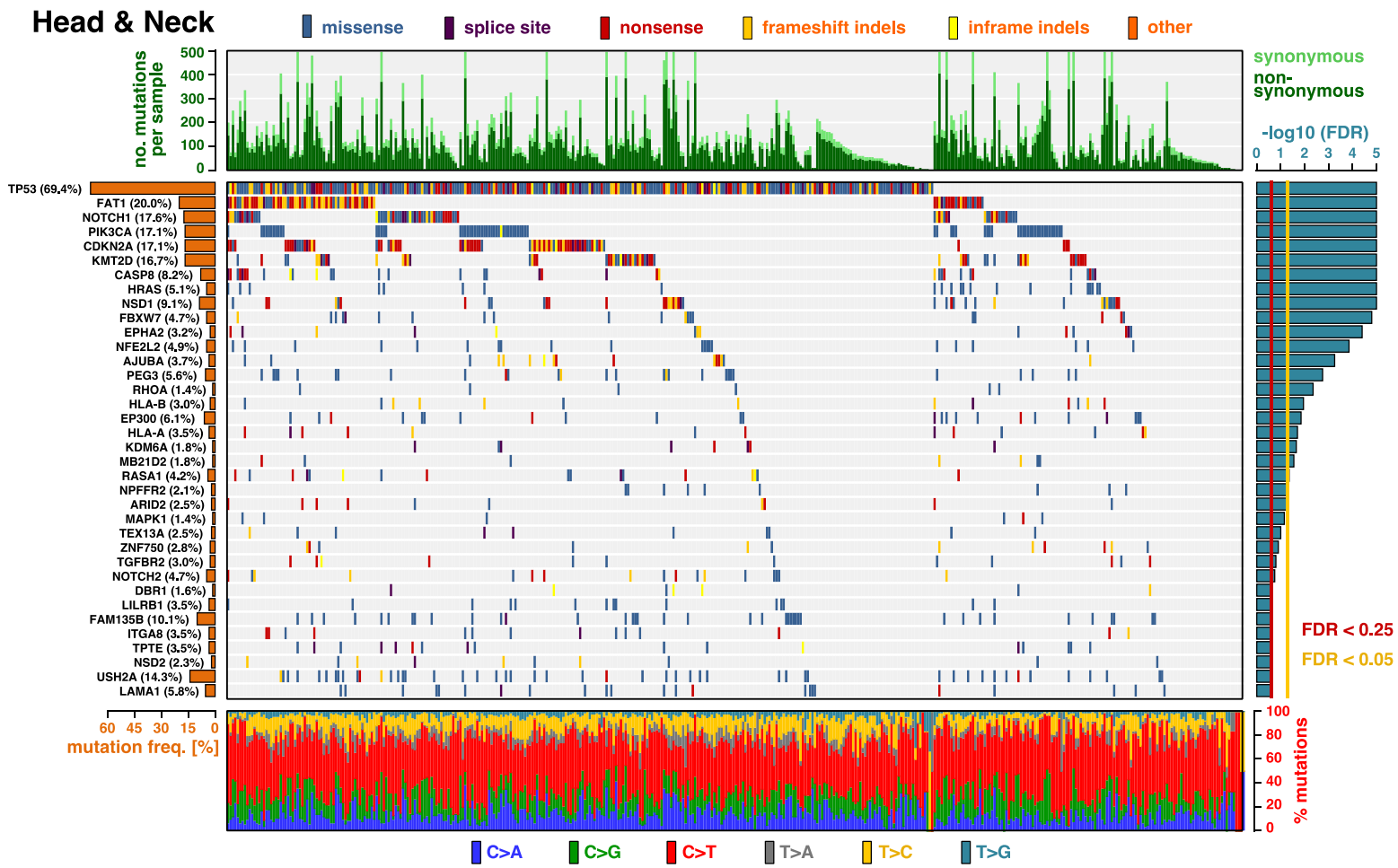
Supplementary Figure 44 | The landscape of driver mutations in colorectal cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 223 colorectal cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.

Supplementary Figure 45 | The landscape of driver mutations in endometrial carcinomas.

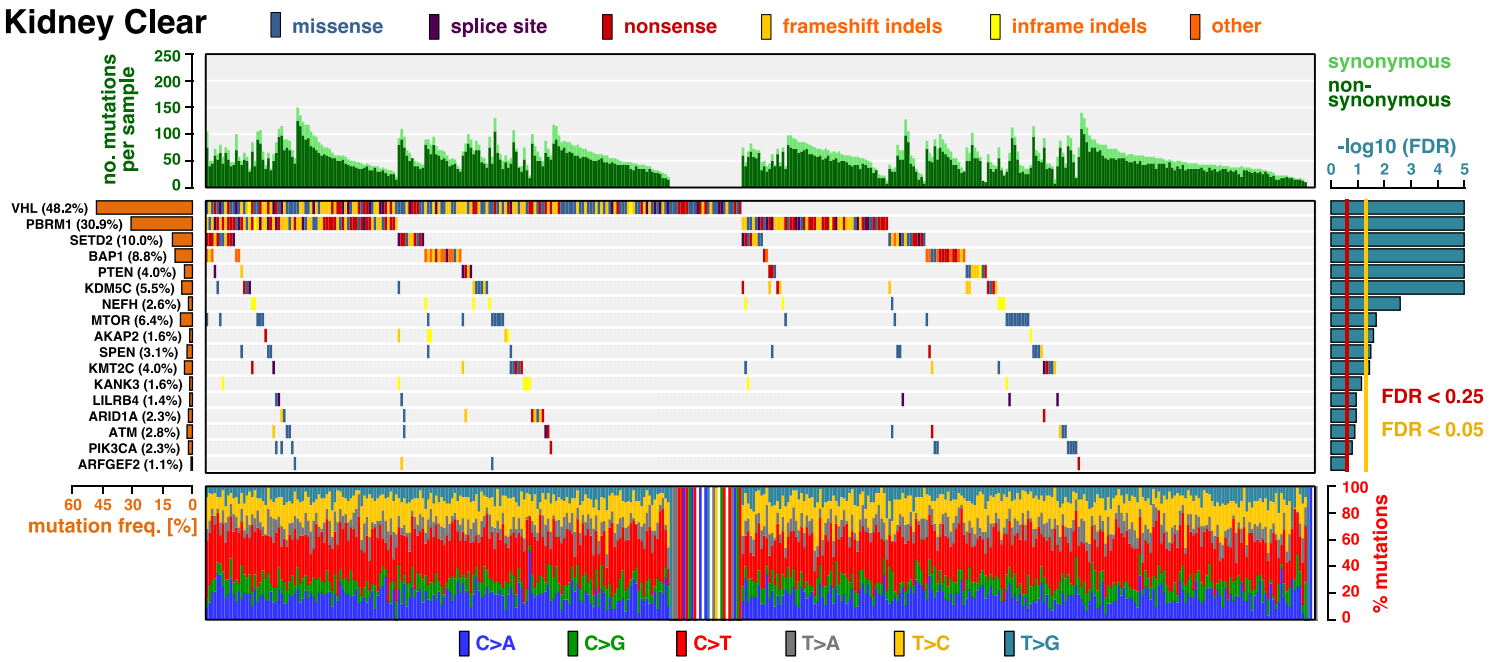
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 327 endometrial carcinoma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



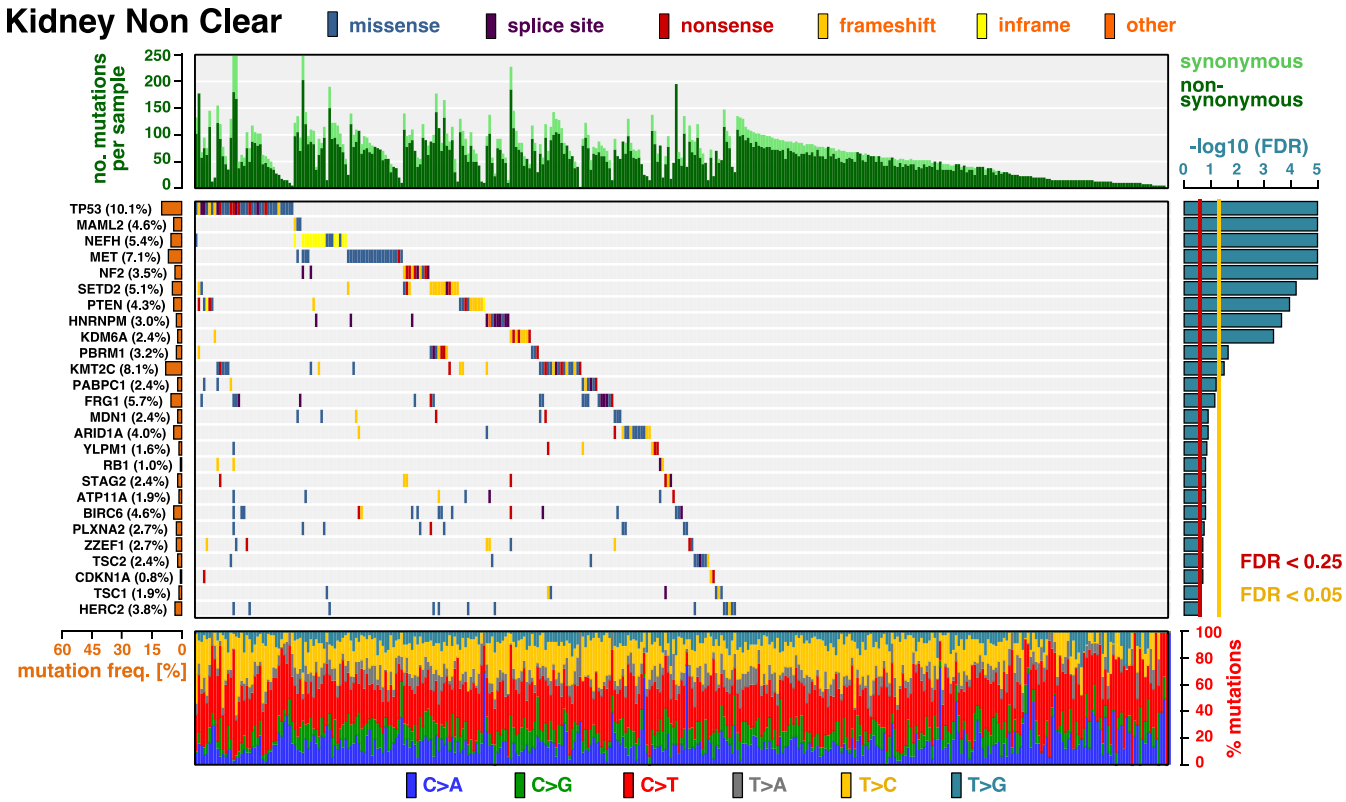
Supplementary Figure 46 | The landscape of driver mutations in gastro-esophageal cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 833 gastro-esophageal cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



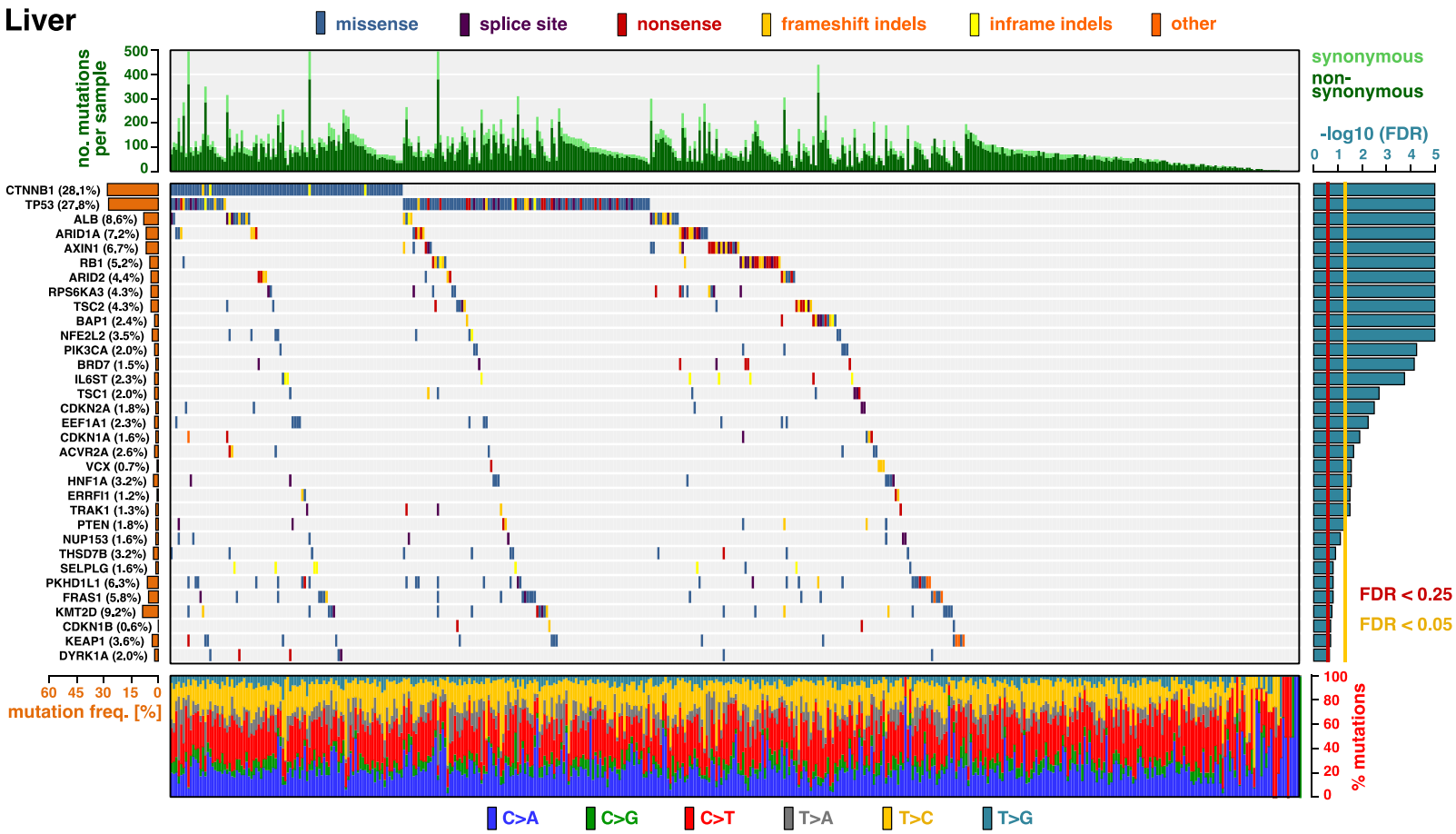
Supplementary Figure 47 | The landscape of driver mutations in head and neck carcinoma.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 425 head and neck carcinoma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



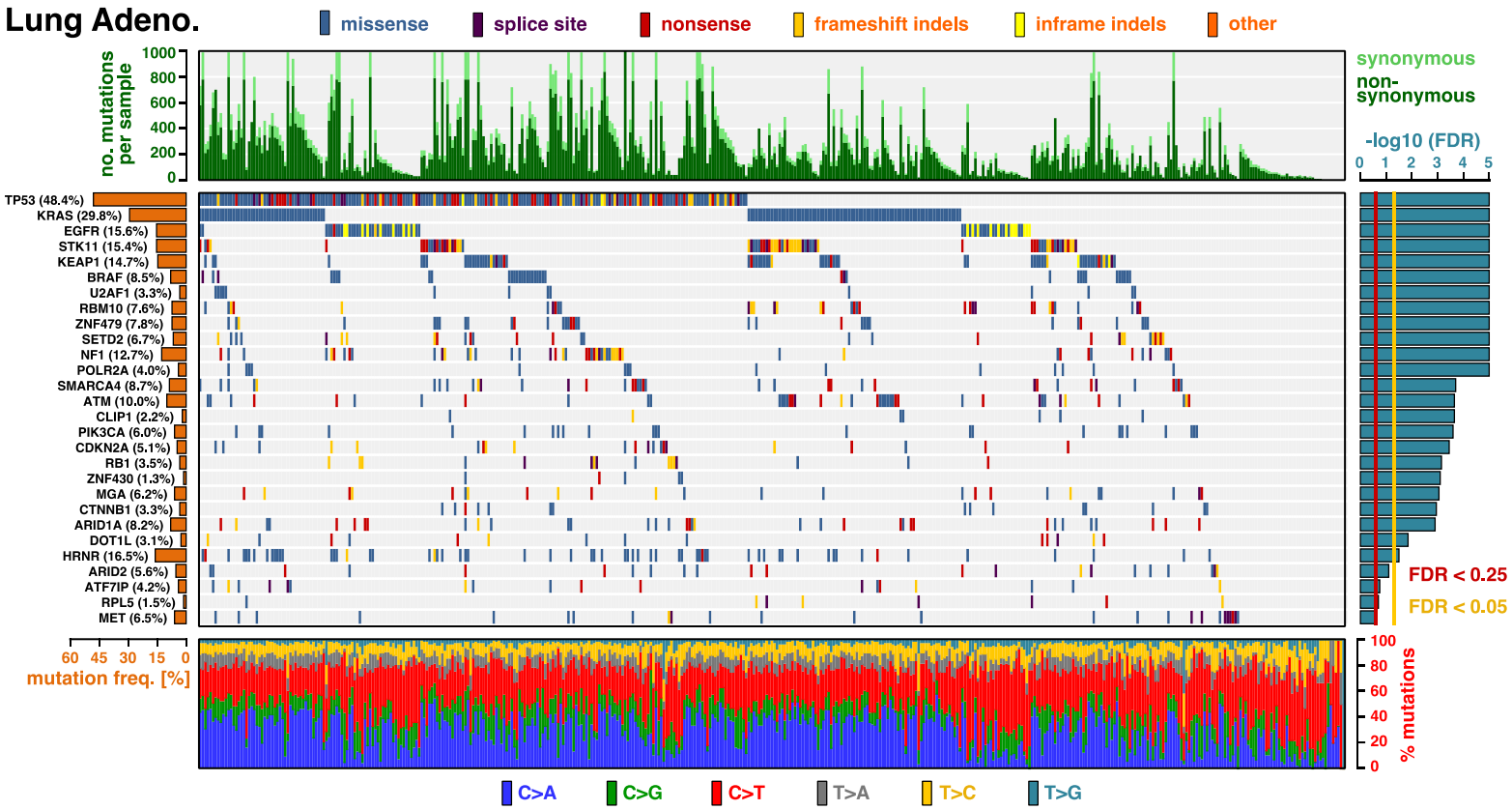
Supplementary Figure 48 | The landscape of driver mutations in clear cell kidney cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 417 clear cell kidney cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



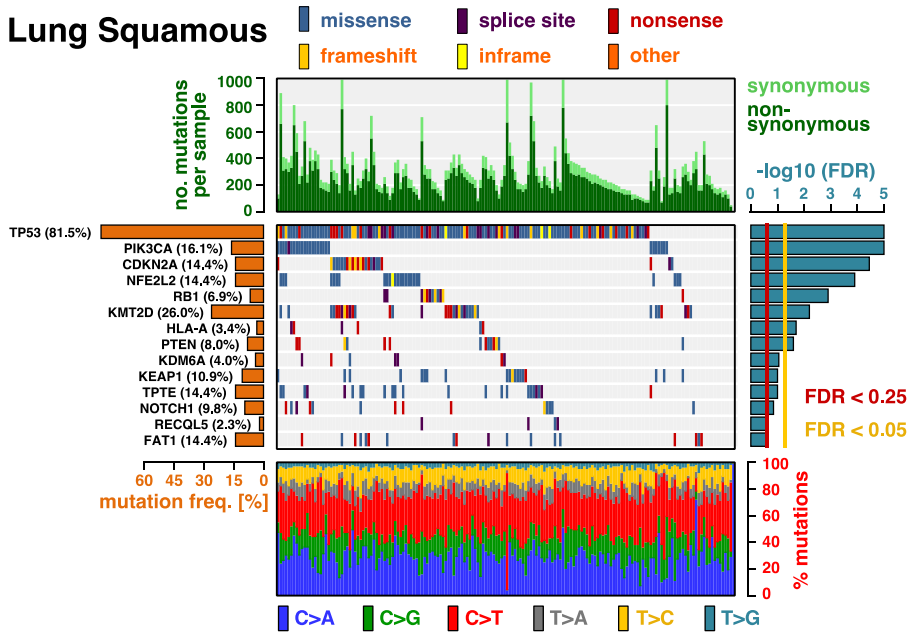
Supplementary Figure 49 | The landscape of driver mutations in non-clear cell kidney cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 366 non-clear cell kidney cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



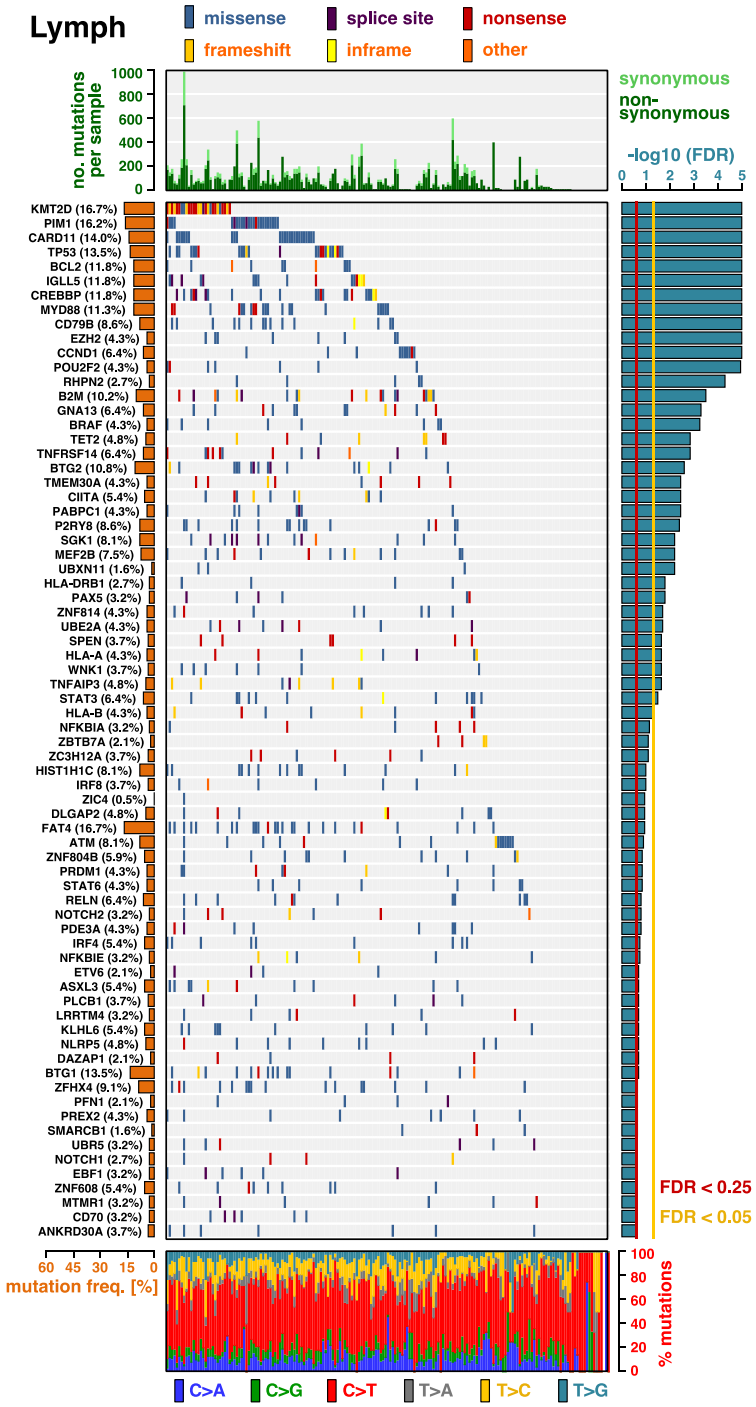
Supplementary Figure 50 | The landscape of driver mutations in hepatocellular carcinoma.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 650 hepatocellular carcinoma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



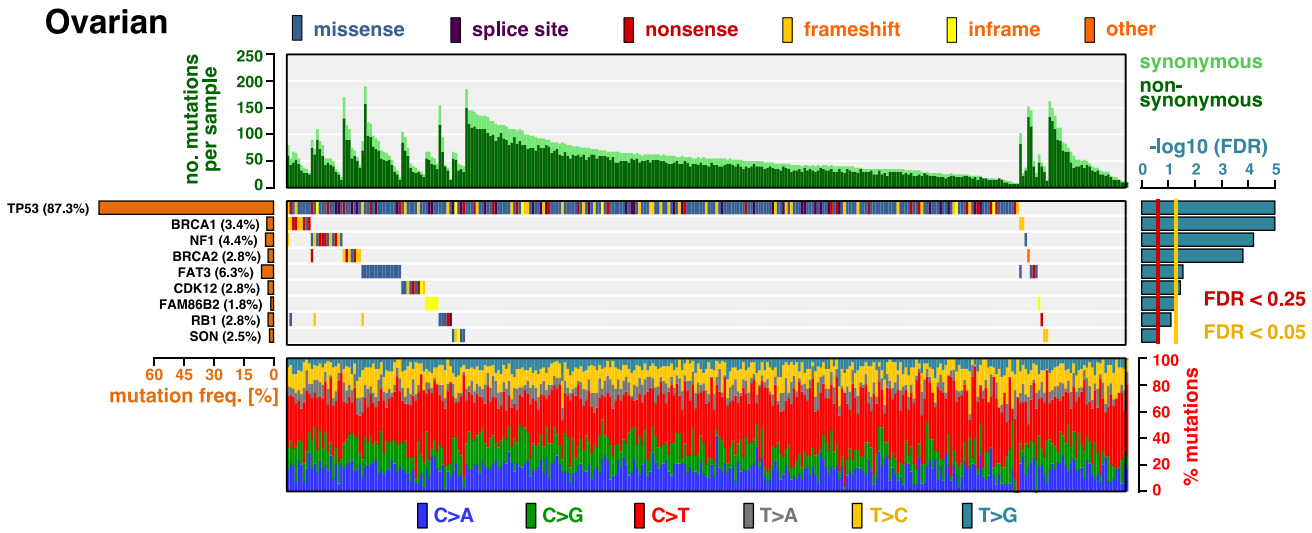
Supplementary Figure 51 | The landscape of driver mutations in lung adenocarcinoma.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 446 lung adenocarcinoma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



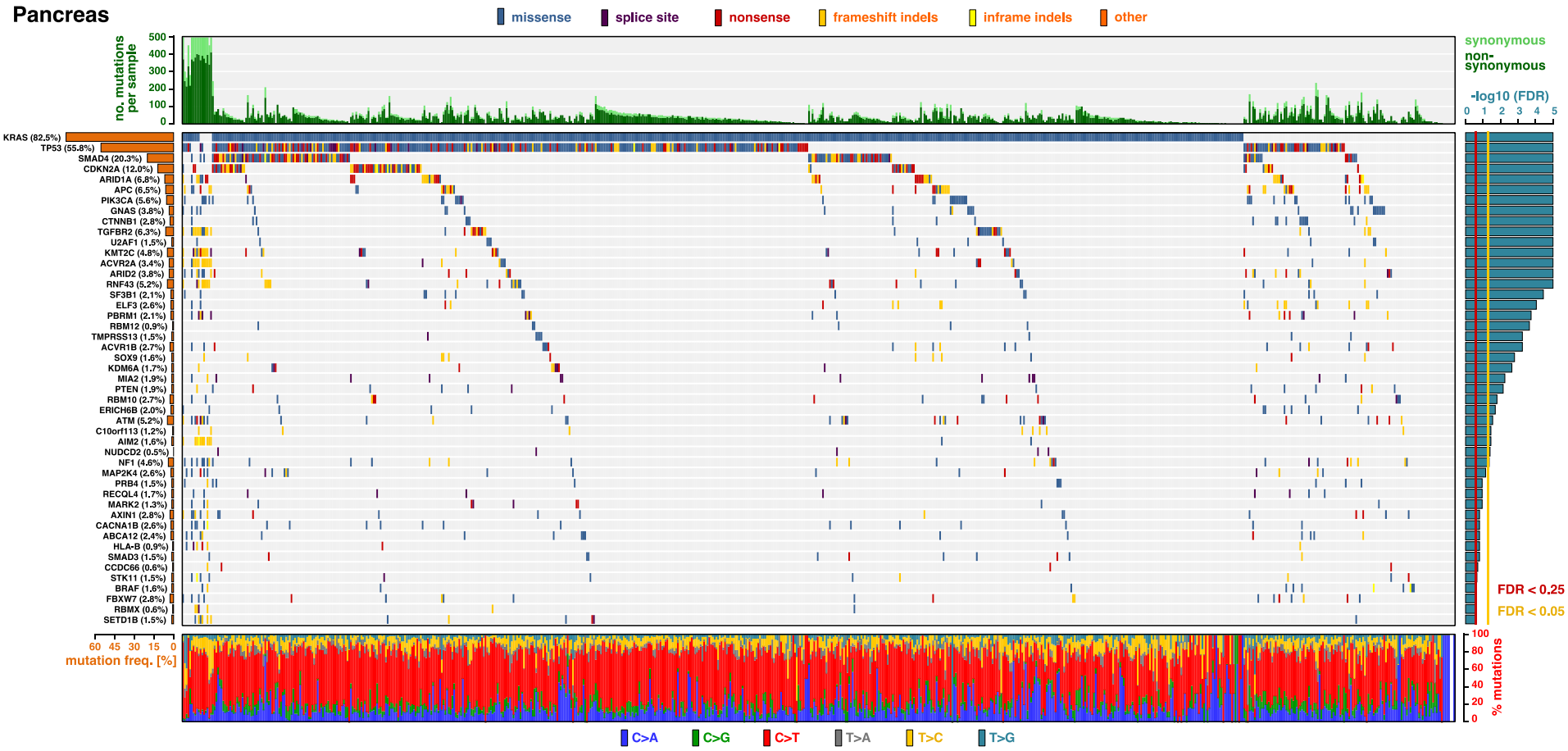
Supplementary Figure 52 | The landscape of driver mutations in squamous-cell lung cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 173 squamous-cell lung cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



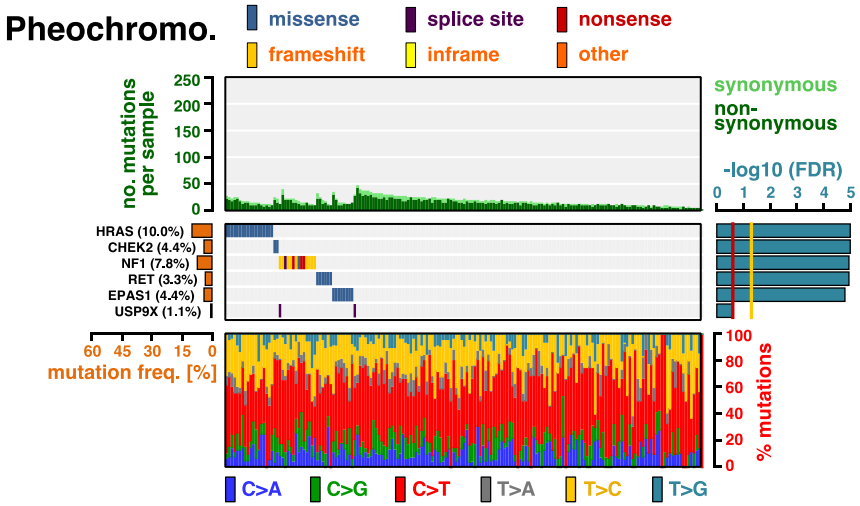
Supplementary Figure 53 | The landscape of driver mutations in lymphomas. The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 185 lymphoma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



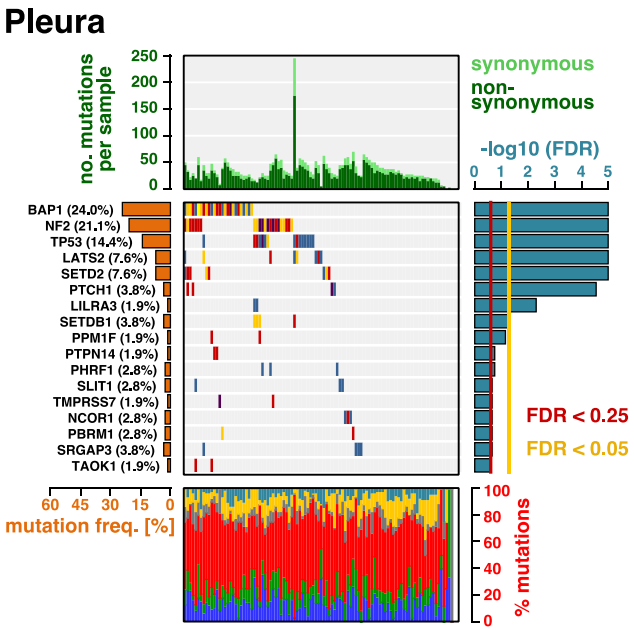
Supplementary Figure 54 | The landscape of driver mutations in ovarian cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 316 ovarian cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



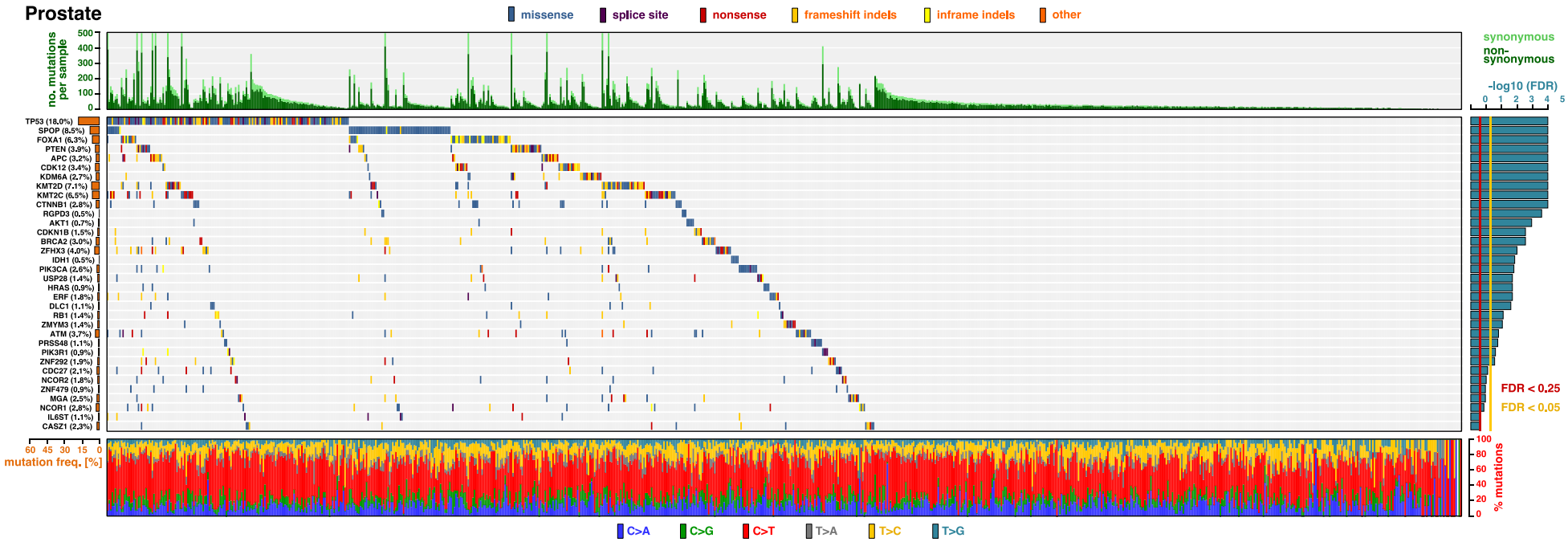
Supplementary Figure 55 | The landscape of driver mutations in pancreatic cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 729 pancreatic cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



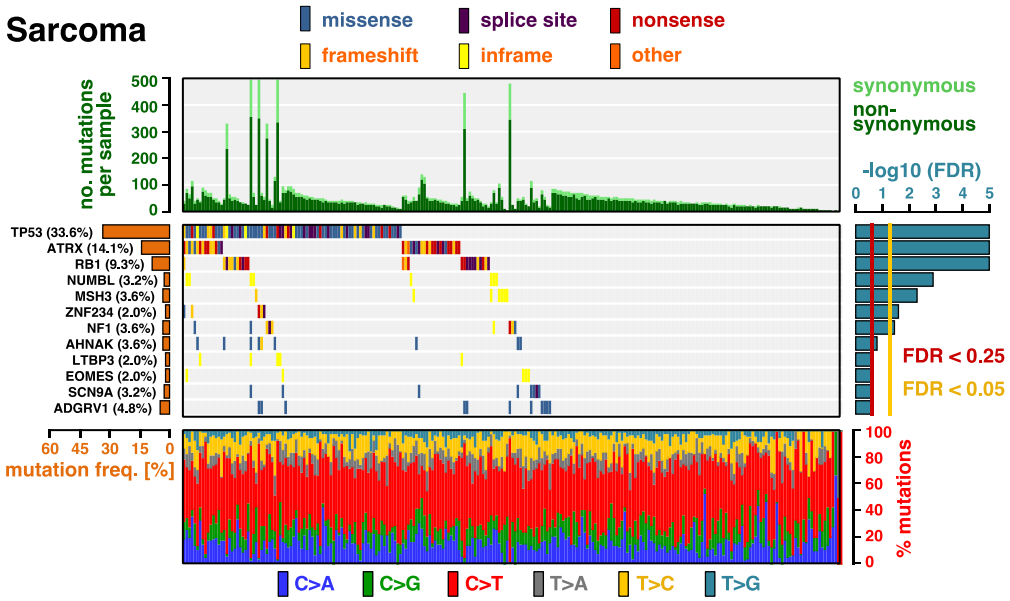
Supplementary Figure 56 | The landscape of driver mutations in pheochromocytomas and paragangliomas.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 179 pheochromocytoma and paraganglioma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at $q<0.25$; the yellow line indicates a threshold at $q<0.05$.



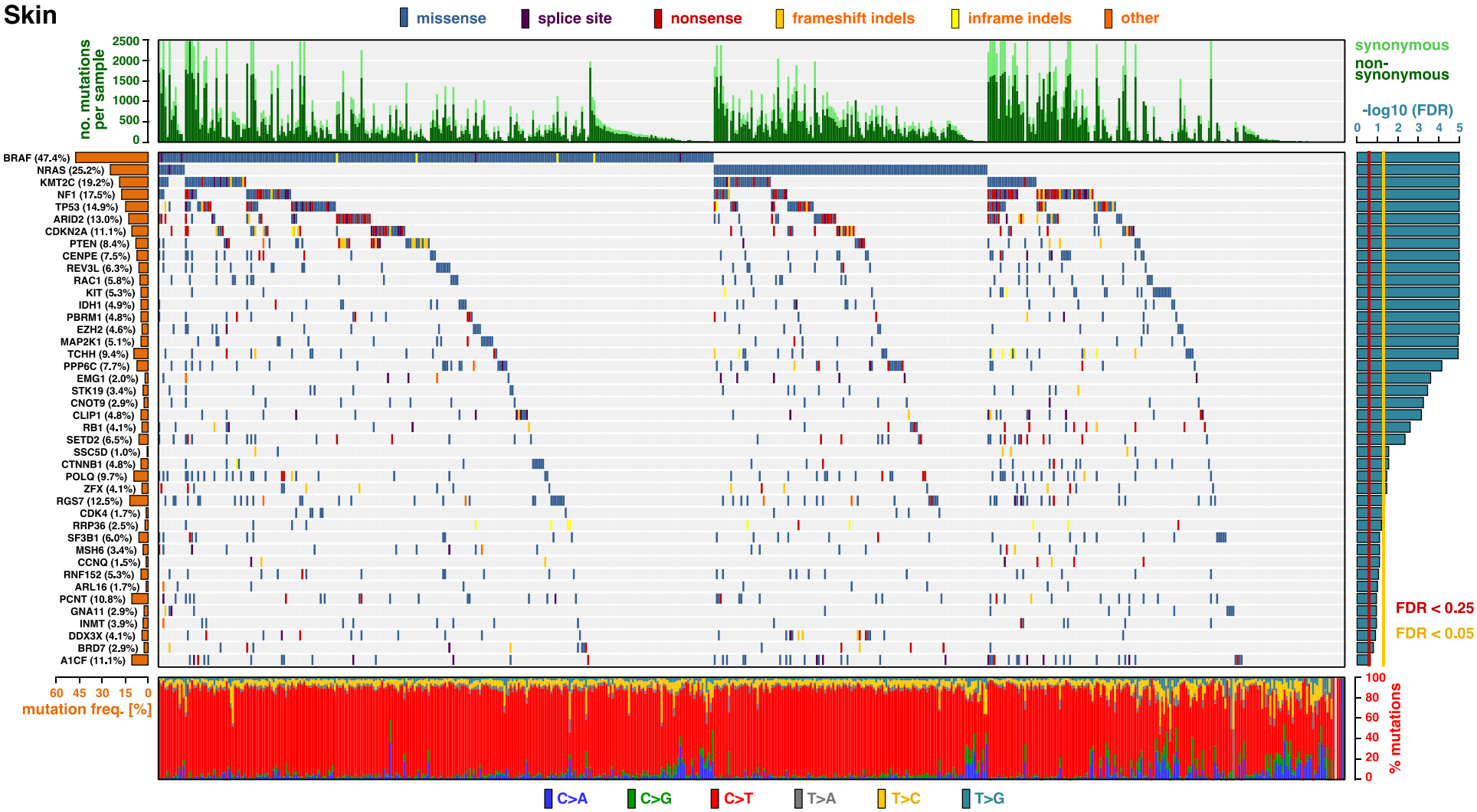
Supplementary Figure 57 | The landscape of driver mutations in mesothelioma.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 104 mesothelioma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



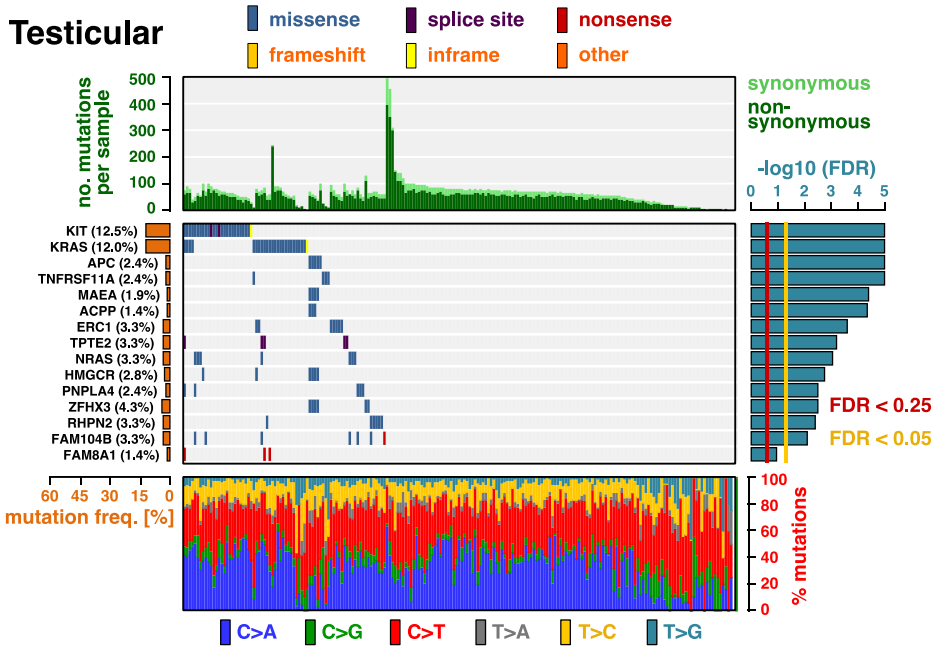
Supplementary Figure 58 | The landscape of driver mutations in prostate cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 880 prostate cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



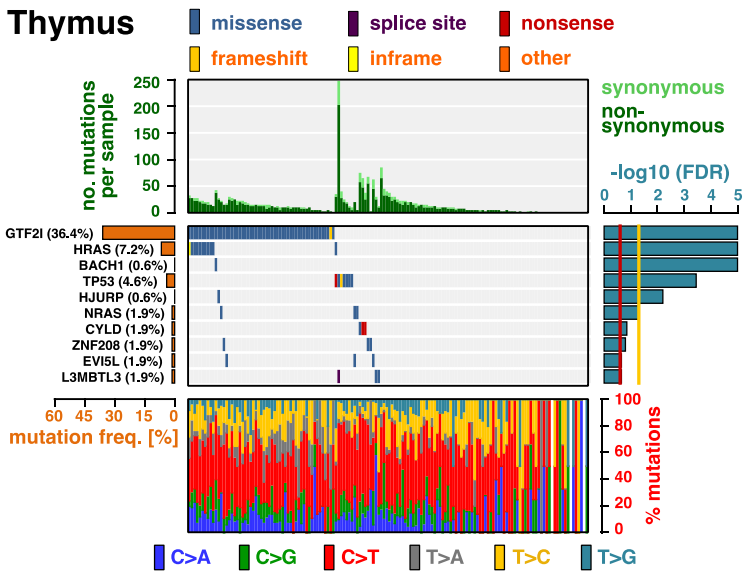
Supplementary Figure 59 | The landscape of driver mutations in adult soft tissue sarcoma.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 247 soft tissue sarcoma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



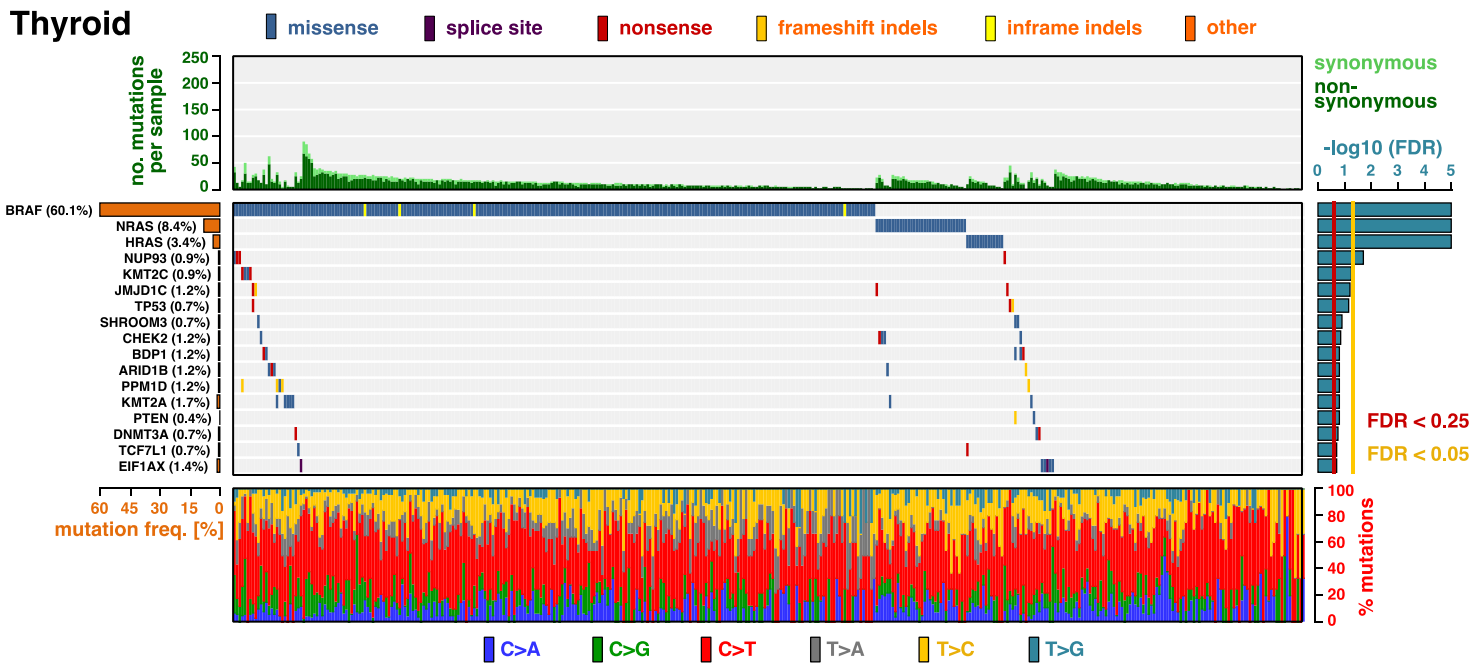
Supplementary Figure 60 | The landscape of driver mutations in melanoma.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 582 melanoma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



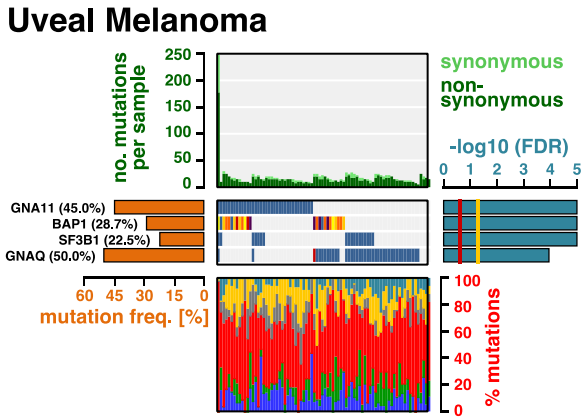
Supplementary Figure 61 | The landscape of driver mutations in testicular germ cell tumors.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 208 testicular germ cell tumor samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



Supplementary Figure 62 | The landscape of driver mutations in thymic epithelial tumors.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 151 thymic epithelial tumor samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



Supplementary Figure 63 | The landscape of driver mutations in thyroid cancer.
The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 402 thyroid cancer samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at q<0.25; the yellow line indicates a threshold at q<0.05.



Supplementary Figure 64 | The landscape of driver mutations in uveal melanoma. The color-coded matrix in the center displays the distribution of non-synonymous mutations across individual tumor samples (x-axis) in the significantly mutant genes (y-axis). Significance values were derived based on 80 uveal melanoma samples by MutPanning and adjusted for multiple testing. The bar graph on top shows the total number of mutations in each sample. Non-synonymous mutations are colored in dark green; synonymous mutations are colored in light green. The relative contribution of the six mutation types (blue: C>A, green: C>G, red: C>T, gray: T>A, yellow: T>C, cyan: T>G) to the mutational burden of each sample is shown in the bar graph below. The cancer type specific mutation frequency of each significant gene is shown on the left (orange). The False-discovery rates, computed by MutPanning (FDR, q-values), are displayed on the right. The red line indicates a cutoff at $q<0.25$; the yellow line indicates a threshold at $q<0.05$.

Supplementary Note

1. Data organization

1.1 Selection of sequencing studies

Sequencing studies were manually selected according to the following criteria:

- **whole-exome sequencing data** only, in particular no whole-genome sequencing data, no targeted sequencing data
- **patient samples** only, in particular no cell lines, mouse models or patient-derived xenograft models
- sequencing data had been aligned against the **Hg19 human reference genome**
- all tumor samples had been sequenced against a **matched normal**, and studies had filtered out germline variants from the matched normal, as well as common germline variants
- sequencing results were available as a standard **mutation annotation file** (MAF) or as a comparable format
- studies had applied filters for common sequencing artifacts, including artifacts introduced by DNA oxidation during sequencing[1], low-confidence mutations with strand bias[2], and low quality variant calls[3]

For studies where only a subset of samples satisfied all these criteria, we manually selected those samples for inclusion in this study. Further, we discarded samples that had been flagged for low quality in either of these studies. Details on how we combined sequencing data from different studies and additional filtering steps can be found in the Online Methods.

1.2 Mutation annotation files

The input data of method that we developed in this study are organized as MAF (**Mutation Annotation Format**) files. In brief, each row in these files corresponds to an individual mutation and these files contain the following columns:

Hugo_Symbol: the gene name (HUGO Gene Nomenclature Committee) in which the mutation was detected

Chromosome: the chromosome (1 to 22, X,Y) on which the mutation was observed

Start_Position: start position (basepair, bp) of the genomic alteration (Hg19)

End_Position: end position (basepair, bp) of the genomic alteration (Hg19)

Strand: + or -, depending on the transcription orientation of the gene

Variant_Classification: classification of the mutation, based on its position (5'UTR, 3'UTR, intron, intergenomic region, splice site) and its effect (inframe insertion, inframe deletion, frame shift insertion, frame shift deletion, missense mutation, nonsense mutation, silent mutation)

Variant_Type: the type of the mutation: insertions (INS), deletions (DEL), substation of a single nucleotide (SNP), substation of two nucleotides (DNP). Mutations, which did not fall into any of these categories, were discarded from the analysis.

Reference_Allele: the reference nucleotide(s) at which the mutation was observed. Empty for insertions

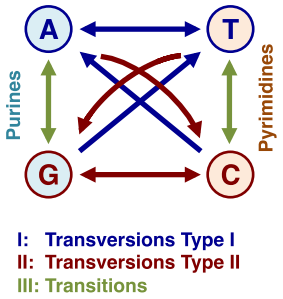
Tumor_Seq_Allele: the nucleotides that were observed as a result of the mutation in the tumor. Empty for deletions

Tumor_Sample_Barcode: the identifier of the sample. By default, we adopted the identifier from the original study. If there were name conflicts between samples from different patients, we appended the ID of the study to the end of the IDs of both samples, in order to obtain unique identifiers.

In addition to this MAF file, our algorithm expects a **sample annotation file** that lists the tumor sample barcodes of all samples together with their tumor type as well as a unique numeric index (usually 1, …, no. samples) that we used in our algorithm internally instead of the tumor sample barcode to reference to the sample.

1.3 Classification of mutations

We used two classification systems for base substitution mutations. Traditionally, six different base substitution types are discriminated (C>A, C>G, C>T, T>A, T>C, T>G), assuming mutational strand symmetry[4]. This classification system will be referred to as **base substitution types** (1, 2, 3, 4, 5, 6) in the following. As a complementary approach, mutations can be classified based on their reference nucleotide (C vs. T), as well as the substitution classes shown in the graphics. This classification system will be referred to as **base substitution classes (I, II, III)** and has been previously associated with biochemical correlates[5]. The relationship between the two classification systems is summarized in the table on the right.



	t(n,c)	n(t)	c(t)
C>A	1	C	I
C>G	2	C	II
C>T	3	C	III
T>A	4	T	I
T>C	5	T	III
T>G	6	T	II

1.4 Mutation-aligned exome files

MAF files associate each mutation with its position in the human genome. However, our statistical framework requires fast access to the nucleotide sequence context around each mutation, and it is time-intensive look up the broad sequence context in the reference exome for every mutation on demand. Hence, our method reorganized the sequencing data to a different file format (Mutation-aligned exome format, Maef) for each chromosome. In brief, each row in a Maef files corresponds to an individual genomic position and lists the indices of the samples that harbor a mutation in this position. The columns are as follows:

position: genomic position (basepair, bp) on the chromosome (Hg19)

reference nucleotide: nucleotide at the position in the Hg19 reference genome

coverage: fraction of samples for which this position had sufficient coverage

class-I substitutions: indices of samples (see below) that have a class-I substitution mutation at this position (transversions: T<>A, C>A, G>T), separated by semicolons

class-II substitutions: indices of samples (see below) that have a class-II substitution mutation at this position (transversions: C<>G, A>C, T>G) , separated by semicolons

class-III substitutions: indices of samples (see below) that have a class-I substitution mutation at this position (transitions: C<>T, G<>A) , separated by semicolons

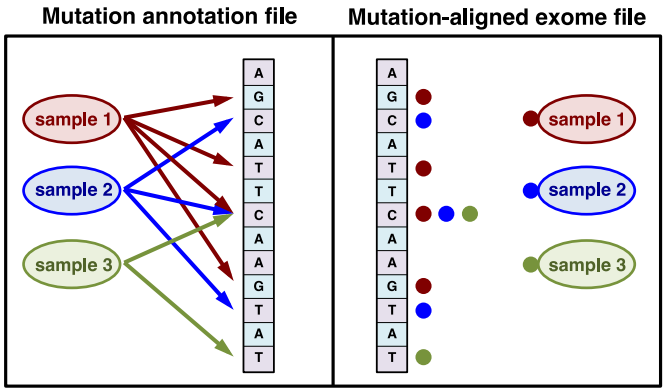
reference amino acid: if in a coding region, the amino acid that is encoded at the position in the reference genome, otherwise empty

reference amino acid number: if in a coding region, we number of the amino acid in the standard variant of the protein

class-I amino acid: if in a coding region, the mutant amino acid, caused by a class-I substitution mutation at this position, otherwise empty

class-II amino acid: if in a coding region, the mutant amino acid, caused by a class-II substitution mutation at this position, otherwise empty

class-III amino acid: if in a coding region, the mutant amino acid, caused by a class-III substitution mutation at this position, otherwise empty



While mutation annotation files are advantageous for the overview on the mutation spectrum of an individual sample, mutation-aligned exome files allow fast access to the sequence context around mutations, which is crucial for the efficient characterization of the sequence context around passenger mutations. The difference between both file formats is shown in the illustration above. We note that the sample indices used to refer to samples in the columns class-I mutations, class-II mutations, and class-III mutations are part of the sample annotation file. If available, these Maef files can be passed to our method directly. Otherwise, our method will automatically convert Maf to Maef files in the first step.

1.5 Additional data used in this study

The Hg19 human reference exome sequence and the Blat alignment tool were downloaded from the UCSC genome browser (https://genome.ucsc.edu). Genomic coordinates of exon/intron boundaries for each gene were annotated using the RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/). The coverage files of all TCGA tumor samples were obtained in a wig file format (http://gdac.broadinstitute.org/runs/stddata_ _2016_01_28/data/). Sequencing data for the TCGA validation was part of our original study cohort and obtained from gdac.broadinstitute.org (cf. above). The data for the MC3 dataset were obtained from Ellrot et al.[2] via https://gdc.cancer.gov/about-data/publications/mc3-2017 (publicly available Maf file). We then excluded samples that were hypermutated and that were flagged based on pathology exactly as described in Bailey et al. ("Data preparation" section of their paper). That way, we arrived at the same Maf file of 9,079 samples as in Bailey et al.[6] Details on the underlying variant calling and filtering pipeline can be found in Ellrot et al., 2018[2]. Details on the underlying variant calling and filtering pipeline of the TCGA dataset can be found on http://gdac.broadinstitute.org/ The source code of the CBASE model to estimate the regional fluctuation of the synonymous mutation density was downloaded (http://genetics.bwh.harvard.edu/cbase/downloads.html), rewritten in Java, and integrated into the MutPanning algorithm.

2. A composite likelihood model to quantify the mutability of genomic positions based on nucleotide contexts

The distribution of passenger mutations across the human exome depends on the regional variation of the background mutation density (mega-base scale), as well as local nucleotide contexts (single-base scale). Nucleotide contexts around passenger mutations reflect the mutational process active in a given tumor[4]. We considered nucleotide contexts in our method to identify driver mutations for two reasons: first, to account for the abundance of passenger mutations in highly mutable nucleotide contexts, which are characteristic of the underlying mutational process ("usual" nucleotide context); and secondly, to identify genomic positions in which passenger mutations are extremely rare ("unusual" nucleotide contexts), and in which mutations are a thus strong indictor of the shift of driver mutations towards functionally important positions.

The first strategy was followed by some of the more recent methods[7-9]. For instance, the dNdScv algorithm[7] aggregates context-specific mutation counts for each gene, and then compares them with a context-specific background expectation. Thereby, the dNdScv method adapts its mutational background model to the mutability of different nucleotide contexts. On the other side of the mutability spectrum, there are "unusual" nucleotide contexts. Since these "unusual" nucleotide contexts strongly deviate from the contexts of the underlying mutational process, passenger mutations occur rarely in these "unusual" nucleotide contexts. In contrast, driver mutations are localized towards functionally important positions, which are not necessarily surrounded by a particular nucleotide context. Hence, an increased number of mutations in these "unusual" nucleotide contexts is a strong indicator of driver mutations shifted towards functionally important positions.

Characteristic nucleotide contexts are typically identified by counting the number of mutations per nucleotide context[4]. Quantifying the absence of passenger mutations (i.e. determining how much a nucleotide context deviates from the characteristic nucleotide context of the underlying mutational process) is intrinsically more challenging, due to the sparsity of passenger mutations in these "unusual" nucleotide contexts. However, this is important to identify positions in the cancer genome that are rarely hit by passenger mutations and that hence serve as a sensitive indicator for the shift of driver mutations towards functionally important positions.

Here, we established a composite likelihood model to address this open problem and quantify the degree of "unusualness" for each possible nucleotide context. In brief, the composite likelihood model compares each nucleotide around a genomic position with the expected nucleotide(s) around passenger mutations. The model accounts for each nucleotide in the surrounding context by a multiplicative factor: if the factor is >1, the surrounding nucleotide increases the similarity to the characteristic nucleotide context of passenger mutations; if the multiplicative is <1, the nucleotide decreases the similarity to the passenger mutation context. As such, the composite likelihood model is well suited to identify both "usual" and "unusual" nucleotide contexts at both ends of the mutability spectrum. Thereby, it can be used to calibrate

the background model to the abundance of passenger mutations in "usual" nucleotide contexts, and to focus the statistical model on mutations in "unusual" nucleotide contexts, in which passenger mutations are rare.

Besides the identification of unusual nucleotide contexts, the composite likelihood model bears two additional advantages. First, the number of possible nucleotide contexts increases exponentially with the number of flanking nucleotides. The composite likelihood model considers the effect of each flanking nucleotide as a multiplicative factor, so that the number of parameters in the composite likelihood model increases linearly with the number of flanking nucleotides. For instance for heptanucleotide contexts, there are 24,576 ($=6 \cdot 4^6$) possible nucleotide contexts, but only 149 ($=5 + 6 \cdot 4 \cdot 7$) parameters in the composite likelihood model. In accordance with previous studies, we observed that a substantial fraction of the context-specificity was mediated by the broad nucleotide context outside the trinucleotide context. The composite likelihood model allows considering the effect of broad nucleotide contexts, which could not be considered when modeling each possible nucleotide context individually, owing to sparsity of mutations per nucleotide context.

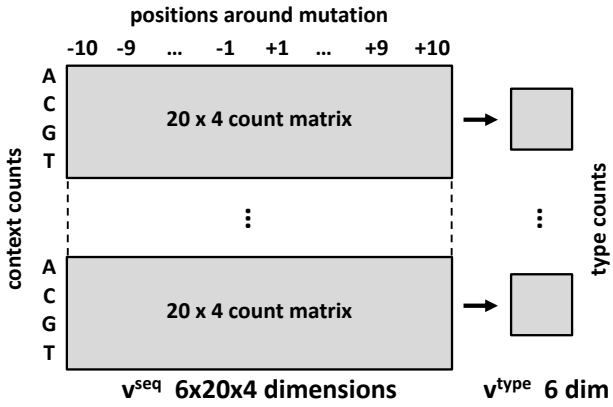
Secondly, determining the mutation probability of each possible nucleotide context individually would likely result in an overfitting of the background mutation signal. For instance, frequent hotspot mutations, such as *KRAS* G12, would likely increase the mutation probability of a single and highly specific nucleotide context in the background model. Overfitting of this outlier would lead to extinction of the signal, mediated by nucleotide contexts, since the background expectation would be high in this particular hotspot / nucleotide context. The composite likelihood model avoids overfitting of the background distribution, since different contexts are not modeled individually but rather decomposed into a product of likelihood factors. Hence, the effect of isolated outliers spreads over several similar nucleotide contexts rather than a single nucleotide context, thus only marginally changing the calibration of the background model.

In this section, we describe the major steps of the composite likelihood model: (i) Samples are grouped into clusters with similar context-dependent mutation patterns, (ii) context-dependent mutation probabilities are factorized into likelihood factors, and (iii) the composite likelihood score is determined for each position in the exome to quantify its mutability based on its surrounding nucleotide context. A summary of this section can be found in the Online Methods (section "Statistical analyses to identify driver genes").

2.1 Bayesian hierarchical clustering of samples

The first step of the composite likelihood model is to identify clusters of samples that share similar context-dependent distribution patterns of their passenger mutations. The primary purpose of this clustering step is to aggregate enough mutations in each cluster to accurately characterize the background signal in each cluster. Particularly for samples with low mutation counts this is an important prerequisite. Our statistical model to detect the excess of mutations in unusual nucleotide context was based on a multinomial distribution (cf. section 3). Hence, we

clustered samples based a Bayesian hierarchical clustering approach[10] that uses a probabilistic distance metrics based on the multinomial distribution. That way, we made our clusters compatible with the statistics used in the subsequent steps.



Bayesian hierarchical clustering approaches have been used previously[10-13]. Their basic idea is to agglomerate data that likely came from the same distribution into the same cluster, but separate data that likely came from different distributions into different clusters. Hence, by choosing a Bayesian Hierarchical Clustering approach, we aimed to generate clusters that were compatible with the multinomial distribution of our statistical model, i.e. that samples that likely came from the same underlying multinomial distribution were gathered in the same cluster.

We first counted for each sample the number of mutations of each base substitution type (C>A, C>G, C>T, T>A, T>C, T>G). For each sample, we summarized these counts into a **type count vector** $v^{\text{type}} \in \mathbb{N}^6$. Each element v_t^{type} in this vector corresponds to the number of base substitutions of type $t \in \{1, \dots, 6\}$. To capture the nucleotide context around passenger mutations, we further counted for each sample the nucleotides that occurred in a 20-nucleotide window around its mutations. We summarized these counts into the nucleotide **context count vector** $v^{\text{seq}} \in \mathbb{N}^{6\times 20\times 4}$. Each element $v_{t,p,n}^{\text{seq}}$ in this vector denotes the count of nucleotide $n \in \{\text{A,C,G,T}\}$ in position $p \in [-10; 10]\backslash\{0\}$ around mutations of type $t \in \{1, \dots, 6\}$.

We next defined a similarity metrics, which quantifies the similarity between the mutation distribution patterns of two samples V (count vectors $v^{\text{type}}, v^{\text{seq}}$) and W (count vectors $w^{\text{type}}, w^{\text{seq}}$) based on their count vectors. More precisely, we quantified the similarity between two count vectors $v, w \in \mathbb{N}^l$ by examining whether updating a distribution prior x by w made the observation of v more likely, compared with the original prior x. Specifically, given a distribution prior $x \in (\mathbb{R}^+)^l$, we determined the probability $P(v \mid x)$ of observing v under prior x using a Dirichlet-multinomial distribution. The Dirichlet-multinomial distribution is a compound distribution, i.e. the observation v is drawn from a multinomial distribution with a probability vector that is randomly drawn from a Dirichlet distribution with prior x. The Dirichlet component inside the compound distribution reflects the uncertainty about the event probabilities of the multinomial distribution (conjugate prior), given the count vector of an independent sample. The multinomial distribution is part of the statistical framework that we used to derive mutational significance (cf. section 3).

As a measure for the similarity between v and w, we determined whether updating the prior x by w would increase the probability to observe v by computing the following Bayes factor:

$$\begin{aligned}\theta(v,w,x) &\coloneqq P(v \mid w+x)/P(v \mid x)\\ &= \left[\frac{\Gamma(|v|+1)\cdot\Gamma(|w+x|)}{\Gamma(|v+w+x|)}\cdot\prod_{1\le k\le l}\frac{\Gamma(v_k+w_k+x_k)}{\Gamma(v_k+1)\cdot\Gamma(w_k+x_k)}\right]\\ &\quad\cdot\left[\frac{\Gamma(|v|+1)\cdot\Gamma(|x|)}{\Gamma(|v+x|)}\cdot\prod_{1\le k\le l}\frac{\Gamma(v_k+x_k)}{\Gamma(v_k+1)\cdot\Gamma(x_k)}\right]^{-1}\\ &= \frac{\Gamma(|v+x|)\cdot\Gamma(|w+x|)}{\Gamma(|x|)\cdot\Gamma(|v+w+x|)}\cdot\prod_{1\le k\le l}\frac{\Gamma(v_k+w_k+x_k)\cdot\Gamma(x_k)}{\Gamma(v_k+x_k)\cdot\Gamma(w_k+x_k)}\\ &= \gamma(|v|,|w|,|x|)^{-1}\prod_{1\le k\le l}\gamma(v_k,w_k,x_k)\end{aligned}$$

v_k denotes the k-th element of the l-dimensional vector v, $|v| \coloneqq \sum_k |v_k|$, and

$$\gamma(a,b,c) \coloneqq \frac{\Gamma(a+b+c)\cdot\Gamma(c)}{\Gamma(a+c)\cdot\Gamma(b+c)}$$

for $a,b,c \in \mathbb{N}$. As $\gamma(a,b,c)$ is symmetrical between a and b, the similarity metrics $\theta(v,w,x)$ is symmetrical between vectors v and w. The prior x of the similarity metrics $\theta(v,w,x)$ is typically derived from a frequency vector $f \in (\mathbb{R}^+)^l$ with $|f| = 1$, e.g. frequencies in the reference genome. Hence, we analogously denoted $\theta(v,w,f) \coloneqq \theta(v,w,\min(|v|,|w|)\cdot f)$, given a frequency vector f. We next applied the similarity metrics $\theta(v,w,f)$ to the type and context count vectors of individual samples. Given two samples with type count vectors $v^{\text{type}}, w^{\text{type}} \in \mathbb{N}^6$, we defined a similarity metrics comparing their base substitution types as

$$\theta_{\text{type}}(v^{\text{type}}, w^{\text{type}}) \coloneqq \theta(v^{\text{type}}, w^{\text{type}}, f^{\text{type}})$$

We chose the prior of the similarity metrics θ_{type} as $f^{\text{type}} \coloneqq \frac{1}{6}\cdot\underbrace{(1,\dots,1)}_{6}$, which reflects our a-priori assumption to observe the six base substitution types C>A, C>G, C>T, T>A, T>C, and T>G at the same frequency. We note that this a-priori assumption holds true in the exome only, where the nucleotides A, C, G, and T occur at a similar frequency of ~25%. To generalize this model to noncoding regions, the imbalance between pyrimidines and purines would need to be integrated into this prior.

Further, we compared the context dependence of the passenger mutation distribution patterns of two samples based on their sequence context count vectors $v^{\text{seq}}, w^{\text{seq}} \in \mathbb{N}^{20\times6\times4}$ by

$$\theta_{\text{seq}}(v^{\text{seq}}, w^{\text{seq}}) \coloneqq \prod_{\substack{p=-10,\dots,10\\ p\neq0}}\prod_{t=1,\dots6}\theta(v^{\text{seq}}_{t,p}, w^{\text{seq}}_{t,p}, f^{\text{ref}}_{n(t),p})$$

where $n(t) \coloneqq \begin{cases}\text{C} & t = 1,2,3\\ \text{T} & t = 4,5,6\end{cases}$ denotes the reference nucleotide of the base substitution type $t \in \{1,\dots,6\}$.

The prior $f^{\text{ref}}_{n(t),p}$ of the similarity metrics θ_{seq} denotes the exonic frequency vector, i.e. each element $f^{\text{ref}}_{n,p,n'}$ denotes the relative fraction of nucleotide n' at position p around nucleotide n for $n,n' \in \{\text{A,C,G,T}\}$. That way, we obtained two similarity metrics between samples that compared their base substitution types (θ_{type}) and the nucleotide sequence context around mutations (θ_{seq}), respectively.

Analogously, we compared the similarity between two disjoint clusters $\mathcal{C}, \mathcal{C}' \subseteq \mathcal{S}$ of samples. We summed up the count vectors across samples in these two clusters, and determined the similarity between these cumulative count vectors as follows:

$$\theta_{\text{type}}(\mathcal{C}, \mathcal{C}') \coloneqq \theta_{\text{type}}\left(\sum\nolimits_{s \in \mathcal{C}} v_s^{\text{type}}, \sum\nolimits_{s' \in \mathcal{C}'} v_{s'}^{\text{type}}\right)$$

$$\theta_{\text{seq}}(\mathcal{C}, \mathcal{C}') \coloneqq \theta_{\text{seq}}\left(\sum\nolimits_{s \in \mathcal{C}} v_s^{\text{seq}}, \sum\nolimits_{s' \in \mathcal{C}'} v_{s'}^{\text{seq}}\right)$$

Note that if both clusters only contain only one sample, we obtained our original similarity metrics θ_{type} and θ_{seq} on individual samples.

We used these similarity metrics to group samples into groups of samples with similar passenger mutation distribution patterns. For this purpose, we started with one cluster per individual sample. In each clustering step, we then merged the two clusters $(\mathcal{C}^*, \mathcal{C}'^*)$ with

$$(\mathcal{C}^*, \mathcal{C}'^*) \coloneqq \arg\max_{\mathcal{C} \neq \mathcal{C}'} \theta_{\text{type}}(\mathcal{C}, \mathcal{C}') > 1$$

or

$$(\mathcal{C}^*, \mathcal{C}'^*) \coloneqq \arg\max_{\mathcal{C} \neq \mathcal{C}'} \theta_{\text{seq}}(\mathcal{C}, \mathcal{C}') > 1$$

if $\max_{\mathcal{C} \neq \mathcal{C}'} \theta_{\text{type}}(\mathcal{C}, \mathcal{C}') > 1$ or $\max_{\mathcal{C} \neq \mathcal{C}'} \theta_{\text{seq}}(\mathcal{C}, \mathcal{C}') > 1$, respectively. We terminated the clustering procedure as soon as $\theta_{\text{type}}(\mathcal{C}, \mathcal{C}') \leq 1$ or $\theta_{\text{seq}}(\mathcal{C}, \mathcal{C}') \leq 1$ for all $\mathcal{C} \neq \mathcal{C}'$. That way, we obtained clusters $\mathcal{C}_1^{\text{type}}, \ldots, \mathcal{C}_K^{\text{type}}$ and $\mathcal{C}_1^{\text{seq}}, \ldots, \mathcal{C}_L^{\text{seq}}$, containing samples with similar base substitution types ($\mathcal{C}_i^{\text{type}}$) and similar mutational sequence contexts ($\mathcal{C}_i^{\text{seq}}$), respectively. To obtain clusters that combined both aspects, we defined

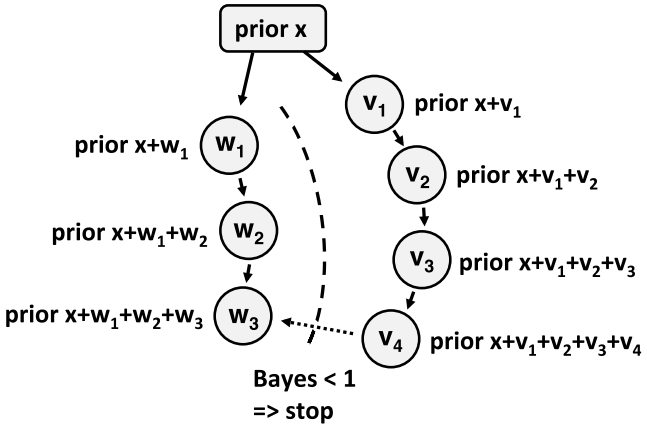
$$\mathcal{C}_{i,j} \coloneqq \mathcal{C}_i^{\text{type}} \cap \mathcal{C}_j^{\text{seq}}$$

for $1 \leq i \leq K$ and $1 \leq j \leq L$.

As every iteration of the hierarchical clustering procedure requires $O(|\mathcal{S}|^2)$ comparisons, the running time would be $O(|\mathcal{S}|^3)$, where $|\mathcal{S}|$ denotes the total number of samples. This is relatively slow when applied to all samples in our study cohort in parallel. Hence, we first performed hierarchical clustering separately on each cancer type in parallel, thus reducing the running time from $O(|\mathcal{S}|^3)$ to $O(|\mathcal{S}'|^3)$, where $|\mathcal{S}'|$ denotes the size of the subcohort per cancer type. Then we continued the hierarchical clustering procedure on the clusters derived from all cancer types with the restriction that clusters from the same cancer type, were not merged in this step (clusters that were separated based on a smaller dataset should also remain separated).

We note that these similarity metrics depend on the total number of mutations, since high mutation counts reduce the variance in the Dirichlet component of the distribution. Therefore pairs of samples with high mutation counts either have a highly positive or a highly negative value according to this similarity measure. Therefore, the Bayesian hierarchical clustering step first clusters according to the samples with high mutation counts since these samples contain the highest signal. These samples serve as the “seeds” of the clusters, around which samples with fewer mutations are clustered in subsequent steps. In contrast to the mutation count per sample, sample purity had no impact on the clustering procedure.

Finally, we removed all clusters that had less than 3 samples or 1,000 mutations. We introduced these minimum thresholds so that in the downstream analyses the distribution underlying each cluster could be estimated with enough certainty (convergence of the Dirichlet-multinomial distribution against the Multinomial distribution for large enough prior), thereby simplifying our statistical model. Samples s that were not assigned to any cluster after this procedure were assigned to the cluster with $\arg\max_{\mathcal{C}} \theta_{\text{type}}(s,\mathcal{C}) \cdot \theta_{\text{seq}}(s,\mathcal{C})$. In this way, we obtained clusters of samples that were calibrated to the multinomial distribution used in the subsequent steps of our model.



2.2 Factorization of the composite likelihood to characterize the mutability of genomic positions based on surrounding nucleotide contexts

We next used the clusters from the Bayesian hierarchical clustering step to aggregate enough mutations to characterize the background signal. For each cluster, we established a composite likelihood model to characterize its context-dependent passenger mutation distribution pattern. We start by describing the factorization of the substitution type-specific mutational likelihoods. Given a cluster $\mathcal{C}$ of samples, denote by $v^{\text{type}} \coloneqq \sum_{s\in\mathcal{C}} v_s^{\text{type}}$ and $v^{\text{seq}} \coloneqq \sum_{s\in\mathcal{C}} v_s^{\text{seq}}$ the sum of their type and sequence context count vectors, respectively. In addition to the classical six different base substitution types t (1: C>A, 2: C>G, 3: C>T, 4: T>A, 5: T>C, 6: T>G), we classified mutations by their reference nucleotide (C vs. T) and three mutation classes c (I: transversions C>A, T>A, II: transversions C>G, T>G, III: transitions C<>T).

For each cluster $\mathcal{C}$, we then defined the following likelihood ratios

$$\lambda_{\text{C}}^{\mathcal{C}} \coloneqq \frac{2}{|v^{\text{type}}|} \cdot \left(v_1^{\text{type}} + v_2^{\text{type}} + v_3^{\text{type}}\right)$$

$$\lambda_{\text{T}}^{\mathcal{C}} \coloneqq \frac{2}{|v^{\text{type}}|} \cdot \left(v_4^{\text{type}} + v_5^{\text{type}} + v_6^{\text{type}}\right)$$

for the reference nucleotides C and T, as well as

$$\lambda_{\text{I}}^{\mathcal{C}} \coloneqq \frac{3}{|v^{\text{type}}|} \cdot \left(v_1^{\text{type}} + v_4^{\text{type}}\right)$$

$$\lambda_{\text{II}}^{\mathcal{C}} \coloneqq \frac{3}{|v^{\text{type}}|} \cdot \left(v_2^{\text{type}} + v_6^{\text{type}}\right)$$

$$\lambda_{\mathrm{III}}^{\mathcal{C}} \coloneqq \frac{3}{|v^{\mathrm{type}}|} \cdot \left(v_3^{\mathrm{type}} + v_5^{\mathrm{type}}\right)$$

for the base substitution classes I, II, and III, as well as

$$\lambda_i^{\mathcal{C}} \coloneqq \frac{6}{|v^{\mathrm{type}}|} \cdot v_i^{\mathrm{type}}$$

for the base substitution types 1, …, 6. Here, v_k^{type} denotes the kth element of the vector v^{type} and $|v| \coloneqq \sum_k |v_k|$ denotes the vector length.

These likelihood ratios describe the ratio between the observed mutation counts (e.g., $v_1^{\mathrm{type}} + v_4^{\mathrm{type}}$) and the expected mutation counts (e.g., $|v^{\mathrm{type}}|/3$). The normalization coefficients 2 and 3 are based on the fact that all nucleotides occur at ~25% frequency in the human exome, in parallel to our definition of the prior f^{type} used for hierarchical clustering.

Based on these definitions, we factorized the composite likelihood as

$$\lambda_t^{\mathcal{C}} = \lambda_{n(t)}^{\mathcal{C}} \cdot \lambda_{c(t)}^{\mathcal{C}}$$

where $n_0 = n(t) = \begin{cases} \mathrm{C} & t = 1,2,3 \\ \mathrm{T} & t = 4,5,6 \end{cases}$ denotes the reference nucleotide and $c(t) = \begin{cases} \mathrm{I} & t = 1,4 \\ \mathrm{II} & t = 2,6 \\ \mathrm{III} & t = 3,5 \end{cases}$ denotes the base substitution class. We note that our classification of base substitutions (C vs. T; transitions vs. transversions) differs from the standard classification of base substitutions (C>A, C>G, C>T, T>A, T>C, T>G). This difference is important to keep in mind for the correct factorization of composite likelihood products. For instance, MSI-related mutation signatures[4] contain both high fractions of C>T and T>C mutations, but low fractions of A>T and G>C mutations. In our model, C>T and T>C base substitutions are factorized as $\lambda_{\mathrm{C}} \cdot \lambda_{\mathrm{III}}$ and $\lambda_{\mathrm{T}} \cdot \lambda_{\mathrm{III}}$, respectively, but not as $\lambda_{\mathrm{C}>} \cdot \lambda_{>\mathrm{T}}$ and $\lambda_{\mathrm{T}>} \cdot \lambda_{>\mathrm{C}}$.

We next describe the factorization of the composite likelihood of the surrounding nucleotide context. We defined the likelihood ratios for the sequence context as

$$\lambda_{t,p,n}^{\mathcal{C}} \coloneqq \frac{v_{t,p,n}^{\mathrm{seq}}}{v_t^{\mathrm{type}} \cdot f_{n(t),p,n}^{\mathrm{ref}}}$$

where $f_{n,p,n'}^{\mathrm{ref}}$ denotes the frequency of nucleotide n' around nucleotide n at position p in the human exome and $n(t) \coloneqq \begin{cases} \mathrm{C} & t = 1,2,3 \\ \mathrm{T} & t = 4,5,6 \end{cases}$ denotes the reference nucleotide of base substitution type t. Hence, $\lambda_{t,p,n}^{\mathcal{C}}$ reflects the ratio between the observed count $v_{t,p,n}^{\mathrm{seq}}$ of nucleotide n in position p around base substitutions of type t, and the corresponding expected count $v_t^{\mathrm{type}} \cdot f_{n(t),p,n}^{\mathrm{ref}}$ of nucleotide n in position p.

We next defined a likelihood ratio $\lambda_{t,(n_{-k},\ldots,n_l)}^{\mathcal{C}}$ that compares the observed number of mutations of type t occurring in a context $(n_{-k}, \ldots, n_l)$ and the corresponding expected number of mutations of type t occurring in that context. More precisely, given a base substitution type $t \in \{1, \ldots, 6\}$ and nucleotide context $(n_{-k}, \ldots, n_l) \in \{\mathrm{A,C,G,T}\}^{k+l+1}$ with $n_0 = n(t) = \begin{cases} \mathrm{C} & t = 1,2,3 \\ \mathrm{T} & t = 4,5,6 \end{cases}$ the reference nucleotide of base substitution type t, we denote by $v_{t,(n_{-k},\ldots,n_l)}^{\mathrm{seq}}$ the number of mutations of type t surrounded by context $(n_{-k}, \ldots, n_l)$ and by $f_{n_0,(n_{-k},\ldots,n_l)}^{\mathrm{ref}}$ the frequency that a nucleotide n_0 is

surrounded by the nucleotide context $(n_{-k}, \dots, n_l)$ in the human reference exome. As all nucleotides occur at ~25% frequency in the human exome, $f^{\text{ref}}_{n(t),(n_{-k},\dots,n_l)} \approx 4 \cdot f^{\text{ref}}_{(n_{-k},\dots,n_l)}$, where $f^{\text{ref}}_{(n_{-k},\dots,n_l)}$ denotes the relative frequency of the nucleotide sequence $(n_{-k}, \dots, n_l)$ in the human exome compared to all possible nucleotide sequences of length $k + l + 1$. Therefore, we expected $4 \cdot v^{\text{type}}_t \cdot f^{\text{ref}}_{(n_{-k},\dots,n_l)}$ mutations of type t to be surrounded by context $(n_{-k}, \dots, n_l)$ and defined

$$\lambda^{\mathcal{C}}_{t,(n_{-k},\dots,n_l)} := \frac{v^{\text{seq}}_{t,(n_{-k},\dots,n_l)}}{4 \cdot v^{\text{type}}_t \cdot f^{\text{ref}}_{(n_{-k},\dots,n_l)}}$$

The composite likelihood model decomposes this ratio as the following product

$$\lambda^{\mathcal{C}}_{t,(n_{-k},\dots,n_l)} = \prod_{\substack{-k \le p \le l \\ p \neq 0}} \lambda^{\mathcal{C}}_{t,p,n_p}$$

which integrates the effect of each flanking nucleotide on the local mutation probability of a position as a multiplicative factor.

We note that although our composite likelihood model considers the effect of each nucleotide as a multiplicative factor, it does not assume that the underlying mutation signatures are symmetric. For instance, the POLE-associated mutation signature[4] is highly asymmetric: C>T mutations predominantly occur in T_G contexts, but not in other T_N or N_G contexts. The reason that the composite likelihood model can capture asymmetrical mutation signatures is that each flanking nucleotide either increases (factor > 1) or decreases (factor < 1) the mutational likelihood. For instance, for C>T mutations in the POLE signature the composite likelihood model would attach strongly decreasing effects to 5' nucleotides A, C, and G, as well as strongly decreasing effects to 3' nucleotides A, C, and T. That way, only the likelihood factor product of T_G contexts would be assigned a substantial mutational likelihood. Hence, allowing decreasing effects enables capturing asymmetrical effects. An exemplary calculation of the composite likelihood is visualized in Supplementary Figure 7.

2.3 Quantification of the mutability of genomic positions based on their surrounding nucleotide context

Using the composite likelihood factorization, we finally quantified the mutability of each genomic position q in the human exome conditional on its surrounding nucleotide context. Given a cluster $\mathcal{C}$ of samples and a position q with reference nucleotide n_0, we defined

$$\lambda^{\mathcal{C}}_{q,c} := \begin{cases} \lambda^{\mathcal{C}}_{n_0} \cdot \lambda^{\mathcal{C}}_{c} \cdot \prod\limits_{\substack{-10 \le p \le 10 \\ p \neq 0}} \lambda^{\mathcal{C}}_{t,p,n_p} & \text{for } n_0 = \text{C, T} \\ \lambda^{\mathcal{C}}_{\overline{n_0}} \cdot \lambda^{\mathcal{C}}_{c} \cdot \prod\limits_{\substack{-10 \le p \le 10 \\ p \neq 0}} \lambda^{\mathcal{C}}_{t,-p,\overline{n_p}} & \text{for } n_0 = \text{A, G} \end{cases}$$

where $(n_{-k}, \dots, n_k)$ denotes the nucleotide context in a $2k$-mer window around position q.

Given a subcohort $\mathcal{S}' \subseteq \mathcal{S}$, we then define

$$\lambda_{q,c}^{\mathcal{S}'} \coloneqq \frac{c_q}{n_{\mathcal{S}'}^{\text{mut}}} \cdot \sum_{i=1,\ldots,k} n_{\mathcal{S}' \cap \mathcal{C}_i}^{\text{mut}} \cdot \lambda_{q,c}^{\mathcal{C}_i}$$

where c_q denotes the fraction of samples in $\mathcal{S}'$ for which the base q is sufficiently covered and $n_{\mathcal{S}'}^{\text{mut}}$ and $n_{\mathcal{S}' \cap \mathcal{C}_i}^{\text{mut}}$ denote the total number of mutations in cohort $\mathcal{S}' \subseteq \mathcal{S}$ and $\mathcal{S}' \cap \mathcal{C}_i$, respectively. These mutational likelihoods $\lambda_{q,c}^{\mathcal{S}'}$ determine how likely we expected a mutation of class c occurs in position q. For $\lambda_{q,c}^{\mathcal{S}'} > 1$ we expected to observe more mutations of class c in position q relative to the average across the exome ("usual" nucleotide contexts). For $\lambda_{q,c}^{\mathcal{S}'} < 1$ we expected to observe less mutations of class c in position q relative to the average across the exome ("unusual" nucleotide contexts). In contrast, mutation probabilities $f_{t,(n_{-k},\ldots,n_l)}^{\text{mut}}$ (e.g. used in COSMIC mutation signatures[4]) characterize the fraction of mutations falling in a particular nucleotide context in aggregate.

3. A statistical framework for the identification of cancer genes

We next developed a statistical framework to utilize nucleotide contexts for the search of driver genes. From the composite likelihood model, which we described in detail in the previous section, we obtained a composite likelihood score λ for each position in the human exome. This score reflects the context-dependent mutability of the position relative to the average mutation density across the exome. Likelihood scores $\lambda > 1$ indicate that mutations occur more likely than on average in the position ("usual" nucleotide contexts); mutability scores $\lambda < 1$ indicate that mutations occur less likely than average ("unusual" nucleotide contexts"). Hence, nucleotide contexts can inform driver gene identification on both ends of the mutability spectrum: accounting for the abundance of passenger mutations in highly mutable nucleotide contexts ($\lambda \gg 1$), and using "unusual" nucleotide contexts ($\lambda \ll 1$) to gauge the localization shift of driver mutations from functionally neutral towards functionally important positions without prior knowledge on their exact location.

In brief, we followed two major steps to compute the p-value of a gene. In the first step, we counted how many mutations we observed in each position of the gene. We then compared this count with the corresponding context-dependent composite likelihood score of the same position, thereby determining which mutations were surprising (unusual sequence contexts) based on their surrounding sequence context. Based on these comparisons, we obtained a joint probability reflecting how closely the mutations of a gene followed the expected passenger mutation distribution pattern. This probability served as a test statistic. In the second step, we compared this probability with simulated "scenarios", in which passenger mutations were randomly distributed along the same gene. Based on a large number of these simulated scenarios, we derived a p-value as the fraction of scenarios in which the probability was lower than the joint probability of the observed mutations, i.e. in which the simulated passenger mutations followed the background distribution pattern less closely than the observed mutations in the gene. These simulations provide a null hypothesis that captures multiple signals for driver gene discovery, including mutational recurrence, mutations in unusual nucleotide contexts, and unusual clustering.

Thereby, our method weighs each nonsynonymous mutation individually, based on its surrounding nucleotide context, as well as in combination with other mutations. Mutations that substantially shift the joint probability of a scenario from "likely" to "unlikely" (e.g. mutations in unusual nucleotide contexts, mutations clustering in mutational hotspots) have a higher impact on the p-value than mutations with marginal effects (e.g. mutations in usual contexts). Furthermore, our test statistics actively downgrades the effect mutations that occur in highly mutable contexts ($\lambda \gg 1$) and amplifies the signal of mutations that occur in "unusual" nucleotide contexts ($\lambda \ll 1$). We observed in this study that particularly the second aspect was particularly important in tumors with high background mutation rates, as it provides an indirect proxy of the shift of driver mutations from functionally neutral towards functionally important positions without prior knowledge of their exact location. Additionally, our test statistic depends

on the number of synonymous mutations observed in the gene, and thereby accounts for the regional variation of the background mutation density.

In contrast to the signals used by existing methods for driver gene discovery (e.g. nonsynonmous mutations, number of hotspots, bioinformatics scores of functional impact), nucleotide contexts have a multidimensional character. Hence, the signal mediated by nucleotide contexts does not provide a direct one-dimensional score that could be used as a test statistic for comparison against a null hypothesis. Instead, we had to first compute a joint probability as a test statistic, which we then compared against random scenarios. Furthermore, another computational challenge was the efficient generation of the large number of random scenarios, as a large number of random scenarios are needed to accurately model the tail of the distribution.

In the first six subsections, we will describe the computation of the joint probability, as well as the efficient generation of the random scenarios. Further, we introduce a score to detect local violations from the null hypothesis of our model (i.e., if passenger mutations are not distributed according to their usual nucleotide contexts), which we used to minimize false-positives. In addition to this p-value, our method computes two additional p-values, which account for positional clustering into mutational hotspots as well as the abundance of deleterious mutations, such as insertions and deletions. We describe the computation of these two p-values in subsection 7. In the last subsection, we describe how to integrate these p-values into a combined p-value. This combined p-value is finally returned by our method and accounts for different signals of driver genes. A summary of this section can be found in the online methods ("Statistical analyses to identify driver genes").

3.1 A joint probability score that accounts for increased mutation counts and mutations in unusual nucleotide contexts

Given a gene $g \in \mathcal{G}$ of length l_g, we denote by $v^g \in \mathbb{N}^{l_g \times 3}$ the mutation count vector, i.e. each element $v^g_{q,c}$ denotes the number of mutations of class $c \in$ {I,II,III} in position $1 \leq q \leq l_g$ in g. Similarly, we defined an $l_g \times 3$-dimensional vector λ^g, for which every entry $\lambda^g_{q,c}$ denotes the mutational likelihood of position $1 \leq q \leq l_g$ in g for mutations of mutation class $c \in$ {I,II,III}, as defined in the previous section. We then decomposed $\lambda^g = (\lambda^{g,s}, \lambda^{g,n})$ and $v^g = (v^{g,s}, v^{g,n})$ into positions (q, c) encoding synonymous (s) and nonsynonymous mutations (n), respectively.

The number of nonsynonymous mutations $|v^{g,n}|$ in a gene g has been modeled using a Poisson distribution previously[14,15]

$$|v^{g,n}| \sim \text{Pois}(\mu)$$

where μ denotes the expected number of nonsynonymous mutations. μ can be derived as the product of the local nonsynonymous mutation probability (i.e., the mutation probability per base pair) and the target size (i.e., the number of genomic positions with sufficient coverage). The local mutation density has been commonly estimated based on the number of synonymous mutations $|v^{g,s}|$ in g, as $|v^{g,s}|$ is not influenced by positive pressure[8,16]. However, we had to consider that also the gene-specific expected number of synonymous mutations $\mu^{g,s}$ is not

explicitly known for g, but has to be estimated indirectly based on the observed number of synonymous mutations $|v^{g,s}|$ and the fluctuation of the mutation density across the genome. Hence, we integrate over all possible $\mu^{g,s}$ to obtain

$$P(|v^{g,n}| \mid |v^{g,s}|) = \int \text{Pois}\left(|v^{g,n}|; \mu^{g,s} \cdot \frac{|\lambda^{g,n}|}{|\lambda^{g,s}|}\right) \cdot P(\mu^{g,s} \mid |v^{g,s}|) \cdot d\mu^{g,s}$$

where we denoted $\text{Pois}(n;\mu) \coloneqq P(X = n)$ for $X \sim \text{Pois}(\mu)$ randomly distributed. Further, we used Bayes' theorem to estimate

$$P(\mu^{g,s} \mid |v^{g,s}|) = \frac{\text{Pois}(|v^{g,s}|; \mu^{g,s}) \cdot P(\mu^{g,s})}{\int \text{Pois}(|v^{g,s}|; \mu) \cdot P(\mu) \cdot d\mu}$$

so that calculating $P(|v^{g,n}| \mid |v^{g,s}|)$ can be reduced to determining the fluctuation of $\mu^{g,s}$ across the genome, reflected by $P(\mu^{g,s})$. According to the CBASE approach described previously[8] the statistical distribution of $\mu^{g,s}$ can be modeled by maximizing a log-likelihood estimator

$$\theta_{\text{opt}}(\{|v^{g,s}| \mid g \in \mathcal{G}\}) \coloneqq \arg\max_{\theta} \sum_{g \in \mathcal{G}} \ln\left(\int \text{Pois}(|v^{g,s}|; \mu) \cdot P(\mu \mid \theta) \cdot d\mu\right)$$

That way, we obtained a statistical model ($P(|v^{g,n}| \mid |v^{g,s}|)$) to estimate mutational excess above the background mutation rate. This model was conditional on the observed number of synonymous mutations in the same gene, in order to account for the fluctuation of the local mutation density across individual genes. When not enough data were available to fit $P(\mu)$, we computed $P\left(|v^{g,n}| \mid l_g\right)$ conditional on l_g instead.

We next aimed to determine a probability score, reflecting whether the mutations in gene g occurred in unusual nucleotide sequence contexts. Given the mutational likelihood $\lambda_{q,c}^{g,n}$ of each position q in g, the probability that an individual mutation hits position q is given by $\lambda_{q,c}^{g,n}/|\lambda^{g,n}|$, where $|\lambda^{g,n}| \coloneqq \sum_{q,c}\left|\lambda_{q,c}^{g,n}\right|$. Hence, we derived the probability of observing the nonsynonymous mutation count vector $v^{g,n}$, given the total number of nonsynonymous mutations $|v^{g,n}|$ and the context-dependent mutational likelihoods $\lambda^{g,n}$ using a multinomial distribution as follows

$$P(v^{g,n} \mid |v^{g,n}|; \lambda^{g,n}) \coloneqq \frac{\Gamma(|v^{g,n}| + 1)}{\prod_{q,c} \Gamma\left(v_{q,c}^{g,n} + 1\right)} \cdot \prod_{q,c} \left(\frac{\lambda_{q,c}^{g,n}}{|\lambda^{g,n}|}\right)^{v_{q,c}^{q,n}}$$

Finally, we composed the probabilities $P(|v^{g,n}| \mid |v^{g,s}|)$ and $P(v^{g,n} \mid |v^{g,n}|; \lambda^{g,n})$, accounting for the nonsynonymous mutation counts and mutational nucleotide contexts, respectively, into a combined statistical model as

$$P(v^{g,n} \mid |v^{g,s}|; \lambda^{g,n}) \coloneqq P(v^{g,n} \mid |v^{g,n}|; \lambda^{g,n}) \cdot P(|v^{g,n}| \mid |v^{g,s}|)$$

for genes g with $|v^{g,n}| > 0$. These probability scores allowed us to discover cancer genes based on their mutational excess and mutational sequence context using a combined null hypothesis model. We will denote $P(v^{g,n} \mid |v^{g,s}|; \lambda^{g,n})$ by $P(v^{g,n})$ in the following.

We note that the normalization step ($\lambda_{q,c}^{g,n}/|\lambda^{g,n}|$) is necessary to transform the composite likelihood scores into mutation probabilities, so that they sum up to 1 across all positions in the gene. The total number of mutations ($|v^{g,n}|$) is then redistributed according to these mutation probabilities, i.e. mutations with a likelihood score ≈1 are expected to have $|v^{g,n}|/|\lambda^{g,n}|$ mutations. This normalization step is in concordance with other models that consider nucleotide

contexts in their background models (e.g. dN/dSCV or CBaSE), which similarly sum context-dependent mutation frequencies up across all positions in a gene and then use this sum to normalize mutation frequencies.

3.2 Defining the mutational significance of a candidate cancer gene

Given the probability $P(v^{g,\mathrm{n}})$ to observe the distribution ($v^{g,\mathrm{n}}$) and nonsynonymous mutation count ($|v^{g,\mathrm{n}}|$) by chance, we next derived a significance value by determining whether the probability value $P(v^{g,\mathrm{n}})$ was high or low, compared with the probability that we expected if a gene g contained passenger mutations only. More precisely, we simulated a large number of mutation count vectors $w^{g,\mathrm{n}}$ in the same gene g, and compared their probability $P(w^{g,\mathrm{n}})$ with the probability $P(v^{g,\mathrm{n}})$ of our observed count vector $v^{g,\mathrm{n}}$. Based on these simulation experiments we then derived the significance as the fraction of simulation experiments yielding a lower probability $P(w^{g,\mathrm{n}})$ than the observed probability $P(v^{g,\mathrm{n}})$.

We note, however, that two mutation probabilities $P(v^{g,\mathrm{n}})$ and $P(w^{g,\mathrm{n}})$ cannot be compared directly if they do not contain the same number of nonsynonymous mutations, i.e. $|v^{g,\mathrm{n}}| \neq |w^{g,\mathrm{n}}|$, as the size of the support space $\{v \in \mathbb{N}^l \mid |v| = k\}$ is strongly dependent on k. As a trivial example, consider uniform distributions on two support spaces of sizes n_1 and n_2. Then the first probability distribution would assign probability $1/n_1$ to each element, whereas the second distribution would assign probability $1/n_2$ to each element. Hence, to compare the probabilities of elements in space 2 with probabilities in space 1, the probabilities have to be rescaled by factor n_2/n_1. Similarly, to compare the probabilities of $v^{g,\mathrm{n}}$ and $w^{g,\mathrm{n}}$ one has to correct for the different sizes of their support spaces. We thus defined $P^*(v)$, which corrects for the different sizes as

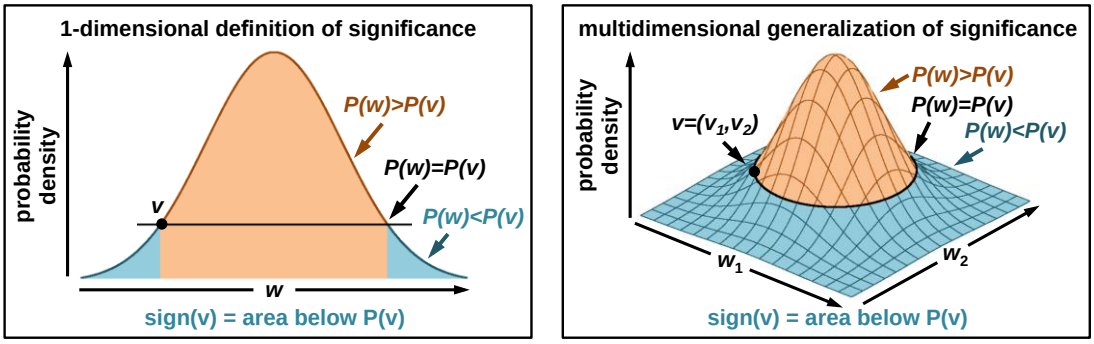
$$P^*(v) \coloneqq \frac{\Gamma(|v| + l) \cdot P(v)}{\Gamma(|v| + 1) \cdot \Gamma(l)}$$

where l denotes the length of vector v.

Based on these considerations, we defined the sequence context-based p-value of a gene g as

$$\mathrm{sign}_{\mathrm{seq}}(g) \coloneqq \sum_{\substack{P^*(w^{\mathrm{n}}) \leq P^*(v^{g,\mathrm{n}}) \\ |v^{g,\mathrm{n}}| \leq |w^{\mathrm{n}}|}} P(w^{\mathrm{n}})$$

reflecting the probability of observing a passenger mutation distribution w^{n} in gene g with a lower $P(w^{\mathrm{n}})$ by chance for genes with $|v^{g,\mathrm{n}}| > 0$. We note that our definition of $\mathrm{sign}_{\mathrm{seq}}(g)$, which is based on comparing $P^*(v^{g,\mathrm{n}})$ with the probability values the rest of the vector space, is a direct generalization of the definition of a two-sided p-value in a one-dimensional space, as illustrated by the graphics below.



As $P^*(w^{\mathrm{n}}) \leq P^*(v^{g,\mathrm{n}})$ is equivalent to

$$P(w^{\mathrm{n}} \mid |w^{\mathrm{n}}|; \lambda^{g,\mathrm{n}}) \leq P(v^{g,\mathrm{n}} \mid |v^{g,\mathrm{n}}|; \lambda^{g,\mathrm{n}}) \cdot \frac{P(|v^{g,\mathrm{n}}| \mid |v^{g,s}|) \cdot \Gamma(|v^{g,\mathrm{n}}| + l) \cdot \Gamma(|w| + 1)}{P(|w^{\mathrm{n}}| \mid |v^{g,s}|) \cdot \Gamma(|v^{g,\mathrm{n}}| + 1) \cdot \Gamma(|w| + l)}$$

$\mathrm{sign}_{\mathrm{seq}}(g)$ can be rewritten as

$$\mathrm{sign}_{\mathrm{seq}}(g) = \sum_{k \geq |v^{g,n}|} F_{k,\lambda^{g,\mathrm{n}}}(\tau_k) \cdot P(k \mid |v^{g,s}|)$$

where τ_k depends on k only and

$$F_{k,\lambda}(\tau) \coloneqq \sum_{\substack{P(w|k;\lambda) \leq \tau \\ |w|=k}} P(w \mid k; \lambda)$$

$$= P(P^*(X \mid k; \lambda) \leq \tau)$$

with $X \sim \mathrm{Multi}(k; \lambda)$ randomly distributed. Hence, computing $\mathrm{sign}_{\mathrm{seq}}(g)$ reduces to computing a limited number of $F_{k,\lambda}(\tau)$. That way, we obtained a significance value reflecting how likely the nonsynonymous mutations in gene g were generated by passenger mutations only.

Finally, we defined

$$\mathrm{sign}_{\mathrm{seq}}^{\mathrm{syn}}(g) \coloneqq \sum_{\substack{P^*(w^{\mathrm{s}}||v^{g,s}|;\lambda^{g,s}) \leq P^*(v^{g,s}||v^{g,s}|;\lambda^{g,s}) \\ |w^{\mathrm{s}}|=|v^{g,s}|}} P(w^{\mathrm{s}} \mid |v^{g,s}|; \lambda^{g,s})$$

which requires the computation of a single $F_{k,\lambda}(\tau)$ and serves as a negative control to flag potential false-positive genes that may result from local violations of the null hypothesis. This component tested whether synonymous mutations followed the context-dependent background distribution of our model.

We note that the strategy of establishing a background distribution model for all genes and then exploiting synonymous mutations to test for local violations of the null hypothesis has also been used in previous approaches. For instance, Lawrence and colleagues[16] established a model that predicts the local background mutation density based on three epigenomic covariates using local regression (MutSigCV). They then used synonymous mutations to test whether their null hypothesis was valid in a given gene; if synonymous mutations indicated a local violation of the null hypothesis, they did not perform the significance test for nonsynonymous mutations, but changed their local regression parameters instead. Analogously, if $\mathrm{sign}_{\mathrm{seq}}^{\mathrm{syn}}(g)$ in our model indicated that the distribution of synonymous mutations deviated from the background model, we consequently did not use the p-value derived from the statistical test of nonsynonymous

mutations. That way, we exploited synonymous mutations to detect local deviations from the null hypothesis of our background model, and thereby minimized the number of false-positive findings resulting from local violations of the null hypothesis (i.e. that passenger mutations follow a context-specific distribution pattern).

3.3 A Monte Carlo simulation approach to compute mutational significance

To obtain $\text{sign}_{\text{seq}}(g)$, we next established a simulation approach to compute $F_{k,\lambda}(\tau) \coloneqq P(P^*(X \mid k;\lambda) \leq \tau)$. As the number of all possible X is too large to simulate each possible X explicitly (all possible mutation distributions in a gene would need to be simulated), we used a Monte Carlo simulation to derive a large number $a \gg 0$ of $w^{(1)}, \dots, w^{(a)} \sim \text{Multi}(k;\lambda)$ and approximate

$$F_{k,\lambda}(\tau) \approx \frac{1}{a} \cdot \left| P\left(w^{(j)} \mid k;\lambda\right) \leq \tau, 1 \leq j \leq a \right|$$

as the fraction of $w^{(j)}$, for which the probability $P\left(w^{(j)} \mid k;\lambda\right)$ fell below the threshold τ. However, as l_g is large, simulating the full vectors w^j is run-time intensive. However, to approximate $F_{k,\lambda}(\tau)$ we did not need the full vectors $w^{(j)}$, but only to evaluate their probability. Hence, instead of simulating the complete vector $w^{(j)}$, we simulated the derived logarithmic probability $\ln P\left(w^{(j)} \mid k;\lambda\right)$ only. Apart from constants, $\ln\left(P(w \mid k;\lambda)\right)$ can be rewritten as

$$\theta(w,\lambda) \coloneqq \underbrace{\sum_{1 \leq i \leq l_g} \ln(\lambda_i) \cdot w_i}_{\theta_1(w,\lambda)} + \underbrace{\sum_{1 \leq i \leq l_g} -\ln \Gamma(w_i + 1)}_{\theta_2(w)}$$

We note that $\theta(w,\lambda)$ can be viewed as a balance between accumulations of mutations in genomic positions with high mutation probability and dissemination of mutations into positions with low mutation probability. More precisely, $\theta_1(w,\lambda)$ measures the dissemination of mutations in positions with low mutational likelihood λ_i. As $\ln \Gamma(w_i + 1) = 0$ for $w_i < 2$, $\theta_2(w)$ measures the number of accumulations of mutations positions with a high mutational likelihood λ_i. Considering that the higher the accumulation of mutations in positions with high mutation probability the lower the dissemination into positions with a low mutation probability and vice versa, we obtain $\text{Cov}(\theta_1, \theta_2) \leq 0$, so that $\text{Var}(\theta) \leq \text{Var}(\theta_1) + \text{Var}(\theta_2)$. Hence, by simulating θ_1 and θ_2 independently from each other, we obtain a conservative approximation of $\text{sign}_{\text{seq}}(g)$.

3.4 Simulation of θ_1

As the expected number of mutations in position i is $\mathbb{E}(w_i) = |w| \cdot \lambda_i / |\lambda|$, the probability of an individual mutation to hit position i is $\lambda_i / |\lambda|$, where $|\lambda| \coloneqq \sum_i |\lambda_i|$. Hence, we can simulate $\theta_1(w,\lambda) = \sum_{1 \leq i \leq l_g} \ln(\lambda_i) \cdot w_i$ as a sum

$$\theta_1(w,\lambda) \sim \sum_{i=1,\dots,|w|} X^{(i)}$$

where the random variables $X^{(i)}$ are distributed independently from each other and each $X^{(i)}$ has a value of $\ln(\lambda_i)$ with a probability of $\lambda_i/|\lambda|$, i.e. $P\left(X^{(i)} = \ln(\lambda_i)\right) = \lambda_i/|\lambda|$. That way, we simulated $\theta_1(w, \lambda)$ in $|w|$ iterations.

3.5 Simulation of θ_2

For the simulation of θ_2 we will assume that the λ_i are sorted in decreasing order in the following. As $\theta_2(w) = -\sum_{1\le i\le l_g} \ln \Gamma(w_i + 1)$ only depends on the elements of the vector w with $w_i \geq 2$, we only needed to simulate the part of w which contains positions with more than two mutations. For this purpose, we define

$$k^{\text{high}} := \min\left\{k \mid \sum_{1\le i\le k} \lambda_i^2 \Big/ \sum_{1\le j\le l} \lambda_j^2 \geq 0.99\right\}$$

$$f^{\text{high}} := \sum_{1\le i\le k_{\text{high}}} \lambda_i \Big/ \sum_{1\le j\le l} \lambda_j$$

Hence, instead of simulating the full w, we can only simulate the first k^{high} of its elements to obtain $\theta_2(w)$ with 99% accuracy. We denote by $\lambda^{\text{high}} := \left(\lambda_i, 1 \le i \le k^{\text{high}} < l\right)$ the restriction of the mutational likelihood vector λ to the first k^{high} elements. Based on these considerations, we simulated θ_2 by drawing

$$\left|w^{\text{high}}\right| \sim \text{Binom}\left(|\text{w}|, f^{\text{high}}\right)$$

$$w^{\text{high}} \sim \text{Multi}\left(\left|w^{\text{high}}\right|, \lambda^{\text{high}}/\left|\lambda^{\text{high}}\right|\right)$$

randomly and approximate

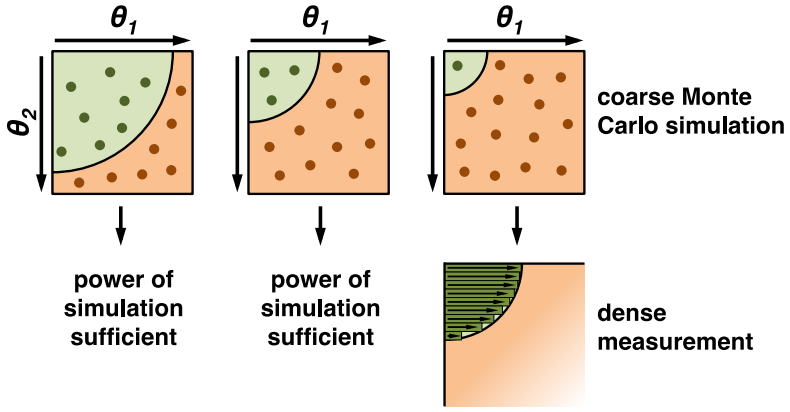
$$\theta_2(w) \approx \sum_{1\le i\le k^{\text{high}}} -\ln \Gamma\left(w_i^{\text{high}} + 1\right)$$

We note that $f^{\text{high}} \approx 0.08$ for most genes, so that we speeded up the simulation procedure almost 10-fold by only simulating the first k^{high} indices.

3.6 Computation of $F_{k,\lambda}$

To obtain an accurate approximation of $F_{k,\lambda}(\tau) \approx \frac{1}{a} \cdot \left|P\left(w^{(j)} \mid k; \lambda\right) \le \tau, 1 \le j \le a\right|$, we simulated for each gene up to 100,000 and 10,000 independent instances of θ_1 and θ_2, respectively. That way, significance values as low as 10^{-9} could be determined, based on their mutational sequence context only. Additionally considering the mutation count yielded even lower significance values. Computing and evaluating all 10^9 instances of (θ_1, θ_2) would be run-time intensive and is not needed for most genes, unless they are significant. Hence, we needed an efficient way to determine how many instances of (θ_1, θ_2) needed to be evaluated to accurately approximate the significance value with sufficient precision. For this purpose, we proceeded in two steps. In the first step, we simulated 1,000 and 100 independent instances of θ_1 and θ_2, respectively. Then we evaluated for 1,000 random pairs (θ_1, θ_2) whether the fell below the threshold τ, thereby obtaining a first approximation $\text{sign}^*_{\text{seq}}(g)$ of $\text{sign}_{\text{seq}}(g)$. If $\text{sign}^*_{\text{seq}}(g) < 0.05$, we entered into the second step, which uses more simulations in order to model the tail of $P(w)$. In this second step

we simulated 100,000 and 10,000 independent instances of θ_1 and θ_2, respectively, and sorted our simulations in increasing order $\theta_1^{(i)} \leq \theta_1^{(i')}$ for $i < i'$ and $\theta_2^{(j)} \leq \theta_2^{(j')}$ for $j < j'$. That way, we obtained a 100,000x10,000-matrix of $\theta_1^{(i)} + \theta_2^{(j)}$. As the matrix entries were increasing by row and columns we started in the upper left corner to evaluate which entries were lower than the threshold. As we knew in the first step that the significance was ≤ 0.05, a maximum of 5% had to be computed and compared to the threshold. This 2-step procedure to approximate $F_{k,\lambda}(\tau)$ is illustrated below.



3.7 Additional statistical components accounting for independent sources of mutational significance, including insertions and deletions

Besides $\text{sign}_{\text{seq}}(g)$, we computed two additional p-values $\text{sign}_{\text{dm}}(g)$ and $\text{sign}_{\text{cum}}(g)$ for each gene g, which reflect whether g contains a significantly increased fraction of protein damaging mutations (indicating a potential tumor suppressor gene) and whether mutations in g exhibit positional clustering (indicating a potential mutational hotspot in an oncogene). In particular, $\text{sign}_{\text{dm}}(g)$ accounts for insertions and deletions in our model, which provide an additional source for the detection of driver genes. The final p-value derived for each gene, combined these three subcomponents.

Denote by N_{dm}^g the number of protein damaging mutations in g (frameshift insertions, frameshift deletions, nonsense mutations, stop codon deletions, stop codon insertions) and by N_{total}^g the total number of mutations in g (including all non-SNV mutations). Then we obtained $f_{\text{dm}} \coloneqq \left(\sum_{g\in\mathcal{G}} N_{\text{dm}}^g\right)/\left(\sum_{g\in\mathcal{G}} N_{\text{total}}^g\right)$ as the genome-wide fraction of damaging mutations and we define

$$\text{sign}_{\text{dm}}(g) \coloneqq P(X^g \geq N_{\text{dm}}^g)$$

with $X^g \sim \text{Binom}\left(N_{\text{total}}^g, f_{\text{dm}}\right)$ randomly distributed.

In order to determine whether g contains significantly more mutation accumulations than expected based on the underlying mutation distribution pattern, we defined

$$\text{sign}_{\text{cum}}(g) \coloneqq \sum_{\substack{|v^{g,\text{n}}|=|w^{\text{n}}| \\ |v^{g,\text{n}}|_\infty \leq |w^{\text{n}}|_\infty}} P(w^{\text{n}})$$

where $|v|_\infty \coloneqq \max_{1\leq i\leq l} |v_i|$ denotes the Chebyshev metric. Assuming independence of all positions, which makes the estimation of this p-value conservative, $\text{sign}_{\text{cum}}(g)$ can be approximated as

$$\text{sign}_{\text{cum}}(g) \approx 1 - \prod_{1 \le i \le l} P(X_i < |v^{g,\text{n}}|_\infty)$$

for $X_i \sim \text{Binom}(|v^{g,\text{n}}|, \lambda_i/|\lambda^{g,\text{n}}|)$ randomly distributed. Similarly, we define $\text{sign}^{\text{syn}}_{\text{cum}}(g)$ as the probability to observe the accumulation $|v^{g,s}|_\infty$ in the synonymous positions, based on the background mutation rate, which serves as an additional filter to detect local violations of the underlying null hypothesis.

3.8 Combining significance values and correcting for multiple hypothesis testing

For each gene g, we obtained the p-values $\text{sign}_{\text{seq}}(g)$, $\text{sign}_{\text{cum}}(g)$, and $\text{sign}_{\text{dm}}(g)$, which reflect different signals for driver gene discovery. We finally aimed to integrage them into a combined p-value to score the mutational significance of a given gene based on multiple signals.

To combine statistically independent p-values $p_1, \dots, p_k$, Fisher's method is commonly used, which assumes that $-2 \cdot \sum_{i=1,\dots,k} \ln(p_i)$ follows a χ^2 distribution with $2k$ degrees for freedom. However, as statistical independence could not be assumed for $\text{sign}_{\text{seq}}(g)$, $\text{sign}_{\text{cum}}(g)$, and $\text{sign}_{\text{dm}}(g)$ (e.g., genes with hotspots typically also deviate from the background mutation distribution pattern), we used the Brown method, which is an extension of Fisher's method[17], and does not assume independence between the different p-values. We computed for each gene g the score

$$\Phi^g := -2 \cdot \ln\big(\text{sign}_{\text{seq}}(g) \cdot \min\big(\text{sign}_{\text{cum}}(g), \text{sign}_{\text{dm}}(g)\big)\big)$$

and assumed that Φ^g followed a scaled χ^2-distribution, i.e.

$$c \cdot \Phi^g \sim \chi^2(k)$$

for two constants c and k.

As mean and variance of the $\chi^2(k)$ distribution are given by k and $2k$, respectively, we estimated

$$c \approx \frac{2 \cdot \mathbb{E}(\Phi^g)}{\text{Var}(\Phi^g)}, \qquad k \approx \frac{2 \cdot \mathbb{E}(\Phi^g)^2}{\text{Var}(\Phi^g)}$$

where $\mathbb{E}(\Phi^g)$ and $\text{Var}(\Phi^g)$ denote the mean and variance of Φ^g over $g \in \mathcal{G}$, respectively. That way, we inferred the combined significance value as

$$\text{sign}(g) := P(X \ge c \cdot \Phi^g)$$

with $X \sim \chi^2(k)$. We observed that the distribution of the combined p-value $\text{sign}(g)$ closely followed a uniform distribution, so that we corrected our significance values for multiple hypothesis testing using the Benjamini-Hochberg procedure[18]. For this purpose, denote by $g^{(1)}, g^{(2)}, \dots, g^{(|\mathcal{G}|)}$ the genes sorted in increasing order by their significance. Then we derive the false-discovery rate as

$$\text{FDR}\big(g^{(k)}\big) := |\mathcal{G}| \cdot \text{sign}(g)/k$$

where $\mathcal{G}$ denotes genes that were tested (genes with ≥1 mutations; >2 point mutations; option 1 by default). Similarly, we combined $\text{sign}^{\text{syn}}_{\text{seq}}(g)$, $\text{sign}^{\text{syn}}_{\text{cum}}(g)$ to a significance value $\text{sign}^{\text{syn}}(g)$ for synonymous mutations and corrected these p-values to $\text{FDR}^{\text{syn}}(g)$.

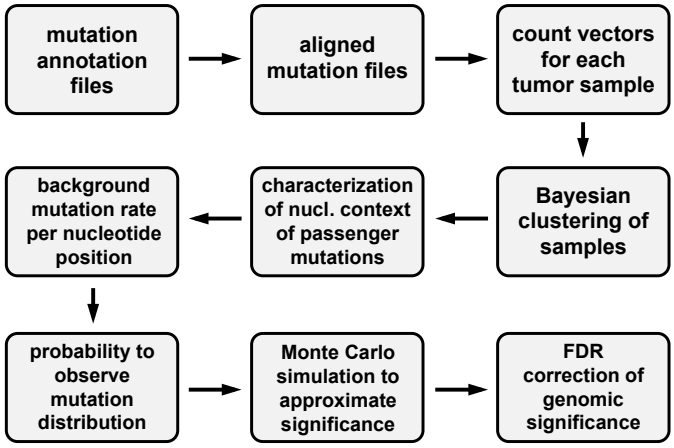
Finally, we filtered out potential false-positive non-CGC genes based on the following filters:

1.) Genes with $\mathrm{FDR}^{\mathrm{syn}}(g) < 0.25$ were filtered out. Their context-dependent distribution of synonymous mutations violates the underlying null hypothesis of the model, so that the test for mutations in unusual nucleotide contexts cannot be applied. That way, we minimized false-positive genes, in which mutations in unusual nucleotide contexts may result from local deviations from the context-specific passenger mutation distribution pattern.

2.) Positions with more than four mutation counts in significantly mutant genes were checked for read misalignments. For this purpose, the mutant sequence was re-aligned against the reference genome using the Blat alignment tool. If the mutant sequence could be aligned error-free to a different region in the reference genome in a 25bp window around the mutation hotspot, the gene was flagged. Non-mutant reads from a different region might have erroneously been aligned to the region thus producing a false-positive mutational hotspot in respective sequence motifs. That way, we minimized potential false-positive genes, in which mutations in unusual nucleotide contexts may result from variant calling errors.

3.) Genes in a blacklist that contains genes in the top 5% of SNP density rate. Due to their high SNP frequency, a substantial number of their mutations are germline variations that slipped through the filtering process, including the matched normal and reference sequencing dataset of the normal population. As germline mutations are distributed independently from the underlying passenger mutation distribution pattern, these SNPs can cause false-positive results. That way, we minimized potential false-positive genes, in which mutations in unusual nucleotide contexts may result from germline variants.

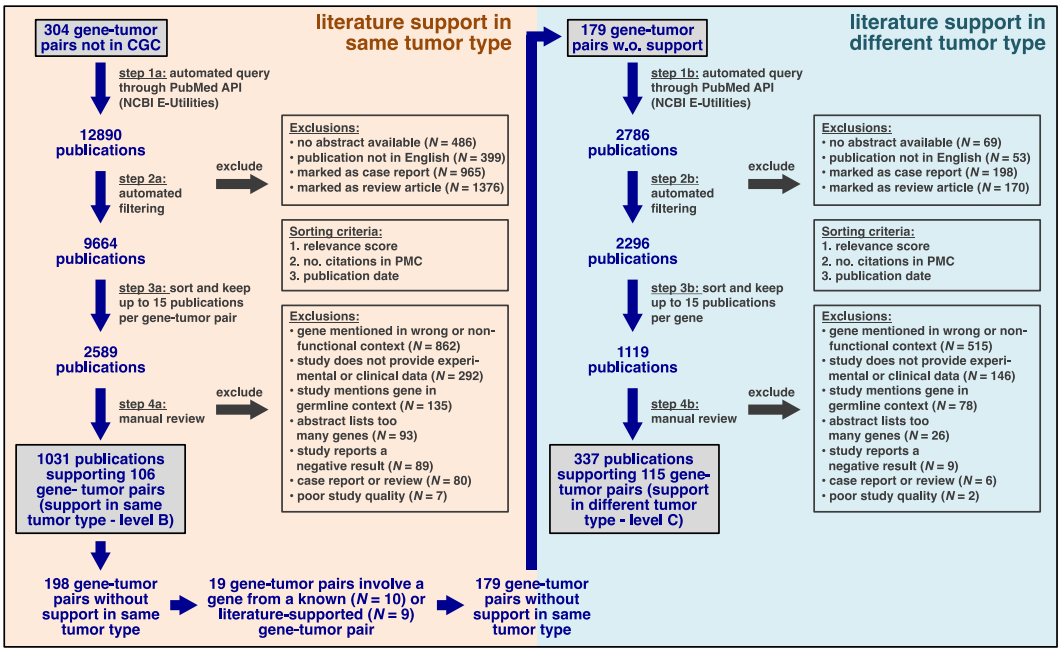
Taken together, our statistical framework can be summarized as follows:



4. Additional analyses used to characterize the driver genes identified by our method

4.1 Stratification of driver genes based on literature support

To examine which non-CGC genes in our driver gene catalog had supporting evidence, we performed a systematic literature search. The major steps of this literature search are visualized below and described in detail in the Online Methods.



In Step 1a, we used the following search terms for the cancer types.

Cancer Type	Search Terms
Adenoid Cystic	adenoid cystic carcinoma
Bladder	bladder cancer
	bladder carcinoma
Blood	leukemia
	ALL
	CLL
	AML
	CML
Brain	brain tumor
	meningioma
	astrocytoma
	glioblastoma
	medulloblastoma
	glioma
Breast	breast cancer
	ductal carcinoma
	lobular carcinoma
Cervix	cervical cancer
	cervical carcinoma
Cholangio	bile duct cancer
	cholangiocarcinoma
	gallbladder cancer
	cancer of the ampulla of vater
	Klatskin tumor
Colorectal	colorectal cancer
	bowel cancer
	colon cancer
	rectal cancer
Endometrium	endometrial cancer
	uterine cancer
	endometroid carcinoma
	endometrial carcinoma

Cancer Type	Search Terms
Gastro-esophageal	gastroesophageal cancer
	esophageal cancer
	esophageal carcinoma
	oesophageal carcinoma
	gastric cancer
	stomach cancer
	gastric carcinoma
Head & Neck	head and neck cancer
	throat cancer
	oral cancer
	laryngeal cancer
	oropharyngeal cancer
Kidney	kidney cancer
	renal cancer
	renal cell carcinoma
Liver	liver cancer
	hepatic cancer
	hepatocellular carcinoma
Lung Ad.	adenocarcinoma of the lung
	lung adenocarcinoma
	pulmonary adenocarcinoma
	non-small cell lung cancer
	lung cancer
Lung Sq.	squamous cell carcinoma of the lung
	squamous cell lung cancer
	squamous cell lung carcinoma
	non-small cell lung cancer
	lung cancer
Lung SCLC	small cell lung cancer
	small cell carcinoma
	oat-cell carcinoma

Cancer Type	Search Terms
Lymph	lymphoma
	lymphatic cancer
	non-hodgkin lymphoma
Ovarian	ovarian cancer
	ovarian carcinoma
Pancreas	pancreatic cancer
	pancreatic adenocarcinoma
	pancreatic ductal adenocarcinoma
	pancreatic neuroendocrine tumors
Pheochromo-cytoma	pheochromocytoma
	phaeochromocytoma
	paragangliomas
Pleura	mesothelioma
	pleural tumor
Prostate	prostate cancer
	prostatic carcinoma
	carcinoma of the prostate
Sarcoma	sarcoma
Skin	melanoma
	cutaneous melanoma
	desmoplastic melanoma
	skin cancer
Testicular Germ Cell	testicular cancer
	germ cell tumor
Thymus	testis tumor
	thymoma
	thymic carcinoma
	thymus cancer
Thyroid	thymic cancer
	thyroid cancer
	thyroid carcinoma
Uveal Melanoma	uveal melanoma

The relevance scores that we used in Steps 3a/b to retain the 15 most relevant mutations were automatically computed based on the table below.

relevance score to prioritize abstracts

mentions	gene name	tumor type*	*cancer*, *tumor*, or *carcinoma*	*mutation*, or *substitution*
title	+4	+2	+2	+2
abstract	+2	+1	+1	+1

*step 3a only

4.2 Method Comparison

We compared the performance of our approach with eight current methods, which are widely used to identify driver genes and cover a wide range of different signals used for driver gene detection.

1.) The MutSigCV algorithm[16,19] primarily detects driver genes based on the excess of nonsynonymous mutations over the regional background mutation rate. To model the mutational background in a given gene, the MutSigCV identifies genes with similar background mutation densities based on epigenomic covariates and synonymous mutations ("gene bagel"). The algorithm then tests whether the number of nonsynonymous mutations in a given gene exceeds the background mutation rate of the associated bagel. Further, the algorithm accounts for the heterogeneity of the background mutation probabilities across tumors and six nucleotide contexts / seven mutation categories (T>C, T>A/G, CpG>T, CpG>A/G, C>T, C>A/G, indels). Consistent with other benchmarking studies[9,20,21], we used the latest publicly available version of this method (MutSigCV[16], version 1.4). However, we note that a newer version of this algorithm has been published (MutSig2CV), which further includes statistical components to account for positional clustering and the abundance of mutations in evolutionarily conserved regions[19].

2.) The dNdScv algorithm[7] primarily utilizes the difference between synonymous and nonsynonymous mutations to detect driver genes under positive selection (normalized dN/dS ratio of nonsynonymous to synonymous mutations $\omega > 1$). Given a gene, the algorithm counts for each trinucleotide context the number of mutations in the following categories: synonymous, missense, nonsense, and essential splice sites. To determine the probability of observing the counts in these four categories, the algorithm models the regional background mutation densitiy by negative binomial regression with the help of epigenomic covariants. The algorithm then determines the probability of the observed mutations counts for $\omega=1$ vs. $\omega\neq1$. To test whether $\omega\neq1$ makes the observed counts significantly more likely, the algorithm uses a likelihood ratio test. Hence, mutational significance is derived by testing whether dN/dS ratio of nonsynonymous to synonymous mutations significantly exceeds 1, after correcting for trinucleotide contexts and epigenomic covariates in the background model[7].

3.) The algorithm OncodriveCLUST[22] exploits mutational hotspots for the search of driver genes. Driver mutations exhibit positional clustering when only few functional sites in a given gene are under positive selection. In brief, the algorithm tests for each gene whether more mutations than

expected fall into such a positional cluster. For this purpose, the algorithm groups mutations into positional clusters (within distance of five of less amino acid residues), and computes the fraction of mutations falling into a positional cluster. This fraction is then normalized by gene length, and mutational significance is derived by determining whether this normalized fraction exceeds the expectation of the background model significantly.

4.) OncodriveFM[23] uses three bioinformatic scores (SIFT[24], PolyPhen2[25], MutationAssessor[26]) to predict which mutations are likely to be of functional importance (FI). For each gene, OncodriveFM determines whether the average functional impact (FI) of mutations falling in the gene. OncodriveFM then derives mutational significance by testing whether the average functional impact is significantly higher than expected by chance (sampling with replacement).

5.) OncodriveFML[9] is an improved derivative of OncodriveFM, and is similarly based on the idea of using bioinformatic scores to predict functional importance. OncodriveFML uses CADD scores[27] to calculate the average functional impact of mutations per gene. The statistical model used by OncodriveFML is more advanced compared with OncodriveFM. To test whether the average functional impact per gene exceeds the background expectation, OncodriveFML accounts for the mutation probabilities of different trinucleotide contexts when sampling random functional impact scores. Thereby, OncodriveFML calibrates its background expectation of the functional impact to the mutability of different nucleotide contexts.

6.) e-Driver[28] is based on the idea that driver mutations are localized towards functionally important protein domains, such as phosphorylation sites. In contrast to OncodriveFM and OncodriveFML, this method does not use a bioinformatic score to quantify the functional impact of each mutation, but tests whether mutations are enriched in protein functional regions (PFR) relative to the regions outside of these domains. Functional regions were retrieved from external databases, such as ENSEMBL.

We executed the most recent versions of the source codes of methods 1-4, and 6 publically available. We further ran the online version of method 5 through an interactive web portal. All methods were applied to the full sequencing dataset underlying our study with standard configurations, which were either specified on the method website or in the method publication. We went through the first 1000 (or 150) significant genes returned by each method in the order of their significance. We then plotted the number of significant genes in CGC[29,30] (or OncoKB[31]) against the number of non-CGC (or non-OncoKB) genes to compare the performance between methods.

7) The RF5[32] method integrates multiple sources of mutational significance to detect driver genes with a robust performance across cancer types. In brief, these tests include a component to detect a mutational excess above the background mutation density (“unaffected residues”, comparable to the strategy of MutSigCV or dNdSCV), to detect an increased number of mutations with functional impact based on VEST3 scores (“VEST mean”, similar to the strategy followed by OncodriveFM and OncodriveFML), a component to detect an increased number of loss-of-function mutations (“Truncation rate”, which are an important source to detect tumor

suppressor genes, as well as components to detect differences in mutation distributions across cancer types (e.g. if a mutation is highly cancer type-specific this would suggest a functional impact in that particular cancer type) and patients (this component has also been accounted for in MutSigCV). Therefore, the tests included in the RF5[32] method are highly compatible with other methods. However, none of the other methods includes this large spectrum of different tests.

Moreover, instead of classical statistical methods for p-value combination (such as Fisher's or Brown's method), the RF5 model utilizes a random forest method to combine p-values. Genes used for model training were excluded when evaluating the random forest performance. Random forests carry the substantial advantage that power dilutions from opposite tests can be minimized. For instance, when combining opposite tests for hotspots (indicating oncogenes) and loss-of-function mutations (indicating tumor suppressor genes), this would result in a slight decrease in power with classical traditional statistical methods (combination of a significant and a non-significant p-value). When selecting the most appropriate test via random forest, this power dilution can be avoided. In addition, the random forest model not only combines p-values, but also provides a dichotomous classification of the significant genes into putative oncogenes vs. tumor suppressor genes.

8) CBaSE[8] is used to model the regional background mutation density in MutPanning. In contrast to other count-based models, such as MutSigCV or dNdSCV, CBaSE does not use a single statistical model to model the regional background mutation density, but selects the background distribution from six possible models. That way, CBaSE has the ability to fit the background mutation model closer to the observed data, which gives CBaSE an additional increase in power. In contrast to the other algorithms used for benchmarking, CBaSE was not a fully independent model, since MutPanning and CBaSE share a similar statistics for the count component. Hence, differences in performance between CBaSE and MutPanning can reflect the impact of the nucleotide context component to identify driver genes, but can also be influenced by differences in the implementation of the underlying statistical models between CBaSE and MutPanning.

4.3 Overview of driver genes that have not been well-implicated in their cancer types previously

We noticed that our catalog contains two groups of candidate driver genes that had not been well implicated in their cancer types. First, known cancer genes in a new tumor type context. Second, genes that had not been reported as mutated in cancer type previously.

Examples for genes in the first group include:

- *ACVR2A* was identified in pancreatic cancer. Previously, mutations in this gene have been primarily associated with colon cancer.
- *APC* was identified in testicular cancer. Previously, mutations in this gene have been primarily associated with colon cancer.

- *ARID1A* was identified non-clear cell kidney cancer. Previously, mutations in this gene have been primarily associated with gastric cancer, liver cancer, breast cancer, pancreatic cancer and bladder cancer.
- *ARID1B* was identified in thyroid cancer. Previously, mutations in this gene have been primarily associated with gastric and colorectal cancer.
- *ATM* was identified in pancreatic cancer. Previously, this gene has been known primarily as a tumor suppressor for leukemia, lymphomas, liver cancer, colorectal cancer, bladder cancer, and neuroendocrine prostate cancer. Furthermore, germline mutations in the gene have been known to be a predisposition factor for pancreatic cancer.
- *AXIN1* was identified in pancreatic cancer. Previously, mutations in this gene have been primarily associated with liver and ovarian cancer. Furthermore, *AXIN1* is part of the WNT pathway and mutations in other WNT pathway genes have been associated with pancreatic cancer.
- *BAP1* was identified in bladder cancer. Previously, this BRCA pathway-associated gene has been associated with breast cancer. Furthermore, germline mutations in *BAP1* have been reported to be associated with an increased risk for bladder cancer.
- *BIRC3* was identified hematological malignancies. Previously changes in expression of this gene have been associated with this cancer type.
- *CDH11* was identified in gastroesophageal carcinoma. Previously, methylation and expression changes in this protein have been associated with this cancer type.
- *CDKN1A*, the gene encoding for p21, was identified non-clear cell kidney cancer. This gene is a well-known tumor suppressor gene in other cancer types.
- *CDKN1B*, the gene encoding for p27, was identified in liver cancer. Previously, mutations in this gene have been primarily associated with breast and prostate cancer.
- *CHD1* was identified in bladder cancer. Previously, mutations in this gene have been primarily associated with prostate cancer.
- *CHEK2* was identified in pheochromocytoma. Previously, mutations in this gene have been primarily associated with breast, prostate and colon cancer.
- *CTNNB1* was identified in pancreatic cancer. Mutations in this gene have been associated with liver cancer, colorectal cancer, lung cancer, breast cancer, ovarian cancer and endometrial cancer, but not with pancreatic cancer Further, mutations in other WNT genes have been associated with pancreatic cancer.
- *CTNND1*, the gene encoding for catenin delta 1, was identified in gastroesophageal cancer. Previously, mutations in this gene have been primarily associated with liver cancer.
- *ERBB4* was identified in bladder cancer. Mutations in *ERBB4* have been previously reported in non-small cell lung cancer. Furthermore, the have been reports of differential expression of *ERBB4* and mutations in *ERBB2* in bladder cancer.
- *FBXW7* was identified in pancreatic cancer. Mutations have previously been reported in colorectal cancer. Reduced expression of *FBXW7* has been found to be associated with chemotherapy resistance of pancreatic cancer.

- *KMT2A* was identified in thyroid cancer. Previously, mutations in this gene have been reported in colorectal, lung, bladder, endometrial and breast cancer.
- *KMT2D* was identified in liver cancer. Previously, mutations in this gene have been primarily associated with colorectal cancer, lung cancer, endometrial cancer, brain tumors, and hematological malignancies.
- *MAP2K4* was identified in pancreatic cancer. Previously, mutations in this cancer type have been primarily associated with prostate and lung cancer.
- *MED12* was identified in blood and brain tumors. Previously, mutations in this gene have been associated with prostate and lung cancer.
- *NCOR1* was identified in bladder cancer. Previously, mutations in this gene have been primarily associated with breast and lung cancer. Further, differential expression has been associated with bladder cancer.
- *NF1* was identified in bladder cancer and sarcoma. Previously, mutations in this gene have been primarily associated with melanoma, lung cancer, ovarian cancer, and brain tumors.
- *PIK3R1* was identified in gastroesophageal cancer. Previously, mutations in this gene have been associated with breast and endometrial cancer.
- *PBRM1* was identified in gastroesophageal cancer and melanoma. Previously, mutations in this gene have been associated with kidney cancer. In addition, differential expression of *PBRM1* has been associated with melanoma. Further, both gastric cancer and melanoma have previously been associated with mutations of other genes in the SWI/SNF complex.
- *POLQ* was identified in melanoma. Previously, mutations in this homologous recombination DNA repair gene have been primarily associated with breast and ovarian cancer.
- *PTEN* was identified in thyroid cancer. This gene is a well-known tumor suppressor gene that is associated with multiple cancer types.
- *RB1* was identified in gastroesophageal and prostate cancer. Previously, focal deletions of this gene have been associated with these cancer types. Further, somatic mutations in this gene have been associated with a wide range of cancer types, including retinoblastoma and small-cell lung cancer.
- *REV3L*, the catalytic subunit of the translesion DNA polymerase zeta, was found in melanoma. This gene has previously been reported to cause chromosomal instability in various cancer types.
- *SETD2* was identified in melanoma. Previously, mutations in this gene have been primarily associated with kidney cancer.
- *SMAD3* was identified in pancreas. Mutations in other TGFb-genes, including *SMAD4*, have been previously associated with pancreatic cancer.
- *SPEN* was identified in hematological malignancies and clear-cell kidney cancer. Previously, mutations in this gene have been primarily associated with breast cancer.
- *TP53* was identified in thyroid cancer. This gene is a well-known tumor suppressor gene that has been associated with multiple cancer types other than thyroid cancer.

We would like to point out that identifying known tumor genes in new tumor type contexts is biologically important, even if these associations have low frequencies. For instance, *TP53* and *PTEN* are frequently mutated in ovarian (87%) and endometrial cancers (55%), but have low mutation frequencies in thyroid cancer (0.75% and 0.49%). Nevertheless, they both emerged as significantly mutated in thyroid cancer, suggesting that they may have a functional role in individual thyroid cancer patients. Our catalog allows systematic exploration of the occurrence of known cancer genes in rare but significant tumor type contexts; this may contribute to our understanding of the biological heterogeneity of these cancer types, such as primary or acquired therapy resistance.

Discovering entirely novel cancer genes, for which somatic mutations have never previously been associated with cancer, is intrinsically more challenging. It is hard to distinguish these genes from false-positives. Hence, we required preliminary biological or experimental evidence that these are functionally relevant in cancer. Our catalog contained some of these genes with strong experimental support. For instance, we identified *POLR2A* (RNA polymerase II subunit A) as significantly mutated in lung adenocarcinoma (Extended Data Figure 10). Mutations in *POLR2A* have been implicated in the development of meningioma in only one study previously (Clark et al., 2016). However, mutations in other polymerases, such as *POLE* or *POLD1*, have a well-established role in cancer. *POLR2A* has been identified as a therapeutic target in colon cancer due to its frequent co-deletion with *TP53* (Liu et al., 2015). Further, we noticed that *POLR2A* contained recurrent mutations in positions that are relevant for the protein-DNA interaction. A crystal structure supporting the functional impact of these recurrent mutations in *POLR2A* is provided in Extended Data Figure 10.

Similarly, we discovered *ANAPC1* as a significantly mutated cancer in gastroesophageal cancer (Supplementary Figure 34). *ANAPC1* is part of the anaphase-promoting complex, which is activated during mitosis after all kinetochores have become properly attached to their centromeres (Peters et al. 2002). An experimental study reported that dysfunction of the anaphase-promoting complex allows more time for the kinetochore attachment, which limits the excess of chromosomal instability in tumors harboring genomic instability (Sansregret et al., 2017). In concordance with the experimental results reported in this study, we observed that *ANAPC1* mutations were negatively associated with chromosomal instability in gastroesophageal cancer (Supplementary Figure 34d, $p = 5x10^{-4}$) (Cancer Genome Atlas Research et al., 2017). A crystal structure supporting the functional impact of mutations in *ANAPC1* is provided in Supplementary Figure 34c.

5. References

1. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* **41**, e67 (2013).
2. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271-281 e7 (2018).
3. Van Allen, E.M. *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* **20**, 682-8 (2014).
4. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
5. Jukes, T.H. Transitions, transversions, and the molecular evolutionary clock. *J Mol Evol* **26**, 87-98 (1987).
6. Bailey, M.H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e18 (2018).
7. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e21 (2017).
8. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat Genet* **49**, 1785-1788 (2017).
9. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol* **17**, 128 (2016).
10. Savage, R.S. *et al.* R/BHC: fast Bayesian hierarchical clustering for microarray data. *BMC Bioinformatics* **10**, 242 (2009).
11. Qin, Z.S. *et al.* Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol* **21**, 435-9 (2003).
12. Giraldo, J.J., Alvarez, M.A. & Orozco, A.A. Peripheral nerve segmentation using Nonparametric Bayesian Hierarchical Clustering. *Conf Proc IEEE Eng Med Biol Soc* **2015**, 3101-4 (2015).
13. Ghosh, S. & Townsend, J.P. H-CLAP: hierarchical clustering within a linear array with an application in genetics. *Stat Appl Genet Mol Biol* **14**, 125-41 (2015).
14. Balin, S.J. & Cascalho, M. The rate of mutation of a single gene. *Nucleic Acids Res* **38**, 1575-82 (2010).
15. Imielinski, M., Guo, G. & Meyerson, M. Insertions and Deletions Target Lineage-Defining Genes in Human Cancers. *Cell* **168**, 460-472 e14 (2017).
16. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
17. Brown, M.B. A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics* **31**, 987-992 (1975).
18. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B* **57**, 289-300 (1995).
19. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).
20. Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A* **113**, 14330-14335 (2016).
21. Hofree, M. *et al.* Challenges in identifying cancer genes by analysis of exome sequencing data. *Nat Commun* **7**, 12096 (2016).
22. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238-44 (2013).
23. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res* **40**, e169 (2012).
24. Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M. & Ng, P.C. SIFT missense predictions for genomes. *Nat Protoc* **11**, 1-9 (2016).
25. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).

26. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39**, e118 (2011).
27. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
28. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30**, 3109-14 (2014).
29. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805-11 (2015).
30. Futreal, P.A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83 (2004).
31. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**(2017).
32. Kumar, R.D., Searleman, A.C., Swamidass, S.J., Griffith, O.L. & Bose, R. Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics* **31**, 3561-8 (2015).