Supplementary Information: A pan-cancer landscape of somatic substitutions in non-unique regions of the human genome

Contents

Machine learning for calling somatic mutations in non-unique genomic regions

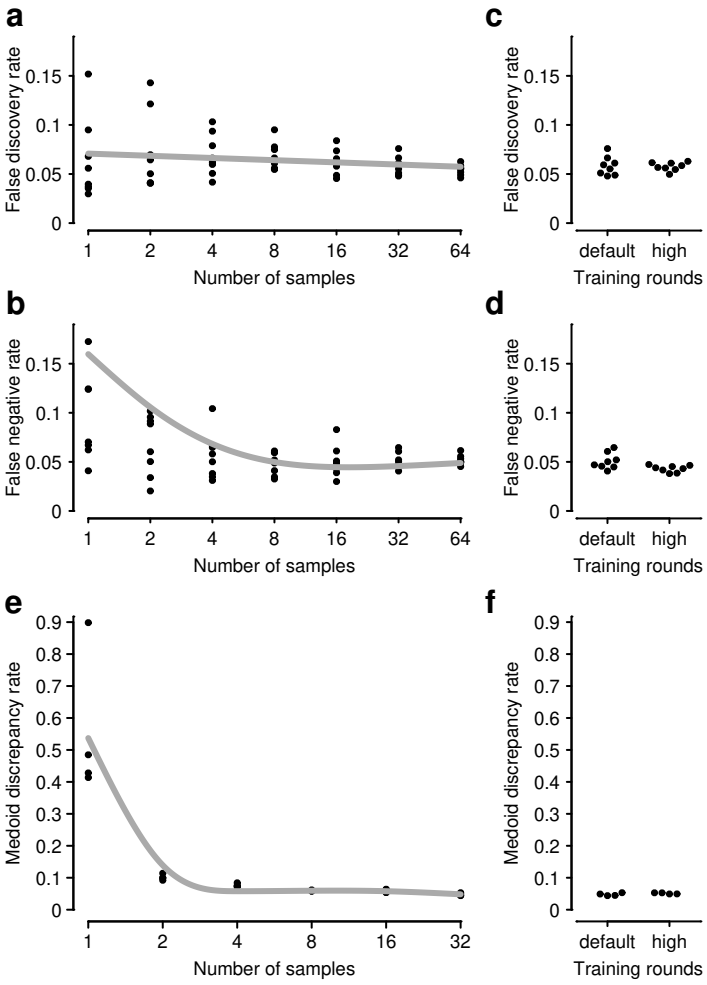


Figure S1. Training machine learning models to call somatic mutations. **(a-d)** A series of models were trained with increasing numbers of samples to explore how volume of training data affects model performance. At each size, 8 models were trained with randomly selected training samples and using 5 xgboost training rounds. All models were evaluated against a fixed test set of 50 samples. At size 32, an additional set of models were trained with 25 xgboost rounds. In all panels, dots correspond to measurements; lines are smoothing curves to summarize trend. **(a)** False positive rate measures new calls with respect consensus calls. **(b)** False negative rate measures propotion of calls in consensus not evaluated by the caller. **(c,d)** Comparison of models of size 32 trained with default and a high number of xgboost rounds. **(e,f)** Another series of models were trained using varying numbers of samples, in batches of size 4, so that each batch used non-overlapping sets of training data. **(e)** Medoid discrepancy rate measures the average difference between calls and the batch medoid; it thus captures consistency between models trained on different samples rather than ability to reproduce a consensus. **(f)** Comparison of models trained using 32 samples with the default and a high number of training rounds.

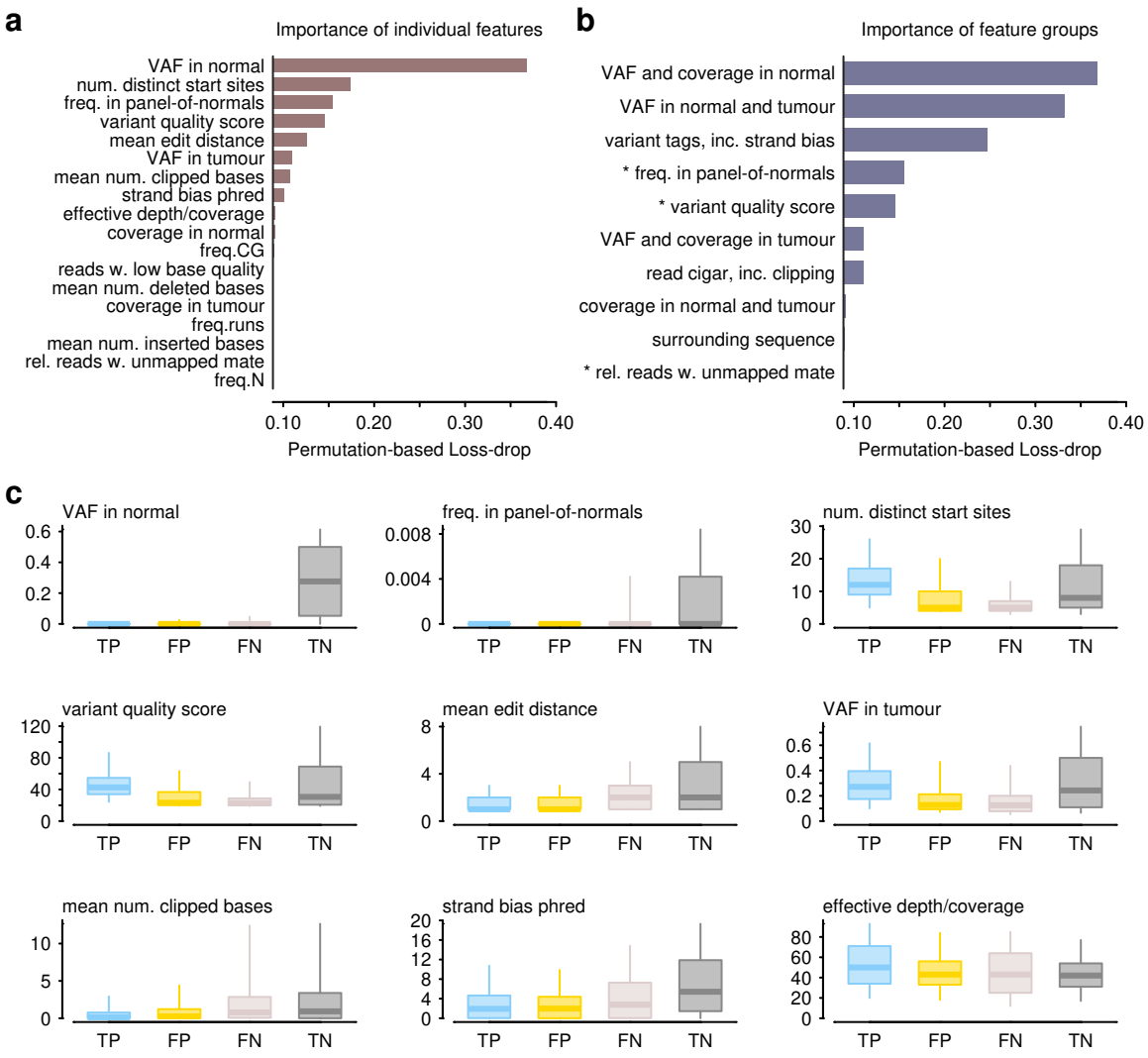


Figure S2. Feature importance for calling somatic mutations. A single classifier was trained using 150 training samples with a high number of xgboost rounds, and then evaluated using 50 held-out samples. Feature performance was evaluated by measuring the change in model performance on test data when certain variables were scrambled. **(a)** Dropout-loss after scrambling individual features indicates reveals the importance of individual variables. The vertical line indicates the baseline value of the loss function when all features are included in the model. **(b)** Analogous to (a) but showing the efect of scrambing several variables at once. Variable groups capture combination of related features, and some groups may overlap. Items marked with a (*) represent groups with a single feature and reproduce results from (a). **(c)** Distributions of feature values at positions in the test set classified as true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) with respect to PCAWG consensus calls. Box bounds, center line, and whiskers represent 25%-75%, 50%, and 5%-95% quantiles, respectively.

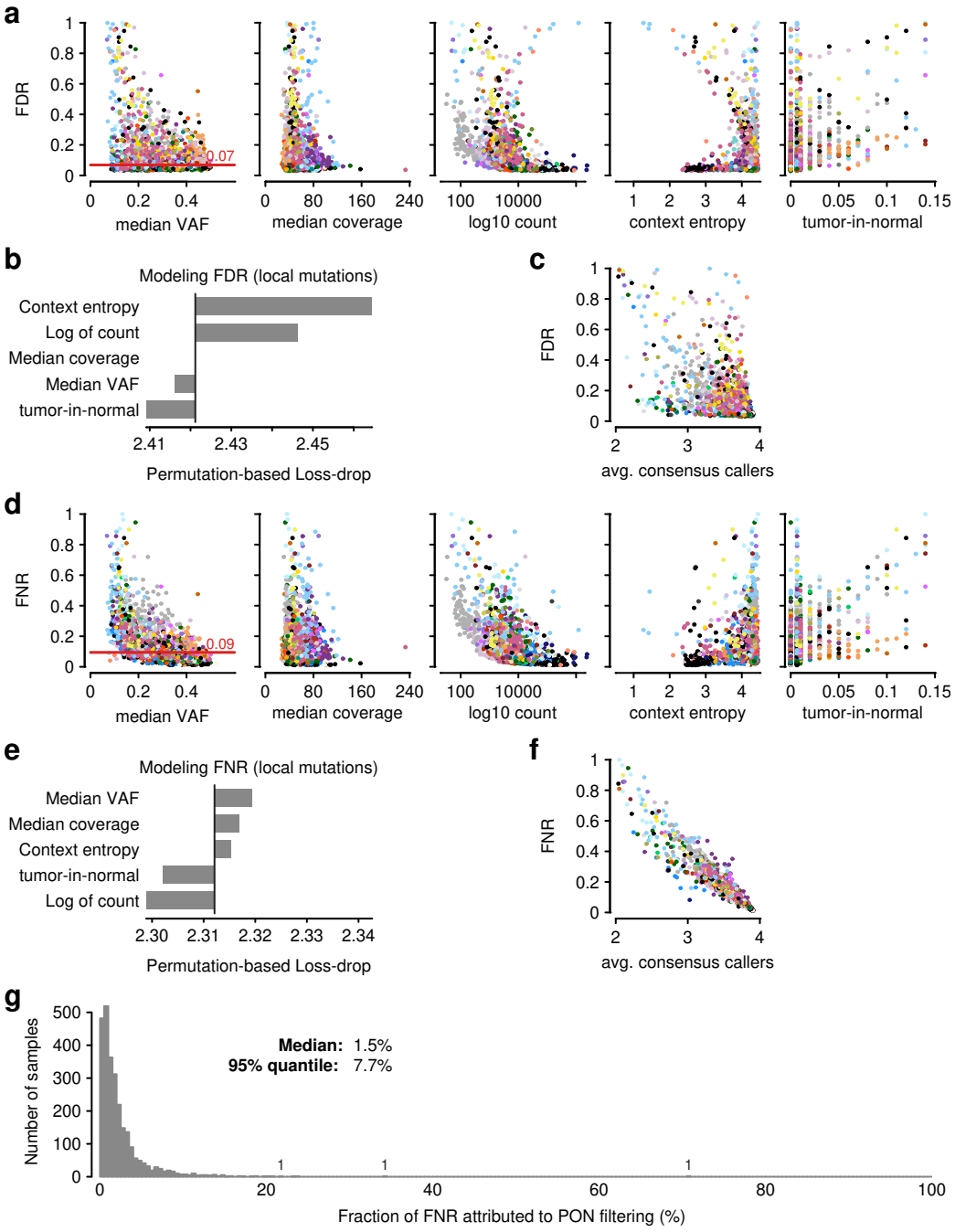


Figure S3. Factors affecting false discovery rate (FDR) and false negative rate (FNR) compared to the consensus. **(a)** Relationships between single features and the FDR. Points represent samples, colored by cancer type. Red line indicates the median FDR in the cohort. **(b)** Feature importance for explaining FDR using a generalized linear model. **(c)** Comparison of FDR against internal concordance in the consensus calls, measured by the averge number of callers that call mutation sites in the sample. **(d,e,f)** Analogous to (a,b,c) for FNR. **(g)** Fraction of FNR that can be attributed to filtering variants because of presence in the panel-of-normals (PON). Numbers along the histograms indicate sample counts in isolated bins.

Characteristics of thesaurus mutations across cancer types

a

consensus	miss	PASS	thesaurus
yes	FN	TP	thesP
no		FP	thesD

TPR = TP / (FN + TP + thesP)
FNR = FN / (FN + TP + thesP)
FDR = FP / (TP + FP + thesP)
thesPR = thesP / (FN + TP + thesP)
thesDR = thesD / (FN + TP + thesP + thesD)

b

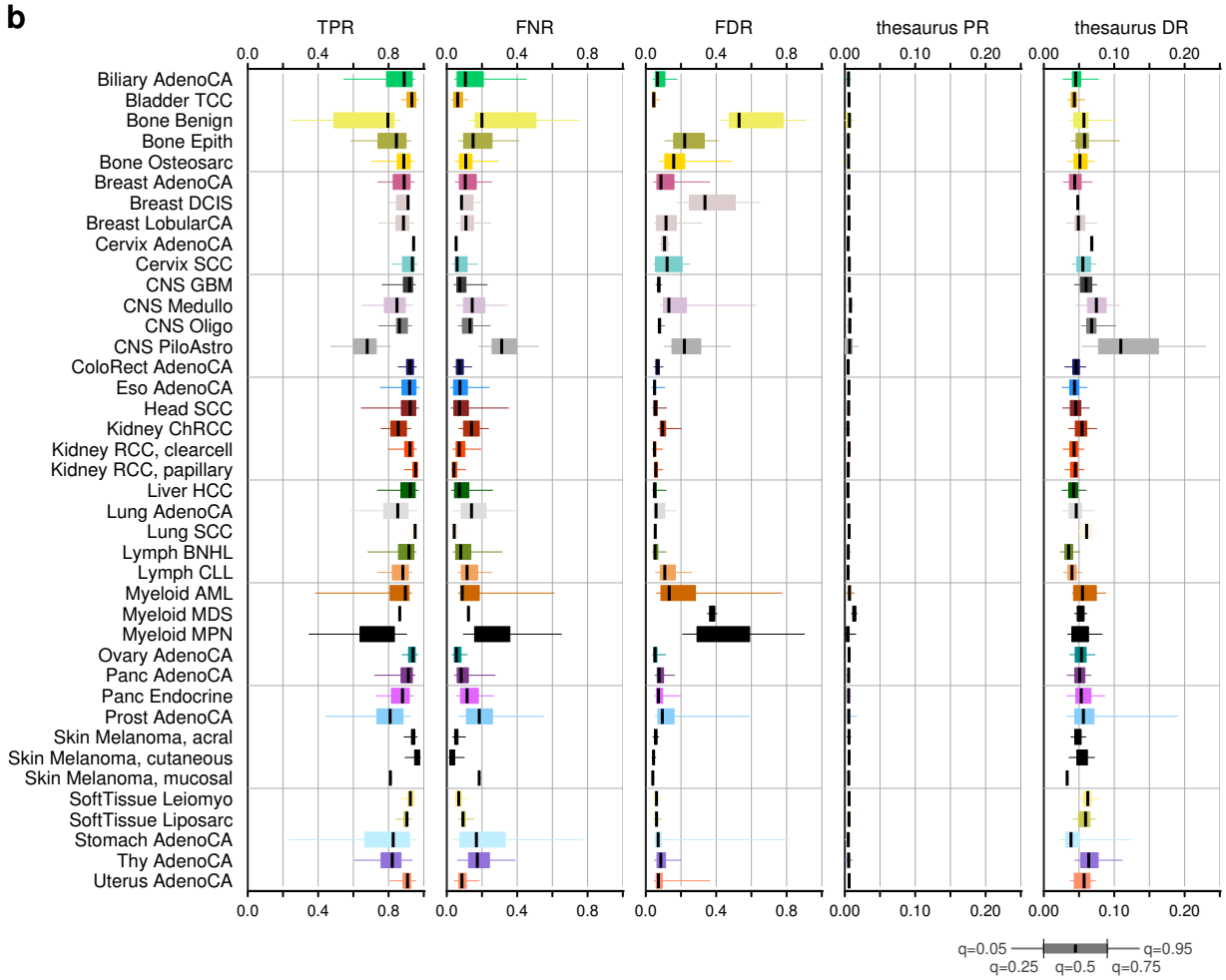


Figure S4. Comparison of mutation calls with consensus. **(a)** Definition of mutation set comparison measures, using the consensus call set as a 'ground truth.' FN, false negative; TP, true positive; FP, false positive; thesP, thesaurus positive, i.e. a site labeled as thesaurus linked that is also in the consensus set; thesD, thesaurus discovery, i.e. a novel site labeled as thesaurus linked; TPR, true positive rate; FNR, false negative rate; FDR, false discovery rate; thesPR, thesaurus positive rate; thesDR, thesaurus discovery rate. **(b)** Summary of mutation measures stratified by histology. Each box summarizes a performance measures across samples in the cohort. Box bounds, center line, and whiskers represent 25%-75%, 50%, and 5%-95% quantiles, respectively.

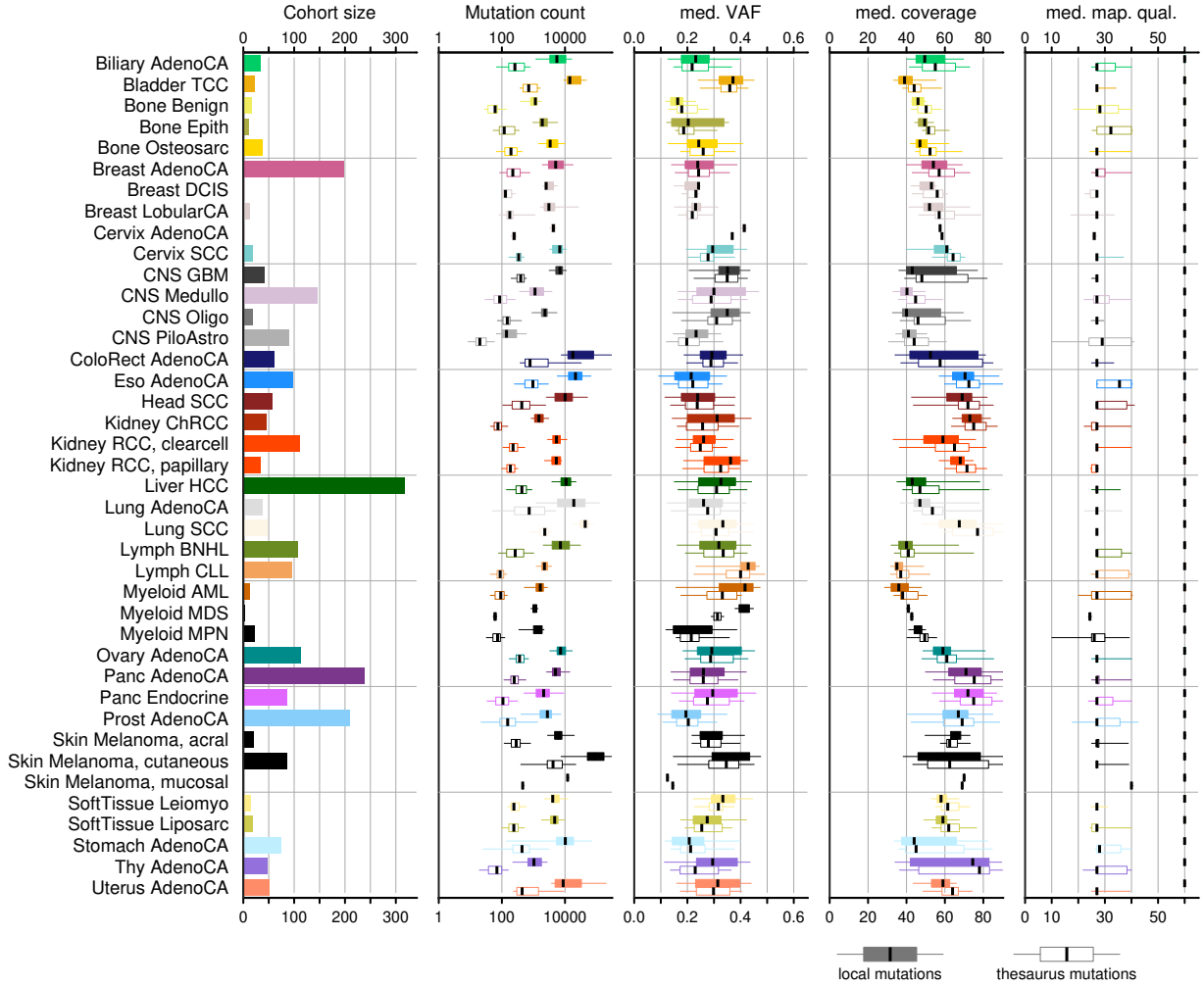


Figure S5. Characteristics of local and thesaurus mutations. The left-most panel summarizes the number of samples in each histology cohort. Subsequent panels (left-to-right) summarize distributions of mutation counts, median mapping quality at mutation sites, median variant-allelic frequency (VAF) of the mutant allele, and median coverage at the mutation site. Box mid-line, edges, and whiskers represent 50%, 25%-75%, and 5%-95% quantiles. As expected, thesaurus mutations are less numerous than local mutations. Distributions of allelic frequency and coverage are similar for the two groups. Thesaurus mutations are supported by reads with lower mapping quality.

Validation of thesaurus mutations with linked-read sequencing

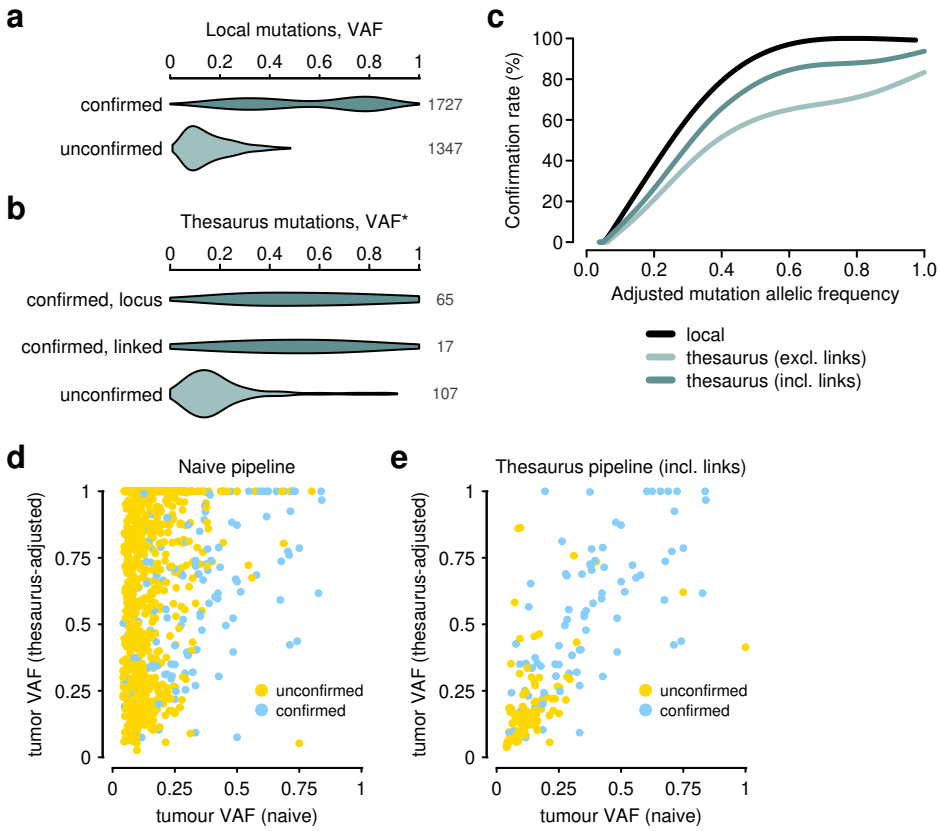


Figure S6. Comparison of mutations detected in short-read and long-read data in an independent dataset. (a) Variant-allele frequency (VAF) of local mutations as measured in short-read data. The mutations are stratified into a group that is confirmed in the long-read data and a group that was not detected in the long-read data. Numbers on the right-hand-side denote the number of mutations in each group. **(b)** Analogous to panel (a), but summarizing thesaurus mutations. One of the stratification groups includes items not detected in the long-read data, but where a linked site was detected in the long-read data. Here, the x-axis is based on thesaurus-adjust allelic frequency. **(c)** Modeling of confirmation rate of mutations as a function of VAF. VAF is measured in the short-read sample using thesaurus adjustment. Lines correspond to spline models with four degrees of freedom. Models consider: local mutations in unique regions, thesaurus mutations with primary flag and evaluated at nominal locus without links (excl. links), thesaurus-filter mutations evaluated using primary flag and links (incl. links). **(d)** Comparison of naive- and theasaurus-adjusted allelic frequency among mutation candidates from a naive pipeline. The validation rate is high among sites with high naive allelic frequency, but the large number of candidates at intermediate frequencies suggest there may be high contamination with germline sites that are 'validated' in the long-read sample. **(d)** Comparison of naive- and thesaurus-adjusted allelic frequncy among candidates from the thesaurus pipeline.

Trinucleotide mutation profiles

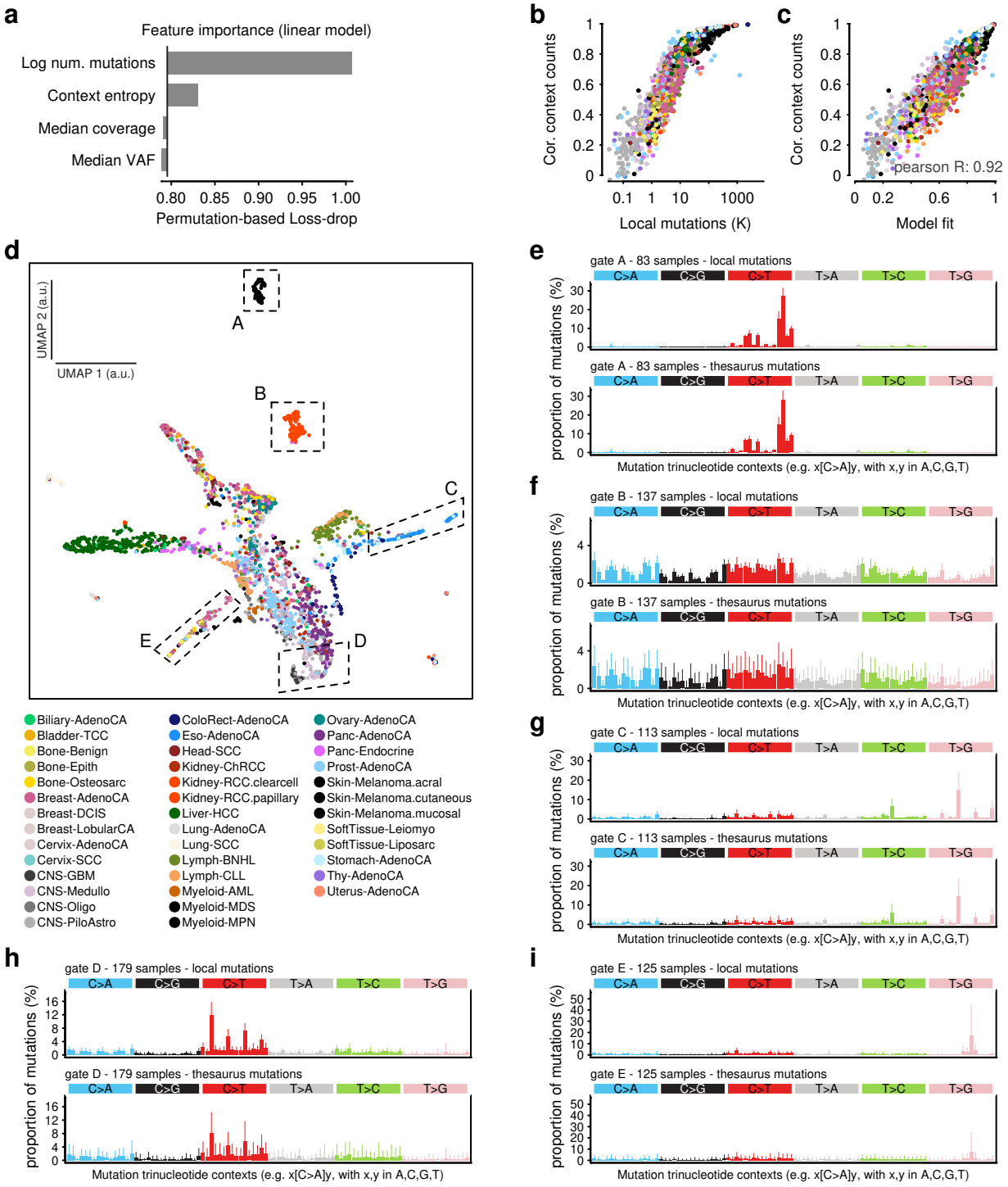


Figure S7. Trinucleotide mutation profiles. Caption on next page.

Figure S7 caption. **(a)** Modeling of correlation (between local and thesaurus mutation profiles) using four explanatory variables. Horizontal axis is the drop in model fit when a variable is omitted from the model. **(b)** Relation between the single most important variable in (a) and correlation. **(c)** Comparison of model fit (based on the single most relevant variable) and observed correlation values. **(d)** UMAP embedding computed using VAF-adjusted mutation profiles of local mutations. Gatings define manually-selected sample groups. **(e-i)** Pairs of panels correspond to gates in the embedding diagram. First panel displays the average mutation profile based on local mutations (PASS); second panel displays thesaurus profiles in the same samples. Error bars denote 5% and 95% intervals.

Mutation burden in functional genomic regions

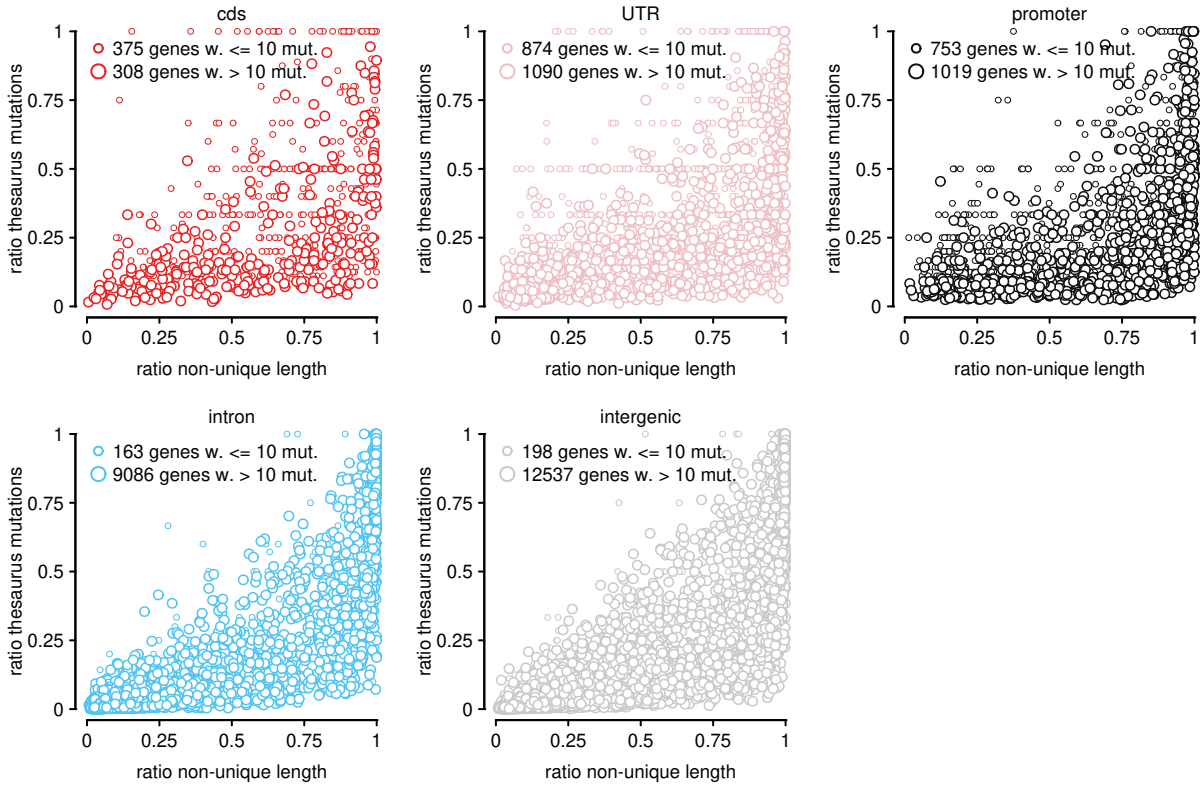
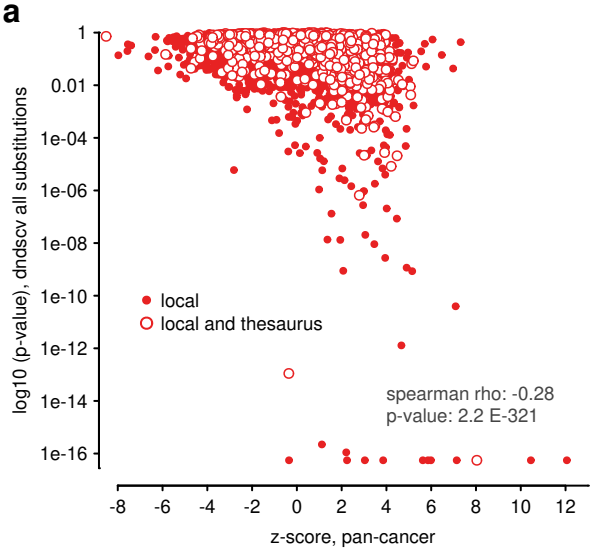


Figure S8. Mutation burden in functional regions with mixed - unique and non-unique - sequences. All dots represent genes, separated by total mutation count into low-count and large-count genes. Panels show properties of coding sequences, untranslated regions, promoters, introns, and intergenic regions. In all panels, x-axes represent the proportion of the gene sequence declared as non-unique. y-axes show the proportion of thesaurus mutations in the gene; mutation counts exclude hyper-mutated samples.

a



b

gene symbol	thes perc	z score	p value	q value
PIK3CA	5.2	8.04	5.55e-17	8.58e-14
KMT2C	11.4	-0.36	1.11e-13	1.49e-10
HLA-A	14.3	2.79	6.55e-07	0.000425
ZNF99	34.1	4.21	8.29e-06	0.00379
TRIM64B	100	4.48	2.08e-05	0.00871
AMY1B	100	3	2.17e-05	0.00874
TRIM49	92	3.92	2.76e-05	0.0101
OR4Q3	15.4	2.89	0.000228	0.0647
POTEM	100	3.43	0.000254	0.071
OR11H1	100	3.48	0.000285	0.0773
PSG6	5.6	3.06	0.000418	0.108
AMY1A	100	2.21	0.000471	0.115
TRIM49C	88.9	3.08	0.000494	0.118
POTEH	97.1	4.4	0.000642	0.143
MUC19	0.7	5.09	0.00427	0.47
ANKRD30A	15.9	5.06	0.00901	0.634
OR4A15	3.4	4.7	0.0108	0.661
ZNF208	19.3	4.53	0.0156	0.737
FLG	23	5.12	0.0658	0.93
TPTE	15.2	5.2	0.0846	0.93

Figure S9. Comparison of gene prioritization schemes. **(a)** Comparison of p-values produced by dndscv (all substitutions) against z-scores. Dots represent genes with coding sequence, stratified according to whether they hold only local mutations, or a combination of local and thesaurus mutations. **(b)** Details on a selection of most-significant genes, ranked according to p-value but including genes with high z-scores. The fraction of thesaurus mutations in each gene is indicated as a percentage in column 'thes perc'.

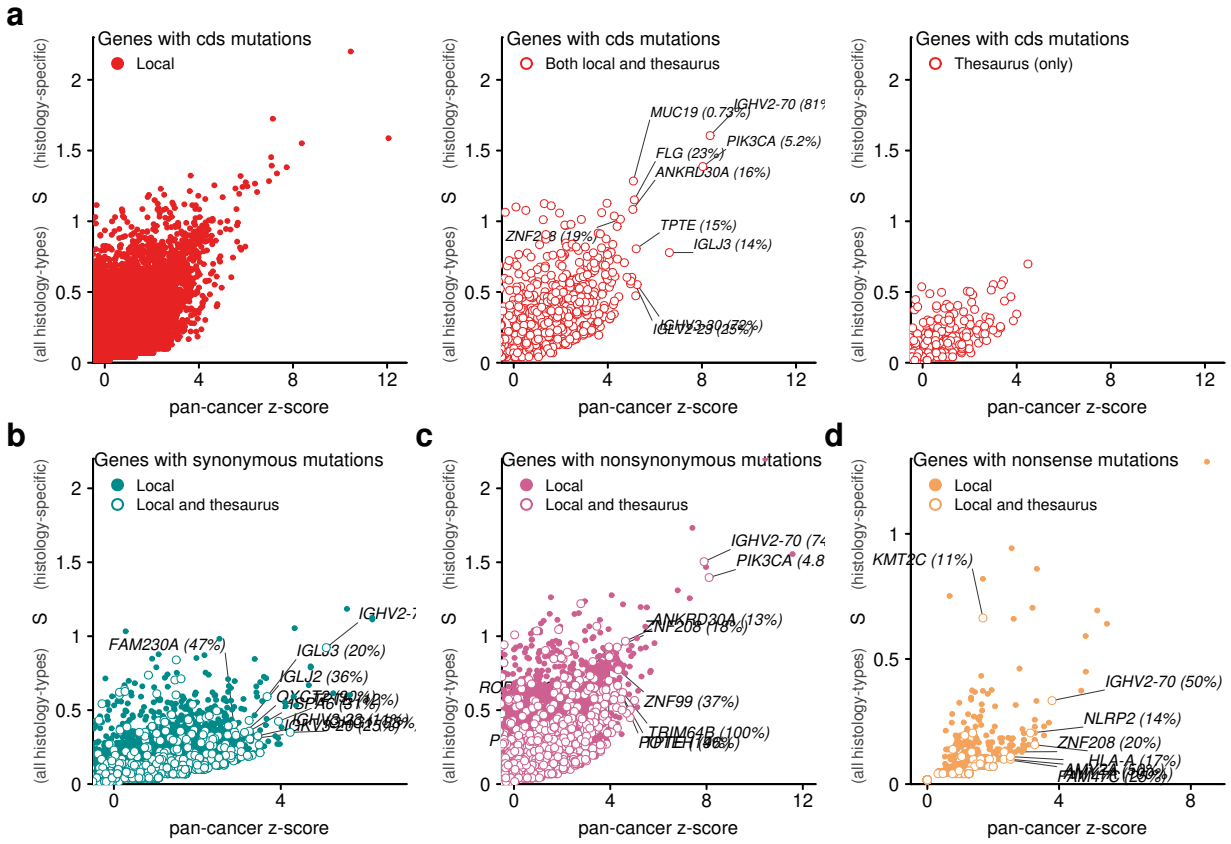


Figure S10. Specificity of mutations in gene coding sequences. All panels show over-representation of mutations in the pan-cancer cohort (z-score, x-axis) and the specificity of the mutation distribution across cancer types (change in entropy, y-axis). **(a)** Views using mutations in gene coding sequences. Sub-panels show the set of mutations with only local mutations, a set of genes that have both local and thesaurus mutations along their gene body, and a set of genes with only thesarus mutations. All panels are displayed on the same scale. **(b-d)** Summary of mutations split by their effect at the protein level: synonymous, nonsynonymous, and nonsense.

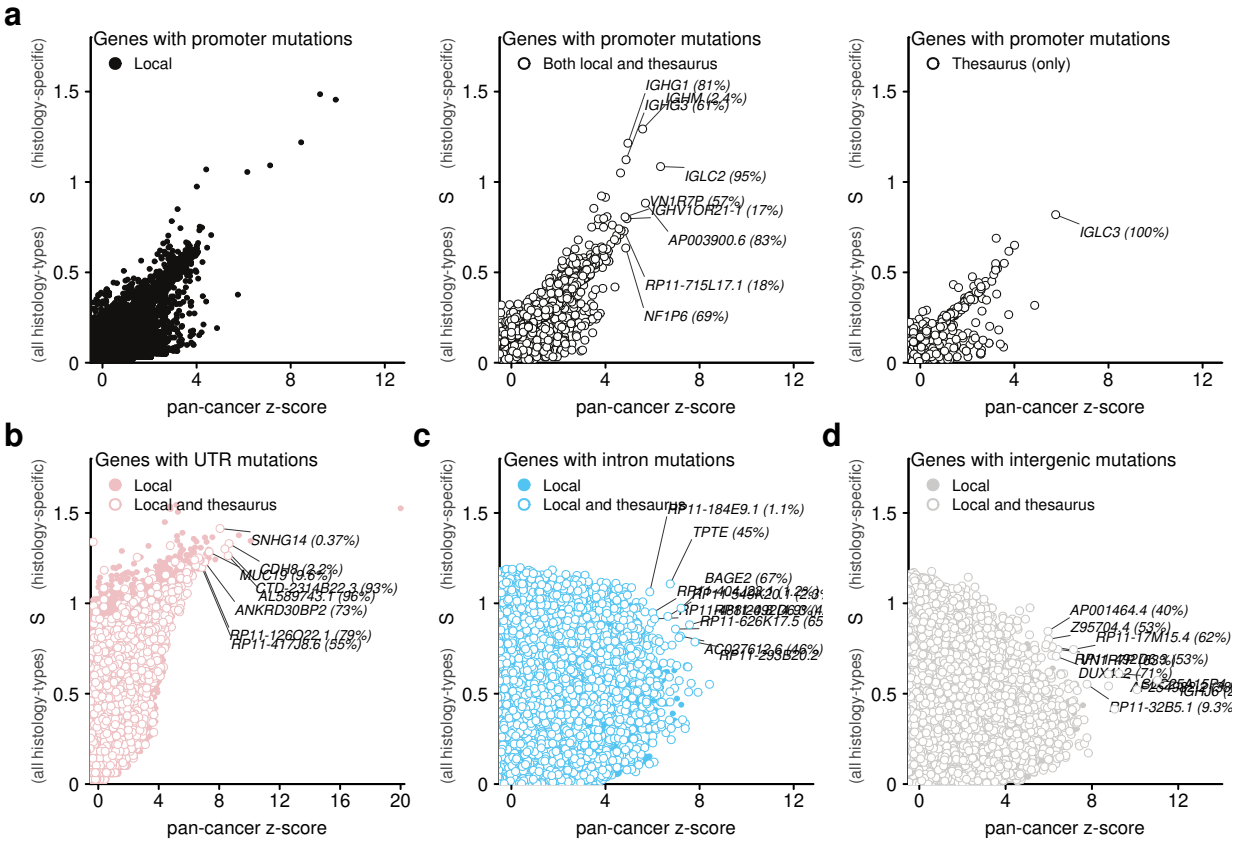


Figure S11. Specificity of mutations. All panels show the over-representation of mutations in the pan-cancer cohort (z-score, x-axis) and the specificity of the mutation distribution across cancer types (change in entropy, y-axis). **(a)** Views of mutations promoter regions. Sub-panels show the set of genes with local mutations only, the set of genes with both local and thesaurus mutations, and the set of genes with thesaurus mutations (only). **(b-d)** Summary of mutations in other genomic regions: untranslated regions (UTR), intronic regions, intergenic regions.

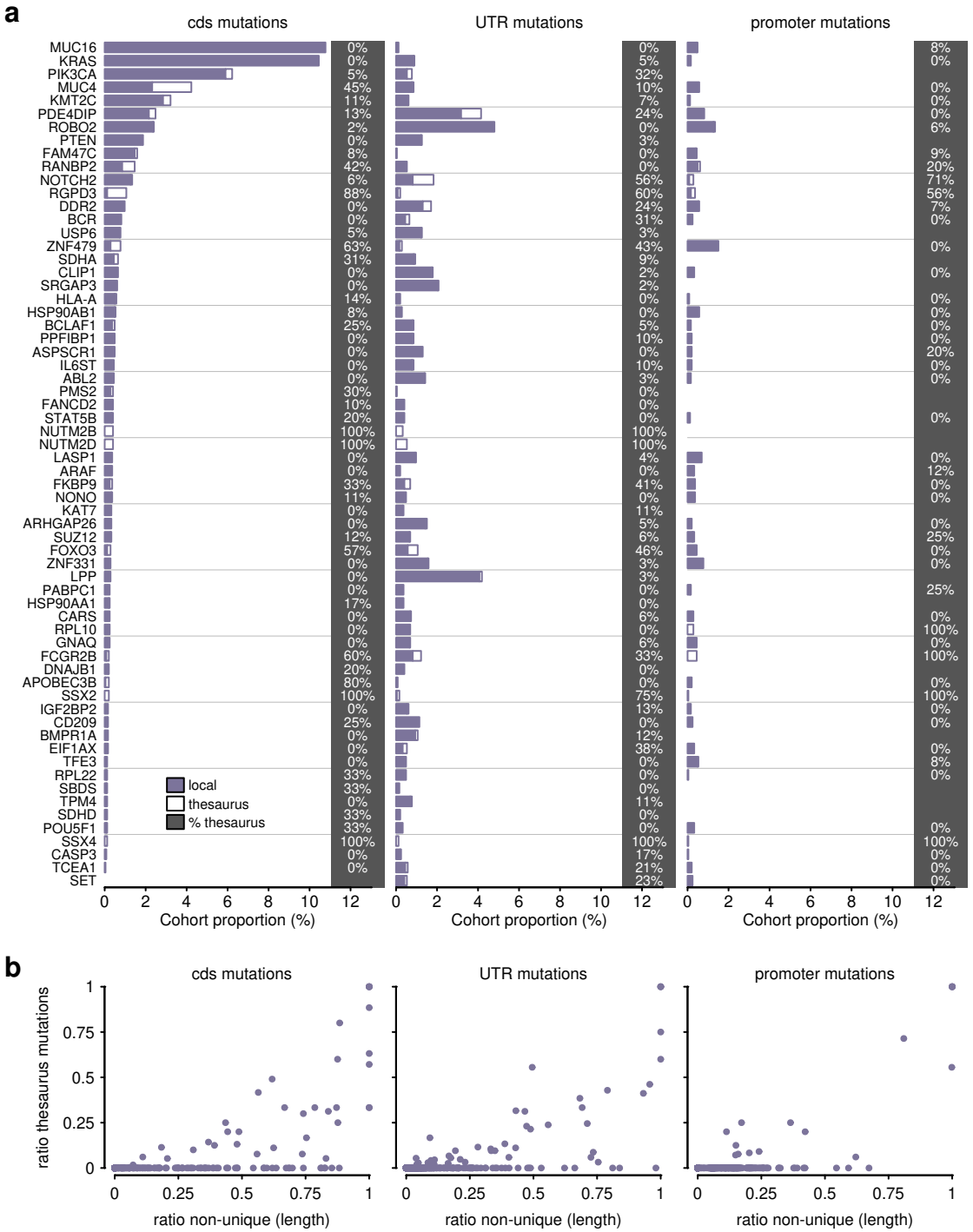


Figure S12. Genes in cancer gene census that carry thesaurus mutations in cds, UTR, and promoter regions. (a) Bars indicate the proportion of patients in cohort that carry mutations in CGC genes (only genes with at least one thesaurus mutation in either cds, UTR or promoter regions are included). Bars are split to separate patients with at least one local mutations from those that carry only thesaurus mutations. Percentages on right-hand-side quantify the proportion of samples that carry exclusively thesaurus mutations. Genes with an indicator of 0% may include thesaurus mutations alongside local mutations in the same patient. **(b)** Relation between the proportion of thesaurus mutations (compared to all mutations) and the size of the non-unique region (compare to entire gene sequence).

Examples of individual genes with thesaurus mutations

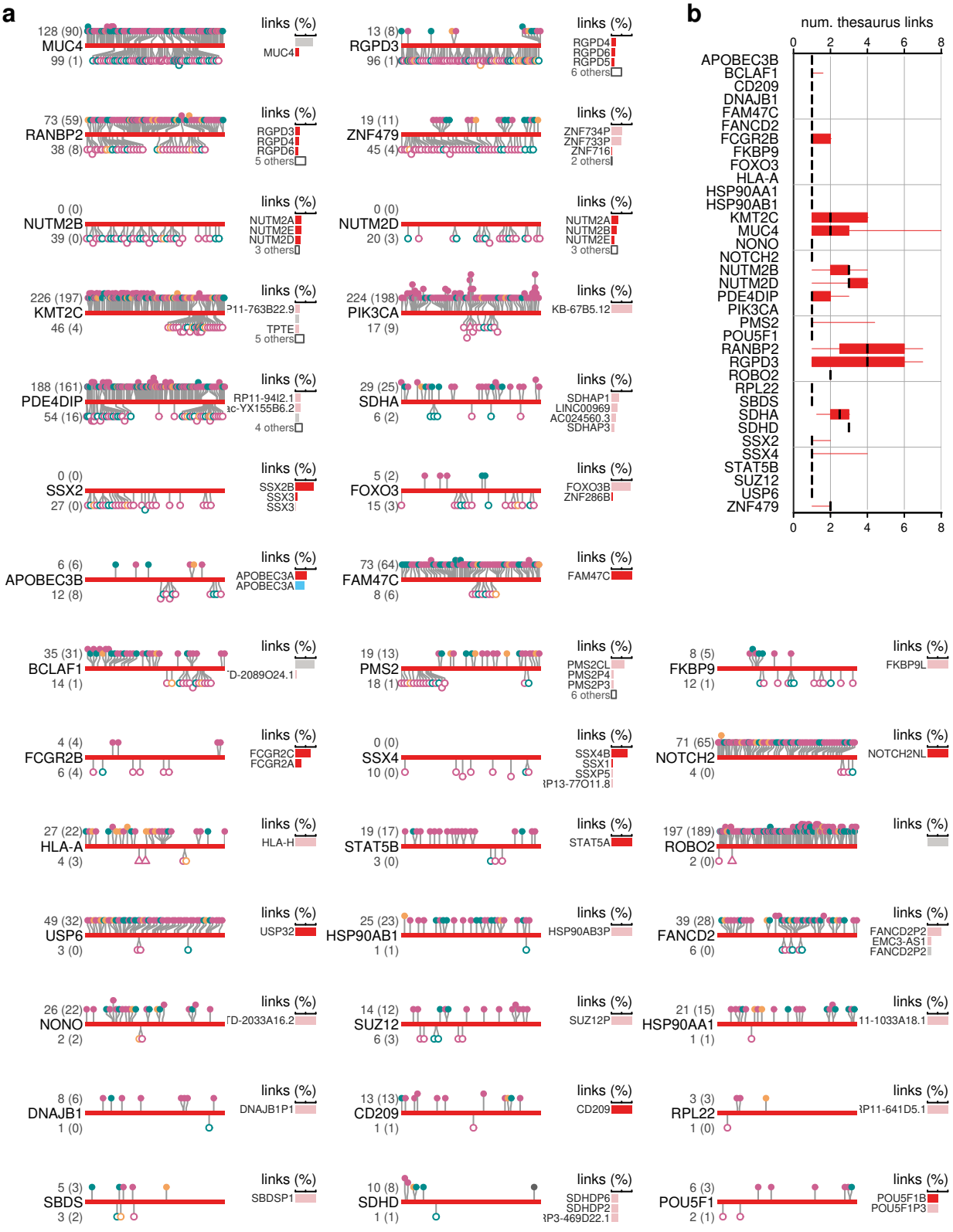


Figure S13. Mutations along coding regions of cancer gene census genes. Caption on next page.

Figure S13 caption. (a) Diagrams denote coding sequences as continuous bars. Lollipops display mutations. Bar charts indicate genomic locations that thesaurus mutations link toward; horizontal scale is 0% to 100%. **(b)** Distributions of the number of thesaurus links originating from each gene in (a). Box bounds, center line, and whiskers represent 25%-75%, 50%, and 5%-95% quantiles, respectively.

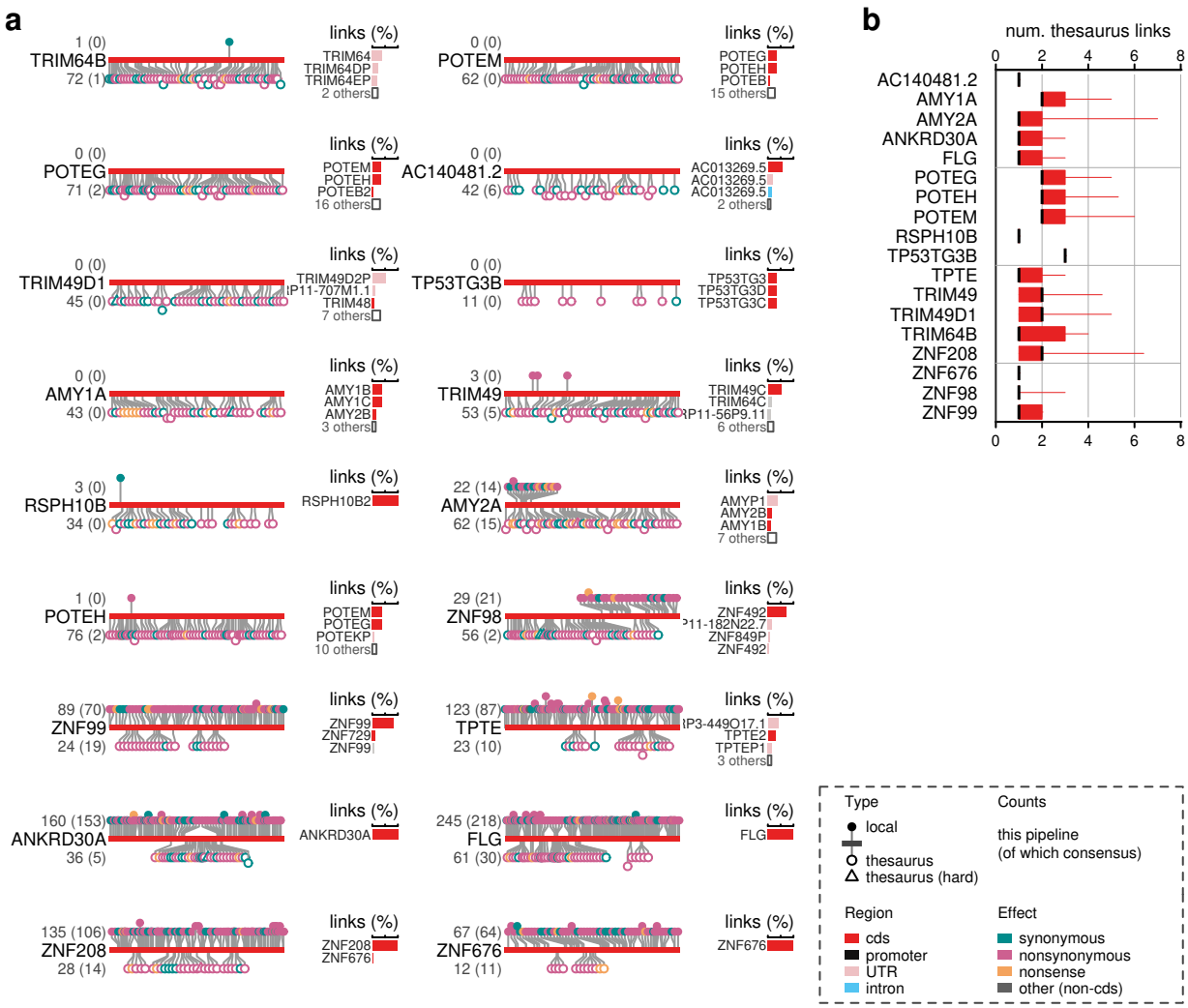


Figure S14. Selected genes with thesaurus mutations in coding regions. **(a)** Examples are structured as in the previous figure. **(b)** Distributions of number of links originating from thesaurus mutations. Genes correspond to examples in (a). Box bounds, center line, and whiskers represent 25%-75%, 50%, and 5%-95% quantiles, respectively.

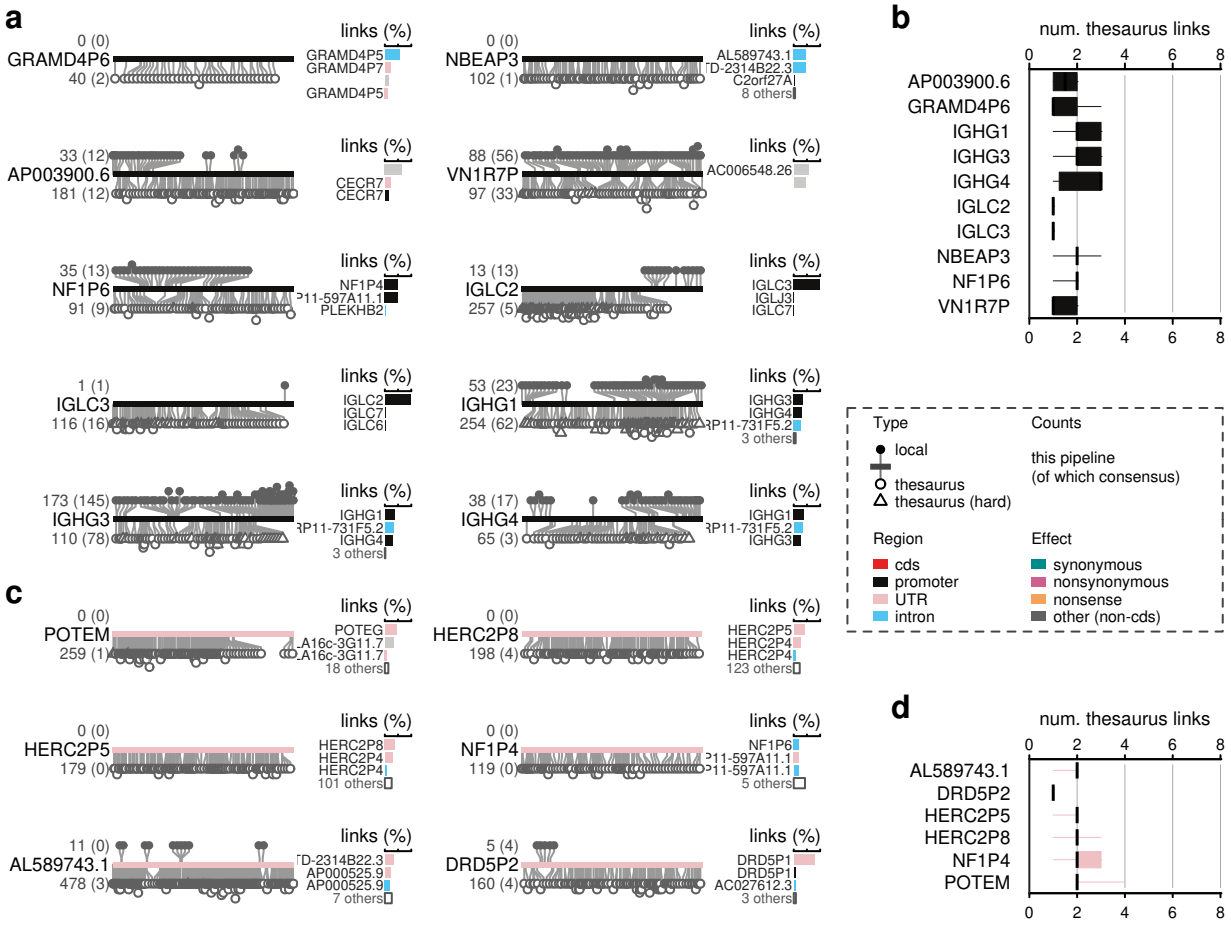


Figure S15. Selected genes with thesaurus mutations in non-coding regions. **(a)** Examples of promoters. **(b)** Distributions of the number of links origating from thesaurus mutations in promoter regions. Box bounds, center line, and whiskers represent 25%-75%, 50%, and 5%-95% quantiles, respectively. **(c,d)** Similar to (a,b) with untranslated (UTR) regions.

Examples of gene sets with thesaurus mutations

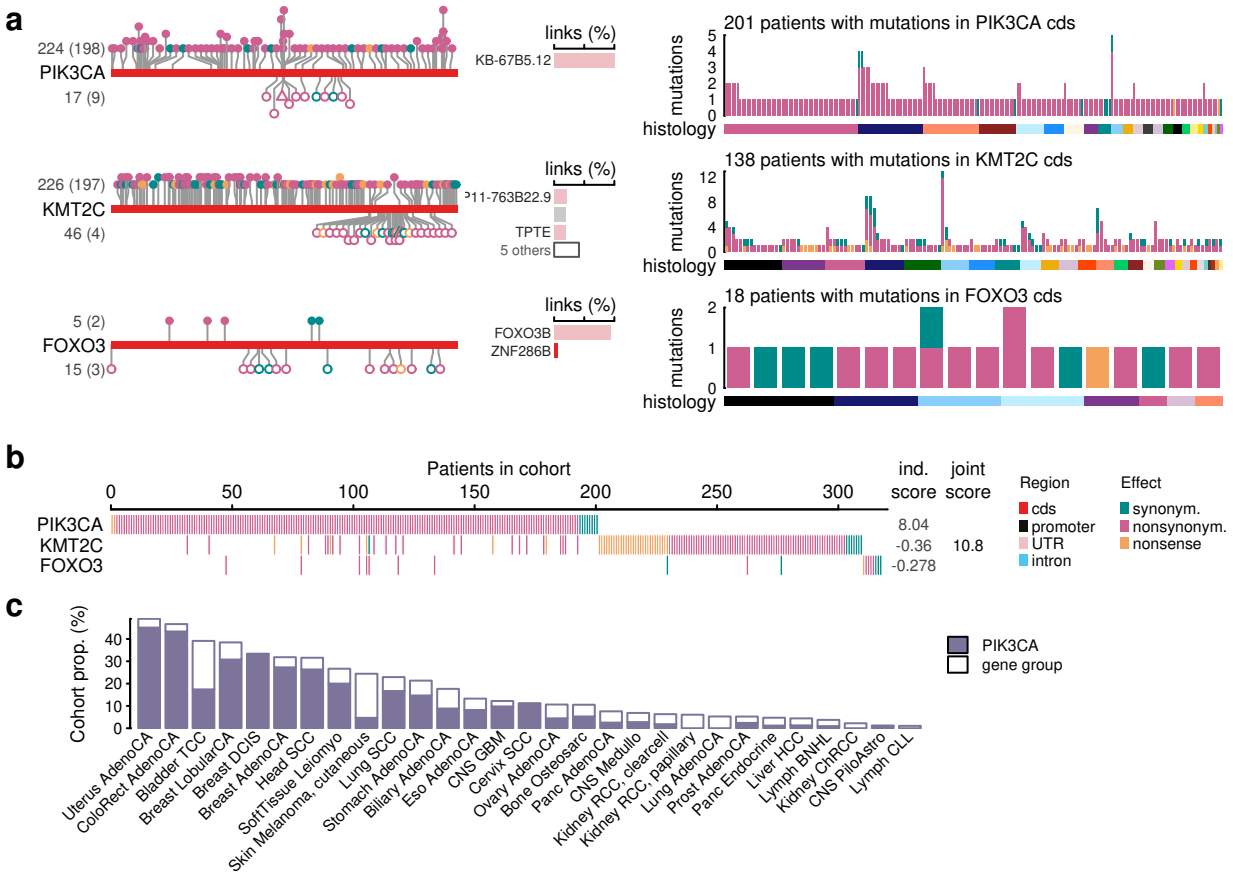


Figure S16. Cohort view of selected cancer genes. **(a)** Distributions of mutations along selected genes. Panels on the right reveal how the mutations are partitioned among patients and cancer types. **(b)** Cohort view of the gene group in (a). For patients with more than one mutation in a gene, only one the most severe consequence is shown (nonsense, nonsynonymous, synonymous). **(c)** Proportions of patients in each cancer type carrying at least one mutation in the leading gene and in the gene group. This summary includes all samples, including hyper-mutator samples.

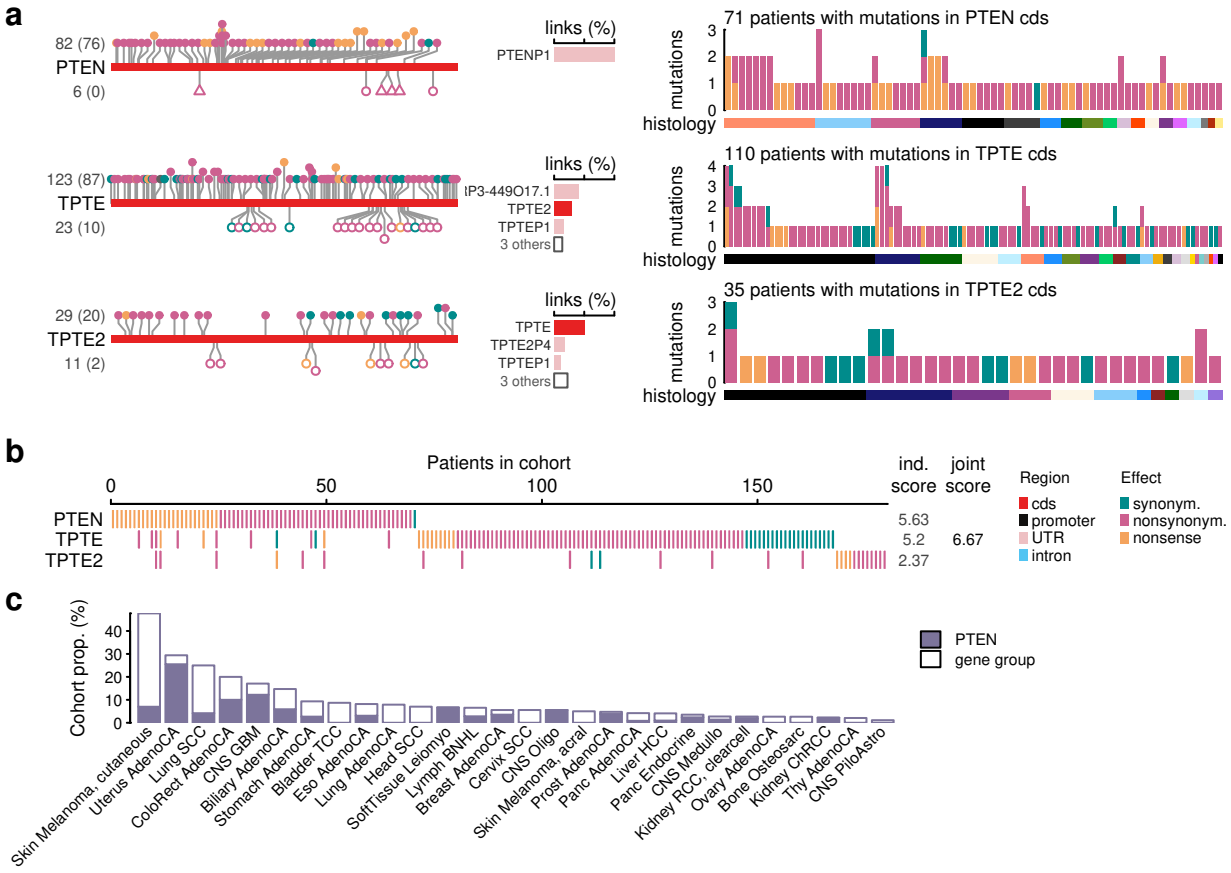


Figure S17. Cohort view of genes pertaining to the PTEN pathway.

Examples of gene families linked via thesaurus annotations

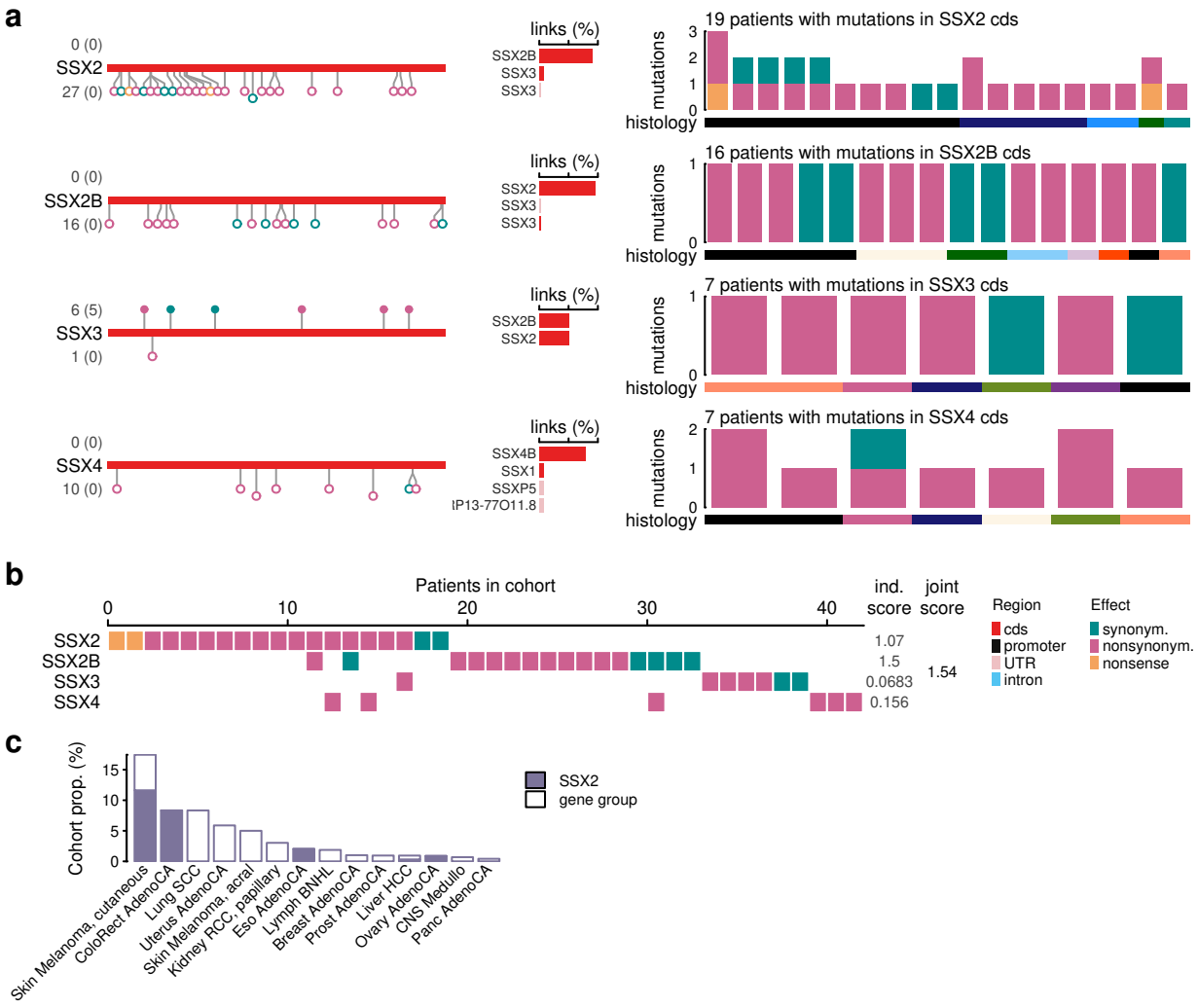


Figure S18. Cohort view of the SSX gene family.

Figure S19. Cohort view of the NUTM2 gene family.

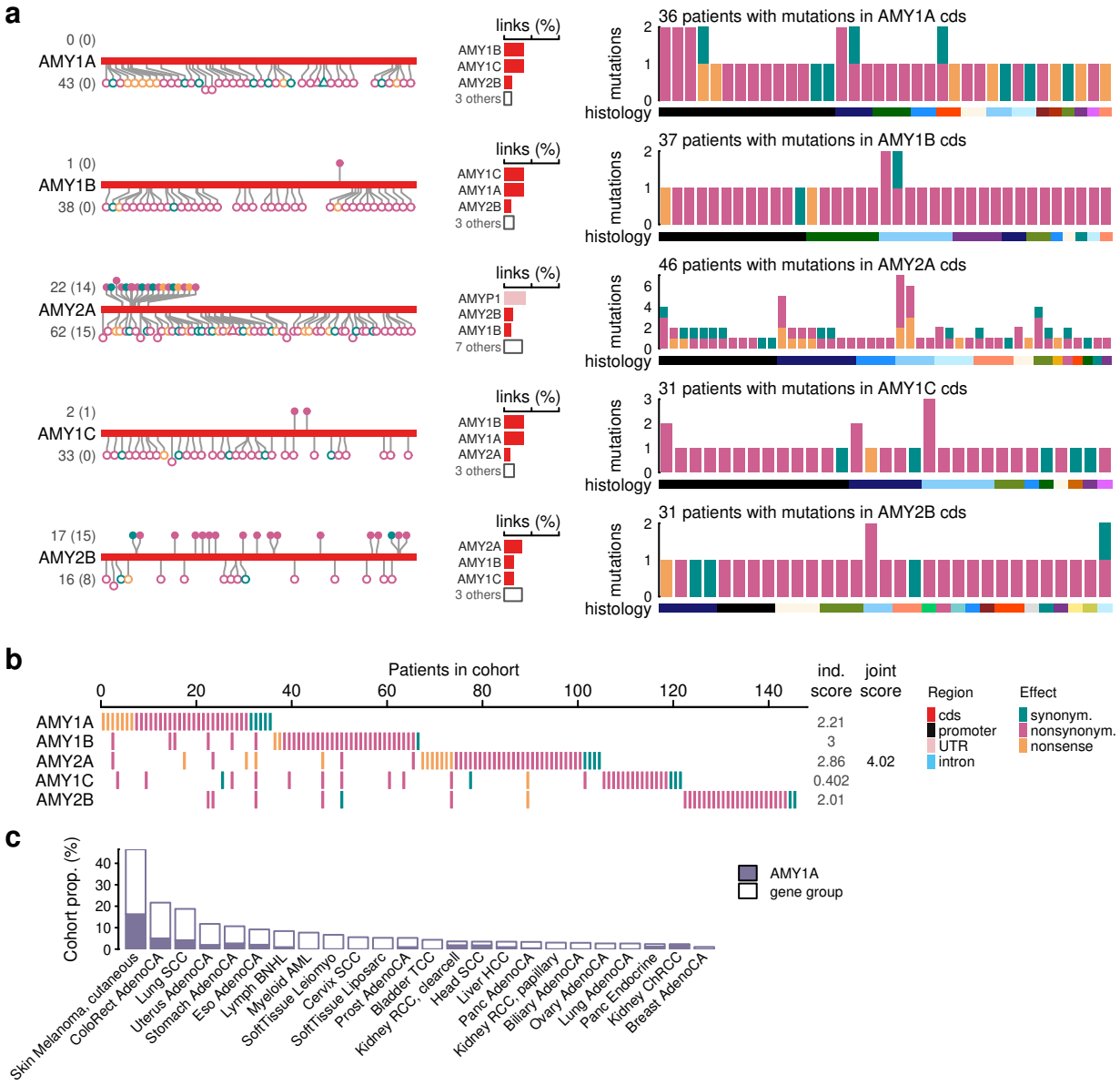


Figure S20. Cohort view of a group of genes from the AMY gene family.

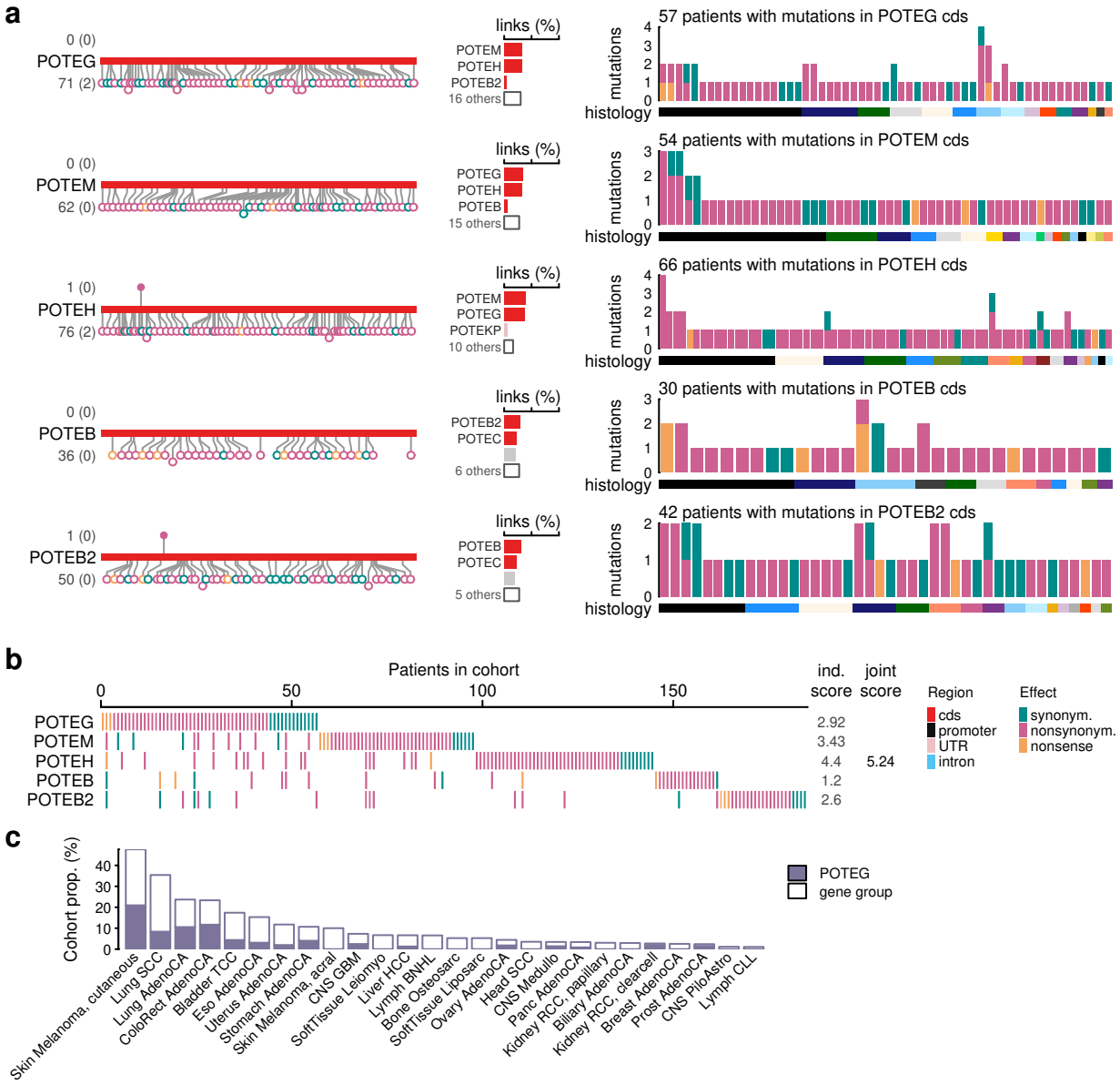


Figure S21. Cohort view of a group of genes from the POTE gene family.

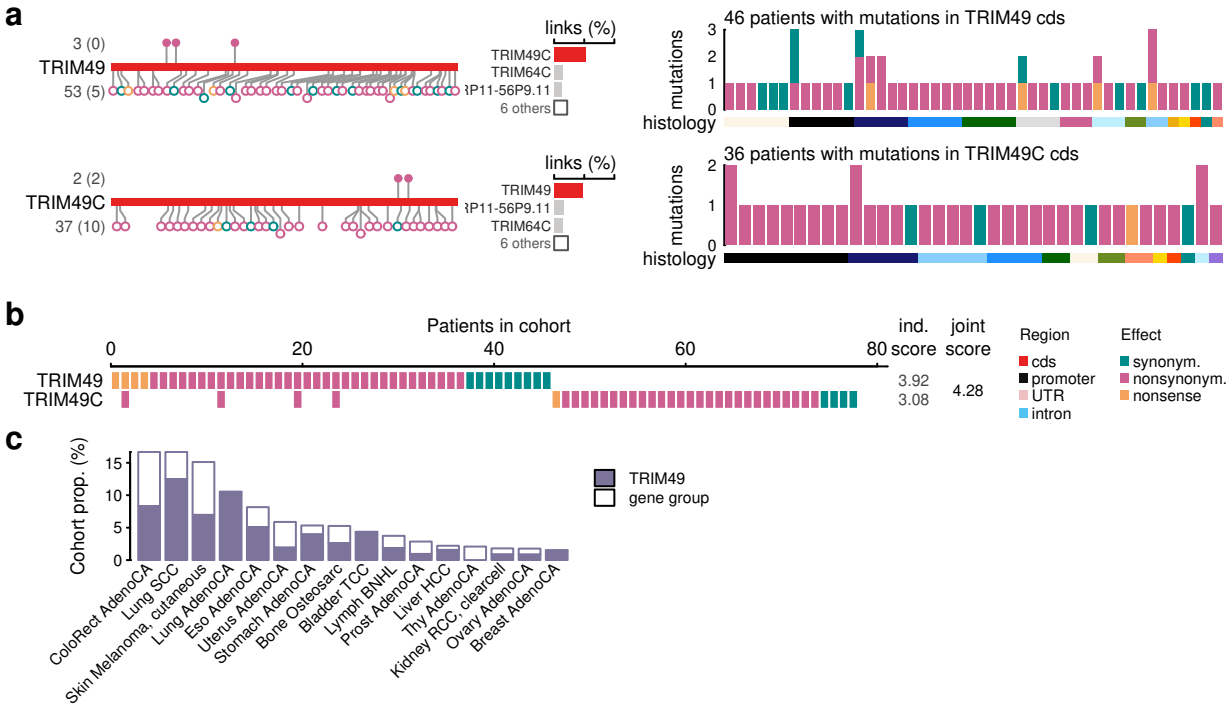


Figure S22. Cohort view of a group of genes from the TRIM gene family.

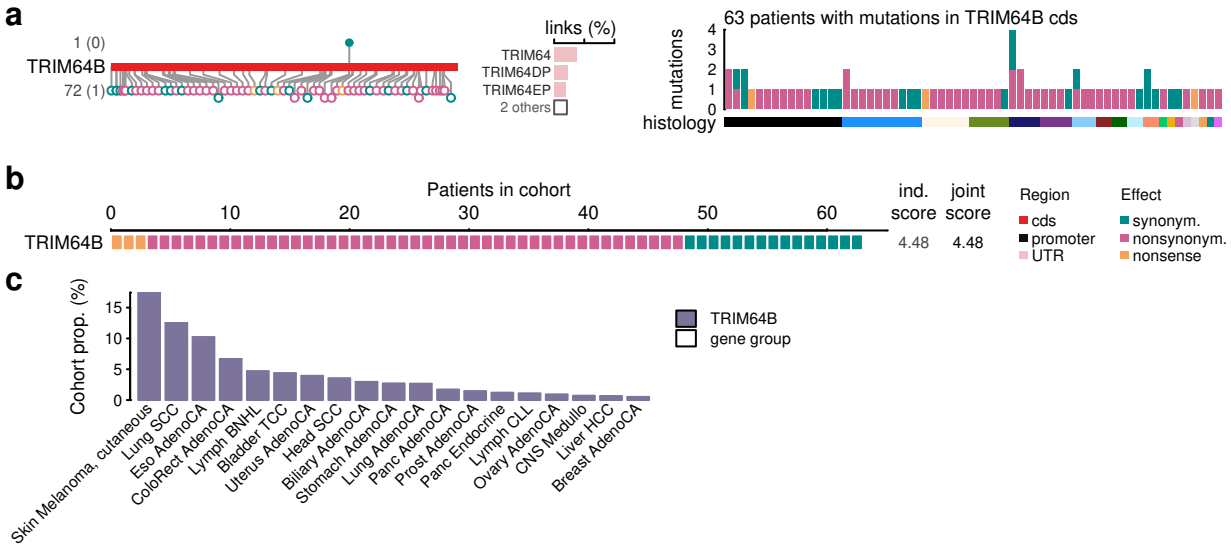


Figure S23. Cohort view of a genes in the TRIM gene family.

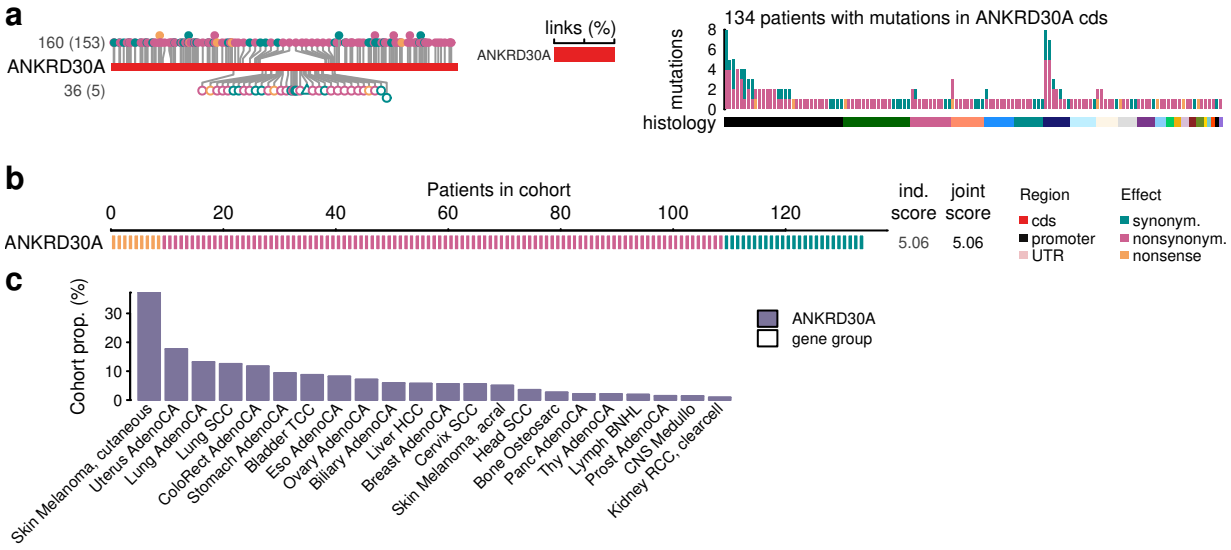


Figure S24. Cohort view of a gene in the ANKRD family.

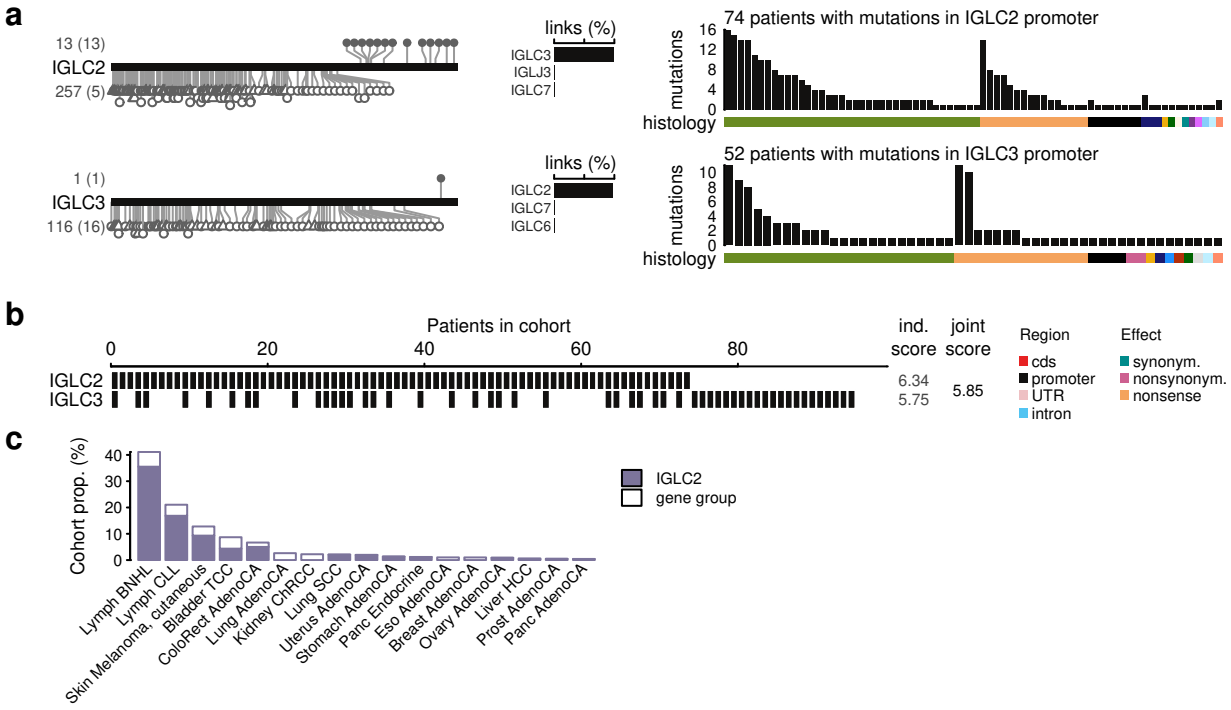


Figure S25. Cohort view of a group of promoters (upstream sequences) of the IGLC family.

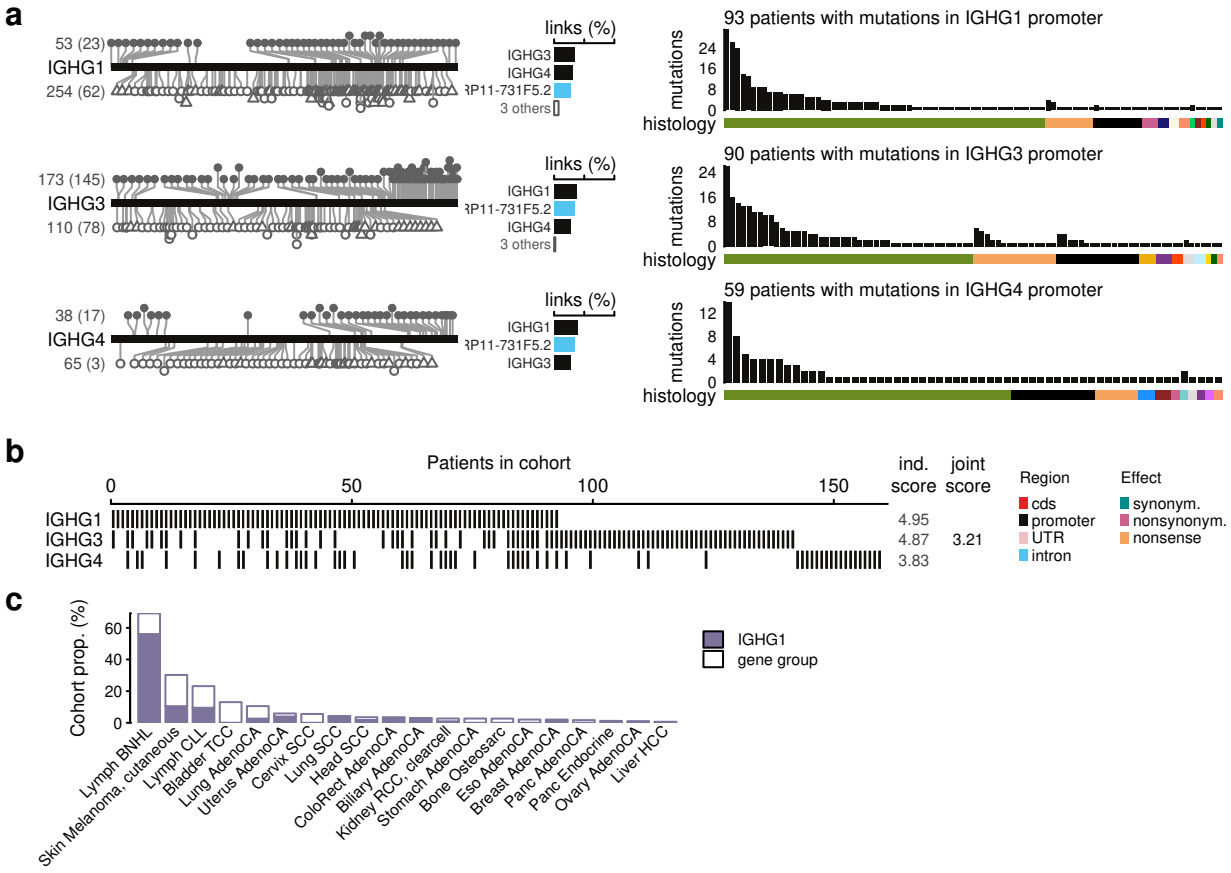


Figure S26. Cohort view of a group of promoters (upstream sequences) of the IGHG family.

Expression of thesaurus mutations in transcriptomic data

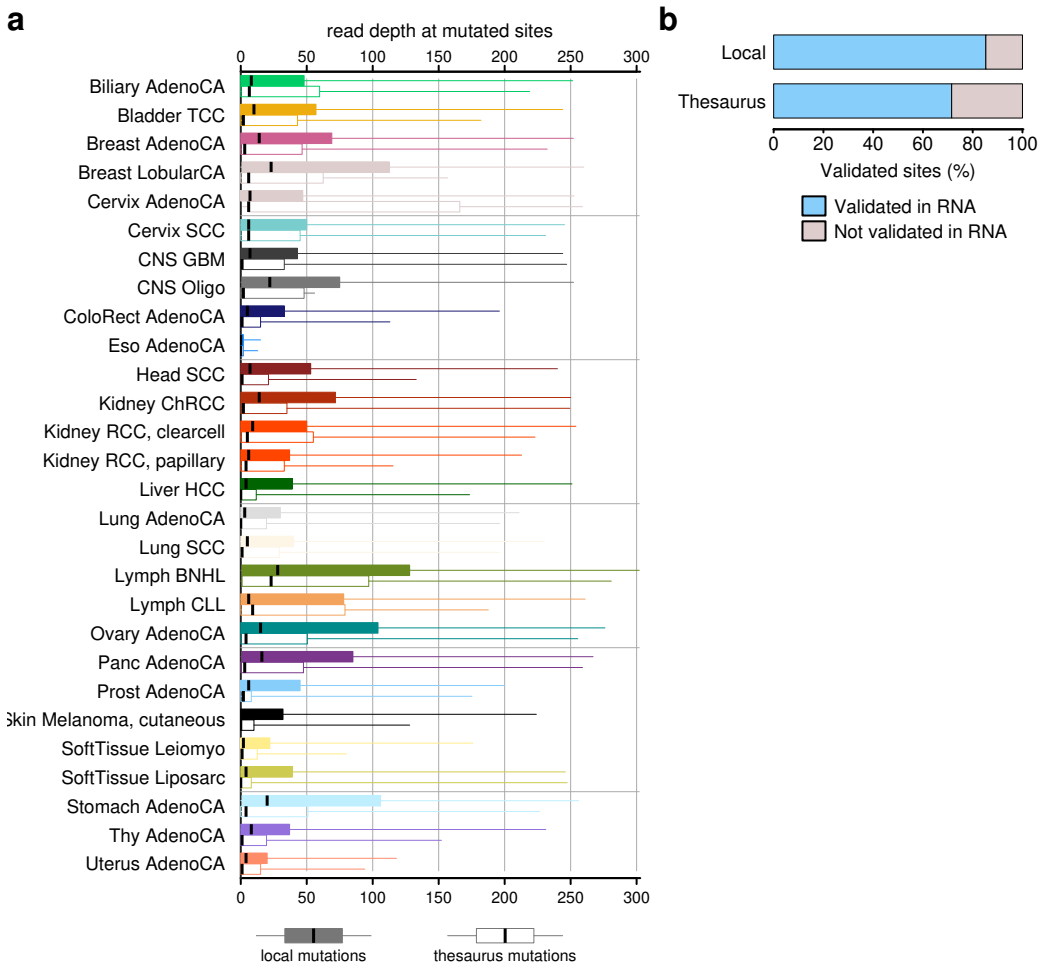


Figure S27. Properties of thesaurus mutations in RNA. (a) Read depth observed in tumour RNA-seq at local and thesaurus mutation sites. Box bounds, center line, and whiskers represent 25%-75%, 50%, and 5%-95% quantiles, respectively. **(b)** Summary of validation of mutated sites in tumour RNA. The summary is based only on sites in genes that carry both local and thesaurus mutations and in samples that have high coverage (expression) in the tumour RNA.

Properties of non-unique regions in hs37d5 and GRCh38 genome builds

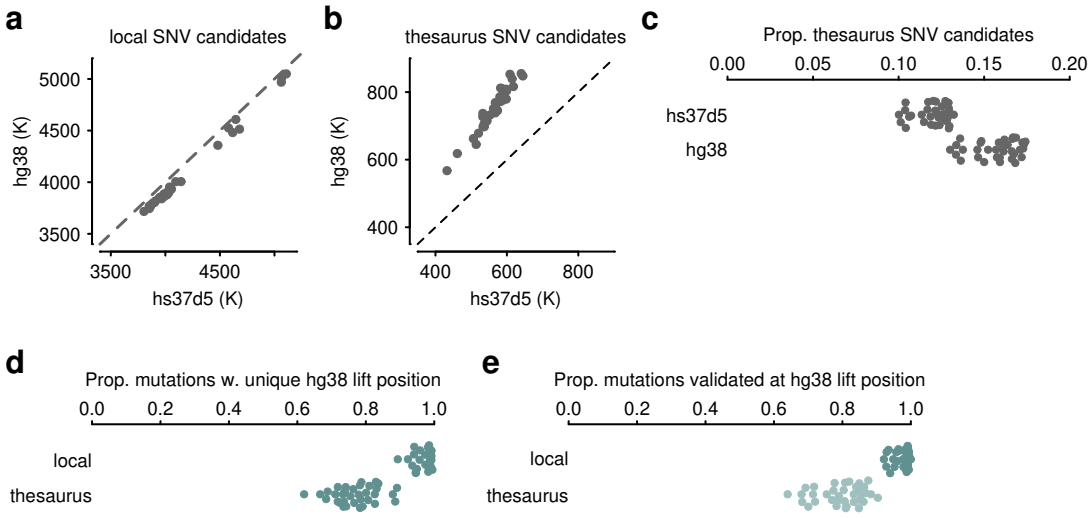


Figure S28. Comparison of 35 samples processed with genome assemblies hs37d5 and hg38. **(a)** Counts of local single-nucleotide variant (SNV) candidates in 35 tumor samples. SNV candidates consist of all positions called in tumour samples, including germline variants, prior to filtering, prioritization, or mutation calling. The dashed line is a diagonal guide showing equal counts in two genome assemblies. **(b)** Similar to (a), showing numbers of SNVs with thesaurus annotations. **(c)** Ratios of SNV candidates with thesaurus annotations to all SNV candidates (local and thesaurus). **(d)** Proportion of somatic mutations in the 35 samples that have a unique lift-over position to hg38. **(e)** Proportion of somatic mutations in the 35 samples where the lift-over position in hg38 also contains a variant. Evaluation of thesaurus variants was performed without considering linked positions.