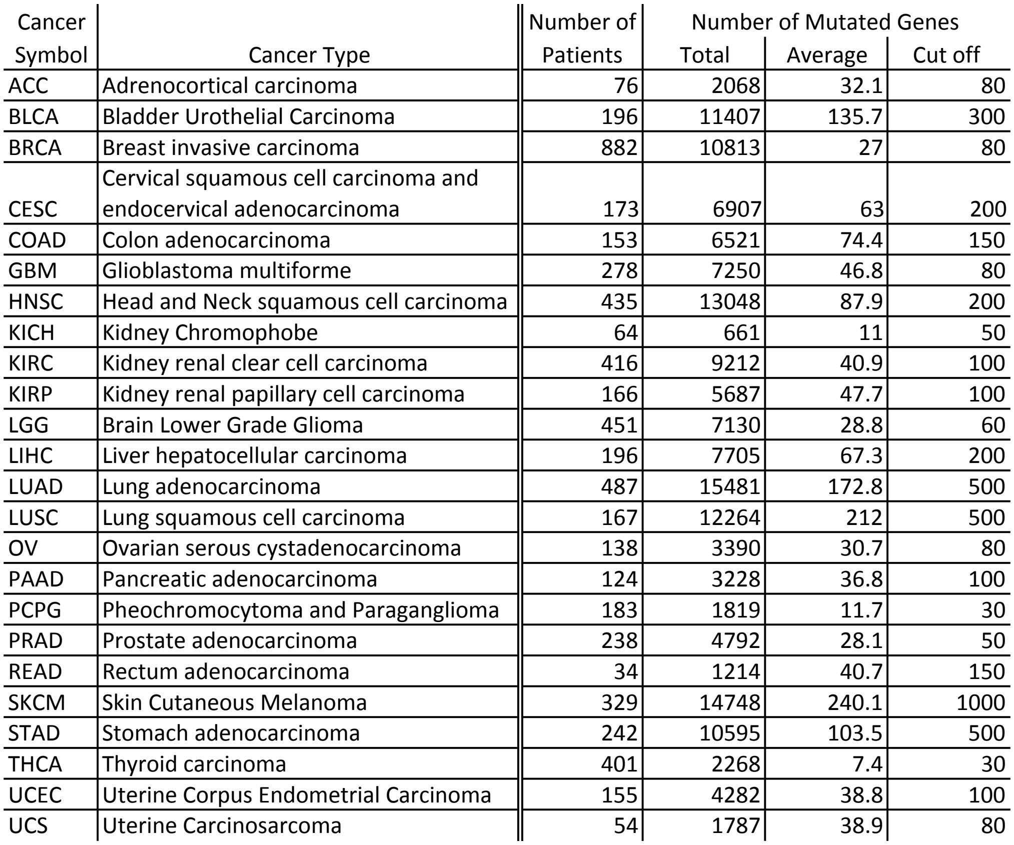
Supplementary Figures and Tables



Cancer		Number of	Number of Mutated Genes		
Symbol	Cancer Type	Patients	Total	Average	Cut off
ACC	Adrenocortical carcinoma	76	2068	32.1	80
BLCA	Bladder Urothelial Carcinoma	196	11407	135.7	300
BRCA	Breast invasive carcinoma	882	10813	27	80
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	173	6907	63	200
COAD	Colon adenocarcinoma	153	6521	74.4	150
GBM	Glioblastoma multiforme	278	7250	46.8	80
HNSC	Head and Neck squamous cell carcinoma	435	13048	87.9	200
KICH	Kidney Chromophobe	64	661	11	50
KIRC	Kidney renal clear cell carcinoma	416	9212	40.9	100
KIRP	Kidney renal papillary cell carcinoma	166	5687	47.7	100
LGG	Brain Lower Grade Glioma	451	7130	28.8	60
LIHC	Liver hepatocellular carcinoma	196	7705	67.3	200
LUAD	Lung adenocarcinoma	487	15481	172.8	500
LUSC	Lung squamous cell carcinoma	167	12264	212	500
OV	Ovarian serous cystadenocarcinoma	138	3390	30.7	80
PAAD	Pancreatic adenocarcinoma	124	3228	36.8	100
PCPG	Pheochromocytoma and Paraganglioma	183	1819	11.7	30
PRAD	Prostate adenocarcinoma	238	4792	28.1	50
READ	Rectum adenocarcinoma	34	1214	40.7	150
SKCM	Skin Cutaneous Melanoma	329	14748	240.1	1000
STAD	Stomach adenocarcinoma	242	10595	103.5	500
THCA	Thyroid carcinoma	401	2268	7.4	30
UCEC	Uterine Corpus Endometrial Carcinoma	155	4282	38.8	100
UCS	Uterine Carcinosarcoma	54	1787	38.9	80

Table S1: **TCGA dataset and statistics.** We list the 24 cancer types studied along with their abbreviations. For each cancer type, we give the total number of patient samples considered after highly mutated samples are filtered out, the total number of mutated genes across these samples, the average number of mutated genes across all samples, and the cutoff on the number of mutated genes within a sample that was used to filter samples. Related to STAR Methods.

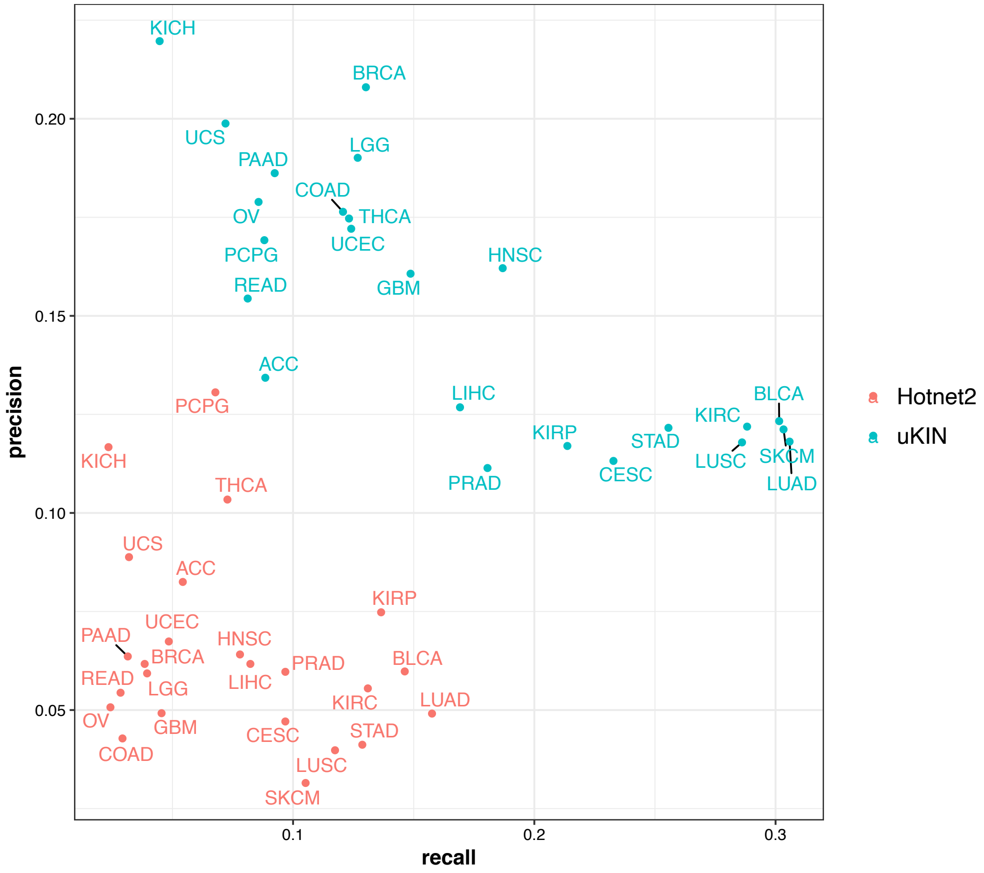


Figure S1: **Comparison between `uKIN` and `Hotnet2`.** For each cancer type, we compute the precision and recall of the genes returned by `uKIN` with α=0.5 and `Hotnet2`. `Hotnet2` is run with default parameters (100 permuted networks, and $\beta = 0.2$ for the restart probability for the insulated heat diffusion process). `Hotnet2` outputs a set of genes predicted to be cancer-relevant, and these genes are not ranked. Thus, for `uKIN`, we consider the same number of top scoring genes as output by `Hotnet2`. `uKIN` exhibits both higher precision and higher recall than `Hotnet2` across all 24 cancer types.

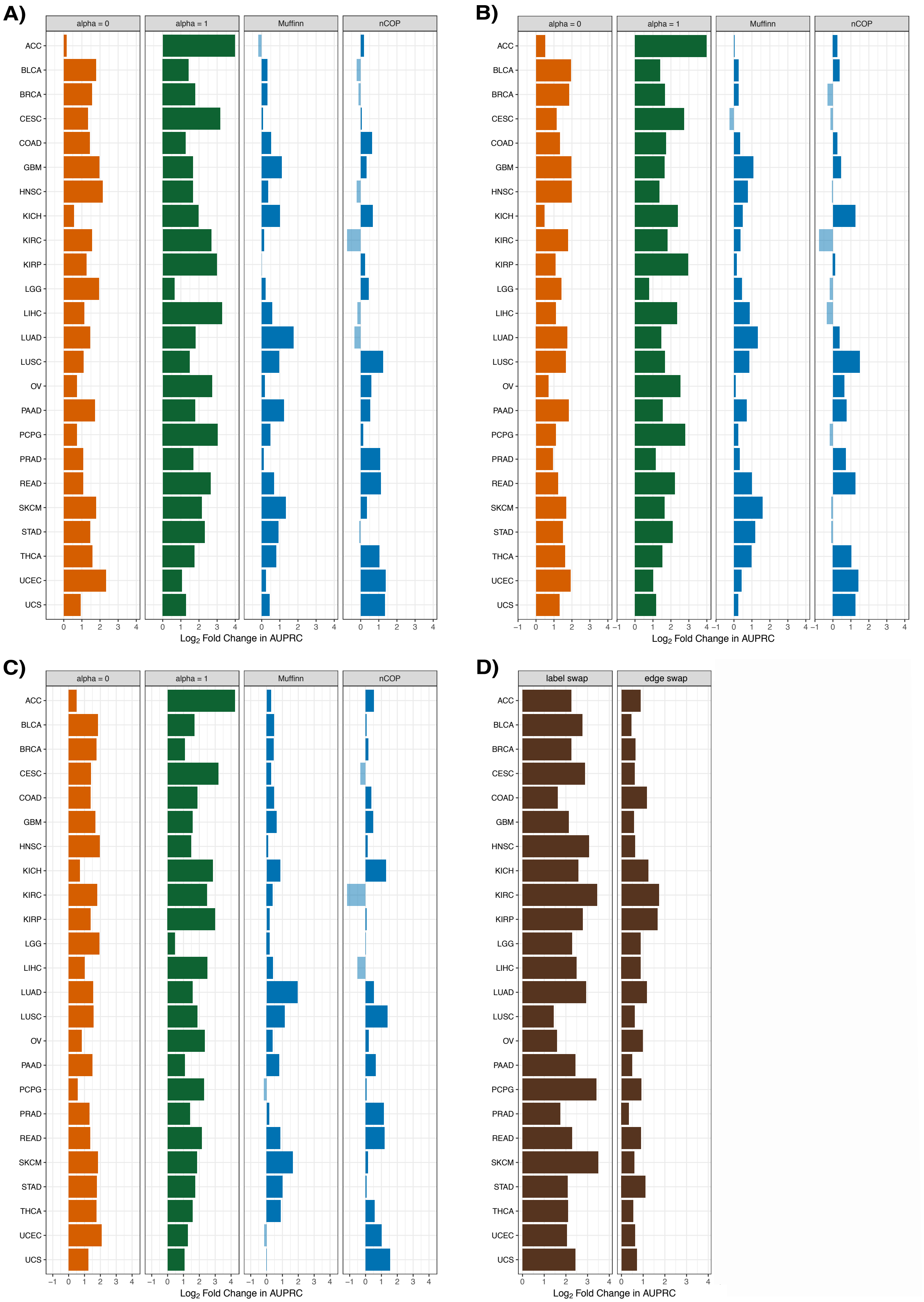

A)
B)
C)
D)
alpha = 0
alpha = 1
Muffinn
nCOP
label swap
edge swap
ACC
BLCA
BRCA
CESC
COAD
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PCPG
PRAD
READ
SKCM
STAD
THCA
UCEC
UCS
−1
0
1
2
3
4
Log$_2$ Fold Change in AUPRC

Figure S2: **Robustness of** `uKIN`**.** **(A)** To make sure that the results reported for `uKIN` are robust with respect to the set of labelled cancer genes $\mathcal{H}$, instead of randomly sampling 400 genes from the Cancer Gene Census (CGC) list, we form $\mathcal{H}$ using genes from other sources. Specifically, we aggregate the cancer genes provided by Hofree et al. in [57] (which they obtained by querying the UniprotKB database for the keyword-terms 'proto-oncogene,' 'oncogene' and 'tumoursuppressor' gene) and Vogelstein et al. [58], excluding any genes present in the set of prior knowledge $\mathcal{K}$. Log-fold AUPRCs are computed as described in the main text. The results are consistent with those shown in Figures 3 and 4 based on the CGC list and show the superior performance of `uKIN` as compared to the other methods in recapitulating known cancer genes. **(B)** To make sure that the results reported for `uKIN` in are robust with respect to number of genes used in evaluation, we compute AUPRCs using the top 50 predicted genes. The results are consistent with those shown in Figures 3 and 4, which use the top 100 predicted genes, and show the superior performance of `uKIN` as compared to the baselines and other methods in recapitulating known cancer genes. The results are also consistent when computing AUPRC's using 150 genes (data not shown). **(C)** To make sure that our method is robust with respect to the specific network utilized, we repeat our entire analysis procedure for `uKIN` with $\alpha = 0.5$ using the Biogrid network. The results are consistent with those shown in Figures 3 and 4, based on the HPRD network. **(D)** To make sure our method utilizes network structure appropriately, we also consider performance of `uKIN` on the real HPRD network as compared to randomized HPRD networks. In the left panel, we use a node label shuffling randomization where the network structure is maintained but gene names are swapped (thereby genes can have very different numbers of interactions in the randomizations). In the right panel, we use a classic degree-preserving randomization (edge swapping). For each of the 24 cancers, we compute the log_2 ratio of the area under the precision recall curve using `uKIN` with $\alpha = 0.5$ on the real network and on the randomized network and show the average over 10 different randomizations. Performance, as expected, is worse for both randomizations across all cancers. We note that significant cancer-relevant information is retained in these randomized networks. In particular, in both types of network randomizations, we maintain the relationships between genes and the samples that they are found to be somatically mutated in. Thus, some highly mutated CGC genes may still be output by `uKIN` when running on randomized networks. Related to Figure 4.

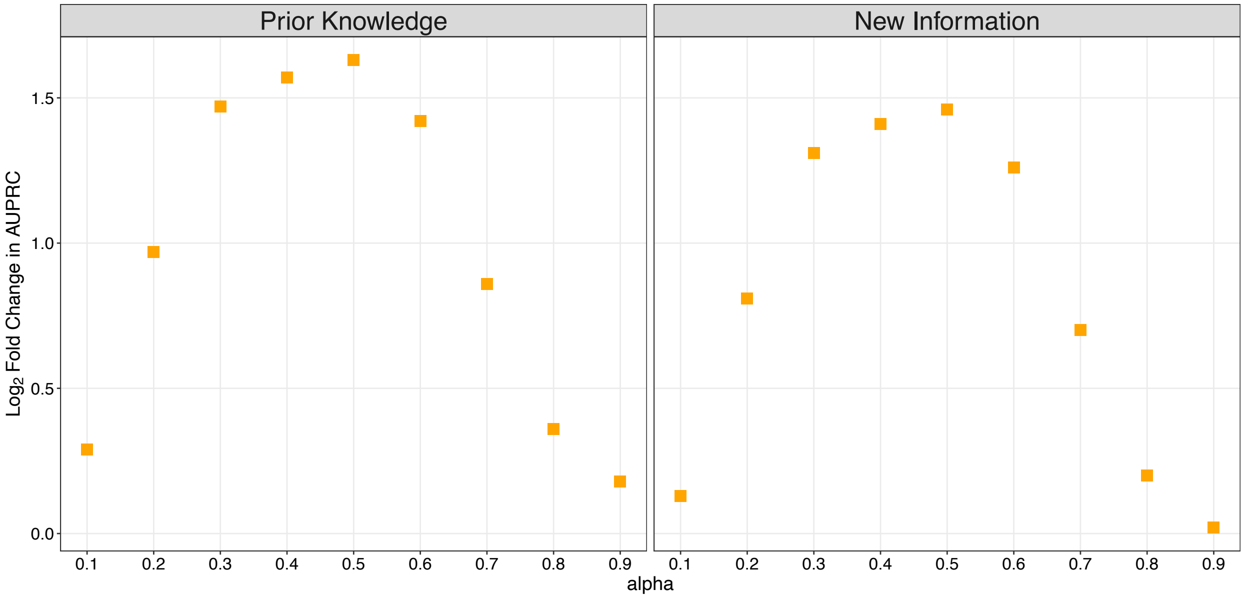


Figure S3: **Integrating both prior and new information via a wide range of α outperforms using only prior data or only new data.** $\texttt{uKIN}$ is run with $0.1 \leq \alpha \leq 0.9$ on the glioblastoma (GBM) dataset as described in the main manuscript, and then compared to the baseline approaches of **(Left)** using only prior knowledge ($\alpha = 0$) or **(Right)** using only new information ($\alpha = 1$). As in the main manuscript, all runs use the HPRD network and for each value of α (x-axis), we show the $\log_2$ ratio, averaged over 100 runs, of the AUPRC of $\texttt{uKIN}$ using that α to the AUPRC for the baseline method. We note that the data in (A) and (B) are linear transformations of each other because $\log_2(\text{AUPRC of } \texttt{uKIN} \text{ with } \alpha = x/\text{AUPRC of } \texttt{uKIN} \text{ with } \alpha = 0) = \log_2(\text{AUPRC of } \texttt{uKIN} \text{ with } \alpha = x/\text{AUPRC of } \texttt{uKIN} \text{ with } \alpha = 1) + \log_2(\text{AUPRC of } \texttt{uKIN} \text{ with } \alpha = 1\ /\text{AUPRC of } \texttt{uKIN} \text{ with } \alpha = 0)$. Related to Figure 4.

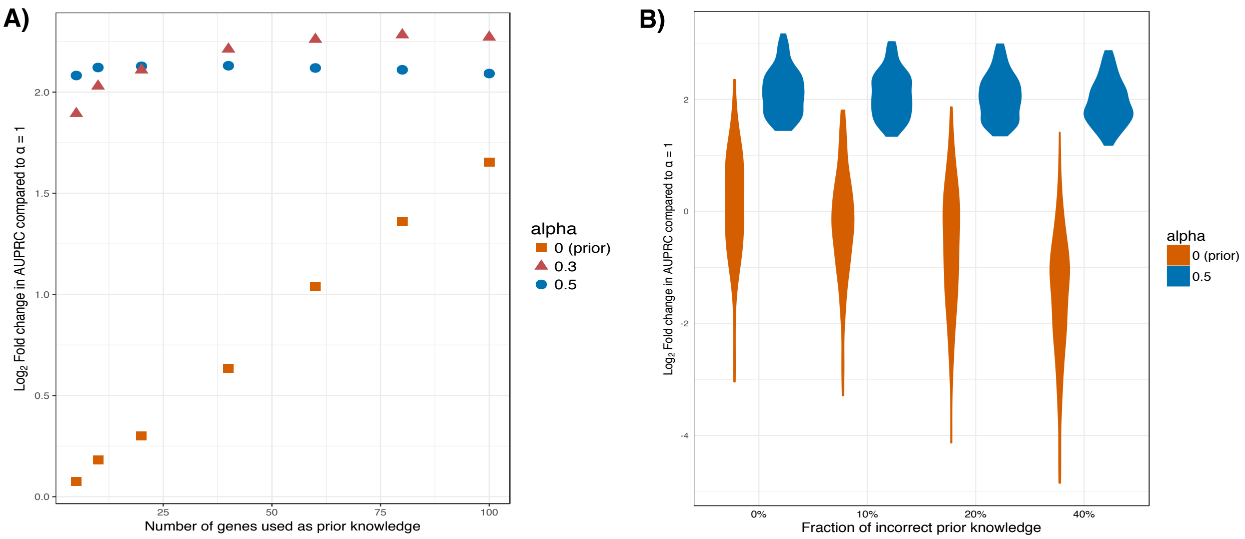


Figure S4: **(A)** `uKIN` **benefits from more knowledge.** As we consider larger numbers of genes comprising the set of prior knowledge ($|\mathcal{K}| = 5, 10, 20, 40, \ldots, 100$), we examine the ability of `uKIN` to uncover CGC genes in the same fixed set $\mathcal{H}$ when using $\alpha = 0.5$ (blue circles), $\alpha = 0.3$ (pink triangles) or $\alpha = 0$ (orange squares). `uKIN` is run on the HPRD network with the kidney renal clear cell carcinoma (KIRC) dataset. We show the $\log_2$ ratio, averaged over 100 runs, of the AUPRC of each version of `uKIN` to the AUPRC for $\alpha = 1$ which is constant across all possible $\mathcal{K}$ (and corresponds to the case where genes are ranked by mutational frequency). For small $\mathcal{K}$, $\alpha = 0$ performs poorly as is expected; as the prior knowledge available increases so does the performance. For both $\alpha = 0.3$ and $\alpha = 0.5$, an increase in the size of $\mathcal{K}$ leads to an initial increase in the performance but eventually performance plateaus. When limited prior knowledge is available ($|\mathcal{K}| < 20$), $\alpha = 0.5$, which uses more of the new information, does better then $\alpha = 0.3$, which relies more on using prior knowledge. When prior knowledge is abundant ($|\mathcal{K}| > 40$), `uKIN` with $\alpha = 0.3$ outperforms $\alpha = 0.5$. As the number of genes comprising the set of prior knowledge increases, spreading information just from those genes ($\alpha = 0$) improves in performance. This is consistent with the observed clustering of CGC genes within biological networks [21]. However, even when propagating information from 100 known cancer genes, the performance is worse than that when integrating it with new information (with either $\alpha = 0.3$ or $\alpha = 0.5$, Figure 3A). **(B)** `uKIN` **is robust to small amounts of erroneous knowledge.** We replace a fraction of the CGCs in the set of prior knowledge genes $\mathcal{K}$ with genes chosen uniformly at random from the set of non-CGC genes in the network. We consider the performance for `uKIN` with $\alpha = 0$ and $\alpha = 0.5$ when 0%, 10%, 20% and 30% of the prior knowledge genes are replaced with non-cancer genes. 100 randomizations are performed at each level of incorrect knowledge. For each run, performance is measured as the $\log_2$ ratio of the AUPRC of `uKIN` (with either $\alpha = 0$ or $\alpha = 0.5$) to the AUPRC for the case where `uKIN` is run with $\alpha = 1$ (which is constant). `uKIN` is run on the HPRD network with KIRC dataset with 20 CGC genes comprising the prior knowledge. Violin plots of this measure are shown are shown for $\alpha = 0$ (orange) and $\alpha = 0.5$ (blue), jittered around the 0%, 10%, 20% and 30% tick marks. At $\alpha = 0.5$, while performance steadily decreases, `uKIN` remains robust to some incorrect knowledge ($\leq 20\%$). As expected, for $\alpha = 0$, the decrease is more notable even when 10% of the prior knowledge is incorrect because in that case `uKIN` uses only prior knowledge. Related to Figure 4.

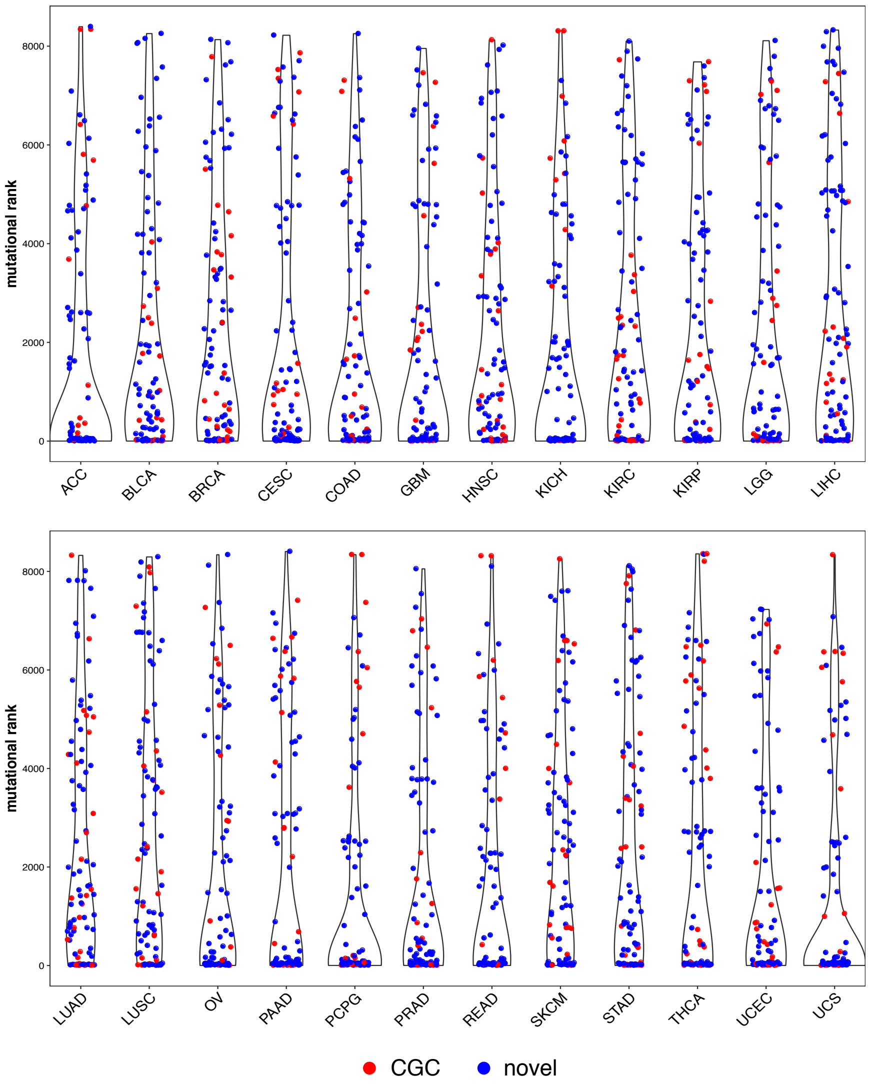


Figure S5: `uKIN` **identifies rarely mutated genes.** To illustrate `uKIN`'s ability to predict genes as cancer-relevant cancer even if they are mutated across fewer numbers of individuals, we consider mutation rates of `uKIN`'s top scoring genes. For each cancer type, we run `uKIN` 100 times with $\alpha = 0.5$ and 20 genes as prior knowledge (see **Methods**). For each gene, its final score is obtained by averaging its scores (arising from the stationary distributions) across the runs; if a gene is in the set of prior knowledge genes $\mathcal{K}$ for a run, this run is not considered for its final score. For each of the 100 genes with highest final scores, we consider the rank of its mutation rate (y-axis). The mutation rate of a gene is computed as the number of observed somatic missense and nonsense mutations across tumors of that cancer type, divided by the number of amino acids in the encoded protein. Then, for each cancer type, genes are ranked by mutation rate where the gene with the highest mutation rate is given the lowest rank. Known CGC genes are in red and novel predictions in blue. The top predictions consist of many heavily mutated genes (i.e., those with low ranks), but `uKIN` is also able to uncover known cancer genes with very low mutational ranks (red dots towards the top). Related to Figure 4.