Cancer Symbol	Cancer Type	Number of Patients	Number of Mutated Genes		
			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. Related to STAR Methods.** 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.

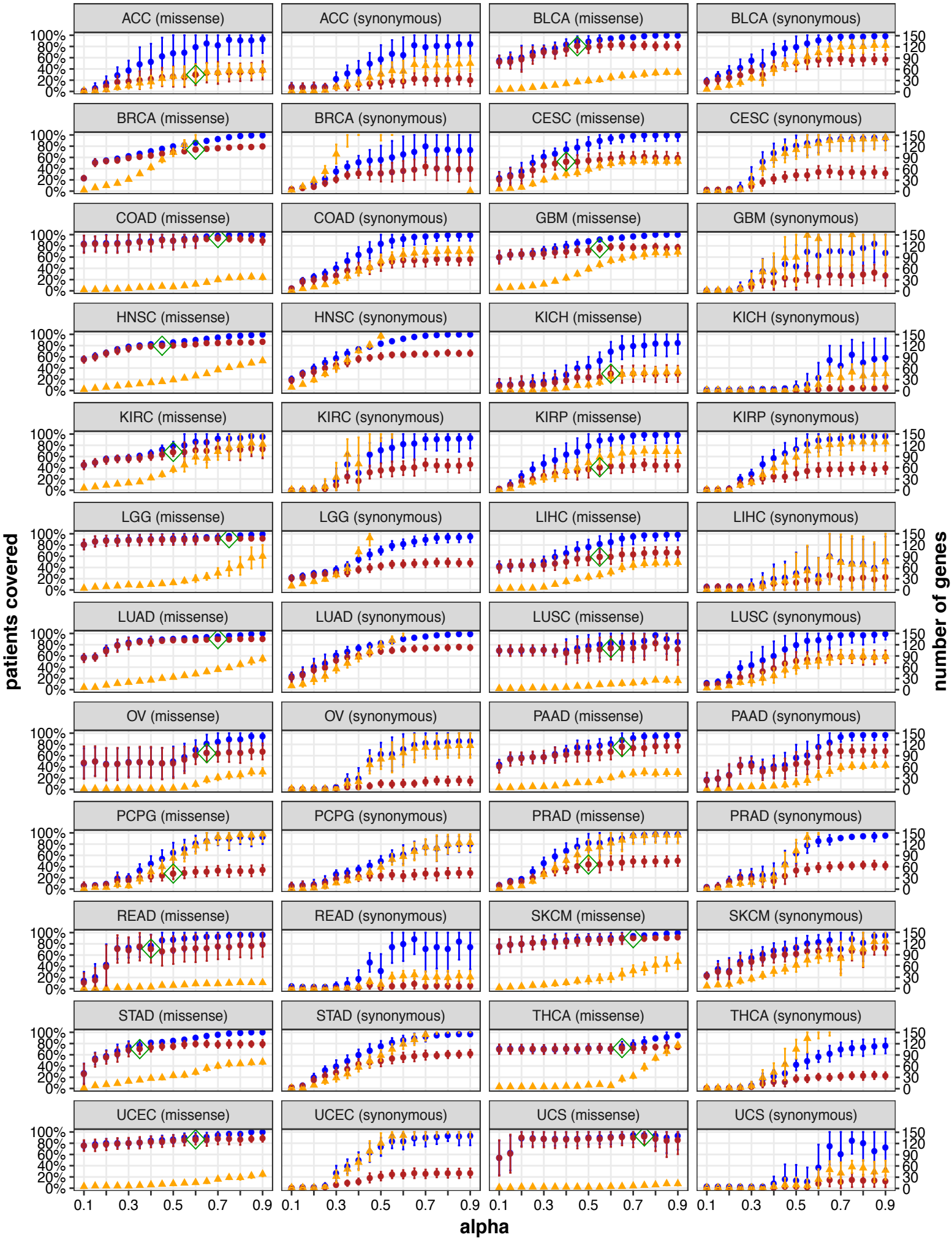
ACC (missense)
ACC (synonymous)
BLCA (missense)
BLCA (synonymous)
BRCA (missense)
BRCA (synonymous)
CESC (missense)
CESC (synonymous)
COAD (missense)
COAD (synonymous)
GBM (missense)
GBM (synonymous)
HNSC (missense)
HNSC (synonymous)
KICH (missense)
KICH (synonymous)
KIRC (missense)
KIRC (synonymous)
KIRP (missense)
KIRP (synonymous)
LGG (missense)
LGG (synonymous)
LIHC (missense)
LIHC (synonymous)
LUAD (missense)
LUAD (synonymous)
LUSC (missense)
LUSC (synonymous)
OV (missense)
OV (synonymous)
PAAD (missense)
PAAD (synonymous)
PCPG (missense)
PCPG (synonymous)
PRAD (missense)
PRAD (synonymous)
READ (missense)
READ (synonymous)
SKCM (missense)
SKCM (synonymous)
STAD (missense)
STAD (synonymous)
THCA (missense)
THCA (synonymous)
UCEC (missense)
UCEC (synonymous)
UCS (missense)
UCS (synonymous)
100%
80%
60%
40%
20%
0%
150
120
90
60
30
0
0.1
0.3
0.5
0.7
0.9
patients covered
number of genes
alpha

Figure S1: **Fraction of individuals covered as α varies across all cancers. Related to Figure 2.** For each random split of the individuals, we run our algorithm on the training sets for different values of α, and plot the fraction of covered individuals in the training (blue) and validation (red) sets. We also give the number of proteins in the uncovered subgraphs G' (orange). For each plotted value, the mean and standard deviation over 100 random splits are shown and the automatically selected α for the missense mutation data is indicated by a green rhombus. The performances on both the training and validation sets are much worse when using synonymous mutations compared to when using missense mutations. Coverage on the validation set for synonymous mutations is consistently lower for the same values of α across respective cancer types than that for missense mutations, with maximum possible coverage on the validation set not exceeding 50% in many cases. Further, it takes significantly more nodes to cover the same fraction of patients when using synonymous mutations.

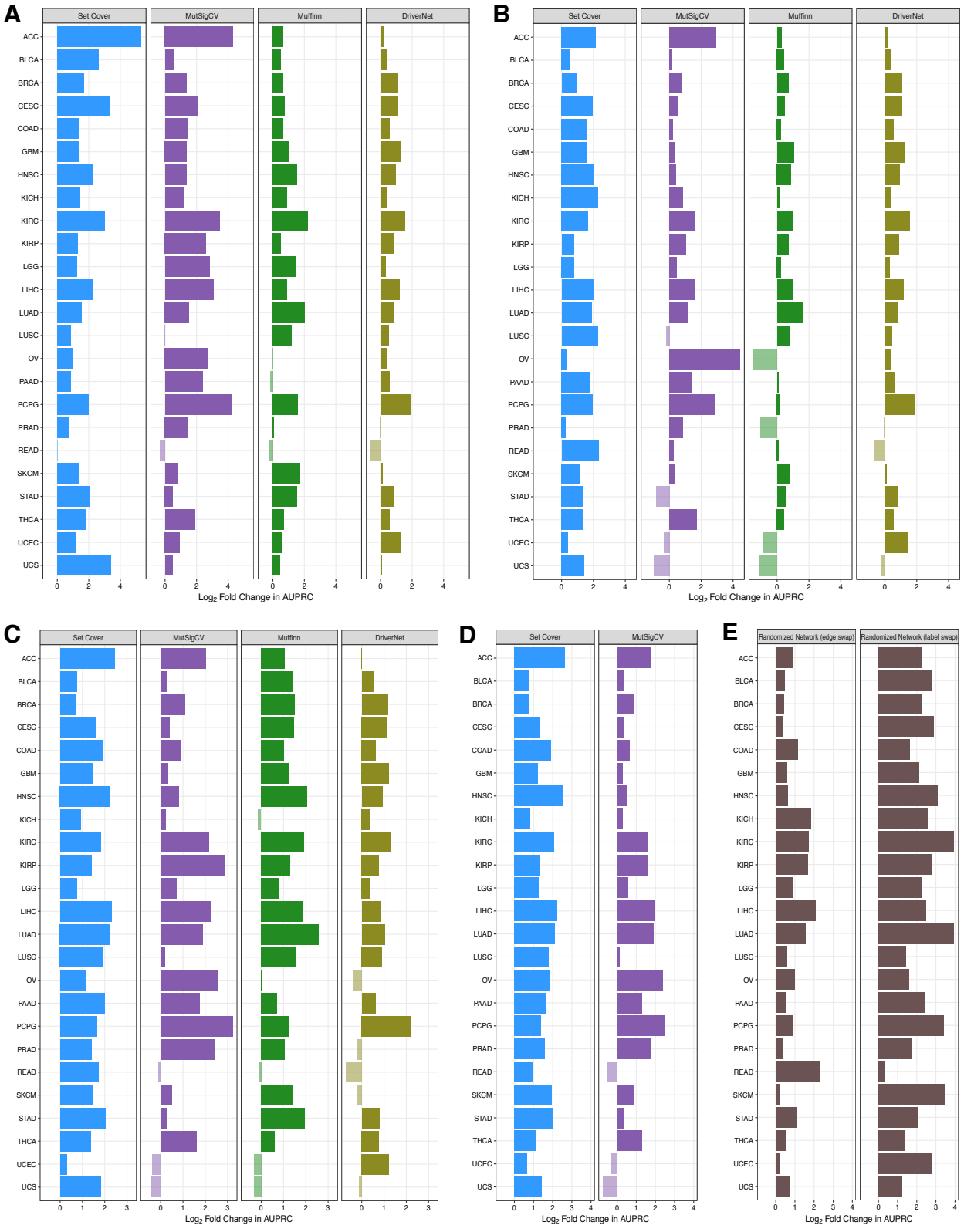
A
Set Cover
MutSigCV
Muffinn
DriverNet
ACC
BLCA
BRCA
CESC
COAD
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PCPG
PRAD
READ
SKCM
STAD
THCA
UCEC
UCS
0 2 4
Log2 Fold Change in AUPRC
B
Set Cover
MutSigCV
Muffinn
DriverNet
ACC
BLCA
BRCA
CESC
COAD
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PCPG
PRAD
READ
SKCM
STAD
THCA
UCEC
UCS
0 2 4
Log2 Fold Change in AUPRC
C
Set Cover
MutSigCV
Muffinn
DriverNet
ACC
BLCA
BRCA
CESC
COAD
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PCPG
PRAD
READ
SKCM
STAD
THCA
UCEC
UCS
0 1 2 3
Log2 Fold Change in AUPRC
D
Set Cover
MutSigCV
ACC
BLCA
BRCA
CESC
COAD
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PCPG
PRAD
READ
SKCM
STAD
THCA
UCEC
UCS
0 1 2 3 4
Log2 Fold Change in AUPRC
E
Randomized Network (edge swap)
Randomized Network (label 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
0 1 2 3 4
Log2 Fold Change in AUPRC

Figure S2: **Robustness. Related to Figure 3. (A)** To make sure that our method is robust with respect to the set of labelled cancer genes, instead of the Cancer Gene Census (CGC) list, we use the list of 413 genes provided by Hofree et al., 2016 which they obtained by querying the UniprotKB database for the keyword-terms 'proto-oncogene,' 'oncogene' and 'tumoursuppressor' gene. Log-fold AUPRCs are computed as described in the main text. The results are consistent with those shown in Figure 3 based on the CGC list and show the superior performance of `nCOP` as compared to the other methods in recapitulating known cancer genes. **(B)** Results using the Vogelstein et al., 2013 list of cancer genes. **(C)** To assess the robustness of our evaluation, we compute AUPRCs using the top 50 predicted genes. The results are consistent with those shown in Figure 3 which use the top 100 predicted genes and show the superior performance of `nCOP` as compared to the other methods in deriving known cancer genes. The results are also consistent when computing AUPRC's using 150 genes (data not shown). **(D)** To make sure that our method is robust with respect to the specific network utilized, we repeat our entire analysis procedure using the Biogrid network. Our approach `nCOP` outperforms the network-agnostic methods in 21 out of 24 of the cancer types. **(E)** Comparison to randomized networks. To confirm the importance of network structure to `nCOP`, we have run `nCOP` on two types of randomized networks. In the left panel, we use a classic degree-preserving randomization (edge swapping) and in the right 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). For each of the 24 cancers, we compute the log_2 ratio of the area under the precision recall curve using `nCOP` 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 patients that they are found to be somatically mutated in. Thus, some highly mutated CGC genes may still be output by `nCOP` when running on randomized networks.

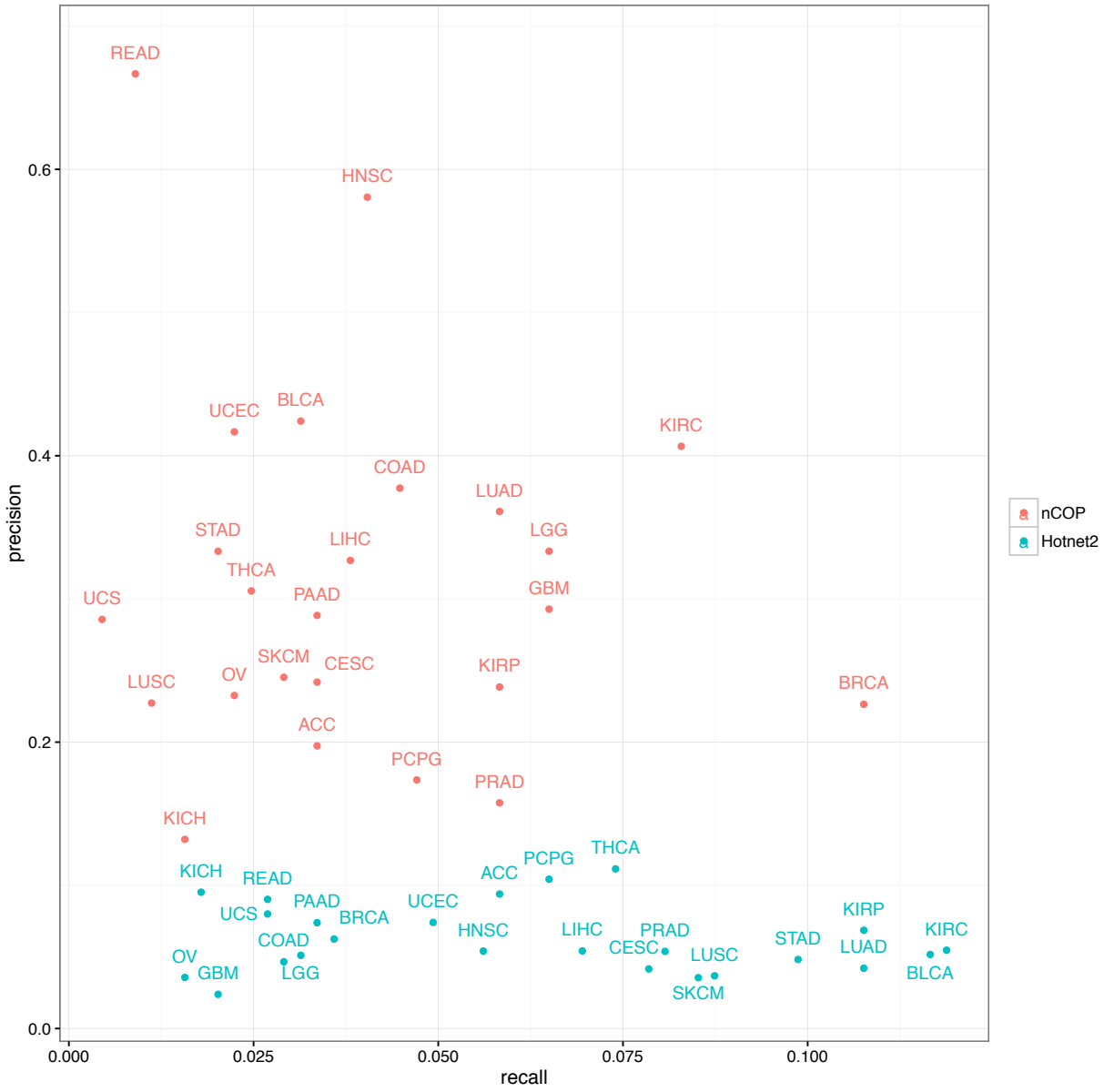

Figure S3: **Comparison between `nCOP` and `Hotnet2`. Related to Figure 3.** For each cancer type, we compute the precision and recall of the genes returned by `nCOP` and `Hotnet2`. For `nCOP`, we choose a single threshold to select predicted cancer genes, corresponding to those genes that occur in at least 25% of the runs. While `Hotnet2` achieves slightly greater recall due to the larger number of genes it highlights, `nCOP`'s precision using this threshold is superior. `nCOP` uncovers fewer but potentially more relevant cancer genes.

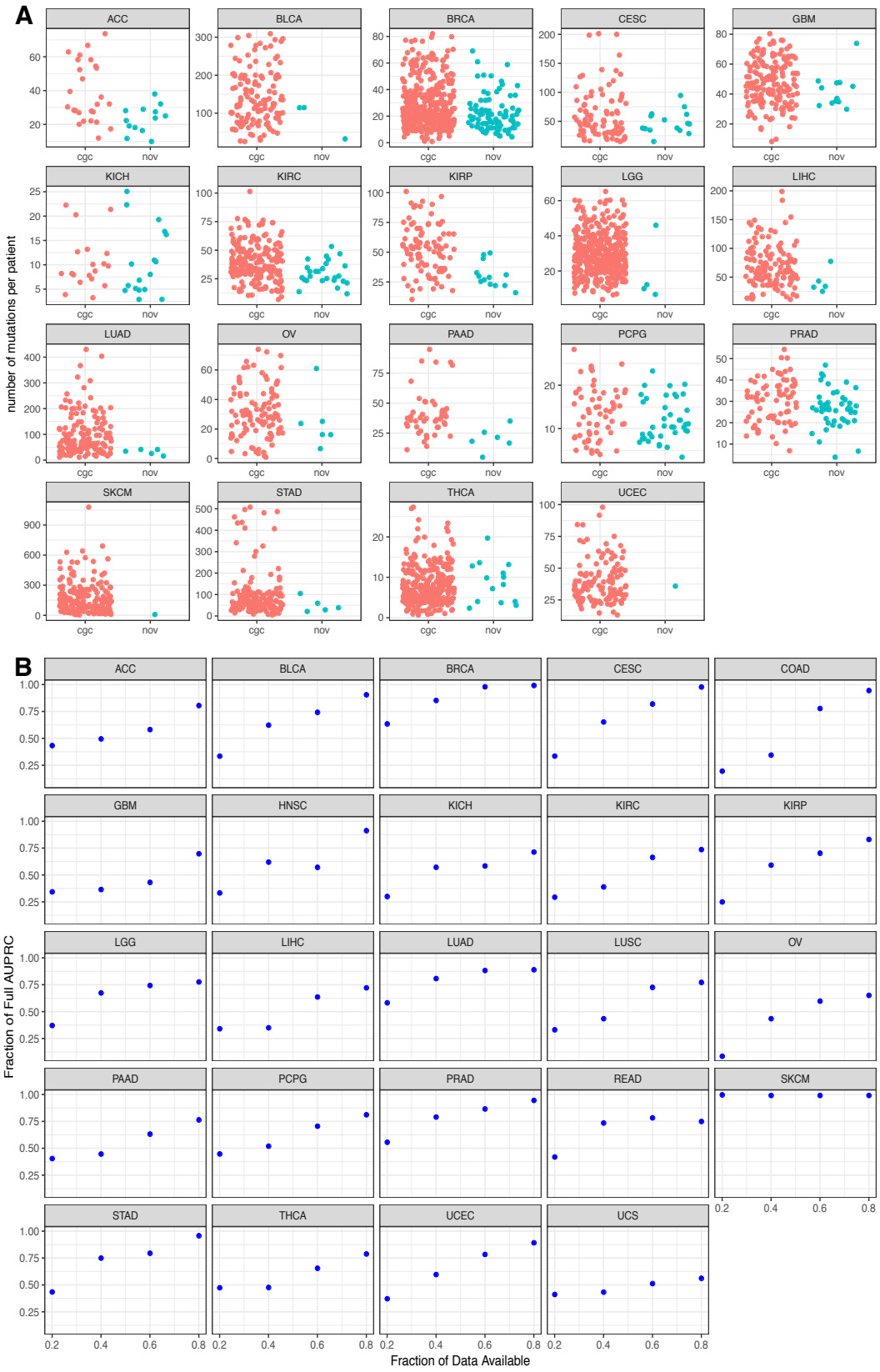
A
ACC
BLCA
BRCA
CESC
GBM
KICH
KIRC
KIRP
LGG
LIHC
LUAD
OV
PAAD
PCPG
PRAD
SKCM
STAD
THCA
UCEC
cgc
nov
number of mutations per patient
B
ACC
BLCA
BRCA
CESC
COAD
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PCPG
PRAD
READ
SKCM
STAD
THCA
UCEC
UCS
Fraction of Full AUPRC
Fraction of Data Available

Figure S4: **Genes uncovered by nCOP. Related to Figure 4. (A)** Novel genes uncovered by nCOP are not due to patients with many mutations. Plotted for each cancer type are the total number of missense mutations for patients having missense mutations only in known CGC genes and not in novel genes (left) and the total number of missense mutations for patients having missense mutations only in novel genes and not in CGC genes (right). Patients with only mutations in novel genes do not harbor more mutations. **(B)** The predictive power of nCOP increases with more data. For each cancer type, we repeatedly sample a fraction of the patients (20%, 40%, 60%, and 80%), rerun our method on the reduced data set, compute the ratio between the AUPRC using the sampled data set and the full data set, and plot the median ratio across 50 samples per fraction. As nCOP uses more data, its predictive power increases and becomes similar to the one on the full data set.