Supplementary Material for "MEXCOwalk: Mutual Exclusion and Coverage Based Random Walk to Identify Cancer Modules"

1 Proof of Theorem 1

Theorem 1. *Cancer driver module identification problem is NP-hard.*

Proof. The transformation is from *Set Packing* which is NP-complete [3]. In the Set Packing problem, given a collection C of finite sets and a positive integer $K \leq |C|$, the problem is to find out whether C contains at least K mutually disjoint sets. The problem is NP-hard even when the size of each set is at most 3, which can easily be extended to the setting where the size of each set is exactly 3. Given an input to the Set Packing problem within this setting in the form of K and C such that for each $S \in C$, $|S| = 3$, we generate G as a complete graph on $|C|$ vertices, corresponding to the such that each finite set in C corresponds to a set of samples S_i for which gene g_i is mutated. We set both *total_genes* and *min_module_size* to K. The answer to the Set Packing problem is Yes, if and only if the maximized score of the cancer module identification problem is exactly $\frac{3\times K}{|\bigcup_{\forall g_i \in V} S_i|}$. □

2 Effects of Mutual Exclusivity Threshold θ

We plot the distribution of MEX_n scores across all the edges in Figure S1. Almost all edges have MEX_n scores larger than 0.5. Therefore, we experiment with θ values greater than or equal to 0.5: $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. Figure S2 show the results of running MEXCOwalk with different θ values. In parts A and B, we observe that changing θ has a minimal effect in recovering known cancer genes with $\theta = 0.7$ giving the largest area under the ROC curve. Figure S2 -C reveals a strong correlation between θ scores and Mutual Exclusivity (MS) scores. Since large values of θ clamp a larger set of edge weights to 0, the resulting modules are only those with really large mutual exclusivity scores. We observe the largest coverage scores when θ is set to 0.7 (Figure S2-D), whereas 0.9 results in significantly lower coverage scores across all *total_genes* values. This is likely due to mutual exclusivity dominating over coverage in edge weights. Finally, we observe that Driver Module Set Scores are mosty in parallel with Coverage Scores; see Figure S2-B.

3 Effects of *min_module_size*

We experiment with a range of values for *min_module_size*; see Figure S3. We see minimal differences in overlaps with CGC database where $min\_module\_size = 3$ results in the largest area under ROC curve. As expected, there is an inverse correlation between *min_module_size* and mutual exclusion scores (Figure S3-C). On the other hand, we observe a positive correlation between *min_module_size* and coverage scores. Again, we observe that Driver Module Set Scores are in parallel with Coverage Scores (Figure S3-B).

4 Static Evaluations

Figure S4-A plots the Receiver Operating Characteristic (ROC) curves of the set of genes in the union of modules each algorithm provides with respect to the CGC genes [2]. In Figure S4-B and Figure S4-C, ROC curves are plotted with a subset of CGC genes with mutation frequency $\leq 1\%$ and $\leq 2\%$ in the pan-cancer cohort, respectively. The same analysis is repeated with the set of druggable genes downloaded from DGIdb 3.0 [1]. Note that only cancer related sources are included in DGIdb 3.0. FigureS5-A shows the ROC curve with all druggable genes; whereas in Figure S5-B and Figure S5-C only those genes with mutation frequency $\leq 1\%$ or $\leq 2\%$ are included, respectively.

5 Output Module Sizes

We investigate the size distribution of the output modules for different methods under consideration. In Figure S8-A we plot the average module sizes for different *total_genes* values. We observe that the average module size remains stable for increasing values of *total_genes* for all the methods except Hotnet2. Additionally, we identify the minimum and maximum module size for each *total_genes* value and for each method. We then plot the percentage of the genes that belong to the modules of minimum or maximum size; see Figure S8-B and Figure S8-C. Strikingly, for the majority of *MEMCover_v1* outputs, more than 70% of genes appear in modules with minimum size which is 1. In the other extreme, Hotnet2 provides a large module of maximum size that contains more than 60% of all the genes, for the majority of *total_genes* settings.

6 Cancer Type Specificity

Figure S9 shows the distribution of best p-values obtained for each module for increasing values of *total_genes* for all the methods.

7 Classification Accuracy

Figure S10 shows the distribution of classification accuracy values across the output modules for increasing values of *total_genes* for all the methods.

8 Hotnet2 and MEMCover output modules

Hotnet2 and MEMcover output modules are shown in Figure S11 when *total_genes* is set to 100. The module colors represent the mostly enriched cancer type. Note that in Figure S11-B, every single gene not surrounded by a rectangle is a module by itself.

9 Sensitivity analysis of MEXCOwalk

9.1 Sensitivity to β

We check the sensitivity of our results to β parameter by employing the settings of 0.2, 0.3, 0.5, 0.6, and 0.7, in addition to the default setting of $\beta = 0.4$. Table S1 shows the percentage of the number of different genes in MEXCOwalk output gene sets at different β settings, with respect to the default $\beta = 0.4$. Figures S12 and S13 compare MEXCOwalk runs employing $\beta = 0.2$ and $\beta = 0.4$ in terms of ROC curves with respect to CGC genes and DGIdb genes. Figure S14 compares MEXCOwalk runs employing $\beta = 0.2$ and $\beta = 0.4$ in terms of modular evaluation metrics: DMSS, CTSS and MCAS. Figure S15 and S16 display the distribution of p-values and accuracy values, that are averaged in CTSS and MCAS, respectively.

9.2 Sensitivity to PPI network

We also assess the sensitivity to the employed PPI. To this end, we repeat all the experiments with the IntAct network [4]. The first column of Table S2 shows the percentage of the number of different genes in MEXCOwalk output gene sets when IntAct network is employed, with respect to the setting where default network, HINT+HI2012 is used. The second column of Table S2 displays overlap sizes with CGC for each of the two runs. For instance, when $total\_genes = 200$, among the genes MEXCOwalk outputs with the IntAct network, 77 are in CGC database. The size of overlap with CGC is 80 when HINT+HI2012 network is used. Finally, the intersection of these 77 and 80 genes is 52. Figures S17 and S18 compare MEXCOwalk runs employing HINT+HI2012 and IntAct in terms of ROC curves with respect to CGC genes and DGIdb genes. Figure S19 compares MEXCOwalk runs employing HINT+HI2012 and IntAct in terms of modular evaluation metrics: DMSS, CTSS and MCAS. Figure S20 and S21 display the distribution of p-values and accuracy values, that are averaged in CTSS and MCAS, respectively.

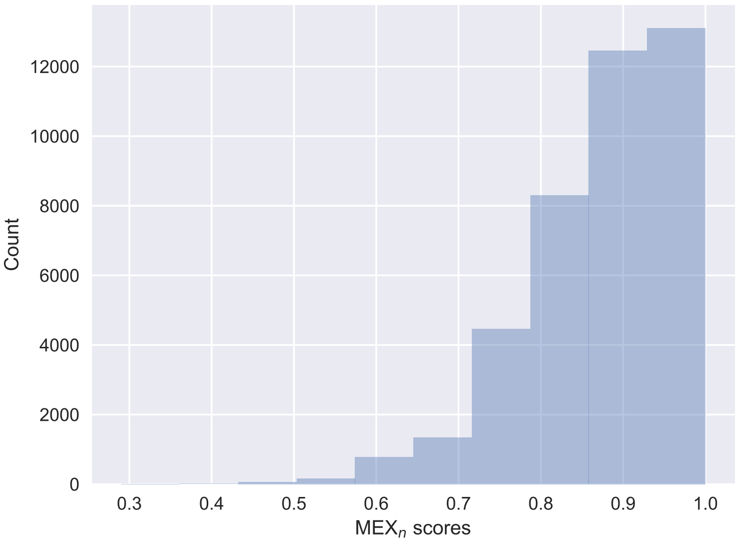

Figure S1: The distribution of MEX_n scores of all edges.

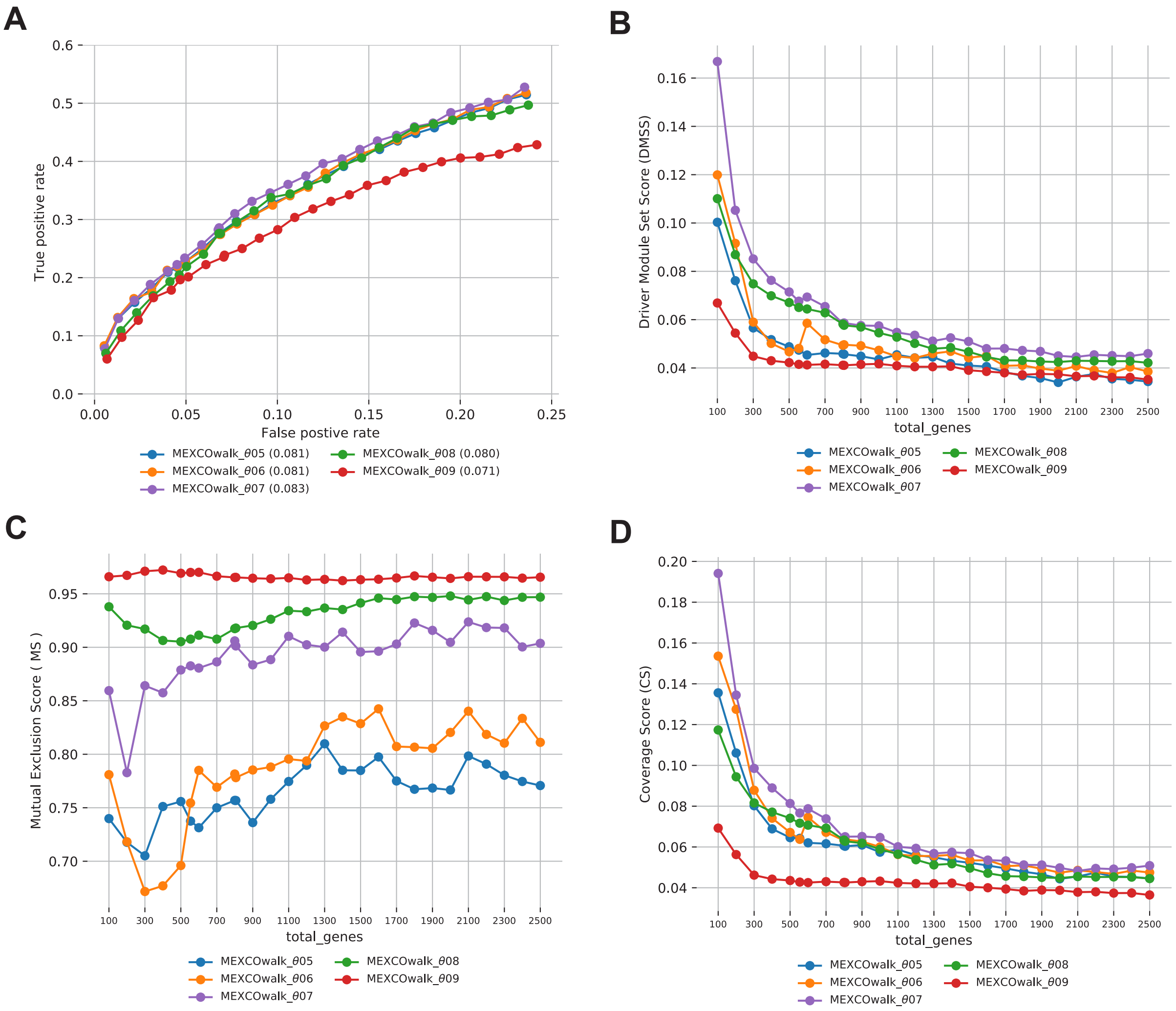

Figure S2: Comparison of MEXCOwalk models with different mutual exclusion score thresholds (θ): 0.5, 0.6, 0.7, 0.8 and 0.9 A) ROC plots and AUROC values written in parentheses. B) Driver Modules Set Score (DMSS). C) Mutual Exclusion Score (MS). D) Coverage Score (CS).

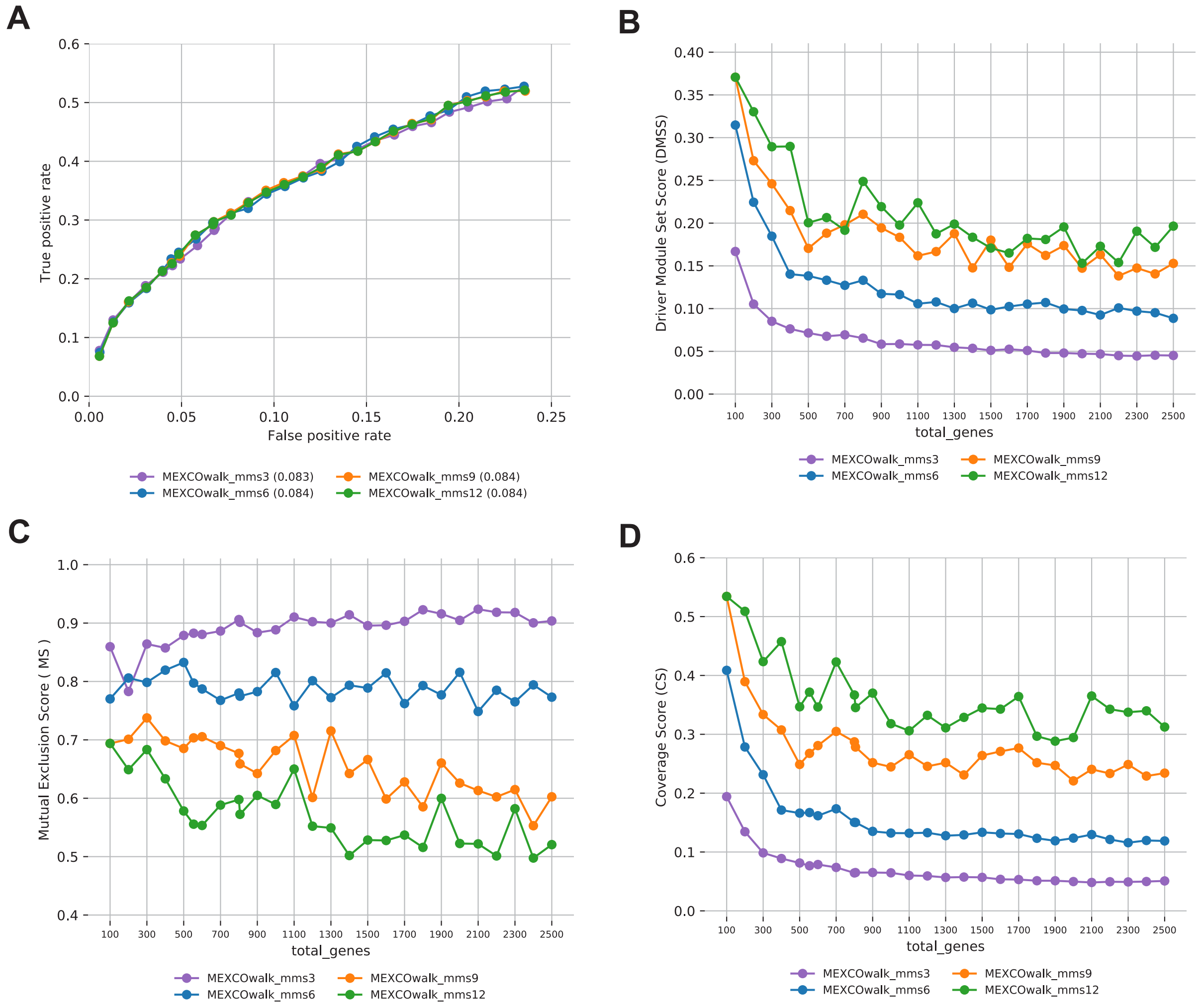

Figure S3: Comparison of MEXCOwalk models with different min_module_size: 3, 6, 9, 12 A) ROC plots and AUROC values written in parentheses. B) Driver Modules Set Score (DMSS). C) Mutual Exclusion Score (MS). D) Coverage Score (CS).

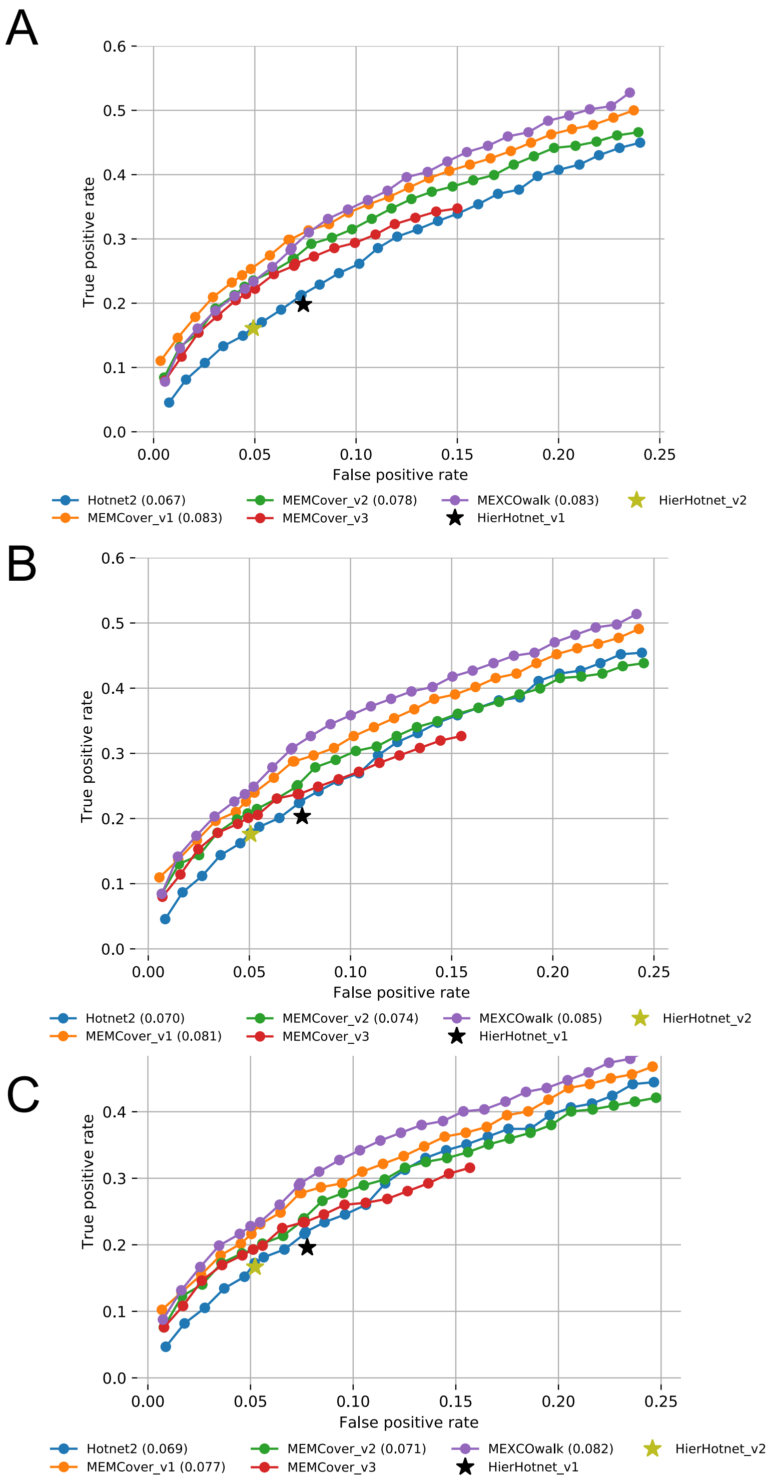

Figure S4: A) The fraction of recovered CGC genes for each *total_genes* value is shown with a ROC plot. B) Same as A, but only those CGC genes with $\leq$ 1% mutation frequency in the pan-cancer cohort are used. C) Same as A, but only those CGC genes with $\leq$ 2% mutation frequency in the pan-cancer cohort are used.

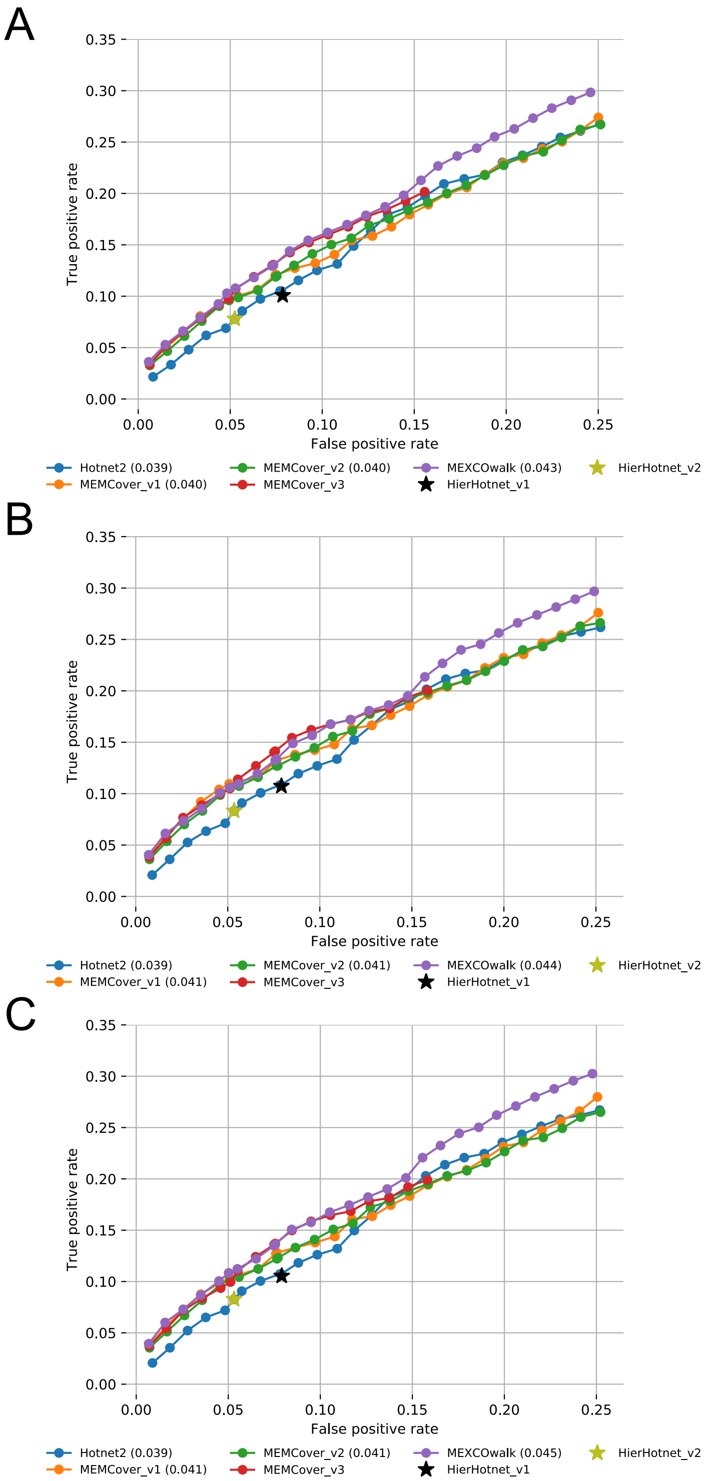

Figure S5: A) The fraction of recovered druggable genes for each *total_genes* value is shown with a ROC plot. B) Same as A, but only those druggable genes with $\leq 1\%$ mutation frequency in the pan-cancer cohort are used. C) Same as A, but only those druggable genes with $\leq 2\%$ mutation frequency in the pan-cancer cohort are used.

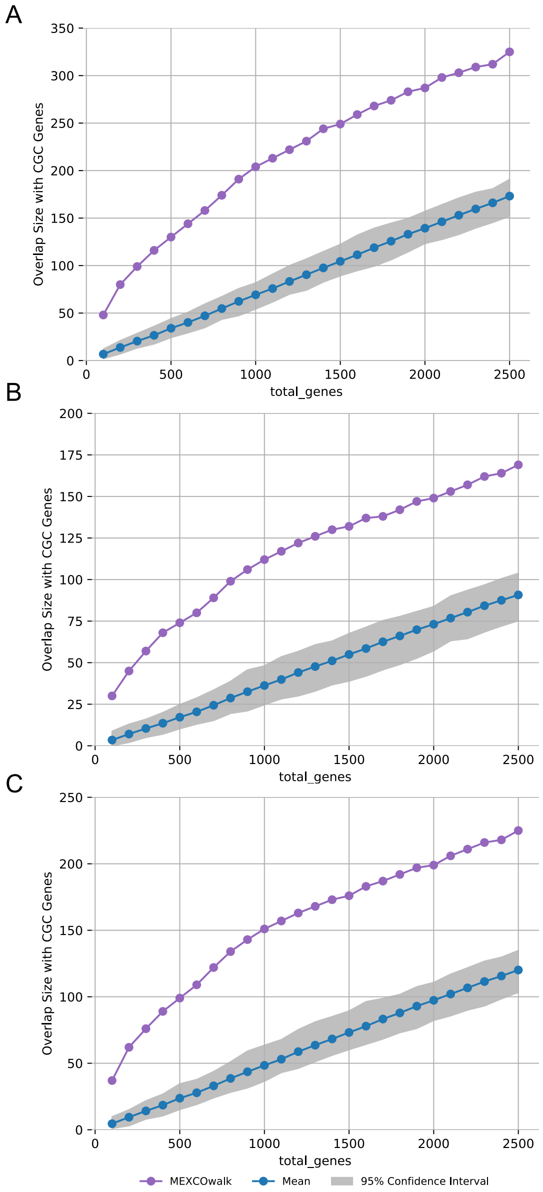

Figure S6: A) The fraction of CGC genes for each *total_genes* value is shown with a ROC plot for MEXCOwalk on original and randomized mutation data. B) Same as A, but only those CGC genes with $\leq 1\%$ mutation frequency in the pan-cancer cohort are used. C) Same as A, but only those CGC genes with $\leq 2\%$ mutation frequency in the pan-cancer cohort are used.

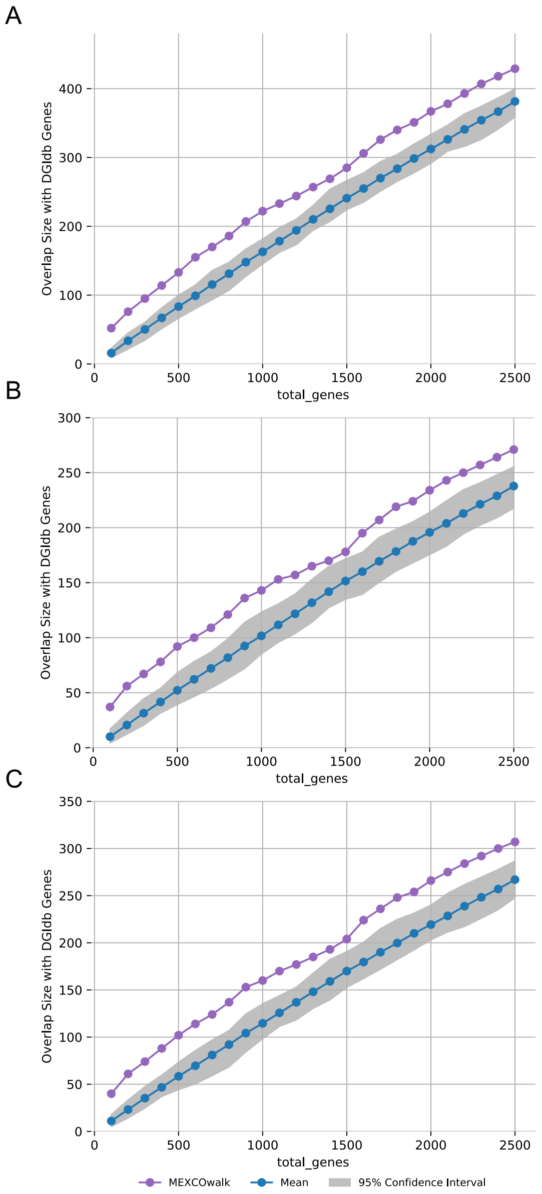

Figure S7: A) The fraction of recovered druggable genes for each *total_genes* value is shown with a ROC plot for MEXCOwalk on original and randomized mutation data. B) Same as A, but only those druggable genes with $\leq 1\%$ mutation frequency in the pan-cancer cohort are used. C) Same as A, but only those CGC genes with $\leq 2\%$ mutation frequency in the pan-cancer cohort are used.

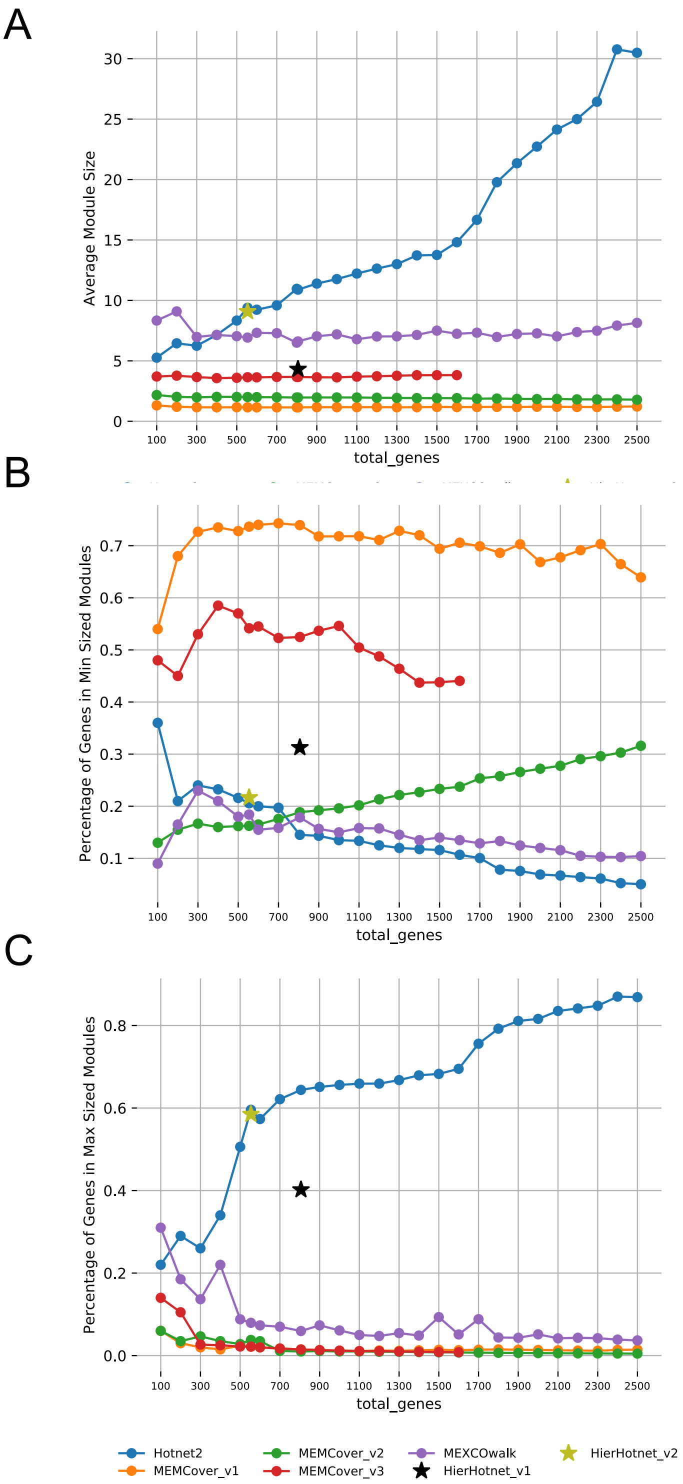

Figure S8: A) Average modules sizes in the outputs of MEXCOwalk, MEMCover, Hotnet2, and Hierarchical Hotnet for increasing values of *total_genes*. B) The percentage of genes across the modules with largest size for MEXCOwalk, MEMCover, Hotnet2, and Hierarchical Hotnet outputs for increasing values of *total_genes*. C) The percentage of genes across the modules with smallest size for MEXCOwalk, MEMCover, Hotnet2, and Hierarchical Hotnet outputs for increasing values of *total_genes*.

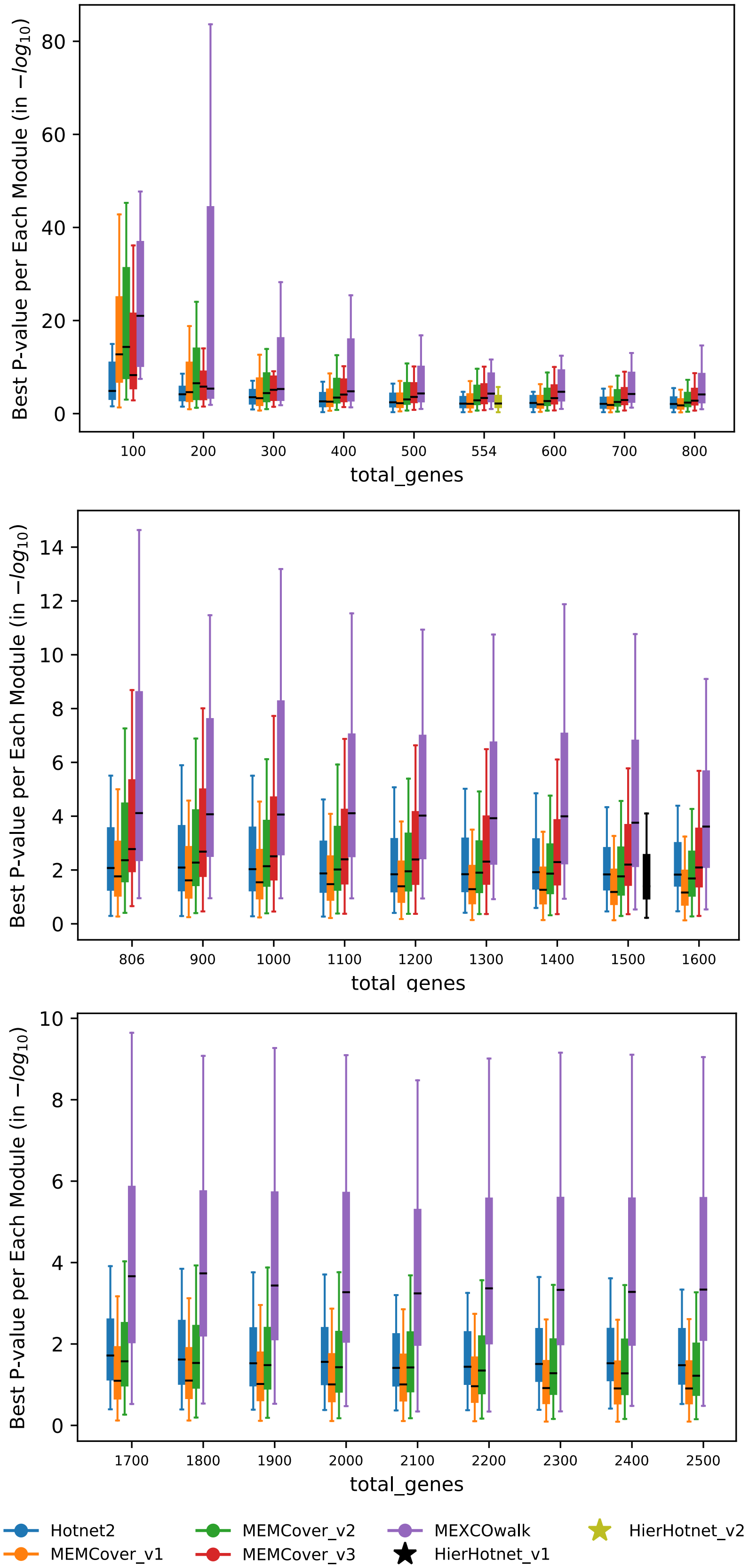

Figure S9: Box plot figure of best p-value for each module obtained for increasing values of *total_genes*. Whiskers represent the interquartile range.

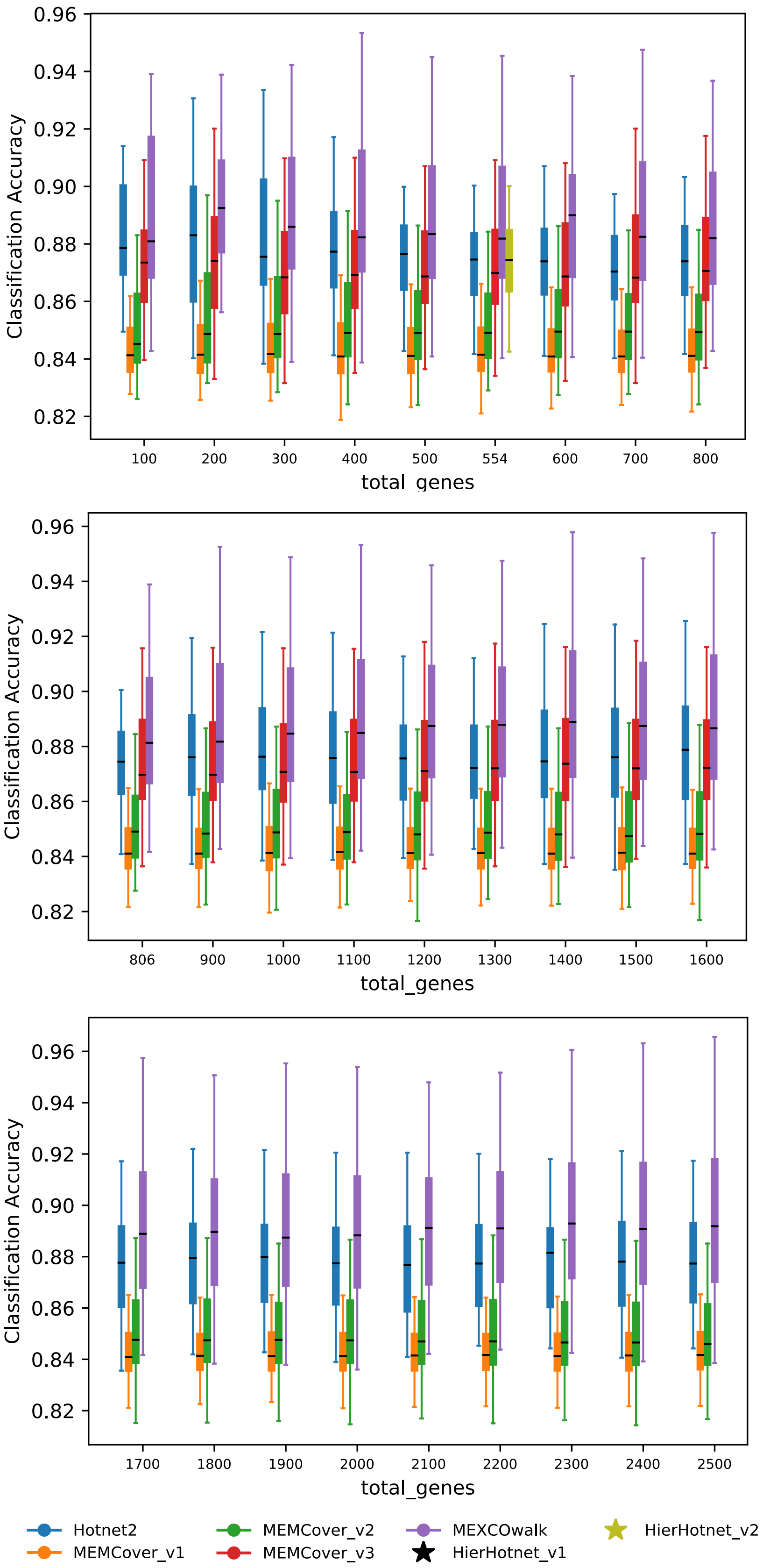

Figure S10: Box plot figure of classification accuracy for each module obtained for increasing values of *total_genes*. Whiskers represent the interquartile range.

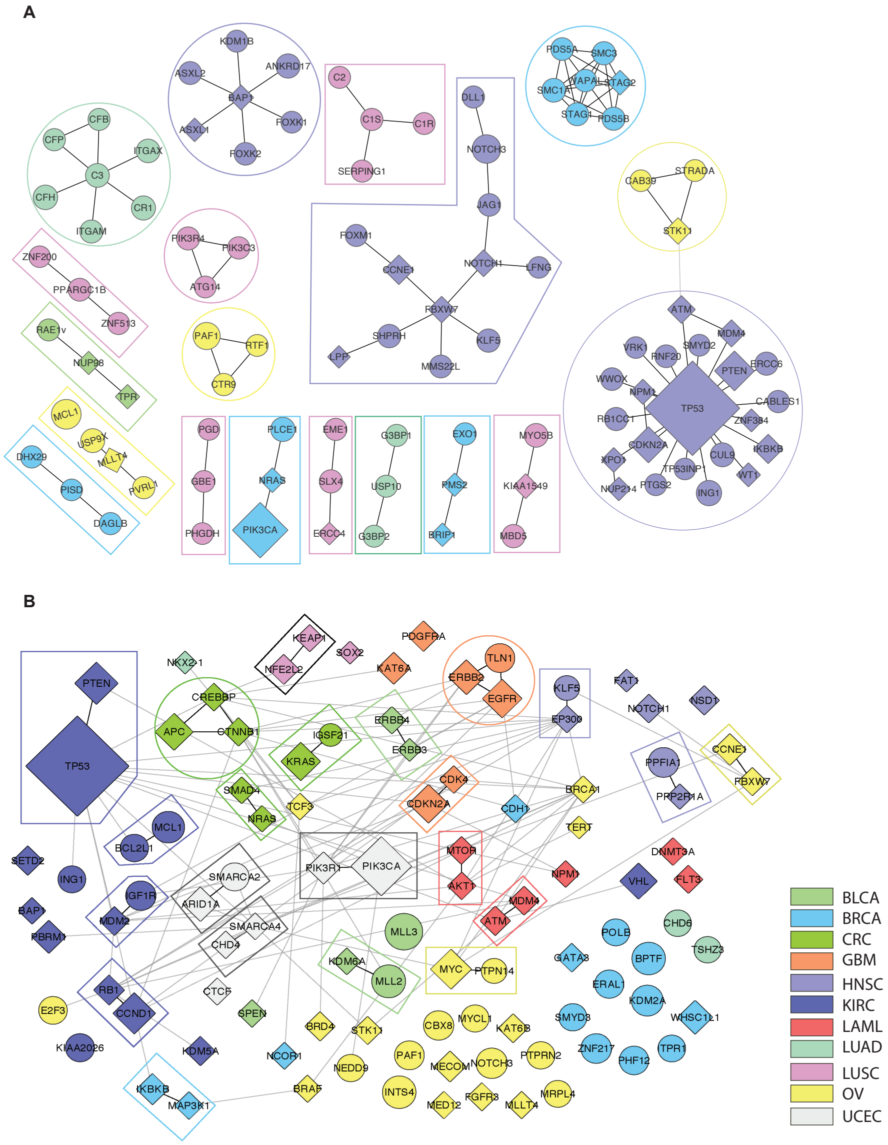

Figure S11: A) Hotnet2 output modules when $total\_genes = 100$. B) MEMCover_v1 output modules when $total\_genes = 100$ Diamond shaped nodes correspond to CGC genes. Sizes of the nodes are proportional with mutation frequencies of corresponding genes. Color of a module denotes the cancer type with the strongest enrichment for mutations in genes of that module. The legend for the color codes are shown on the bottom right.

Table S1: Percentage of the number of different genes in MEXCOwalk output gene sets as value of parameter β varies from selected value = 0.4.

total_genes	0.2	0.3	0.5	0.6	0.7
100	8.00	3.00	3.00	4.00	4.00
200	7.50	1.50	2.00	3.00	5.00
300	8.33	3.00	2.33	5.33	6.67
400	7.54	3.27	1.26	4.02	5.03
500	10.00	4.20	2.60	4.40	5.00
600	8.00	3.17	2.83	4.50	6.50
700	7.86	3.86	2.86	4.71	6.00
800	7.13	2.75	1.88	3.50	6.01
900	5.89	2.22	2.11	4.11	5.67
1000	6.90	2.50	1.60	3.90	5.30
1100	7.45	3.00	2.27	3.73	4.91
1200	6.83	2.67	1.58	3.00	4.50
1300	6.77	3.00	1.69	3.39	5.08
1400	6.93	3.36	2.07	3.65	4.43
1500	6.87	2.73	1.87	3.40	4.07
1600	6.25	2.75	1.75	3.25	3.94
1700	6.01	2.71	2.06	3.24	4.18
1800	5.83	2.50	2.11	3.44	4.72
1900	5.89	2.53	2.37	3.79	4.58
2000	5.66	2.80	2.70	3.70	4.65
2100	4.81	2.24	2.19	3.38	4.62
2200	4.77	2.23	2.05	3.73	4.91
2300	4.52	2.17	1.91	3.91	4.96
2400	5.21	2.54	1.54	2.96	4.33
2500	5.04	2.36	1.76	3.04	4.36

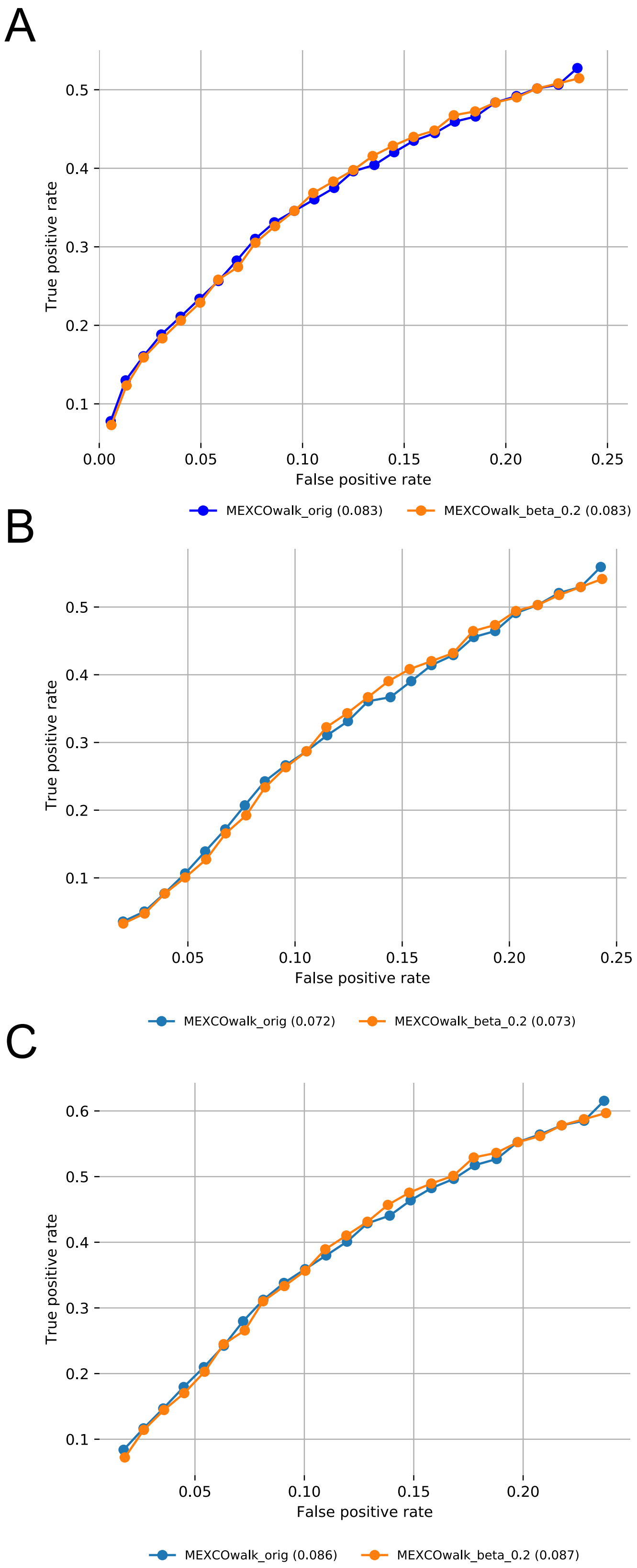

Figure S12: A) The fraction of CGC genes for each *total_genes* value is shown with a ROC plot for MEXCOwalk original run ($\beta = 0.4$) and a version of MEXCOwalk where $\beta = 0.2$. B) Same as A, but only those CGC genes with $\leq$ 1% mutation frequency in the pan-cancer cohort are used. C) Same as A, but only those CGC genes with $\leq$ 2% mutation frequency in the pan-cancer cohort are used.

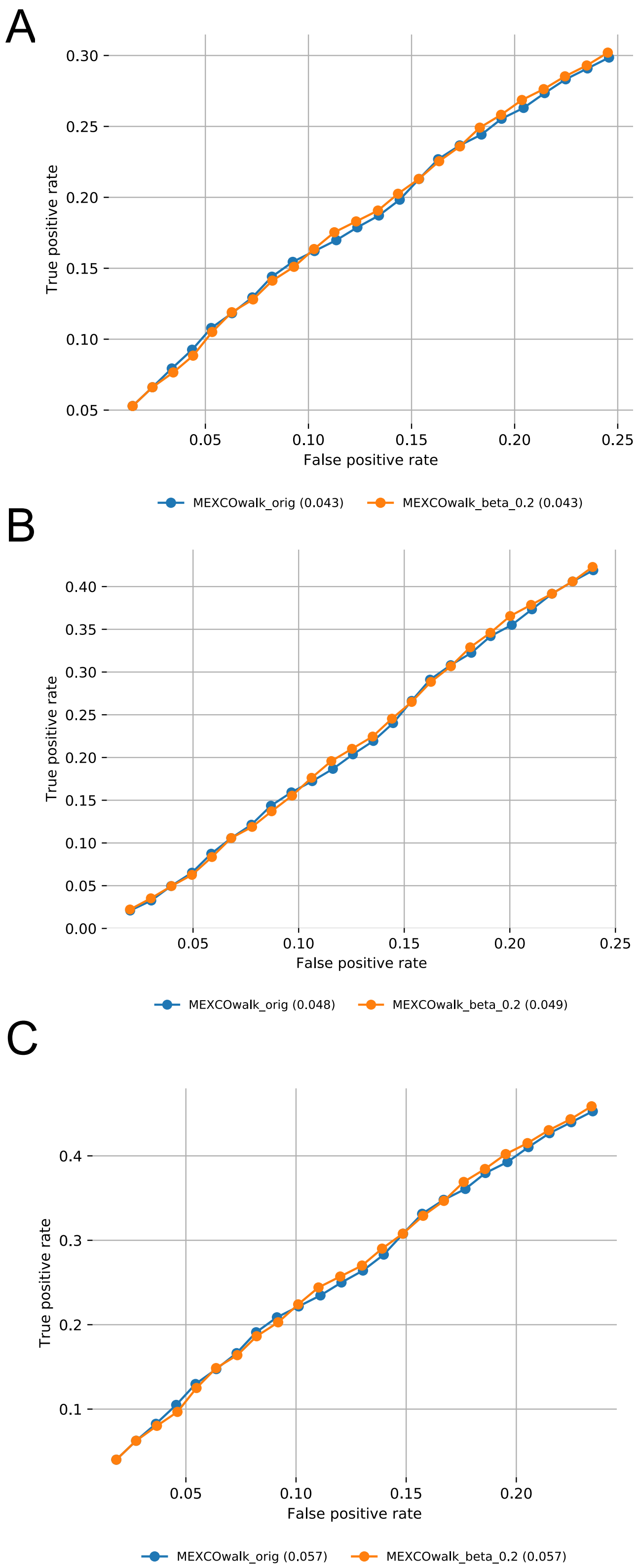

Figure S13: A) The fraction of recovered druggable genes for each $total\_genes$ value is shown with a ROC plot for MEXCOwalk original run ($\beta = 0.4$) and a version of MEXCOwalk where $\beta = 0.2$. B) Same as A, but only those druggable genes with $\leq 1\%$ mutation frequency in the pan-cancer cohort are used. C) Same as A, but only those druggable genes with $\leq 2\%$ mutation frequency in the pan-cancer cohort are used.

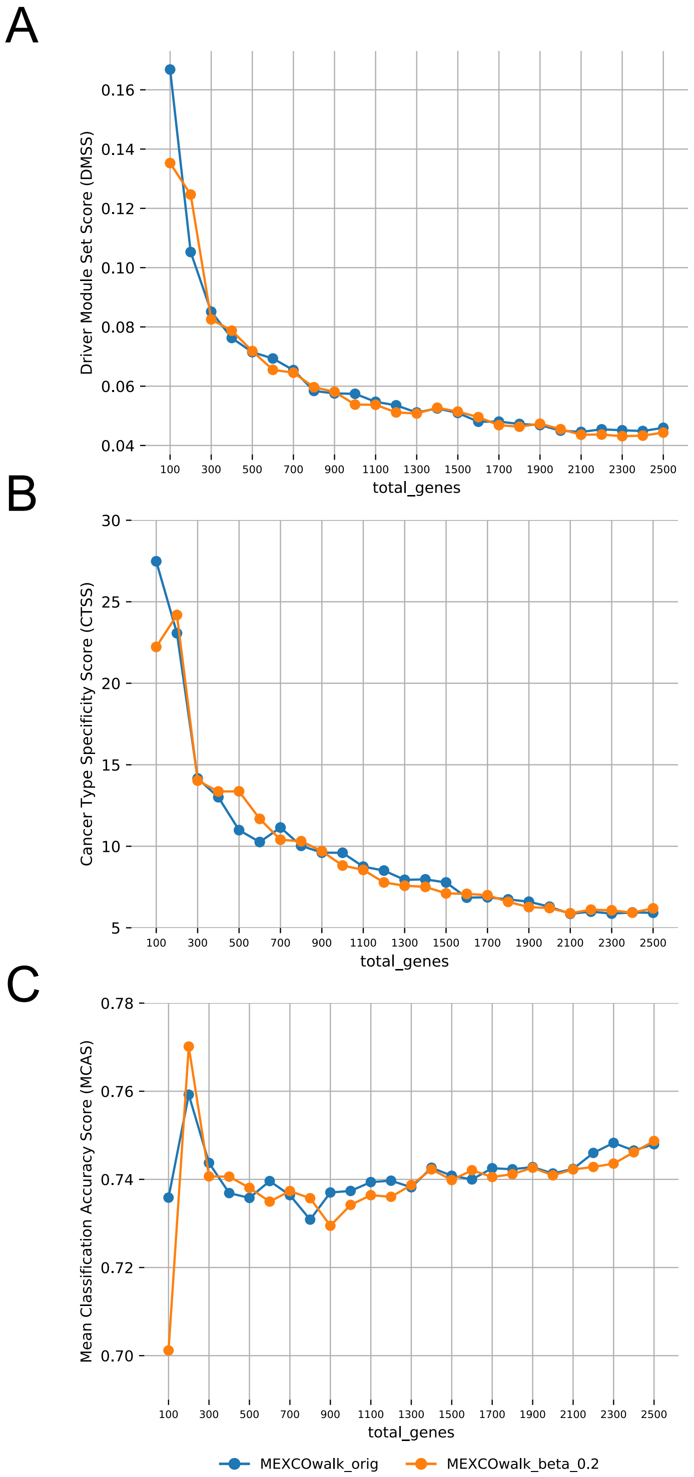

Figure S14: A) DMSS evaluations of output modules of MEXCOwalk original run ($\beta = 0.4$) and a version of MEXCOwalk where $\beta = 0.2$. B) CTSS evaluations of output modules of MEXCOwalk original run ($\beta = 0.4$) and a version of MEXCOwalk where $\beta = 0.2$. C) MCAS evaluations of output modules of MEXCOwalk original run ($\beta = 0.4$) and a version of MEXCOwalk where $\beta = 0.2$.

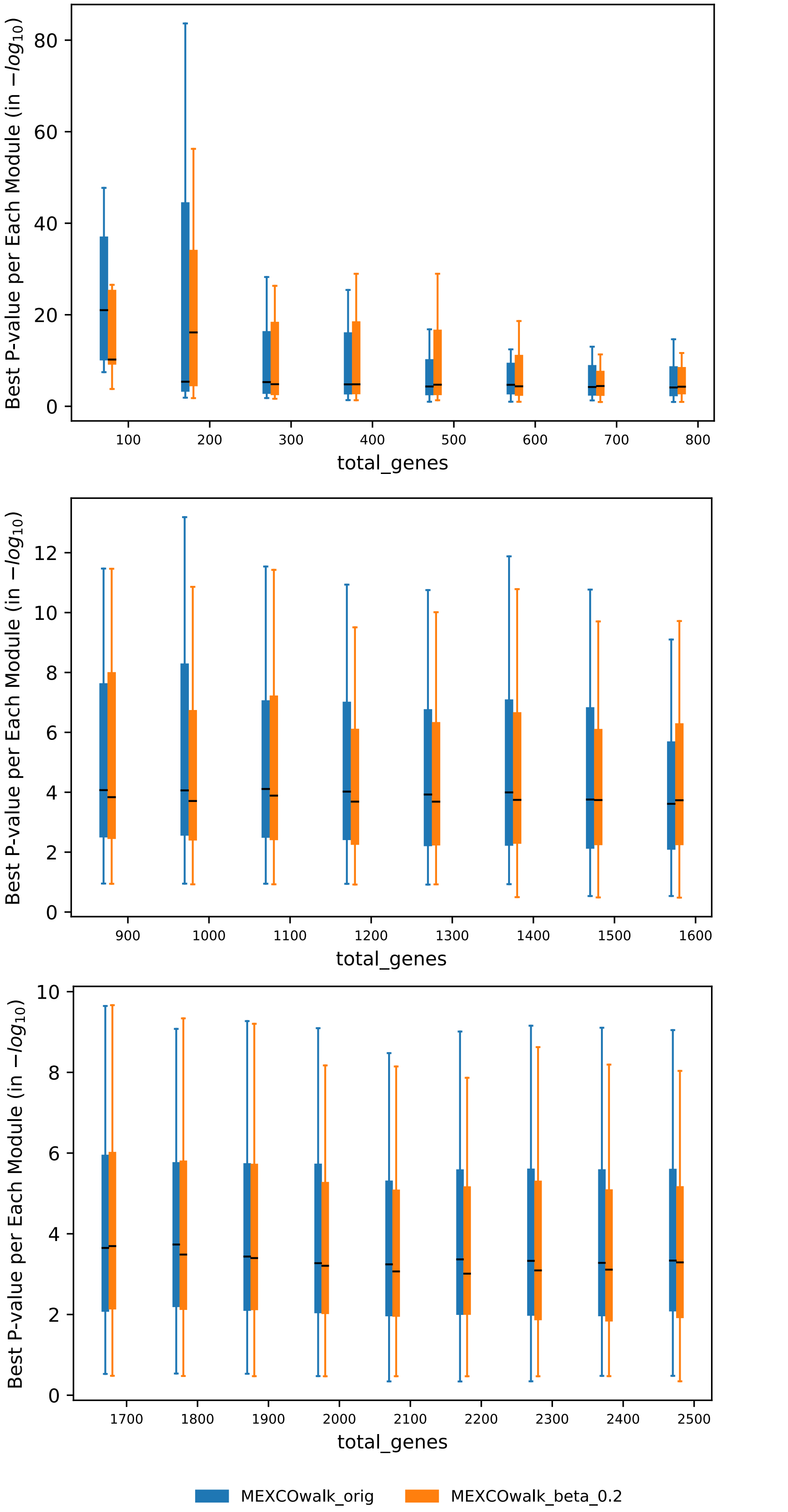

Figure S15: In reference to Figure S14-B, box plot distribution of best p-value for each module obtained for increasing values of *total_genes*. Whiskers represent the interquartile range.

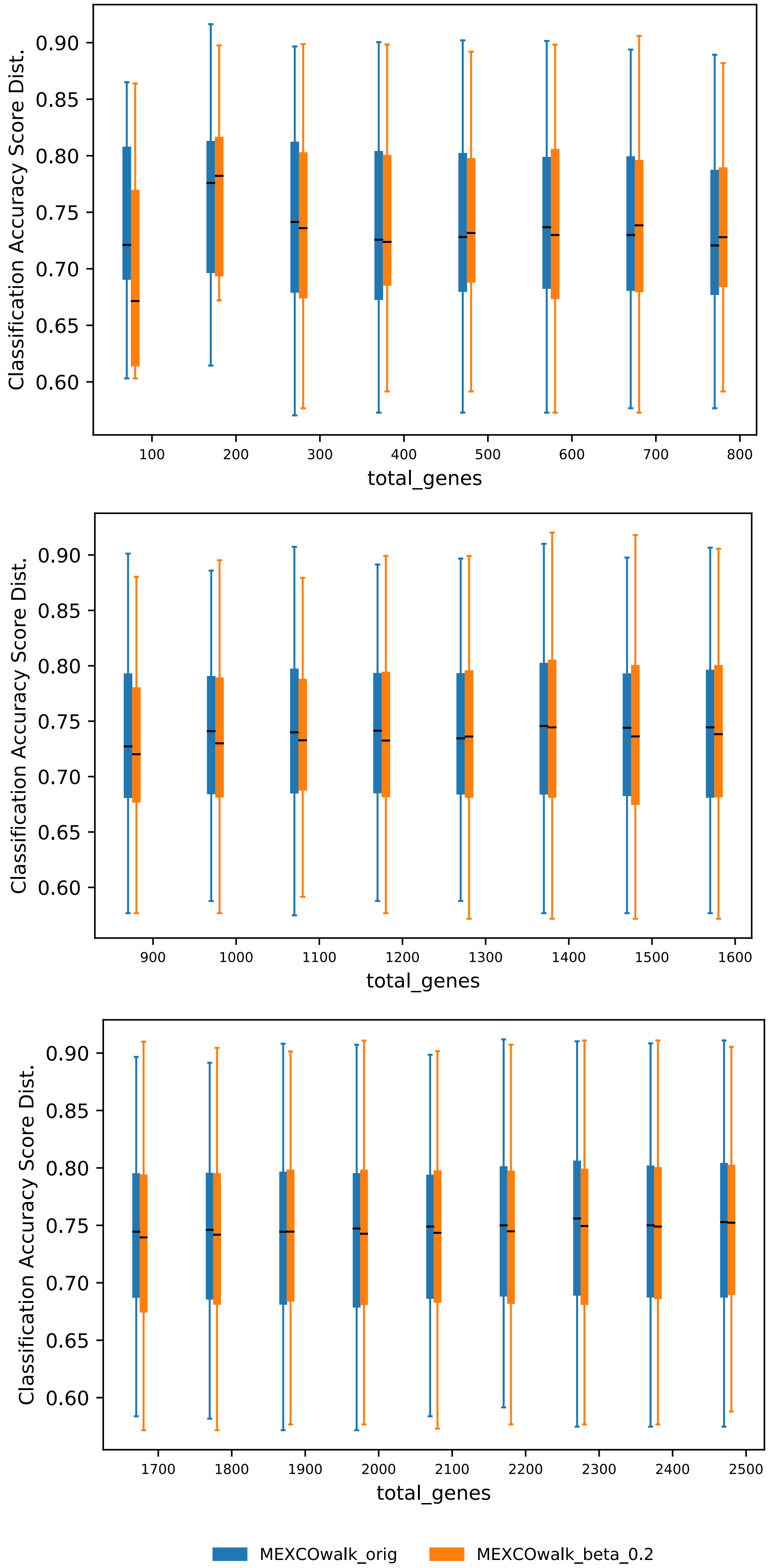

Figure S16: In reference to Figure S14-C, box plot distribution of classification accuracy for each module obtained for increasing values of *total_genes*. Whiskers represent the interquartile range.

Table S2: **Left:** Percentage of the number of different genes in MEXCOwalk output gene sets when the input PPI is changed from HINT+HI2012 to IntAct. **Right:** Overlaps with CGC.

total_genes	difference(%)	CGC overlap size		
		IntAct	HINT+HI2012	common
100	55.78	48	48	31
200	51.00	77	80	52
300	50.00	100	99	66
400	47.37	113	116	80
500	48.40	132	130	97
600	47.67	146	144	106
700	47.57	163	158	117
800	48.19	174	174	124
900	45.89	190	191	139
1000	45.75	199	204	145
1100	45.73	209	213	157
1200	44.67	220	222	166
1300	43.44	230	231	178
1400	42.58	239	244	188
1500	41.07	249	249	196
1600	39.69	256	259	205
1700	38.89	262	268	212
1800	37.83	269	274	219
1900	37.89	279	283	228
2000	38.12	282	287	233
2100	36.41	287	298	240
2200	35.23	299	303	249
2300	34.57	305	309	258
2400	33.97	318	312	268
2500	33.15	323	325	281

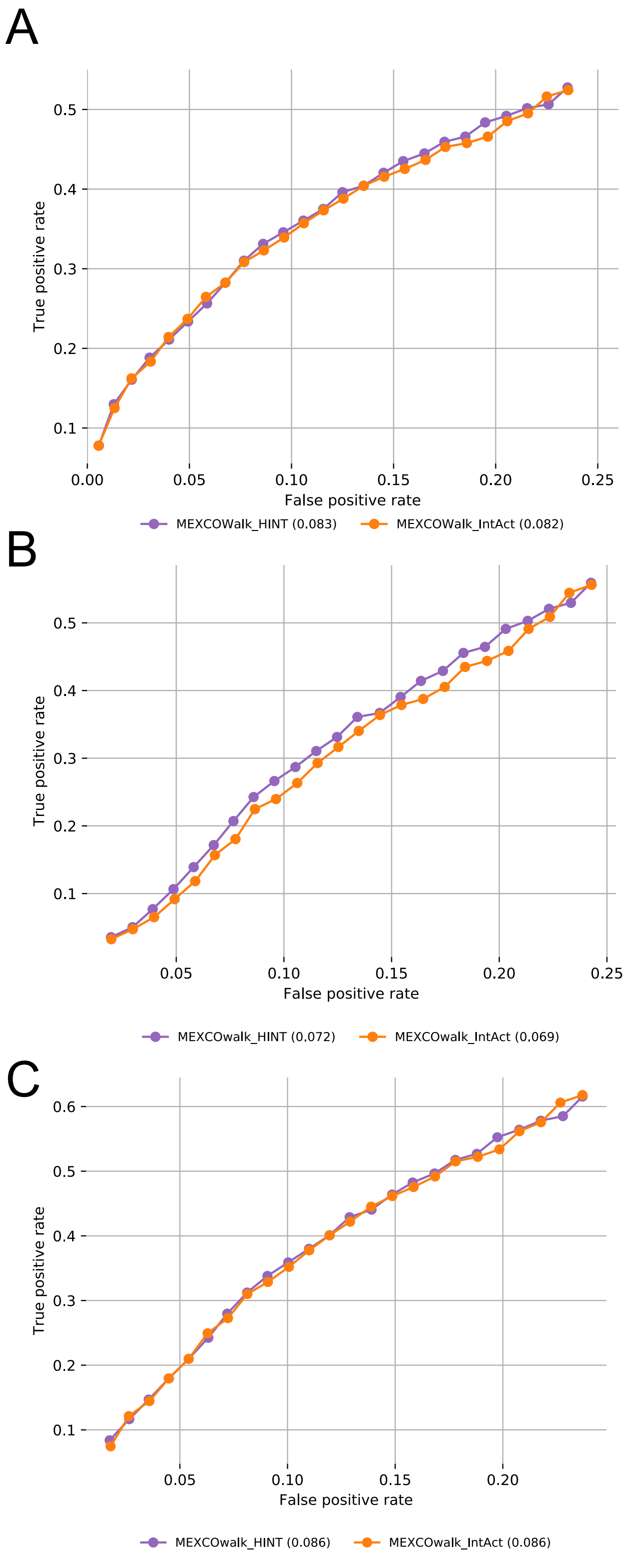

Figure S17: A) The fraction of recovered CGC genes for each *total_genes* value is shown with a ROC plot for MEXCOwalk original run and a version of MEXCOwalk where IntAct is used as the PPI network. B) Same as A, but only those CGC genes with ≤ 1% mutation frequency in the pan-cancer cohort are used. C) Same as A, but only those CGC genes with ≤ 2% mutation frequency in the pan-cancer cohort are used.

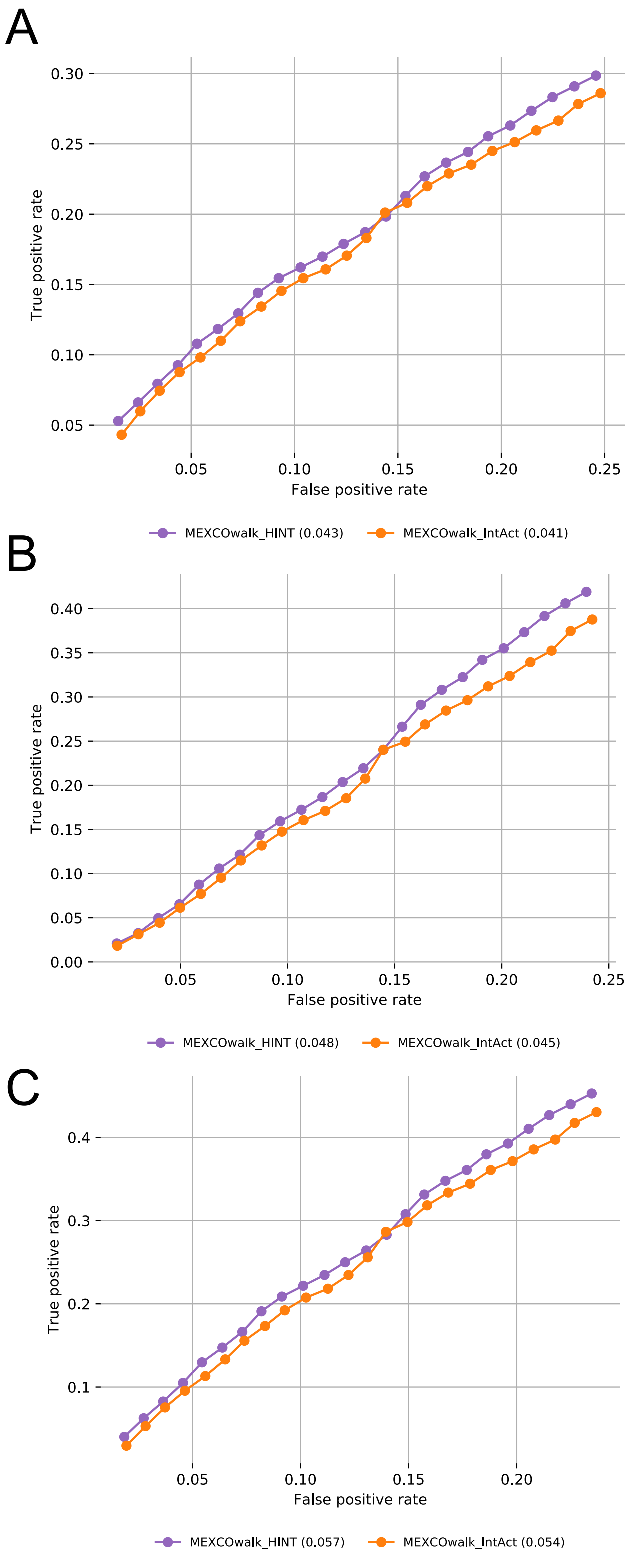

Figure S18: A) The fraction of recovered druggable genes for each *total_genes* value is shown with a ROC plot for MEXCOwalk original run and a version of MEXCOwalk where IntAct is used as the PPI network. B) Same as A, but only those with recovered druggable genes with $\leq 1\%$ mutation frequency in the pan-cancer cohort are used. C) Same as A, but only those with recovered druggable genes with $\leq 2\%$ mutation frequency in the pan-cancer cohort are used.

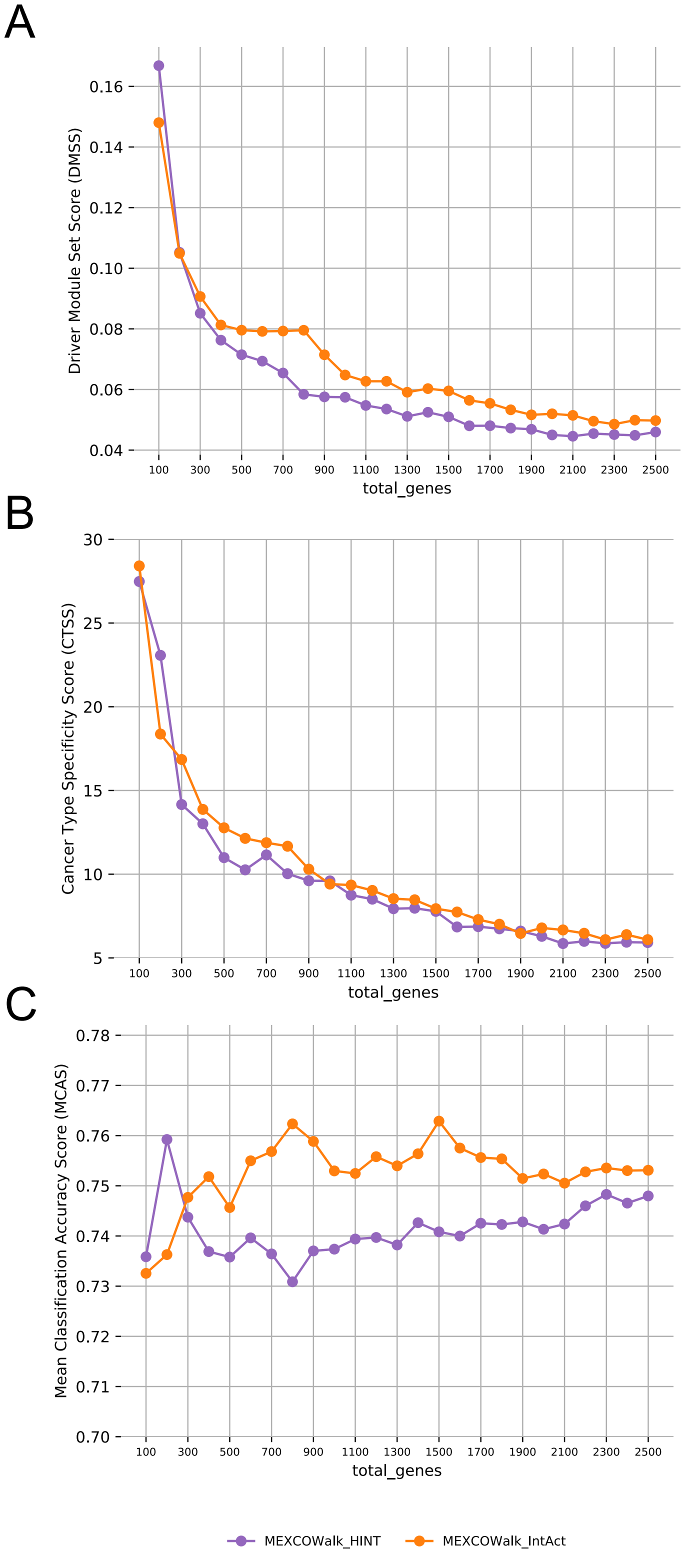

Figure S19: A) DMSS evaluations of output modules of MEXCOwalk original run and a version of MEXCOwalk where IntAct is used as the PPI network. B) CTSS evaluations of output modules of MEXCOwalk original run and a version of MEXCOwalk where IntAct is used as the PPI network. C) MCAS evaluations of output modules of MEXCOwalk original run and a version of MEXCOwalk where IntAct is used as the PPI network.

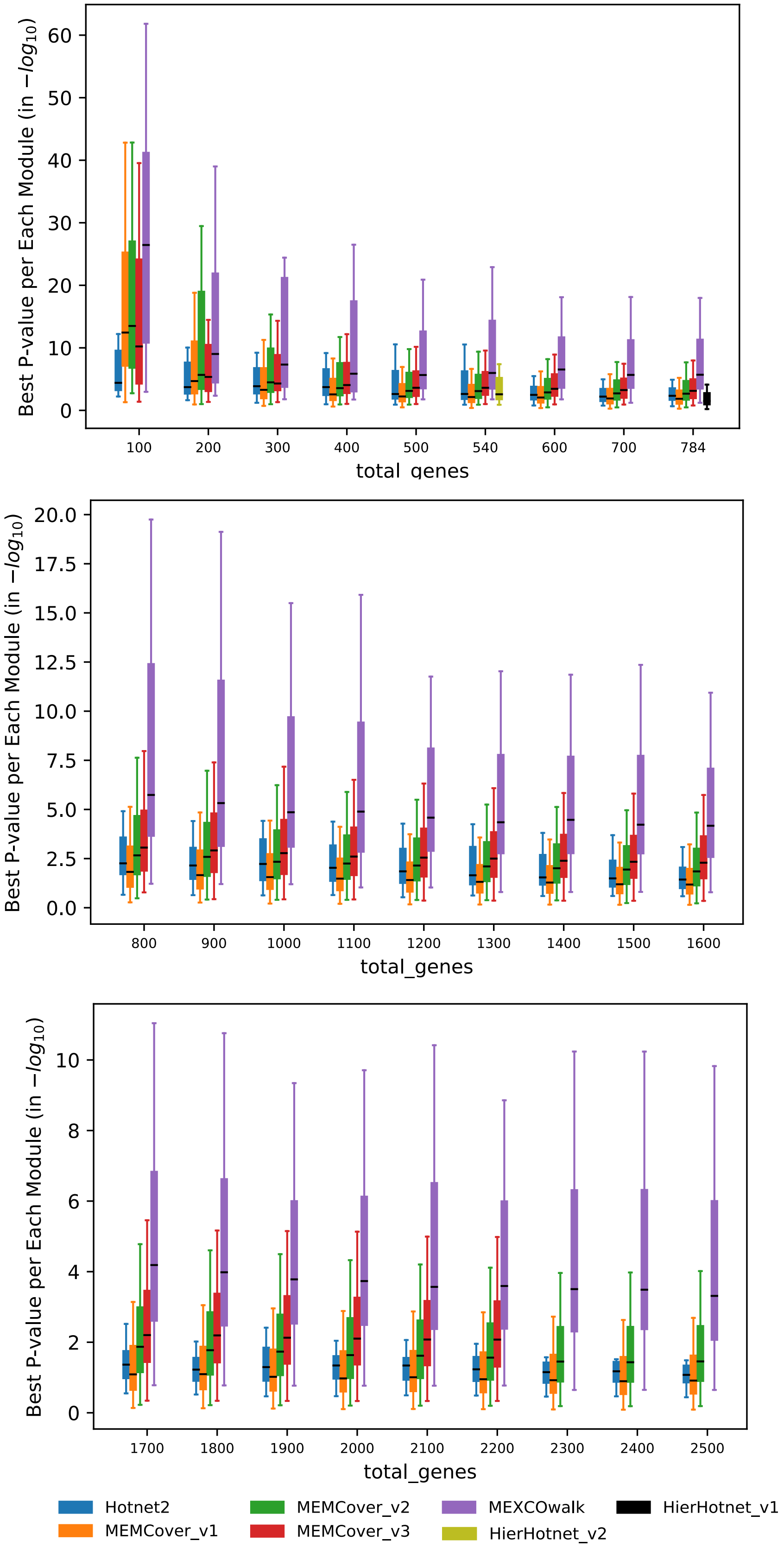

Figure S20: In reference to Figure S19-B, box plot distribution of best p-value for each module obtained for increasing values of *total_genes*. Whiskers represent the interquartile range.

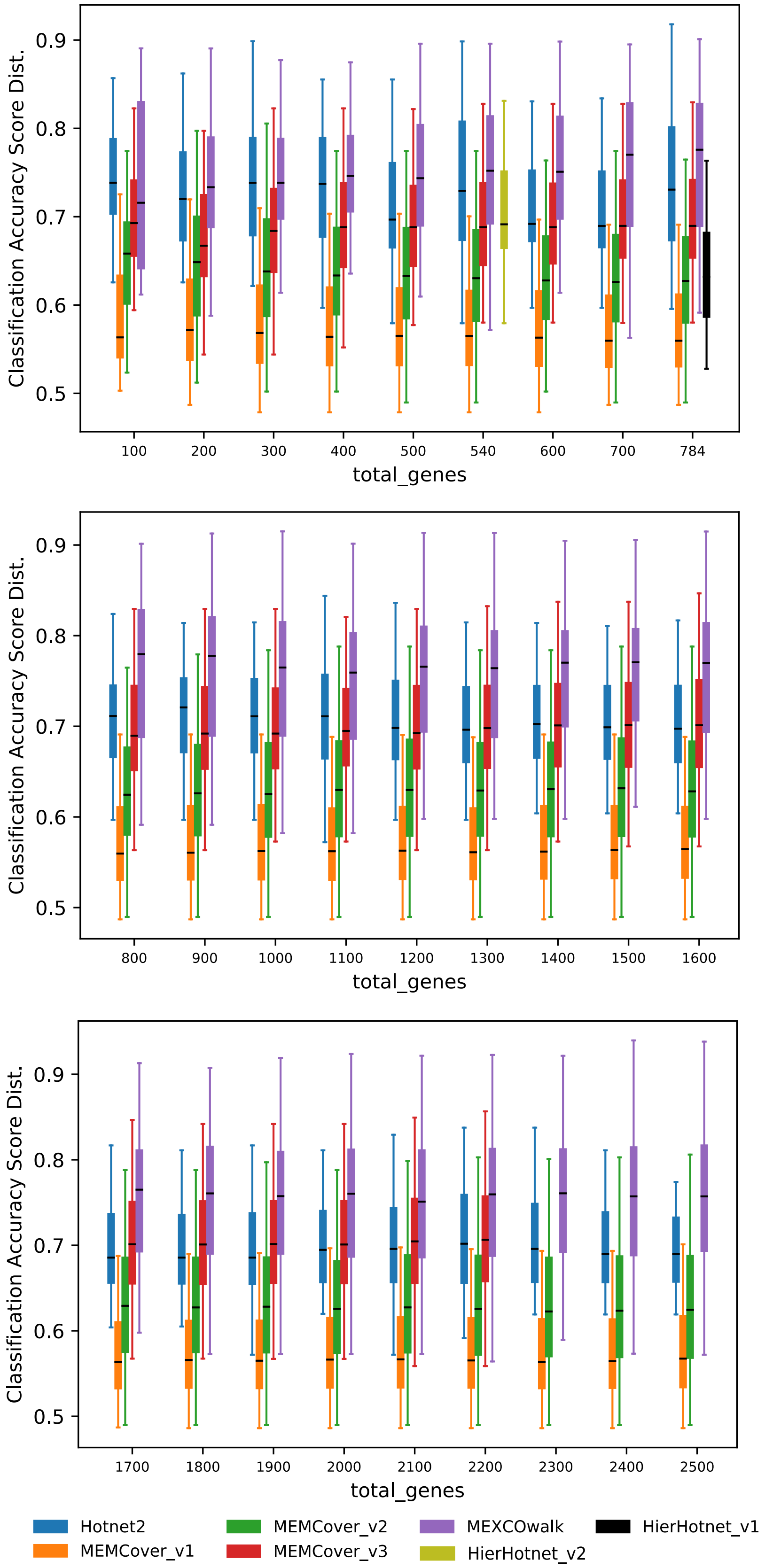

Figure S21: In reference to Figure S19-C, box plot distribution of classsification accuracy for each module obtained for increasing values of *total_genes*. Whiskers represent the interquartile range.

References

[1] Adam C Coffman, Alex Wollam, Gregory Spies, Kelsy C Cotto, Nicholas C Spies, Susanna Kiwala, Yang-Yang Feng, Alex H Wagner, Malachi Griffith, and Obi L Griffith. DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic Acids Research*, 46(D1):D1068–D1073, 2017.

[2] S.A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C.G. Cole, S. Ward, E. Dawson, and L. et al. Ponting. Cosmic: somatic cancer genetics at high-resolution. *Nucleic Acids Res*, 45:D777–D783, 2017.

[3] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1979.

[4] S. Orchard, M. Ammari, B. Aranda, L. Breuza, and L. et al. Briganti. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*, 42(Database issue):D358–63, 2013.