The American Journal of Human Genetics, Volume *98*

Supplemental Information

MEGSA: A Powerful and Flexible Framework for Analyzing Mutual Exclusivity of Tumor Mutations

Xing Hua, Paula L. Hyland, Jing Huang, Lei Song, Bin Zhu, Neil E. Caporaso, Maria Teresa Landi, Nilanjan Chatterjee, and Jianxin Shi

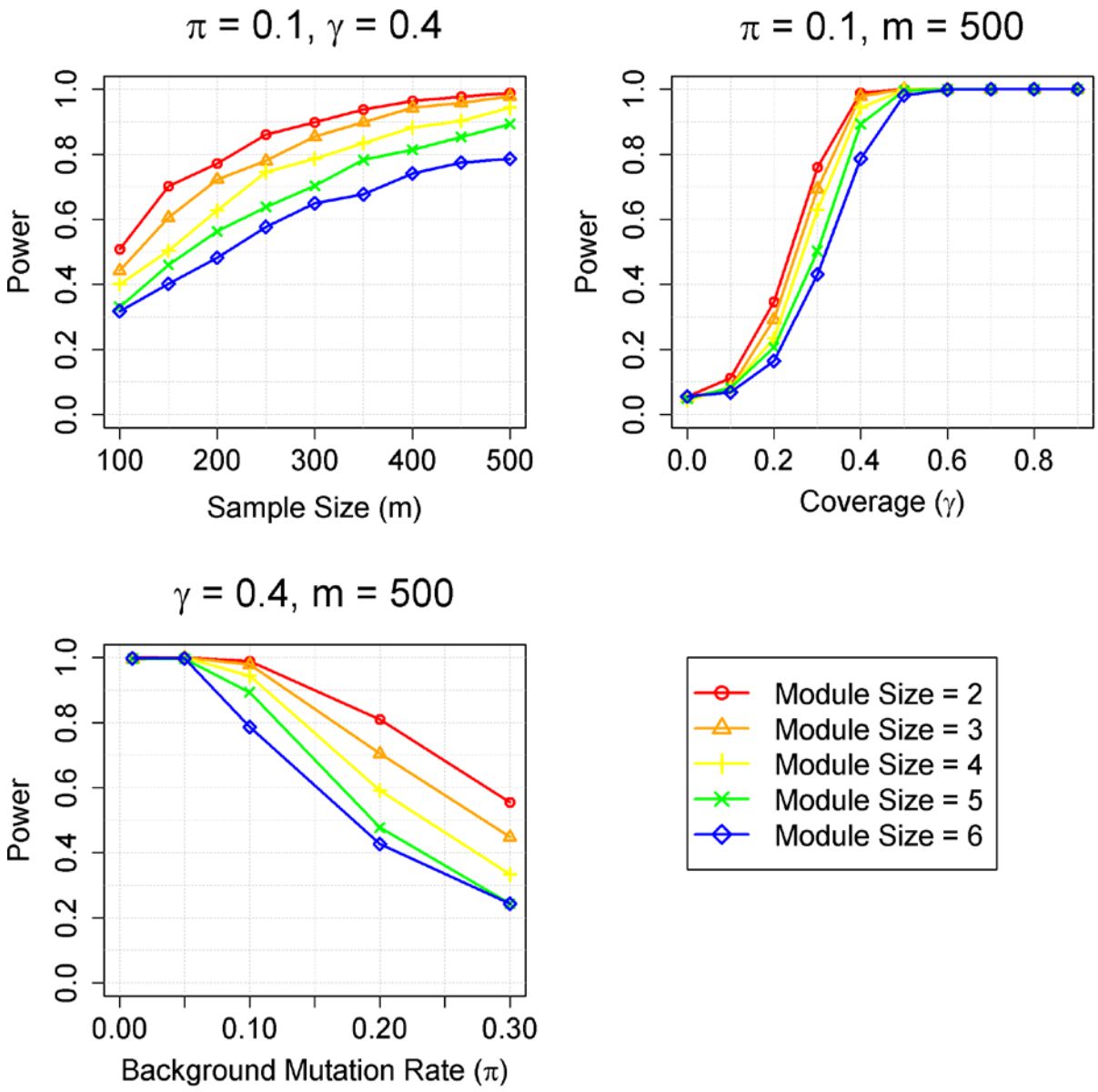


Figure S1: Power behavior of our likelihood ratio test. Power was estimated at level α=0.05 based on 1000 simulations. Power increases with sample size and coverage but decreases with background mutation rate. π is the background mutation rate for all genes. γ is the coverage of the simulated MEGS. m is the sample size.

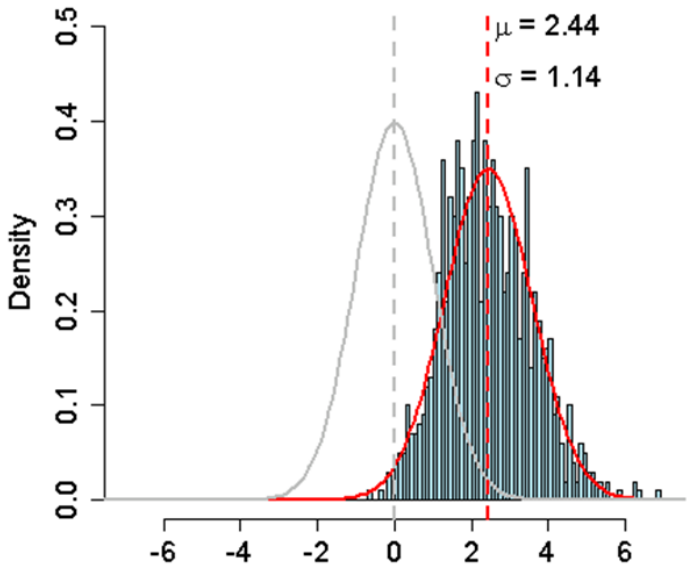


Figure S2: Null distribution of LRT-SB. Simulations were performed under the null hypothesis with 1000 samples and four genes. All four genes had the same mutation frequency 20%. The histogram was based on simulations. The red curve is fitted to the histogram with mean 2.44 and standard deviation 1.14. In the original paper, the null distribution was claimed to be *N(0,1)*.

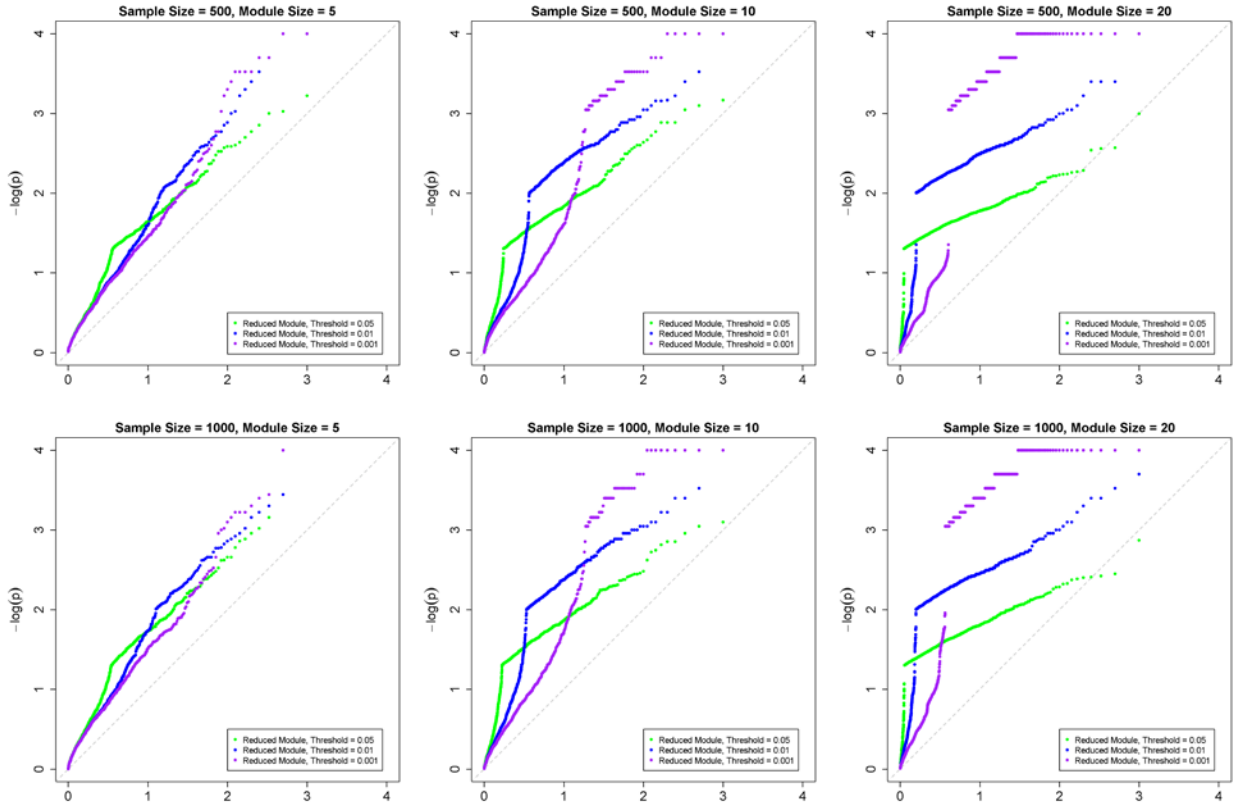


Figure S3: Quantile-quantile (QQ) plot of *P*-values for "cliques" under global null hypothesis, generated by the algorithms in MEMo. Mutation frequency is 10% for each gene. MEMo uses a default threshold P_0=0.05 in their procedure. Here, we tried three different thresholds 0.05, 0.01 and 0.001. For each simulation, we make QQ plot against the uniform distribution U(0,1). The x-coordinate is log(p) with p~U(0,1). The y-coordinate is the log(p) for observed overall P-values produced from the MEMo algorithm. *P*-values deviate U(0,1).

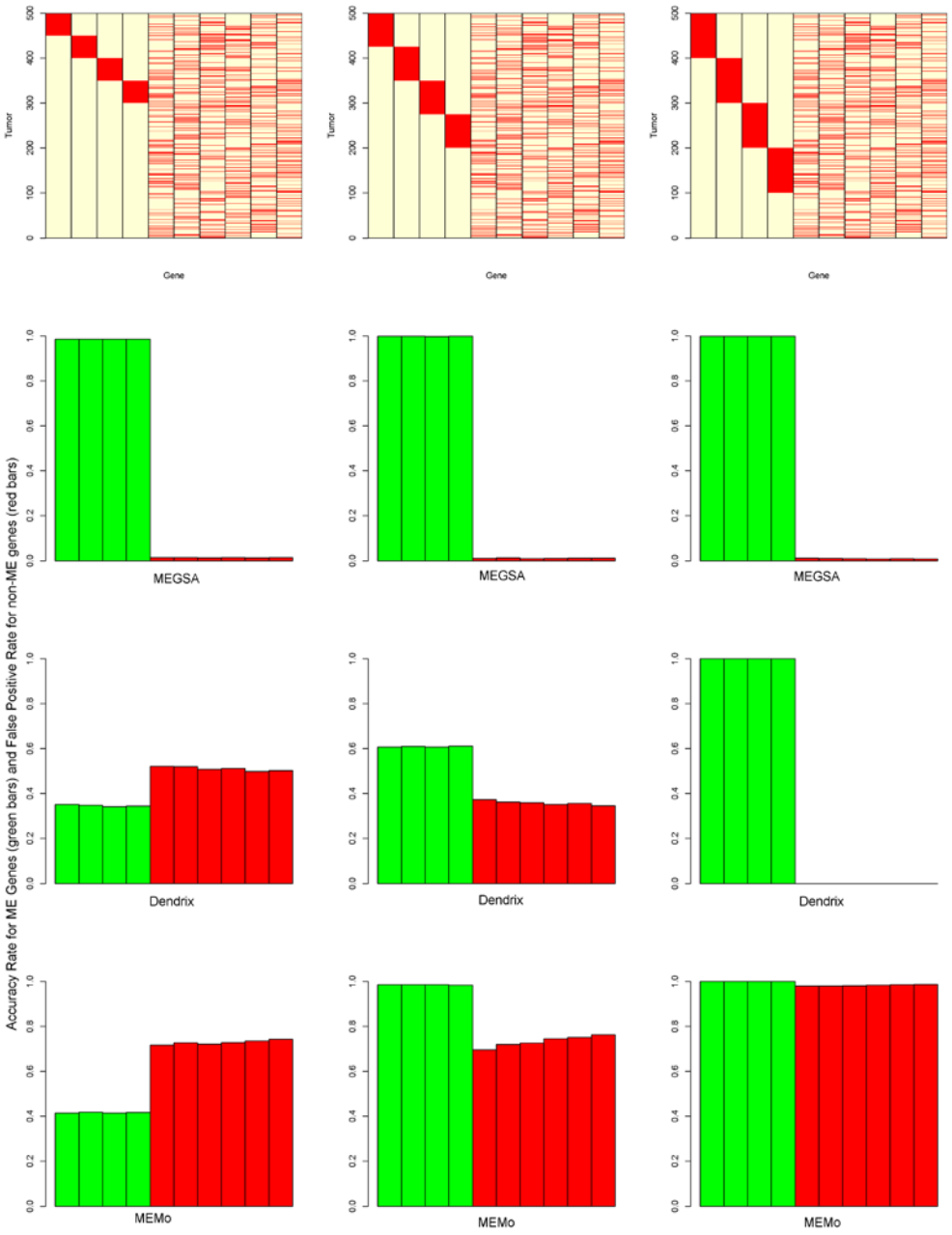


Figure S4: **Number of true positive and false positive genes.** For MEGSA and Dendrix, numbers are calculated based on the top MEGS candidate. For MEMo, the numbers are calculated based on the algorithm described in Supplementary Note. We performed simulations using three sets of parameters with coverage ranging from 0.4 to 0.8. The figures in the first row show the simulated pattern with four genes (left four) in MEGS and six genes randomly mutated (right six). The left four bars (green) show the probability of choosing the true MEGS genes. The right six bars (red) show the probability of choosing false positive genes. The simulation was based on 4 MEGS genes and 6 non-MEGS genes.

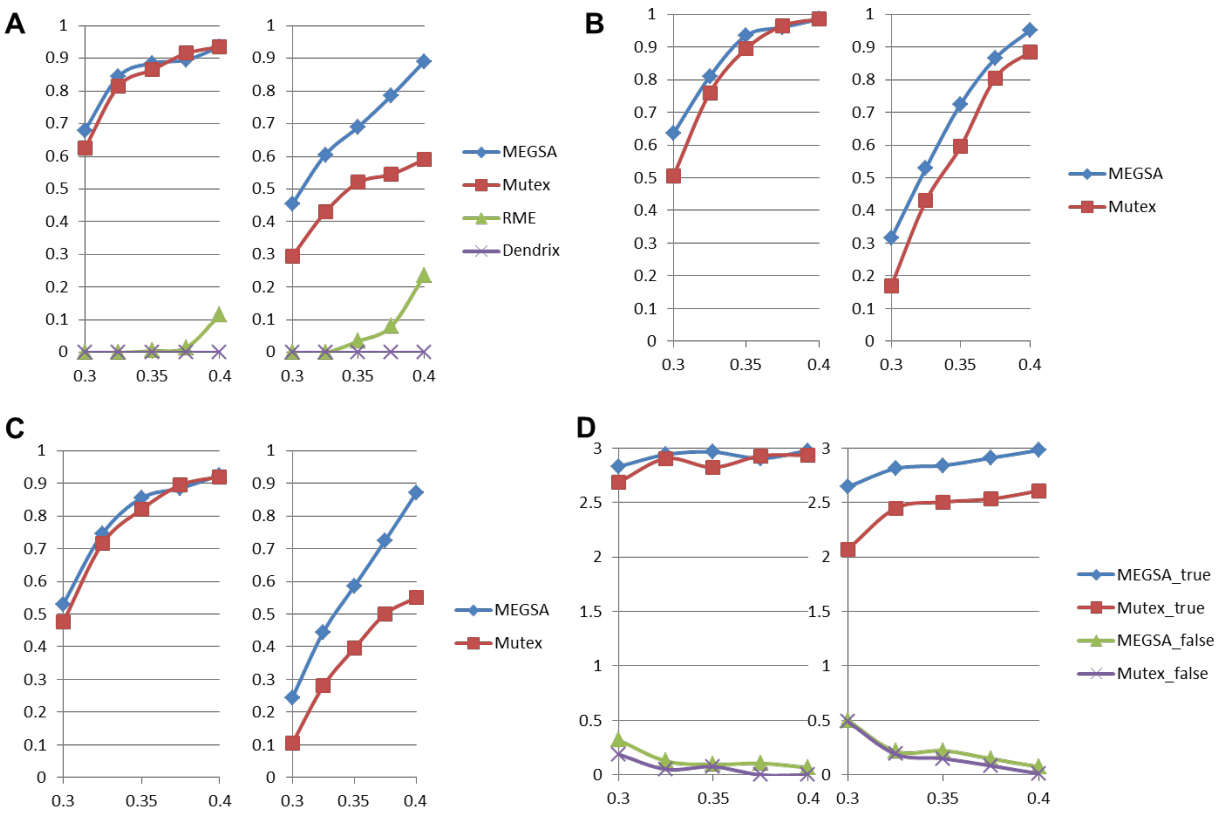


Figure S5: Performance comparison of methods for detecting mutually exclusive gene sets. In all simulations, we have 53 genes with 50 being randomly simulated with specific mutation frequencies and 3 genes as MEGS. For each comparison, the left panel is for balanced MEGS with mutation frequency ratio 1:1:1; the right panel is for imbalanced MEGS with mutation frequency ratio 4:1:1. In all figures, the x-coordinate is the coverage γ, ranging from 0.3 to 0.4. (A) Probability of ranking the true MEGS as top candidate. (B) Power of detecting true MEGS using MEGSA and Mutex. (C) Probability that the identified top MEGS is statistically significant and identical to the true MEGS. (D) The numbers of detected true positive genes (out of 3) and false positive genes.

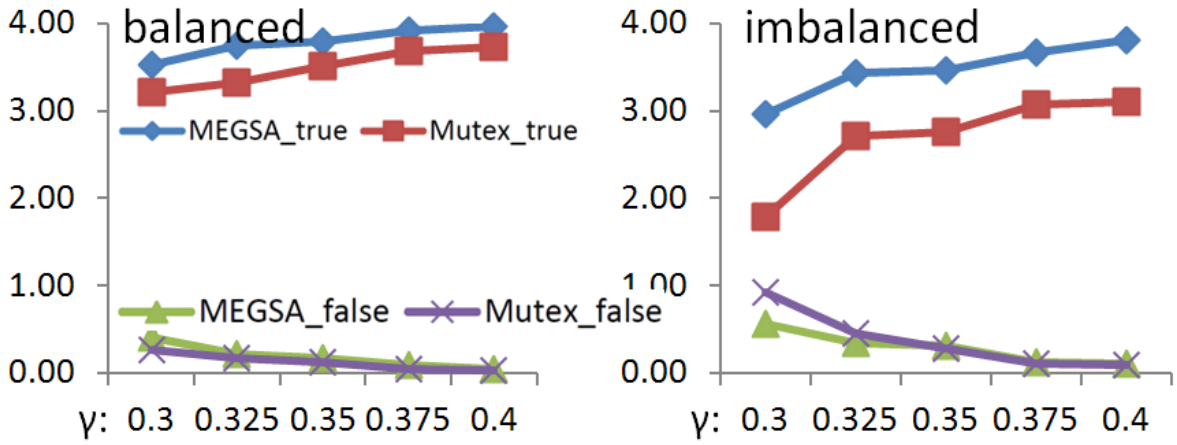


Figure S6: The number of true positive and false positive genes detected by MEGSA and Mutex. In all simulations, we have 54 genes with 50 being randomly simulated with specific mutation frequencies and 4 genes as MEGS. In imbalanced MEGS, the mutational frequencies have a ratio 3:1:1:1.For each simulation, we counted the number of selected true genes out of 4 and the number of falsely selected genes not belonging to the simulated MEGS. MEGSA has similar false positive rate but higher number of selected true genes, particularly for imbalanced MEGS.

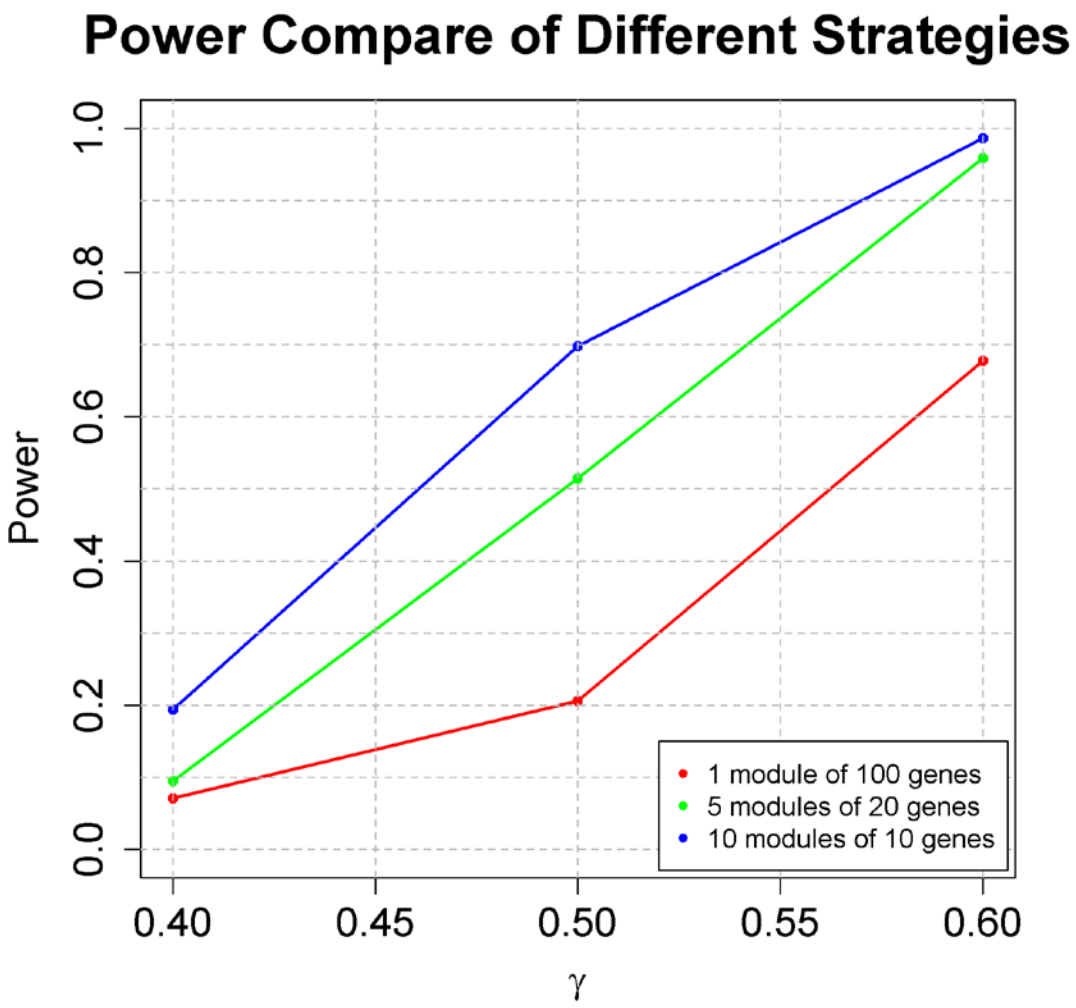


Figure S7: Compare statistical power of two search strategies using MEGSA. The simulation was based on M=100 genes and N=500 samples. A balanced MEGS had four genes with background mutation rate 10%. The rest 96 genes were randomly mutated with frequency 10%. We compared the statistical power of rejecting the global null hypothesis for two analysis strategies: *de novo* discovery and pathway-guided search. For *de novo* discovery of MEGS (red curve), we applied MEGSA to the whole set of 100 genes and rejected the global null hypothesis if the overall *P*-value is less than 0.05. For pathway-guided search, we assume that 100 genes are split to *L*(*L=5 or 10*) non-overlapping modules, each of which has *100/L* genes. The simulated MEGS were in one of the modules. We applied MEGSA to each of the modules and rejected the global null hypothesis if any of the *L P*-values were less than 0.05/*L* based on the Bonferroni correction. The power was calculated as the proportion of simulations rejecting the global hypothesis. If the MEGS is within any of the cliques, pathway-guided search may substantially improve the power due to the reduced multiple testing burden.

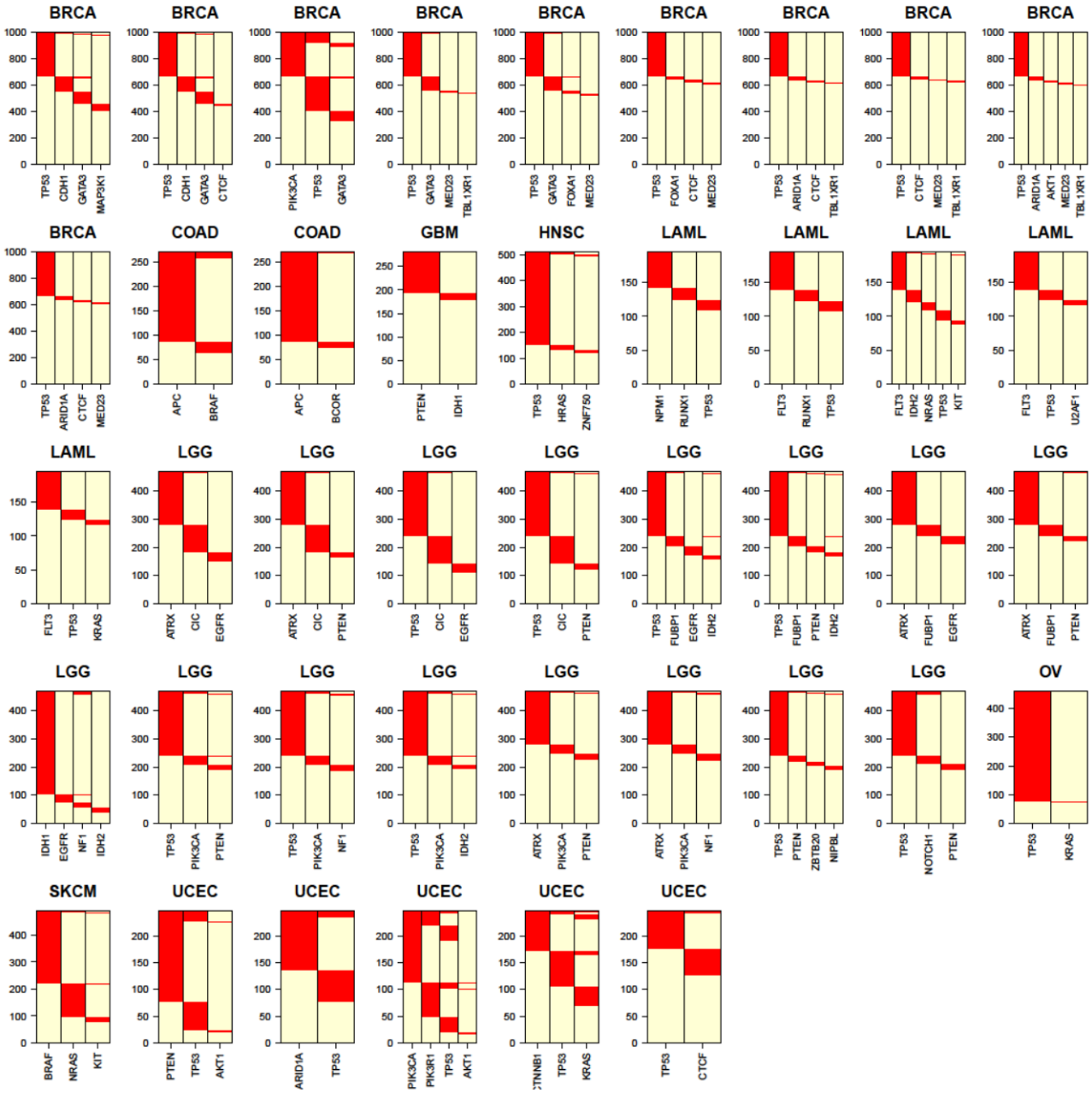


Figure S8A: Significant MEGS **(**multiple testing corrected P<0.05**)** identified by MEGSA for different cancers using the whole exome sequencing data in TCGA.

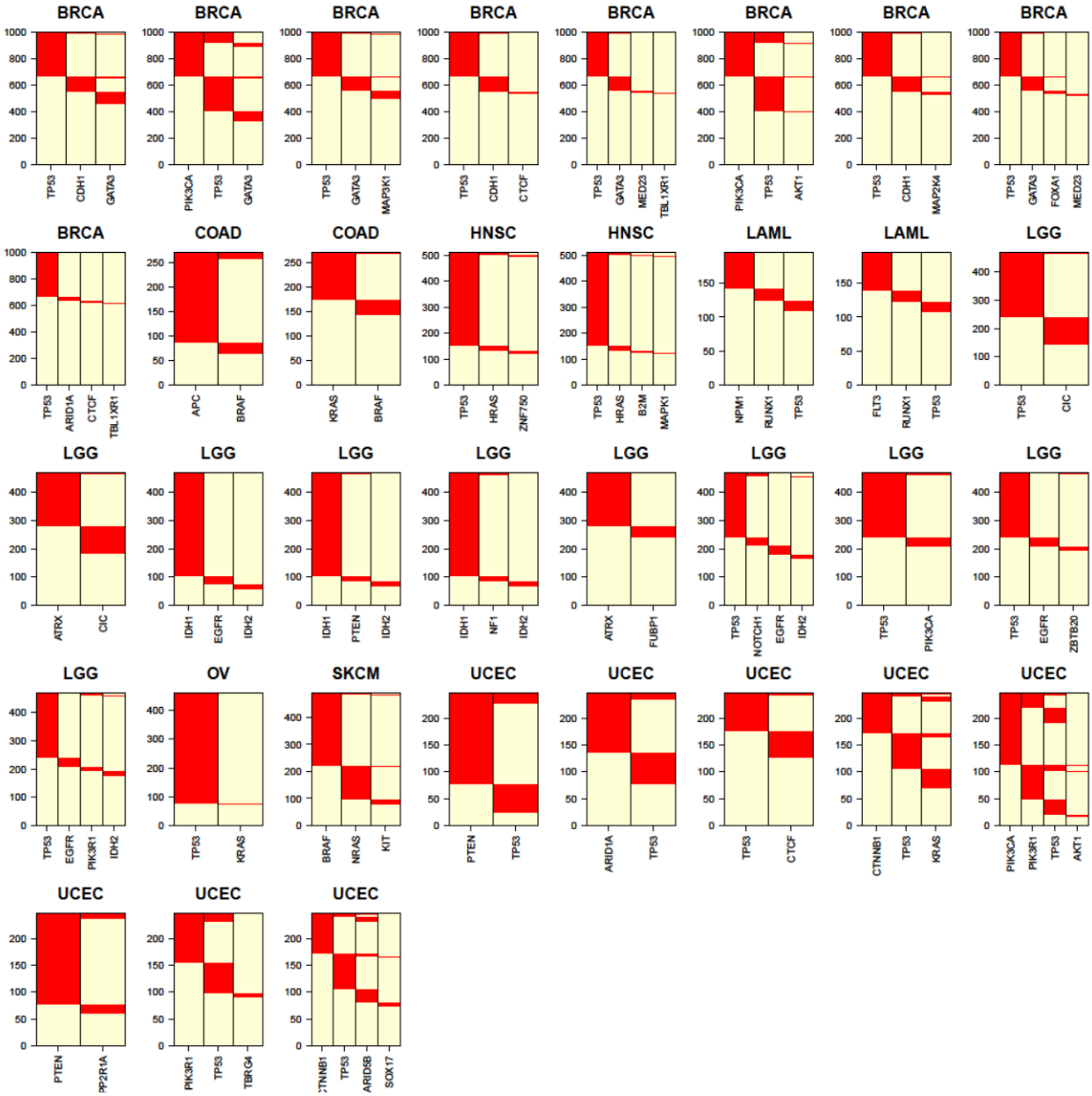


Figure S8B: Significant MEGS **(**FDR q-value<0.05**)** identified by Mutex for different cancers using the whole exome sequencing data in TCGA.

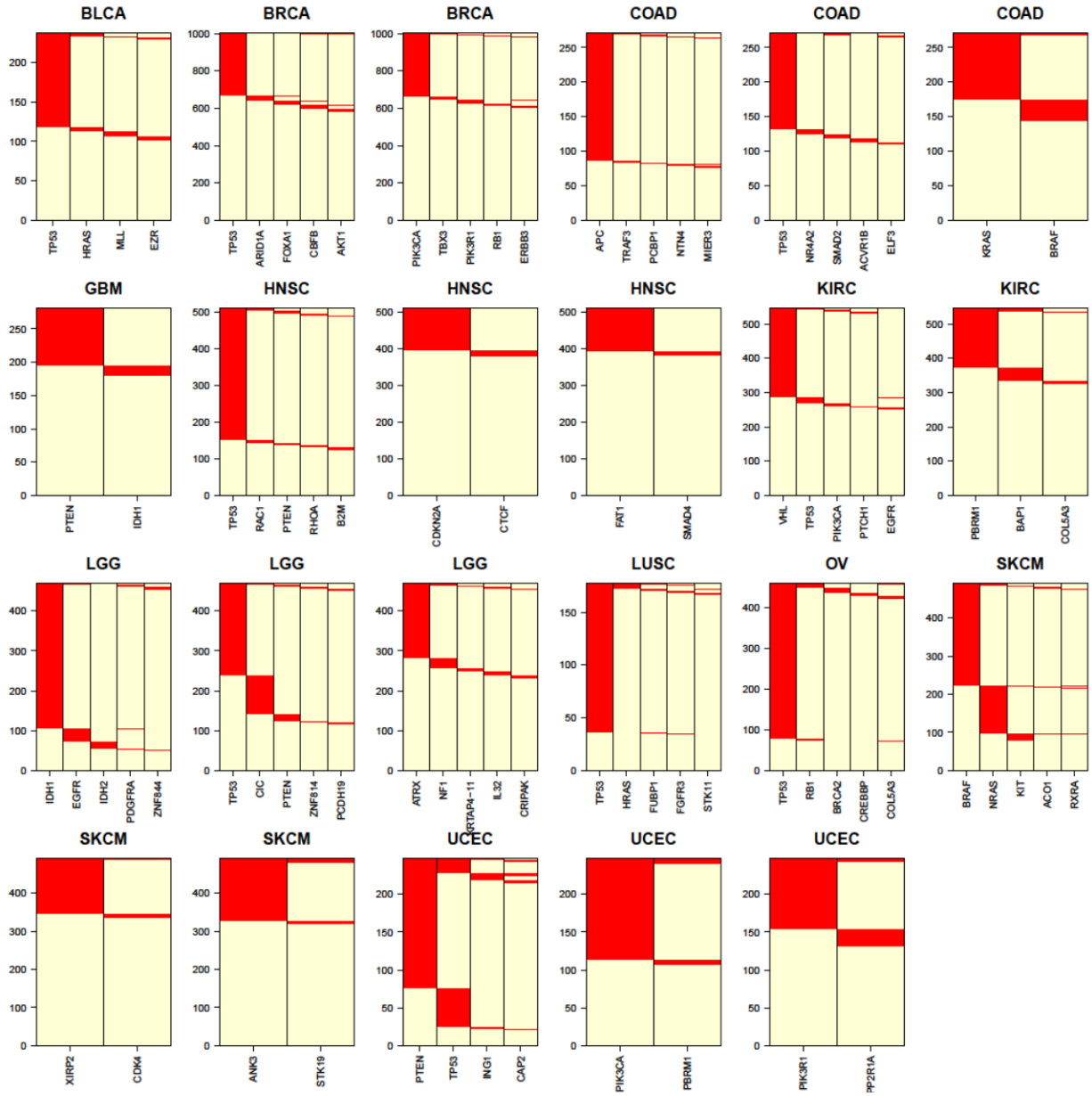


Figure S8C: MEGS identified by RME for different cancers using the whole exome sequencing data in TCGA. LAML results were not available because RME did not run successfully on the LAML data.

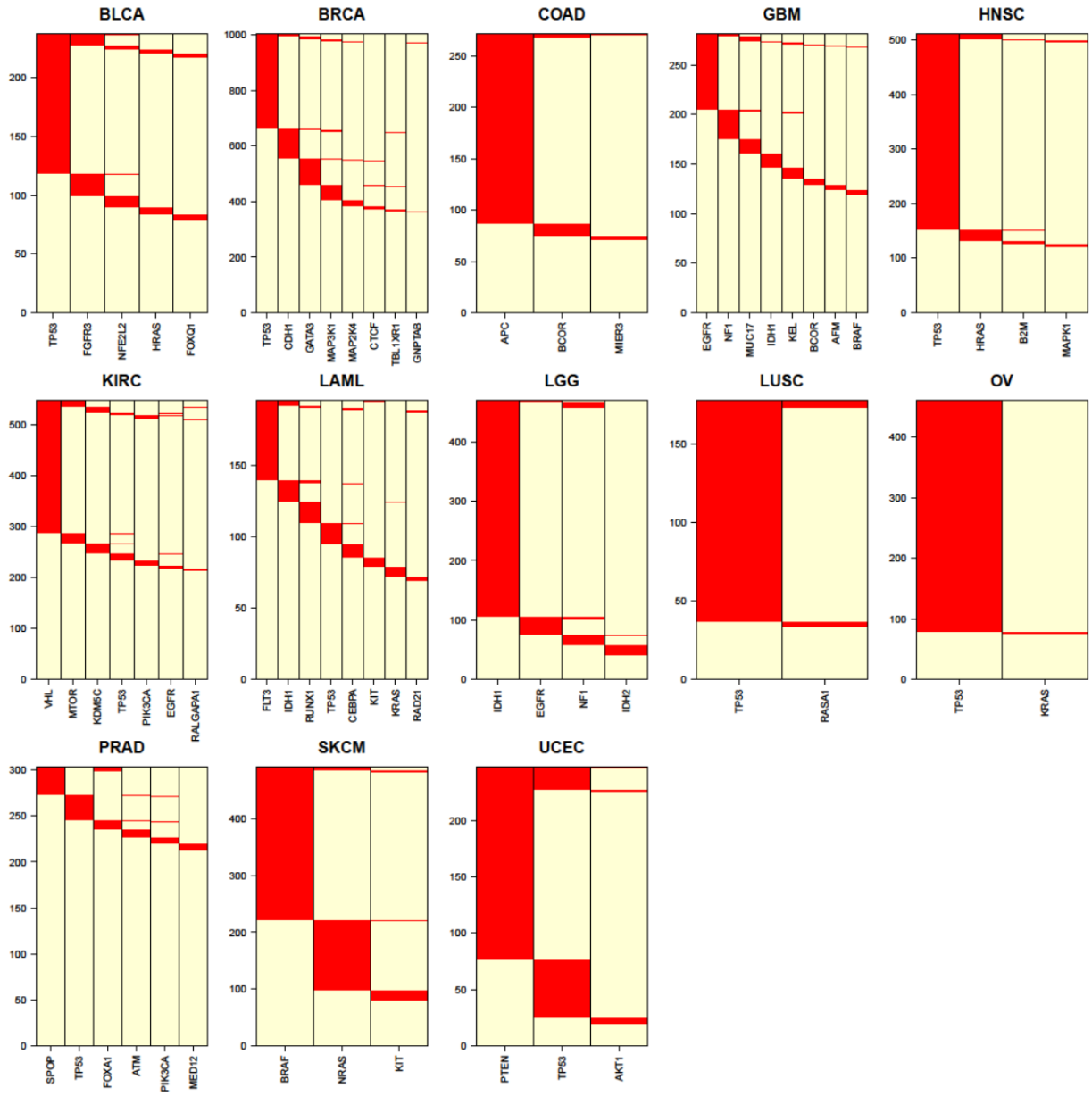


Figure S8D: The top MEGS identified by Dendrix for different cancers using the whole exome sequencing data in TCGA.

α	π	k=2		k=3		k=4		k=5		k=6	
		m=100	m=500	m=100	m=500	m=100	m=500	m=100	m=500	m=100	m=500
0.05	0.1	0.031	0.061	0.068	0.054	0.056	0.052	0.054	0.051	0.054	0.052
	0.2	0.061	0.053	0.053	0.051	0.053	0.051	0.053	0.051	0.053	0.051
0.001	0.1	0	0.0021	0.0004	0.0012	0.001	0.0011	0.0012	0.00107	0.0011	0.00109
	0.2	0.00145	0.0011	0.0012	0.00108	0.00108	0.0010	0.0011	0.0010	0.0011	0.00105
0.0001	0.1	0	0.000021	0.000011	0.000102	0.00007	0.000124	0.000113	0.000099	0.000117	0.0001
	0.2	0.000074	0.00012	0.00013	0.00011	0.00011	0.00010	0.00012	0.000092	0.00012	0.00010

Table S1: Type I error rate for the likelihood ratio statistic. Here, α is the specified significance level; π is the gene mutation rate, which was assumed the same for all genes in each simulation; k is the number of genes for simulations; m is the number of subjects for simulation.

Cancer type	Number patients	Number of genes	How to select genes
BLCA	238	39	Tumor Portal
BRCA	989	39	Tumor Portal
COAD	269	39	Tumor Portal
GBM	282	34	Tumor Portal
HNSC	511	40	Tumor Portal
KIRC	451	25	Tumor Portal
LAML	196	26	Tumor Portal
LGG	465	45	MutSigCV
LUSC	178	25	Tumor Portal
OV	462	15	Tumor Portal
PRAD	300	7	Tumor Portal
SKCM	428	39	Tumor Portal
UCEC	248	94	Tumor Portal

Table S2: Cancer types, the number of subjects and the number of genes included for analysis of mutual exclusivity. Genes were selected by the TumorPortal website: http://cancergenome.broadinstitute.org/, which were derived based on the TCGA tumor sequencing data. The TumorPortal website does not report results for LGG; thus we downloaded mutation data from the TCGA website and ran MutSigCV to identify significantly mutated genes.

Supplementary note: Summary of some methods for detecting mutually exclusive gene sets

1. MEMo

MEMo partitions all input genes into *N* cliques (fully connected gene networks) based in external biological information. The key is how to derive an overall *P*-value for testing the global null hypothesis for a clique with *M* genes. For a given gene set with *m* genes, MEMo defines the coverage (the proportion of patients with at least one mutation in these genes) as statistic for assessing mutual exclusivity and approximates the significance by permutations. MEMo uses the following procedure to derive an overall *P*-value for the clique with *M* genes.

(1) Run permutations for the whole clique with *M* genes. If the *P*-value (P_M) is less than a specified threshold P_0 (default 0.05 in MEMo), the algorithm stops, the overall *P*-value is P_M and the best MEGS has *M* genes. Otherwise, go to step 2;

(2) Delete the gene with the smallest mutation frequency and test the significance for the remaining *M-1* genes using permutations. Let *P*-value be P_{M-1}. If $P_{M-1}<P_0$, record the overall *P*-value as P_{M-1} and the best MEGS has *M-1* genes. Otherwise, repeat step (2) until only two genes remain. In this case, the overall *P*-value is P_2.

After deriving overall *P*-values for all cliques, MEMo selects statistically significant cliques by controlling FDR using these overall *P*-values. However, controlling FDR requires that *P*-values under null hypothesis follow a uniform distribution U(0,1). Obviously, the overall *P*-values based on the above procedure do not follow U(0,1). To numerically demonstrate this, we have re-implemented their algorithm and summarize overall *P*-value results in **Supplementary Figure S3**. Clearly, these overall *P*-values under null hypothesis dramatically deviate from U(0,1), suggesting that MEMo does not control type-I error correctly.

Another problem is the way of selecting "optimal" MEGS. If the true MEGS has 3 genes that are very strongly mutually exclusive, then it is very likely the permutation test has a *P*-value < P_0 for the whole clique with *K* (*K*>>3) genes, i.e. MEMo includes too many false positive genes. This is confirmed by simulations, reported in **Supplementary Figure S4**.

2. Dendrix/MDPFinder/Multi-Dendrix

Dendrix, MDPFinder and Multi-Dendrix are designed for *de novo* discovery of MEGS. All algorithms use the same criterion for ranking gene sets motivated by two requirements: (1) most patients have at least one mutation in MEGS (high coverage) and (2) most patients have no more than one mutation in MEGS (approximate exclusivity). For a subset of *m* genes with mutation matrix A_0, let $T(A_0)$ denote the total number of mutations carried by all patients and $\Gamma(A_0)$ denote the number of patients with at least one mutation. The "weight" statistic is defined as:

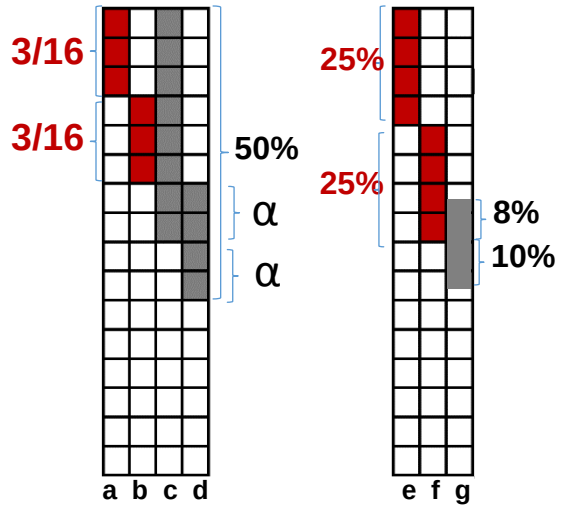
$$W(A_0)=\Gamma(A_0)-(T(A_0)-\Gamma(A_0))=2\Gamma(A_0)-T(A_0).$$

These algorithms rank and identify MEGS by maximizing $W(A_0)$ over all possible subsets. However, this criterion is neither appropriate for comparing putative MEGS with the same size nor different sizes.

In the left panel with four genes, genes *a* and *b* are mutually exclusive. Each gene is mutated in 3/16 patients. $\Gamma(a,b)=3/16+3/16=3/8$ and $T(a,b)=3/8$, thus $W(a,b)=3/8$. Genes *c* and *d* are randomly mutated. Gene *c* covers 50% patients and gene *d* covers 2α patients. In this case, it is easy to verify that $W(c,d)=1/2$ for any value of α when sample size is infinite. Thus, the weight statistic will choose *(c,d)* but not *(a,b)*. Note that, the current calculation is based on proportion assuming infinite sample size, suggesting that increasing sample size will not solve the problem.

In the right panel, genes *e* (mutated in 1/4 proportion of patients) and *f* (mutated in 1/4 patients) are mutually exclusive but *g* (mutated in β proportion of patients) is randomly mutated. Thus $W(e, f) = 0.5$. Because g is randomly mutated, so it has 50% probability to be less overlap with (*e,f*), which gives $W(e, f, g) > 0.5$, i.e. the probability of choosing (*e, f, g*) is 50%, even when sample size goes to infinity.

In summary, the weight statistic is not suitable as a criterion for choosing MEGS, even when sample size is infinite. However, when the coverage of MEGS is higher than 50%, the performance of the weight statistic increases. But even in this case, it may include many false positive genes according to our simulations. Moreover, Dendrix only performs permutations for evaluating the nominal significance of the selected MEGS without correcting for multiple testing.



3. Mutex

Mutex has two innovations: a new metric for measuring MEGS and a search strategy for search genes with common downstream targets. Mutex also uses permutations to derive null distribution and thus it has correct type-I error rates. Here, we briefly describe their basic idea of defining mutual exclusivity, although they used complicated permutations. Consider a set of *K* genes. For each gene *k*, Mutex merges the mutations of the other *K-1* genes as a "super-gene" and tests the mutual exclusivity between gene *k* and the super-gene to derive P_k. The metric for testing the set of *M* gene is defined as the weakest signal, i.e. maxP_k. Using the weakest signal for each set of genes likely excludes false positive genes. Although this statistic is intuitively attractive, it has low accuracy, particularly for imbalanced MEGS, as we show in simulations. In addition, our MEGSA can be adapted to search for MEGS using their databases for common downstream targets.

4. LRT-SB

Szczurek and Beerenwinkel (2014) derived a likelihood ratio statistic based on a data generative model different from us. To discuss, we assume that there is no measurement error in the mutation data. In their data generative model, they assumed that background mutations could only happen for patients who had MEGS mutations. Thus, for a patient with one mutation in the *m* genes in consideration, this mutation must not be background mutation. This is apparently not reasonable.

This assumption leads to a likelihood function $L(\gamma, \pi; A_0)$:

$$\log L(\gamma,\pi;A_0) = q_0\log(1-\gamma) + \sum_{k=1}^{m} q_k(\log(\gamma)+\log(k/m)+(k-1)\log(\pi)+(m-k)\log(1-\pi)),$$

where γ is the coverage of MEGS, π is the background mutation rate (for all genes), A_0 is observed mutation matrix, *m* is the number of genes in the MEGS and q_k is the number of rows in A_0 that have k mutations ($k = 0,1,...,m$).

Here, $\gamma = 0$ corresponds to the null hypothesis that these genes are not in MEGS. However, this likelihood degrades when $\gamma = 0$. Thus, they cannot define a likelihood ratio test in a standard way. To solve this problem, they derived a likelihood function under the null hypothesis, denoted as $L_0(\Pi;A_0)$:

$$\log L_0(\Pi;A_0) = \sum_{g=1}^{m}\left(k_g\log(\pi_g)+(m-k_g)\log(1-\pi_g)\right),$$

where $\Pi = (\pi_1,...,\pi_m)$, π_g is the background mutation rates of gene g, k_g is the number of samples that have mutations in gene g, $g = 1,...,m$.

Then, they derived the likelihood ratio statistic as

$$\text{LRT} = 2[\log L(\hat{\gamma}_1,\hat{\pi}_1;A_0) - \log L_0(\hat{\Pi}_0;A_0)],$$

where $\hat{\gamma}_1$ and $\hat{\pi}_1$ are estimated under H_1 and $\hat{\Pi}_0$ is estimated under H_0. Because the two models are not nested, LRT does not follow χ_1^2 asymptotically. The authors applied the Vuong's method to derive a statistic (referred as LRT-SB)

$$\text{LRT-SB} = \frac{1}{2\sqrt{n\sigma}}\left(LRT - \log(n)(2-m)\right).$$

Here, n is the number of patients and m is the number of genes in consideration. The authors claimed that LRT-SB followed N(0,1) under H_0 and calculated the *P*-value as 1-Φ(LRT-SB). However, they interpreted the Vuong's method incorrectly.

The Vuong's method was originally developed for comparing two models that were not nested. The "null hypothesis" for Vuong's statistic is that the two models (H_0, H_1 in this case) are equally likely given the data (the mutation matrix in this case). If data are generated under H_0 (null hypothesis for ME), LRT-SB does not follow N(0,1), as is shown in **Supplementary Figure S2**.

Moreover, the authors assume equal background mutation rate across genes. Thus it is expected to have low power for imbalanced MEGS even if the null distribution can be fixed.