Supplementary Materials

CEA: Combination-based gene set functional enrichment analysis

Duanchen Sun[1,2], Yin-Liang Liu[1,2], Xiang-Sun Zhang[1], Ling-Yun Wu[1,2,*]

[1]IAM, MADIS, NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[2]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

1. The performance of CEA using different d and T

In the CEA method, we introduced a randomization parameter d, which could help the algorithm to escape the local minimum. Larger d will increase the variance of solutions identified by the algorithm. That is, the probabilities to find better solutions as well as worse solutions are both increased. Therefore, the algorithm need repeat sufficient times in order to find better solutions. Larger d often requires more repeat times of algorithm. In this section, we conducted a simulated experiment to explore the effects of parameters d and T to the final result of CEA.

The simulated datasets were extracted from the biological process (BP) domain of GO. We simulated the active gene lists using a more appropriate approach (see below). Our simulation was based on the biological assumption that the active gene list derived from a specific biological experiment usually has close relationship with several biological processes.

The original annotation matrix of BP domain, derived from the Bioconductor R package *org.Hs.eg.db*, contains 14614 genes and 13226 terms. We first filtered the terms and kept the terms that annotate 50 to 100 genes as candidate terms. Namely, too general or specific terms were filtered out. This preprocessing could avoid the final results have a large variance and reduce the total computation time.

In the experiment, the following values of parameters d and T were considered:

$$d = \{0, 0.01, 0.1, 1, 10\},$$

$$T = \{1, 10, 50, 100, 200, 500\},$$

The detailed procedure of our exploration is as follows:

1) Randomly select one term from the candidate terms into the current term combination, until the number of annotated genes is no less than 200.
2) Randomly select 100 annotated genes from the genes annotated by the current term combination as the active gene list.
3) For each d and T, execute the CEA algorithm to compute the enriched term combinations using the given active gene list.
4) Sort the identified term combinations based on the p-values of the Fisher's exact test.
5) Record the mean value of $-log_{10}(p)$ of the top 30 enriched term combinations.
6) Repeat the above procedure for 100 times to achieve a robust result.

The final results were shown in Figure S1.

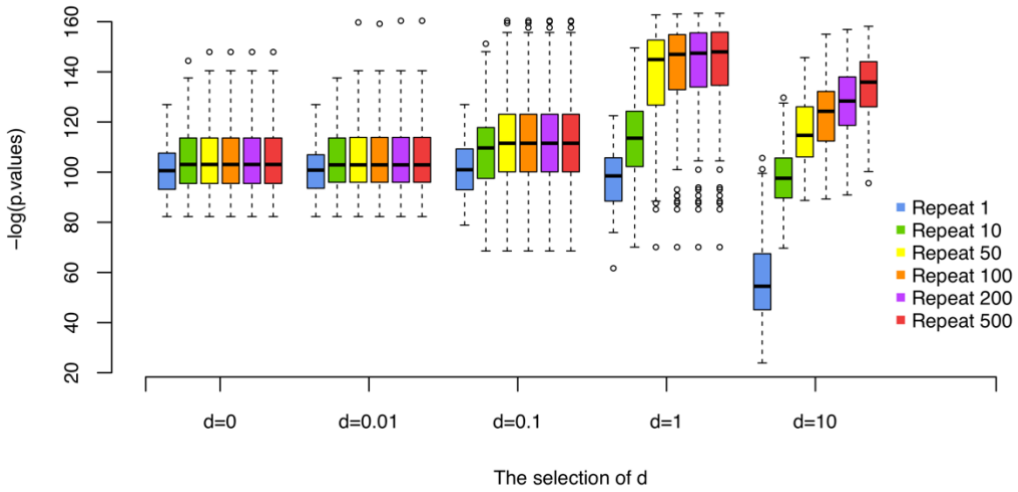

Figure S1. The performance of CEA using different d and T. For each d, a group of boxplots for each repeat times T was plotted. The performance was evaluated by the negative logarithmic transformation of p-values.

As expected, the results clearly showed that the performance of CEA can be significantly improved by introducing the randomization parameter d and CEA needs more repeat times to achieve a desired performance when d increasing.

Generally, for a fixed d, the performance of CEA will be improved if more repeat times is executed. The performance of CEA would be very poor when the repeat times is insufficient (e.g. $d = 10$ and $T \leq 10$). However, the performance cannot be improved infinitely by increasing the repeat times T. For enough large T, the marginal improvement becomes very small. Therefore, the users should balance the trade-off between the performance and the computation time.

We can roughly estimate an appropriate repeat times $\tilde{T}$ for a given d from the above results. It seems that $\tilde{T} = 500d$ would be enough for good performance. In this paper, we selected $d = 1$ and $T = 500$ as the default values to execute the CEA algorithm.

2. The preprocessing of gene expression datasets

In this work, we used real gene expression datasets of human complex diseases to test whether the term combinations identified by CEA are meaningful and closely related to the corresponding disease. The selection of gene expression datasets is based on the following criteria:

1) *Homo sapiens* organism disease;
2) Published (submission date) in recent ten years;
3) A balanced number of case and control samples and the total number is at least 50;

According to these criteria, four gene expression microarray datasets of human complex diseases were selected from the Gene Expression Omnibus repository [1] (http://www.ncbi.nlm.nih.gov/geo/), with accession number GSE4115, GSE11223, GSE9750, GSE36895, respectively, for real datasets analyses.

As for lung cancer dataset (GSE4115 [2]), we combined the original primary and prospective datasets, which made a total of 97 and 90 smokers with and without lung cancer, respectively. For ulcerative colitis dataset (GSE11223 [3]), we only used the uninflamed samples in each cohort, which made a total of 66 ulcerative colitis patients and 69 healthy control donors. All samples of the cervical carcinogenesis dataset (GSE9750 [4]) were kept. As for renal cell carcinoma (GSE36895 [5]), the paired expression profiles of 23 clear-cell RCC patients and their related normal cortex were used for further analysis.

For all expression datasets, we averaged the expression values of the probesets mapping on the same gene. The summaries of the preprocessed datasets are shown in Table S1.

Table S1: The summary of gene expression datasets used in our work.

Dataset	accession number	#disease	#normal	#genes
lung cancer	GSE4115	97	90	12493
ulcerative colitis	GSE11223	66	69	10506
cervical carcinogenesis	GSE9750	33	24	12494
renal cell carcinoma	GSE36895	23	23	20108

3. Active gene lists used in the real data analysis

For each microarray dataset, we generated a representative active gene list after preprocessing the original dataset. The detailed procedure of generating the active gene list is introduced in the main text.

The active gene lists used as the input of each enrichment analysis method are listed as follows:

Lung cancer (81):

SLC5A1 PRUNE ATP8B1 NSUN3 HDGFRP3 STK38 AGPS TRIM36 DCLRE1C BTD RPL35A SOX9 DND1 C6 TSR1 NNT ZNF160 TFE3 HTRA1 ADH6 PDE8B ZNF611 U2AF2 ECD TMEM110 GOSR2 GTF2H3 SUGP2 MOCS2 PPP2R2D RPL18 P2RX4 NEDD9 SLC4A4 ADK PGF CRY1 EXT2 NOTCH2NL EIF2B3 CORO2A FGF14 DMD DLAT DIP2A USP46 HAUS2 ALPK1 MAN1A2 PPM1D CEP57 DAPP1 PRDX2 NPFFR1 STX3 LAT FBXO9 WWC3 TGDS ARID5A UBQLN4 GNPDA1 RHOQ TNFRSF1A CPE ODF2 PYGB FUT8 ZFR NUDT4 TXN DNAJC6 MTPAP RRAGB ABHD17B IL13RA1 MSH6 MYO1C UNC93B1 MFSD11 KDELR3

Ulcerative colitis (56):

PLCB3 ELL MAPKAPK2 DOCK7 DOHH STK25 TBXA2R INPPL1 C6orf120 APOC1 CEP290 STK35 LARP1 GTF2H5 PPP1R14B SBF1 DIRC2 BRD4 AXIN1 INSR SKIV2L PRCP B3GALT5 TAF12 VPS52 RPS29 ZNF304 C14orf2 ITGA3 GAS6 ARF6 SPSB1 USP54 SLC2A8 GCA CCL11 SERPINF1 FBXL12 TBC1D2B MAN2A1 HIST1H2BN GNB2 ACYP2 ARAF BLVRA HOMER3 PUS1 ACSM1 ADAL C3orf33 GBE1 COMP OXSR1 MVD MLXIP DDX6

Cervical carcinogenesis (94):

PITPNA ZDHHC3 GJA1 SYNGR1 KCTD15 ESR1 AHNAK TRPS1 CDKN2A KANK1 KRT13 KIF18B SYPL1 NAGK MCM6 LMBRD1 UBE2E1 CHMP2B SPRR3 USO1 GINS2 RPL10A NEK2 MCM2 ZNF586 DNMT1 POLD1 RAD54L GOLGA4 CRYL1 GINS1 RPS12 SKP1 SLC24A3 UBE2C MAP2K4 CHAF1B PLCD1 KNTC1 PRDM2 MCM5 ZNF415 TK1 KIF4A KIF2C AURKA CAPN7 TP53AIP1 CCNF LPAR6 SNX3 RPS6KA1 ATP6V1F LAPTM4A PPP2R5A ITM2B DUSP1 NUP62 ATP13A2 RPL29 ATP10D CENPF USP46 LIG1 ARHGAP10 STX7 BBOX1 KLF4 CLCA4 SPAG5 TMEM9B DSC2 RYR1 LANCL1 SYNGR3 AVPR1B TPX2 PSMC3IP SASH1 MAPK10 CDC20 CDT1 CDC45 GIGYF2 TRIM13 TIMELESS GALR3 SLC15A3 IL17RC CDC6 CLCN3 RALB DTL PERP

Renal cell carcinoma (85):

NPHS2 SPAG4 UMOD SFRP1 FGF1 SLC12A1 EGLN3 IGFBP3 ATP6V0D2 HK2 CALB1 GGT6 CWH43 CLDN8 HILPDA HEPACAM2 LPPR1 ATP6V0A4 ACSF2 ANGPTL4 SCNN1G PTH1R CLIC5 FAM3B CLCNKB ENO2 SLIT2 PPAPDC1A PRKCDBP FUT11 CRHBP TMPRSS2 PLCXD3 SAP30 SLC47A2 PTGDS HS6ST2 FXYD4 ATP6V1G3 TYRP1 TCEAL2 TNNC1 DMRT2 CNTN1 HPD SER INA5 KNG1 GPD1L STAP1 C5 CAV1 PDK1 PTPRO RASL11B SLC26A7 GAS1 CAV2 TFAP2B LDHA NPHS1 TCF21 DDB2 SLC2A12 PACRG KCNJ10 DIO1 DACH1 ARHGEF26 GPC3 BMPR1B SEC61G NRK ALDOA VEGFA MUC15 EIF4H CA10 MAN1C1 COL4A6 SOSTDC1 SOST ATP6V1C2 ATP6V1B1 ANGPTL1 FABP5

Reference

1. Barrett T, Edgar R: **Gene expression omnibus: microarray data storage, submission, retrieval, and analysis**. *Methods Enzymol* 2006, **411**:352-369.
2. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, Gilman S, Dumas YM, Calner P, Sebastiani P *et al*: **Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer**. *Nat Med* 2007, **13**(3):361-366.
3. Noble CL, Abbas AR, Cornelius J, Lees CW, Ho GT, Toy K, Modrusan Z, Pal N, Zhong F, Chalasani S *et al*: **Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis**. *Gut* 2008, **57**(10):1398-1405.
4. Scotto L, Narayan G, Nandula SV, Arias-Pulido H, Subramaniyam S, Schneider A, Kaufmann AM, Wright JD, Pothuri B, Mansukhani M *et al*: **Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression**. *Genes Chromosomes Cancer* 2008, **47**(9):755-765.
5. Pena-Llopis S, Vega-Rubin-de-Celis S, Liao A, Leng N, Pavia-Jimenez A, Wang S, Yamasaki T, Zhrebker L, Sivanand S, Spence P *et al*: **BAP1 loss defines a new class of renal cell carcinoma**. *Nat Genet* 2012, **44**(7):751-759.