Supplementary manuscript of

Multi-modal optimization to identify personalized biomarkers for disease prediction of individual patients with cancer

Jing Liang1, Zong-Wei Li1, Cai-Tong Yue 1, Zhuo Hu1, Han Cheng2, Ze-Xian Liu3, Wei-Feng Guo1,3\*

1 School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China

2 School of Life Sciences, Zhengzhou University, Zhengzhou 450001, China

3 State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou 510060, China

\* To whom correspondence should be addressed.

[Tel:86-0371-677810](Tel:86-21-54920100)18; Fax: [86-0371-677810](Tel:86-21-54920100)18; Email: [guowf@zzu.edu.cn](mailto:liangjing@zzu.edu.cn,)

## Part A: The analysis of MMPDNB for identifying cancer tissue special driver and biomarker genes.

According to the Cancer Gene Census, there are respectively 23 ,18 genes are annotated as driver genes and 74 ,55 genes are annotated as biomarker genes of BRCA and LUNG cancer tissue. For identifying driver genes, MMPDNB found two breast cancer driver genes and fifteen lung cancer driver genes. For identifying biomarker genes, MMPDNB found six type breast cancer biomarker gene and six type lung cancer biomarker genes by analyzing all PDNBs. These well-estimated cancer tissue specific driver and biomarker genes and those genes identified by MMPDNB were provided in the **Tables S1-S2**.

### The calculation of F-score for different algorithms from two aspects

Due to multi-modal EAs could provide multiple PDNBs while DNB algorithms and network controllability algorithms only provide one PDNB, thus we calculated the F-score of PDNB for identifying driver and biomarker genes from two aspects. One the one hand, we calculated the average F-score value of PDNB, which has F-score and the largest warning score among the all PDNBs of individual patient. On the other hand, we calculated the average F-score value of PDNBs, which has F-score among the all PDNBs of individual patient.

As show in the **Fig. S1**, we found the results from the two aspects have almost the same values. It should point out that DNB algorithms and network controllability algorithms only provide one PDNB, thus their F-score about two aspects have same value and the different between two aspects is among multi-modal EAs. Therefore, we showed that their distribution of F-score for the PDNBs detected by MMPDNB and distribution of *P*-value for those F-score. From the **Fig. S2-S3,** we can find that the distribution of F-score and *P*-value from two aspects are close on BRCA, LUSC and LUAD. Therefore, the average F-score value of PDNB with the maximum score and the average F-score value of the set of Pareto optimal PDNBs have almost the same values, thus we only compared all PDNBs provided by each algorithm to identify driver and biomarker genes in the submitted manuscript.

### The analysis of F-score for PDNB identified by MMPDNB

For the F-scores of PDNB in identifying driver and biomarker genes are around 0.07 and 0.03, we calculated the average precision, average recall and average F-score of all PDNB modules by considering cancer tissue specific driver genes and biomarkers as gold standards (**Figs.S4-S5**). From the results of **Figs. S4-S5**, it can be seen that for the four multi-modal EAs, the average recall of the multi-modal evolutionary algorithm is smaller than DNB algorithms and network controllability algorithms, but the precision is higher. According to the analysis, we found that it was caused by a smaller recall.

Specifically, for MMPDNB and MP-MMEA, due to the objective function 2, the number of genes for PDNB generally is small, thus the probability of identifying the correct driver and biomarker gene is also small. Therefore, MMPDNB and MP-MMEA have higher precision but lower recall. For MO\_Ring\_PSO\_SCD and DN-NSGA-II, their optimization ability is not enough for this problem, so they cannot find PDNBs with small number of genes but large early warning score. In addition，PDNBs identified by DNB algorithms and network controllability algorithms have larger gene numbers, which leads to driver and biomarker genes are selected with the higher probability. But in general, MMPDNB and MP-MMEA which have optimization ability, perform better in identifying driver and biomarker genes (**Figs. S4C-S5C)**.

### The reason why MMPDNB can generate the biggest F-score in some cancer

According to the **Figs.S4C-S5C**, we found that MMPDNB cannot generate the biggest F-score for identifying driver and biomarker genes in some cancer. The reason why MMPDNB cannot generate the largest F-score in some cancers is the number of genes in most PDNBs provided by MP-MMEA and DFVS was smaller than MMPDNB’s (i.e., the number of genes in BRCA patients’ PDNB identified by DFVS usually are 3-5, while the number of driver gene are same as MMPDNB). Thus, their precisions are higher than MMPDNB’s under the same recall. However, due to the number of genes was small, the genes in those PDNB have no interaction and the warning scores of those PDNBs do not exist. Therefore, MP-MMEA and DFVS cannot predict the early warning signal for some cancer individual patients. This result was also consistent with **Fig. 2B** in themanuscript.

If we deleted those PDNB without early warning score, the results are shown in **Fig. S6-S7**. We can conclude that: the PDNBs identified by MMPDNB has the highest F-score for driver and biomarker genes almost among all algorithms.



**Fig. S1 Performance comparisons of MMPDNB and other methods for discovering cancer tissue specific driver genes.** The mean value of patients with F-Score value was used to evaluate the algorithm's performance on in identifying cancer driver genes**. (A)** Comparisons result in terms of F-score of the PDNB with maximum early warning signal score among the set of Pareto optimal PDNBs. **(B)** Comparisons result in terms of F-score of all multiple PDNBs among the set of Pareto optimal PDNBs.



**Fig. S2** **The box-plot of F-score about PDNB with the maximum score and the set of Pareto optimal PDNBs identified by MMPDNB.** (**A**) The distribution of F-score of driver genes in PDNB from two aspects. (**B**) The distribution of F-score of biomarker genes in PDNB from two aspects.



**Fig. S3** **The box-plot of *P*-score about PDNB with the maximum score and the set of Pareto optimal PDNBs identified by MMPDNB.** (**A**)The statistical significance of PDNB to identify driver gene from two aspects. (**B**) The statistical significance of PDNB to identify biomarker gene from two aspects. The red dotted line denotes that the significant threshold value is –log10(0.05). If the value is larger than this threshold value, we think that the enrichment result is significant.



**Fig. S4 (A-C)** The average precision, recall, and F-score of different algorithms for discovering cancer tissue specific driver genes.



**Fig. S5 (A-C)** The average precision, recall, and F-score of different algorithms for discovering cancer tissue specific biomarker genes.



**Fig. S6** Performance comparisons of MMPDNB and other methods for discovering specific driver genes after deleted PDNBs without early warning score.

****

**Fig. S7** Performance comparisons of MMPDNB and other methods for discovering biomarker genes after deleted PDNBs without early warning score.

**Table S1. Breast and lung cancer driver genes and Driver genes identified by MMPDNB**

|  |  |  |
| --- | --- | --- |
|  | **Breast cancer** | **Lung cancer** |
| **Driver genes** | |  | | --- | | AKT1\ARID1A | | ARID1B\BAP1 | | BRCA2\ CASP8 | | CDH1\CDKN1B | | CDKN2A\ CTCF | | ERBB2\ ESR1 | | FOXA1\ GATA3 | | MAP2K4\ MAP3K1 | | NCOR1\ PIK3CA | | RB1\ SALL4 | | SMARCD1\ TBX3 | | TP53 | | |  | | --- | | BRAF\ CDKN2A | | EGFR\ KDR | | KEAP1\ KRAS | | LRIG3\ MAP2K1 | | RBM10\ RET | | SMARCA4\ STK11 | | TP53\ DROSHA | | FGFR2\ NFE2L2 | | NOTCH1\ RAD21 | |
| **Driver genes contained in all PDNBs** | AKT1、ARID1B | SMARCA4、KDR  TP53、RET  CDKN2A、FGFR2  MAP2K1、NFE2L2  NOTCH1、RBM10  STK11、BRAF  CDKN2A、EGFR  KRAS、 |

**Table S2. Breast and lung cancer biomarker genes and Biomarker genes identified by MMPDNB**

|  |  |  |
| --- | --- | --- |
|  | **Breast cancer** | **Lung cancer** |
| **Biomarker**  **genes** | |  | | --- | | CD24\ALDH1\P63\CD44\CD133\ALDH1A1 | | ALDH\ALDH1A3\EPCAM\ESAGD2\GLI1 | | LGR5\ SOX2\CD117\CD13\CK18\CD105 | | CD29\CD56\CD227\CD49F\33A10\LY6D | | CK14\ACTA2\ANKRD30A\KRT19\CSN2\CITED1 | | GNG11\CD10\CD15\PROCR\LSD1\CD166 | | EMA\SMA\CD14\MUC\CD73\K19 | | CD61\CK17\CD34\KRT14\SYTL2\PROM1 | | ETV5\GLYCAM1\ESR1\IGFBP4\VEGF\GCDFP15 | | CD209\CD90\CK5\CK19\KRT5\SLPI | | LALBA\PGR\KRT4\CD45\MYLK\WAP | | PRLR\OXTR\ZEB2\TP63\PDPN\S100A6  CD83\CD86 | | |  | | --- | | CD133\ALDH1A1\ABCB1\ALDH\ALDH1\Nestin | | CD31\CD68\CD25\CK19\CD24\MET | | CD146\Endocan\ALDH-1\CD44\GLI1\ABCG2 | | ABCG5\c-Myc\CD166\CD44v\CD163\CD66b | | CD3\CD1a\Bmi1\CC10\CD117\ CD133 | | ALDH1A1\ Beta-catenin\ CD44\ FOXP3 CD31\ CD44 variant6 | | P63\ ALDH1\ ALDHA1\ CD117\ CD83\ c-Myc | | CD90\ CD166\ VWF\ Podoplanin\ BMI1\ KLF4 | | EpCAM\ CD34\ NANOG\ Oct4\ ESA\ SOX2\CD24 | |
| **Biomarker**  **genes contained in all PDNBs** | ALDH1A3\CD44  KRT19\MYLK  SLPI\SYTL2 | ABCB1\ABCG2  GLI1\MET  CD44\GLI1  MET |

## Part B: The biologic significance of multi-modal PDNB for LUSC and LUAD patients.

For LUSC patient samples, there are multi-modal PDNBs of twelve patient samples contained (41% of early patient sample with multi-modal PDNBs) drug targets for Lung Squamous Cell Carcinoma. The differential genes of multi-modal PDNBs belonging to the early stage of LUSC include 14 drug target genes, as shown in **Table S3**. Drug target genes include but not limited to PMS2, MAPK11 and RHOA. Mutations in PMS2 have been associated with hereditary nonpolyposis colorectal cancer and cancer caused by mutate in PMS2 is also distribute in the lungs. Gene MAPK11 encodes a protein that involved in the integration of biochemical signals for a wide variety of cellular processes, including cell proliferation, differentiation, transcriptional regulation, and development. Overexpression of gene RHOA is associated with tumor cell proliferation and metastasis. According to the drugs-gens networks, drug Fluorouracil acting on above three drug target genes has shown sensitivity in the clinical responsion of LUSC.

For LUAD patient samples, there are multi-modal PDNBs of sixteen patient samples (70% of early patient sample with multi-modal PDNB) contained drug targets for Lung Adenocarcinoma. The differential genes of multi-modal PDNBs belonging to the early stage of LUAD include 18 drug target genes, as shown in **Table S4**. Drug target genes include but not limited to FGF5, FGF4 and MAX. Gene FGF5 is identified as oncogene. Proteins encoded by FGF5 are involved in biological processes, including cell growth and tumor growth. Drugs Dasatinib and PD-0325901 can show sensitivity in the clinical responsion of LUAD by acting on FGF5. The function of proteins encoded by FGF4 is similar to FGF5. But, gene FGF4 and FGF3, another oncogenic growth factor, are located closely on chromosome 11. Co-amplification of both genes was found in various kinds of human tumors. Drugs Sorafenib and PD-0325901 can show sensitivity in the clinical responsion of LUAD by acting on FGF4. Mutations of gene MAX have been reported to be associated with hereditary pheochromocytoma. Proteins encoded by MAX are able to form homodimers and heterodimers with other family members, which include Mad, Mxi1 and Myc. Myc is an oncoprotein implicated in cell proliferation, differentiation and apoptosis. Drug Gefitinib and Thapsigargin acting on target genes MAX has shown sensitivity in the clinical responsion of LUAD.

To further verify the effectiveness of target genes and corresponding drugs, we queried the drug response datasets (GDSC). For LUAD data, we found that one target genes (e.g., SIRT1) and one type drugs (e.g., Selisistat) are effective in the drug response dataset for various LUAD cell lines. **Fig. S8A** shows that the sensitivity of the drug Selisistat, which acts on the drug target SIRT1, is significantly correlated with STK11 mutation cell line in LUAD cancer tissues. **Fig. S8B** showed that LUAD cancer cells with STK11 mutation was significantly inhibited by Selisistat compared with wild type cell line.

For survive analysis, none of drug targeted genes of LUSC patients cannot serve as survival risk markers. However, we found that there are two drug targeted genes in 18 drug targeted genes of LUAD which can actually divide all patients into discriminative high-risk and low-risk groups (**Fig. S9**). Therefore, those two genes not only can serve as drug target but also can serve as survival risk biomarker.



**Fig. S8** The sensitivity of drugs acting on drug targets of LUAD. (**A**) The volcano plot of drugs acting on drug targets SIRT1. (**B**) The box-plots of IC50 on specific genomic changes cell line and wild type cell line.



**Fig. S9.** The survival analysis by drug targeted genes on TCGA LUAD data.

**Table S3.** Drug targeted genes and effective drugs obtained by MMPDNB in the early stages of LUSC cancer.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cancer** | **Cancer Stage** | **Sample with MMDNB** | **Differential Genes in MMDNB** | **Drug Target Gene** | **Drug (iGMDR)** | **Drug (GDSC)** |
| **LUSC** | Stage IA | TCGA-85-7710 PDNB | UPF2\SMG1 | SMG1 | PIK-93 |  |
| TCGA-43-7657 | MMP1\CMA1 | CMA1 | Fluorouracil |  |
| TCGA-43-7658 | GRHL3\CUL5 | CUL5 | Fluorouracil |  |
| Stage IB | TCGA-60-2709 | RGS6\GTPBP4\NF2\STMN2 | NF2 | Afatinib |  |
| TCGA-22-5471 | PMS2\MLH3 | PMS2 | Fluorouracil |  |
| TCGA-43-6771 | SIRT1\MEF2D | SIRT1 | CX-5461 |  |
| TCGA-22-5482 | CDC42\IQGAP3 | IQGAP3 | Dasatinib |  |
| TCGA-22-5472 | CTPS2\COPS2\COPS3\WNK4\SLC12A3\GMPS\- | WNK4 | Fluorouracil |  |
| Stage IIA | TCGA-56-7823 | MAPK11\DUSP6 | MAPK11\DUSP6 | Fluorouracil |  |
| TCGA-56-7730 | ABCG2\HIF1A\NOD1\NOD2 | ABCG2 | Fluorouracil |  |
| TCGA-22-5483 | CHN1\RHOA | RHOA | Fluorouracil |  |
| TCGA-92-7340 | HLA-DOB\HLA-DQA2\PLXNB3\MICAL1\- | PLXNB3\MICAL1 | Dasatinib、 Crizotinib | Selisistat |

**Table S4.** Drug targets and effective drugs obtained by MMPDNB in the early stages of LUAD cancer

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cancer** | **Cancer Stage** | **Sample with MMDNB** | **Differential Genes in MMDNB** | **Drug Target Gene** | **Drug (iGMDR)** | **Drug (GDSC)** |
| **LUAD** | Stage IA | TCGA-44-2661 | PRPS1L1\PRPSAP2 | PRPS1L1 | Tipifarnib、Dasatinib、BLEOMYCIN |  |
| TCGA-50-5935 | FGF18\FGF5 | FGF5 | Mitomycin C、Dasatinib、  PD-0325901 |  |
| TCGA-44-6145 | CYP4A11\SIRPG\PPARA\CD47 | CYP4A11 | Bortezomib |  |
| TCGA-44-2655 | NEDD4L\SCN5A | NEDD4L | GSK-650394 |  |
| TCGA-91-6835 | NUDT21\CAPRIN1 | CAPRIN1 | Bicalutamide |  |
| TCGA-55-6980 | PLRG1\CDC40\HSPA6\HSPA4 | HSPA6 | Bortezomib |  |
| TCGA-44-6778 | TH\PSMA8 | PSMA8 | Bortezomib |  |
| TCGA-91-6828 | SOX2\FGF4 | FGF4 | Sorafenib、  PD-0325901 |  |
| Stage IB | TCGA-91-6847 | P4HB\MTTP | P4HB | Thapsigargin、Dacomitinib |  |
| TCGA-91-6836 | ARNTL2\SIM1 | ARNTL2 | Dasatinib |  |
| TCGA-91-6831 | F2RL1\ST14 | F2RL1 | Bortezomib |  |
| TCGA-55-6985 | SNAI1\MAX | SNAI1\MAX | Gefitinib、  Thapsigargin |  |
| TCGA-44-6777 | PRPH\LMNB1\RPL22\RPL11 | RPL11 | Neratinib |  |
| TCGA-55-6986 | RPL4\RPL37 | RPL37 | Thapsigargin |  |
| TCGA-44-2668 | VPS13B\EXOC2 | VPS13B | SB-216763 |  |
| TCGA-44-2662 | KCNN4\KCNN2\GRWD1\RFWD2\ SIRT1 | KCNN4\GRWD1\ SIRT1 | CX-5461\Thapsigargin |  |

## Part C: The parameter sensitivity analysis of the compared algorithms

In the experiments, the parameters of all the compared MOEAs are tuned based on the settings suggested in their original papers. To be specific, common parameters in four evolutionary algorithms are the population size and the maximal number of function evaluations, which need to be consistent. It should note that MP-MMEA have no particular parameters to validated. The particular parameters of DN-NSGA-II and MO\_Ring\_PSO\_SCD are shown in **Table S5.**

For , , , , , and, we have done parameter sensitivity analysis to investigate the influence of each parameter to the algorithm performance. We selected respectively three patients from each cancer as subjects, and 30 independent runs were performed on each subject. For each test, the particular parameters were fluctuated around the original setting and the other parameters are fixed to the values used in the experiments.

**Figs. S10** depict the performance of multi-modal EAs with different particular parameters for detecting early warning score in three type cancer. Firstly, it can be observed that the performance of DN-NSGA-II is not very sensitive to the parameters  and , and DN-NSGA-II can obtain a relatively good performance when the parameter settings are consistent with their original paper or close to their original paper. Secondly, although the performances of MO\_Ring\_PSO\_SCD are different with different parameters in the subjects, the performance of the algorithm also obtain a relatively good performance when some parameter settings are consistent with their original paper or close to their original paper. Except that parameters *w* and *C1* have an effect on MO\_Ring\_PSO\_SCD in detecting PDNB s’ early warning scores for LUSC patients, the changes in PDNBs’ scores were not significant in other cases. It is hard to give an optimal value of parameter for different type of cancer. In this case, the parameters are set to their original paper might be a better solution because the parameters in the original paper ensure optimal performance of the algorithm in the test problem sets. Finally, we compare the pareto front of different algorithm using the optimal parameters with MMPDNB’s. As shown in **Fig. S11**, we find that MMPDNB has better results.

In addition, we chose the parameters corresponding the best results for these algorithms and compared the results of other algorithms with those of MMPDNB on three cancer datasets. **Figs. S12-14** shows that MMPDNB still has the optimal results from the three aspects of average IGD value, average HV value and the detection of early warning signals. Therefore, the parameter setting of the algorithm has no effect on the result of the comparison between algorithms and the setting of parameters in the original paper is a suitable solution for this problem.

**Table S5** The parameters of two multi-modal EAs

|  |  |  |
| --- | --- | --- |
| Algorithms | DN-NSGA-II | MO\_Ring\_PSO\_SCD |
| Parameters | The crossover probability  The mutation probability | Inertia weight  Acceleration constants and  The Ring topology size |



**Fig.** **S10** **The early warning score of other algorithms with different particular parameter settings.** It should be pointed out that the vertical axis is . (**A-B**) The average early warning score of PDNB identified by DN-NSGA-II with different parameters  and . (**C-F**) The average early warning score of PDNB identified by MO\_Ring\_PSO\_SCD with different parameters , , , and .



**Fig. S11** The Pareto front of different algorithms under optimal parameter setting.



**Fig. S12 The comparison results of other algorithms using optimal parameters and MMPDNB on IGD values.** (**A-C**) The box-plot of IGD index of three multi-modal EAs on (**A**) BRCA, (**B**) LUSC, and (**C**) LUAD cancer datasets.



**Fig. S13** **The comparison results of other algorithms using optimal parameters and MMPDNB on HV values.** (**A-C**) The box-plot of HV index of three multi-modal EAs on (**A**) BRCA, (**B**) LUSC, and (**C**) LUAD cancer datasets.



**Fig. S14** **The comparison results of other algorithms using optimal parameters and MMPDNB for detecting the early warning signals.** **Figs. A-C** respectively show the results of MMPDNB and other multi-modal EAs for detecting the average early warning score of cancer individual patients at different stages.

## Part D: The parameter sensitivity analysis of the compared algorithms

We compared the calculation time of all methods on three cancer data. The experimental environments are MATLAB R2021a, processor: 11th Gen Intel(R) Core (TM) I5-11400, and RAM:16G. As shown in **Fig. S15,** the calculation time of MMPDNB is shorter than that of the other three multi-modal evolutionary algorithms. Although the calculation time of DNB algorithms and network controllability algorithms is generally shorter than that of MMPDNB, these methods cannot identify early warning signals of cancer, and the most important thing is that those algorithms cannot find multi-modal PDNB. As computer configurations improve, the computing time of MMPDNB will decrease. Therefore, MMPDNB can obtain better experimental results in an acceptable time.



**Fig. S15** The average calculation time of each algorithm on three cancer data.