nature research

https://doi.org/10.1038/s41596-020-00430-z

Supplementary information

OCTAD: an open workspace for virtually screening therapeutics targeting precise cancer patient groups using gene expression features

Supplementary text

Clustering. All the compounds in LINCS L1000 database (11,313 compounds with known chemical structures) were clustered into 2695 compound groups based on their chemical structural similarities. The similarity score of two compounds was calculated as the Tanimoto coefficient of their 2048-bit FCFP8 fingerprints in RDKit [1]. The Butina Algorithm in RDKit was employed to perform chemical structural clustering. As the L1000 molecule dataset is highly diverse and covers a large chemical space of both traditional drug-like molecules and novel complicated scaffolds, a simple threshold of intra-cluster distance was not competent to this task. Thus, different intra-cluster distances from 0.4 to 0.8 were used to gather the molecules at different levels. All compounds were first attempted to be clustered with a cutoff of 0.4, then those classes with only one or two members were re-clustered with a cutoff of 0.5. These steps were repeated until a cutoff of 0.8 was achieved. Finally, compounds with similar chemical structures (or substructures) were grouped together. The detailed clusters and their member compounds are listed on the web portal.

Drug Enrichment analysis. After structural clustering, the enrichment of compounds belonging to each group in the prediction was calculated using ssGSEA [2]. The p value was derived from the frequency of lower ssGSEA scores of the 1000 randomly altered sRGES rankings. A default p value of 0.05 was set for further analysis.

Purchasability fetching. The purchasability of the compounds in LINCS L1000 database were determined through searching "in stock" subset of compounds in ZINC [3]. Exact structure searching and similar structure (Tanimoto coefficient > 0.9) searching were performed.

Synthetic accessibility scoring. The synthetic accessibility of the compounds in LINCS was estimated by the SA_score in RDKit [4]. This score ranges from 1 (easy to make) to 10 (very difficult to make).

Distributions of chemical clustering, purchasability, and synthetic accessibility of L1000 compounds

In this work, 2695 compound groups were obtained after clustering. Among them, 2078 groups contained no less than 3 members. As shown in the t-SNE plot (Fig. 3a), the L1000 dataset covers a broad chemical space. The novel complicated chemotypes (green dots) were far from the traditional drug-like molecules (purple dots). There are 2161 purchasable molecules, and 742 molecules with their analogues purchasable. About half of the L1000 compounds were predicted with low or moderate synthetic difficulty.

OCTAD Feature collection

Somatic mutation, and copy number data were compiled from cBioportal (https://www.cbioportal.org/) via the R package cgdsr. For mutation data, only those genes that were altered in at least 50 samples were kept. Putative copy-number was computed from GISTIC 2.0 (Gain: score >= 1 and loss: score <= -1). Tumor subtype information was retrieved from the R package TCGAbiolinks (choosing Subtype_Selected). Other phenotypic data were collected from the treehouse project (https://treehousegenomics.soe.ucsc.edu/public-data/).

Supplementary Figures

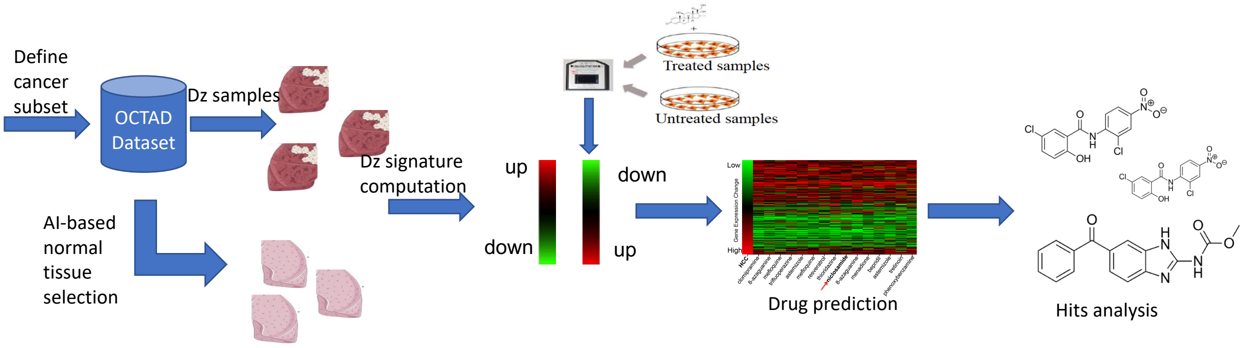

Figure S1: Illustration of the systems-based therapeutic discovery approach. AI: Artificial Intelligence; Dz: Disease.

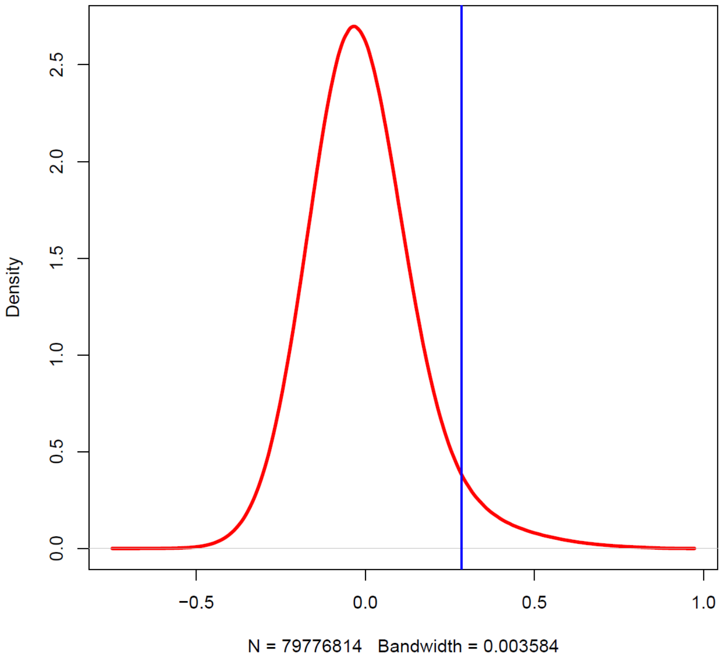

Figure S2. Distribution of correlation values between primary cancer samples and normal tissue samples in OCTAD dataset. This figure relates to finding appropriate control samples for a given set of case samples. To obtain the default correlation cutoff, we applied computeRefTissue function for every primary cancer sample in the OCTAD package and analyzed the correlation distribution. The significant correlation threshold is 0.285 (P < 0.05) based on the distribution.

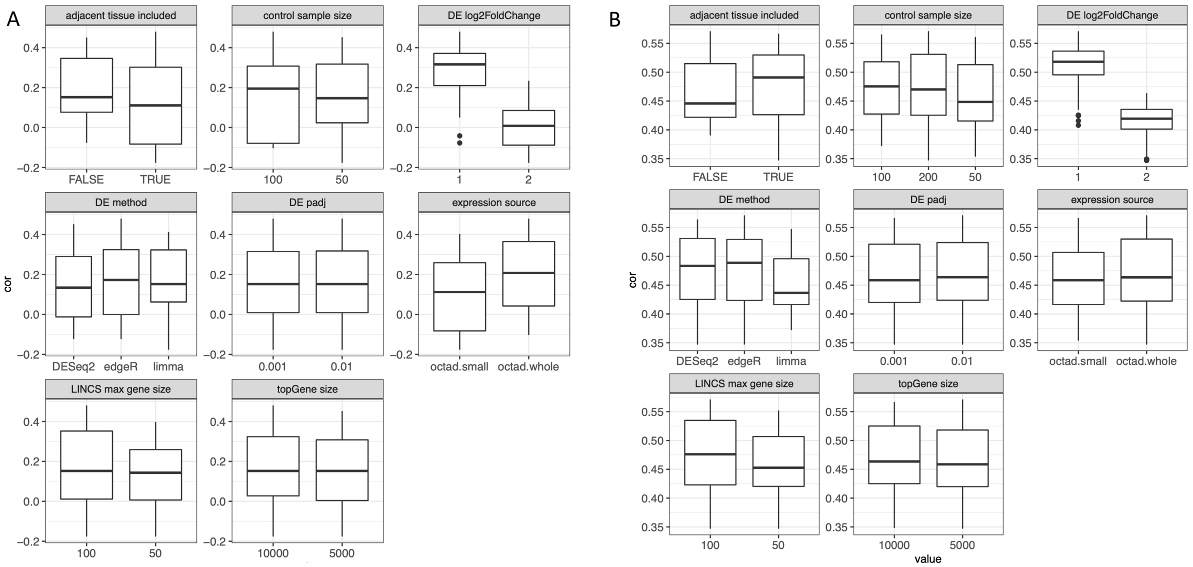

Figure S3: Correlation between sRGES and efficacy data under different parameter values in (A) colon adenocarcinoma and (B) breast invasive carcinoma. Y shows correlation values and X shows the values commonly used. The following parameters were examined: adjacent tissue included, control sample size, DE log2foldchange threshold, DE method, DE padj threshold, expression source, max gene size in LINCS prediction, and topGene Size in DE analysis. For each value, we enumerated the values of other parameters and reported the correlation for each combination. In order to reduce the computing time, we randomly chose 200 samples when the number of case samples is too large (> 1000) and we reduced the number of options for those parameters that were not significant in liver cancer prediction (Figure 7).

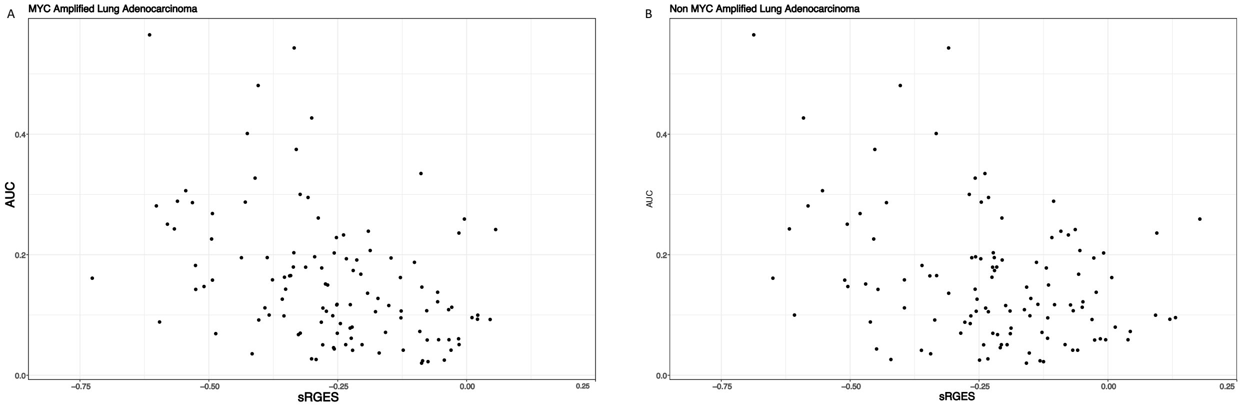

Figure S4: Correlation between sRGES and AUC in (a) MYC amplified lung adenocarcinom/a, and (b) non-MYC amplified lung adenocarcinoma. The AUC data was taken from recomputed AUC in the PharmacoGx package [5].

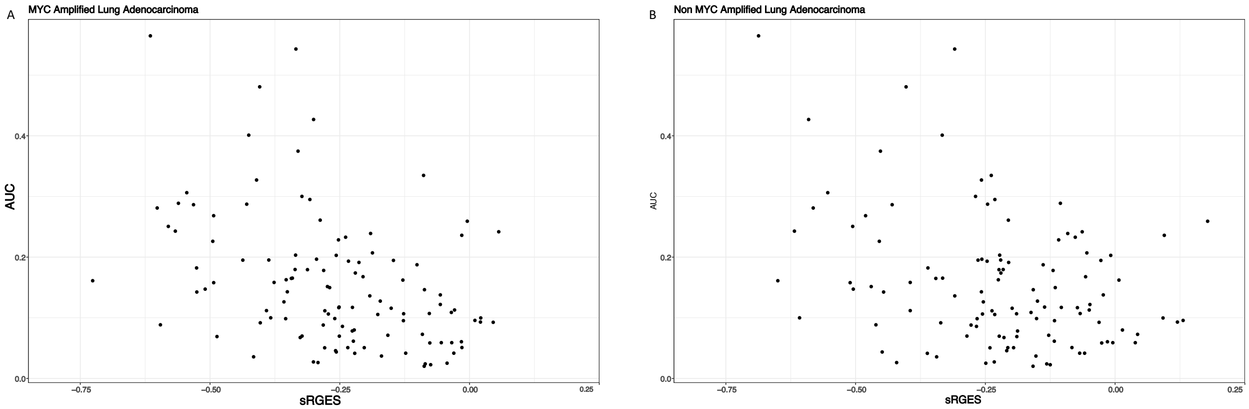

Figure S5: Correlation between sRGES and AUC in (a) PIK3CA mutated breast cancer, and (b) PIK3CA wild type breast cancer. The AUC data was taken from recomputed AUC in the PharmacoGx package [5].

References

1. RDKit. https://www.rdkit.org/.
2. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 14, 7 (2013).
3. Sterling, T. & Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J Chem Inf Model* 55, 2324–2337 (2015).
4. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminform* 1, 8 (2009).
5. Smirnov, P. *et al.* PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* 32, 1244–1246 (2016).