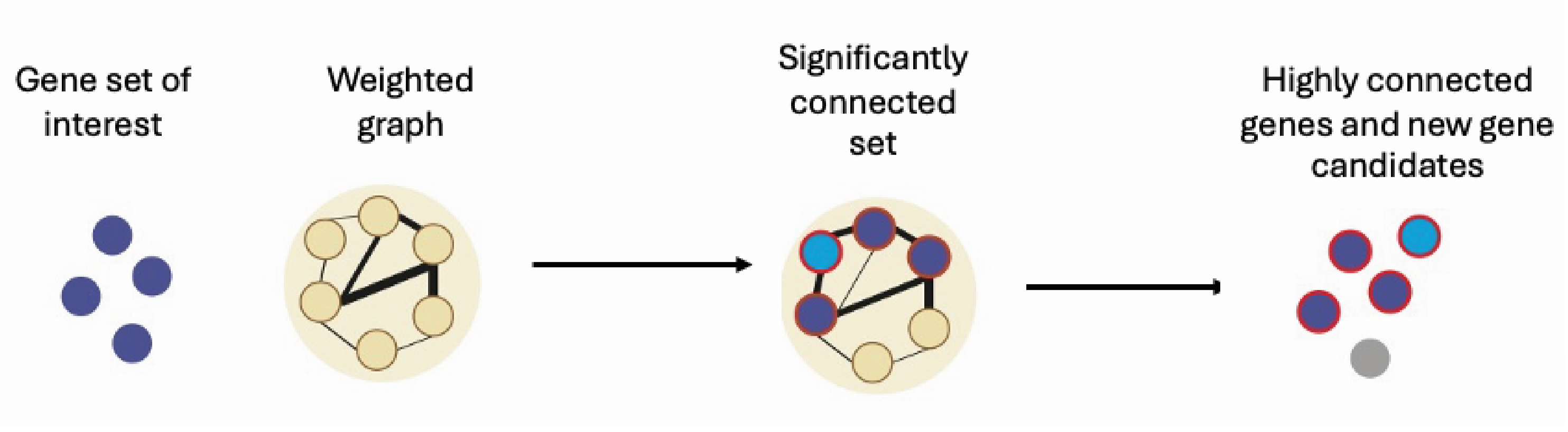
CTD2 now deployed on the Cancer Genomics Cloud to interpret cancer genes in network & pathway context (Abstract # 2330)

Varduhi Petrosyan(1), Vladimir Kovacevic(1), Cera Fisher(2), Predrag Obradovic (1), Jack DiGiovanna(2),
Brandi Davis-Dusenberry(2), Aleksandar Milosavljevic(1)
1 - Baylor College of Medicine, Houston, TX
2 - Velsera, Charlestown, MA

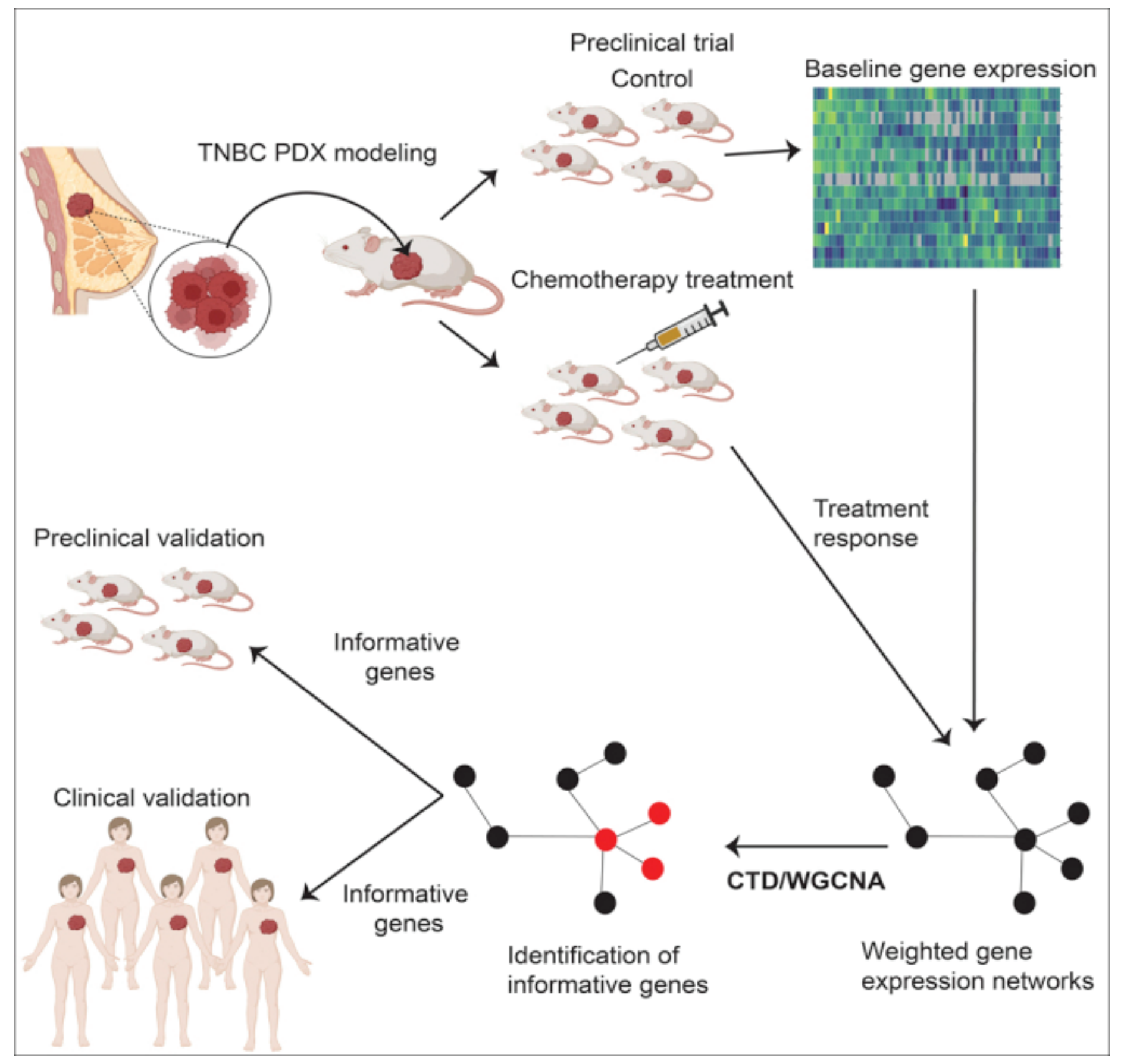
Introduction

CTD is a method that interprets sets of genes from cancer "omics" experiments in the context of biological networks. By identifying subsets of genes that are significantly connected in a specific disease context, CTD allows for the identification of biological informative gene sets.We have previously applied CTD to identify pertubations in breast cancer [1], diagnose rare metabolic disorders [2], and identify multigene biomarkers that are predictive of response in Triple Negative Breast Cancer (TNBC) [3].

Despite its utility in identifying biological signal, the computational cost of running CTD on large datasets has been a barrier to its adoption. To address this gap, we developed CTD2 which is both faster than CTD, and has the new capability to perform guilt-by-association analysis and deployed it as a tool on the Cancer Genomics Cloud (CGC)[4]. A gene implicated via guilt-by-association is illustrated by the light blue node in the graph below.


Gene set of interest
Weighted graph
Significantly connected set
Highly connected genes and new gene candidates

Biomarker discovery with CTD/CTD2



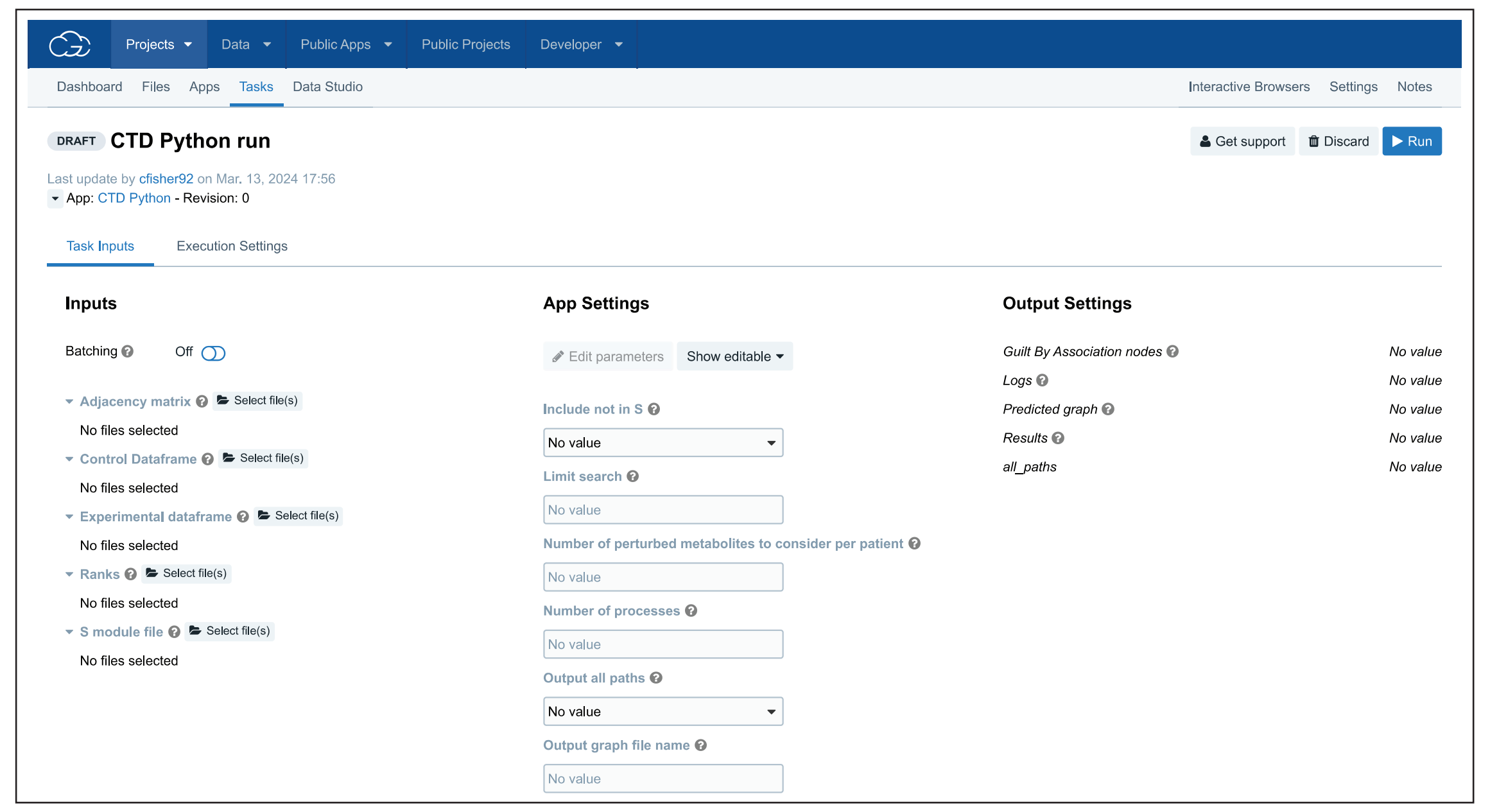
In a previous study [3], we leveraged CTD to identify small multigene biomarkers of response to both taxane- (docetaxel or paclitaxel) and platinum-based (carboplatin or cisplatin) chemotherapy in Triple Negative Breast Cancer (TNBC).

Baseline gene expression and response to both docetaxel and carboplatin treatement was obtained from a large set of TNBC Patient Derived Xenograft (PDX) models. We then built network models of both carboplatin and docetaxel response. A CTD/WGCNA approach was used to identify highly connected sets within these networks that served as biomarkers of response. 9 genes were identified as biomarkers of carboplatin response and 6 genes were identified as biomarkers of docetaxel response.

The biomarkers identified by the CTD/WCGNA approach outperformed biomarkers identified with WGCNA as well as other commonly used feature selections methods. The small biomarker sets identified by the CTD/WGCNA approach were informative across platforms (RNA-seq and array) and across species (PDX and patients). Additionally, these biomarkers were informative across drugs of the same class.The platinum biomarkers were informative for both carboplatin and cisplatin and the taxane biomarkers were informative for both docetaxel and paclitaxel.

Cancer Genomics Cloud Tools

The Cancer Genomics Cloud is an NCI-funded cloud platform that enables the analysis of large cancer datasets in a user-friendly portal. With CTD2 as an app on the CGC, any user with an account can upload or generate an adjacency matrix and analyze it with CTD2 and its "guilt by association" feature with just a few mouse clicks.



DRAFT CTD Python run

Last update by cfisher92 on Mar. 13, 2024 17:56
App: CTD Python - Revision: 0

Inputs

Batching Off

Adjacency matrix: No files selected
Control Dataframe: No files selected
Experimental dataframe: No files selected
Ranks: No files selected
S module file: No files selected

App Settings

Include not in S: No value
Limit search: No value
Number of perturbed metabolites to consider per patient: No value
Number of processes: No value
Output all paths: No value
Output graph file name: No value

Output Settings

Guilt By Association nodes	No value
Logs	No value
Predicted graph	No value
Results	No value
all_paths	No value

https://www.cancergenomicscloud.org

Above is an example draft task of the CTD2 app on the Cancer Genomics Cloud, showing the interface researchers can use to input data and select runtime parameters.

CTD2 is significantly faster than CTD

While the original CTD package was deployed in R, CTD2 is a python package with additional functionality that was not available in CTD. By deploying CTD2 in python, we have increased its computational speed by ~20X for large datasets. This increase in speed allows for the analysis of much larger datasets including thousands of genes.

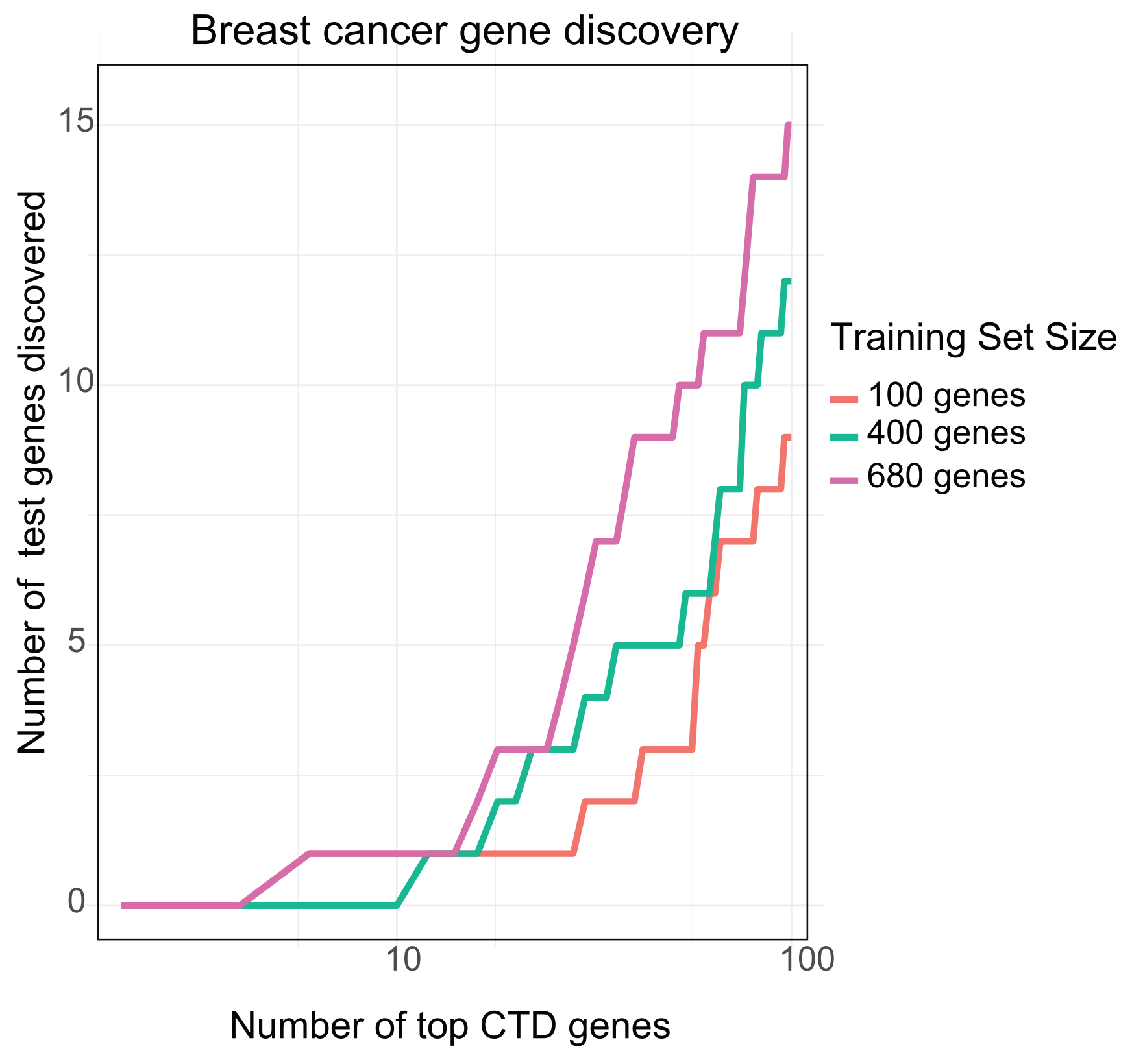
Because of CTD2's increased speed, it can be utilized to identify connections in both large experimental derived graphs and knowledge graphs such as WikiPathways [5] and STRING [6].

CTD2 also allows users to rank how connected other genes in a graph are to a users gene set of interest to identify novel gene candidates.

CTD2 Speed Improvements

S module size	R	Python [num_processes=1]	Python [num_processes=2]	Python [num_processes=4]	Python - [num_processes=8]	Python - [num_processes=12]
5	53.943s	14.841s	17.609s	16.116s	16.737s	18.134s
10	243.951s	30.815s	29.348s	30.091s	26.151s	27.423s
20	207.083s	42.551s	31.424s	24.865s	33.169s	33.466s
50	546.891s	104.899s	68.077s	46.717s	45.608s	48.42s
100	1462.092s	234.333s	153.061s	94.486s	76.07s	76.159s
200	3046.301s	510.385s	306.24s	191.94s	141.417s	130.566s

CTD2 "guilt by association" analysis prospectively predicts breast cancer genes



To test the functionality of CTD2, we investigated if it could rediscover known breast cancer genes. Using The Cancer Genome Atlas (TCGA) [7], we built breast cancer/control expression graphs over the 5,000 most variable cancer genes for each breast cancer subtype.

We subsequently utilized DisGeNET [8] to identify genes associated with breast cancer, categorizing them into a training set discovered before 2015 (n=680) and a test set discovered after 2015 (n = 257). Training sets consisting of 100, 400, and all 680 pre-2015 genes were then utilized to investigate the impact of the amount of information in the training set on the ability of CTD to recapitulate post-2015 genes.

For each training set, CTD ranked all other genes in the graph. The results are depicted in the plot, where the number of post-2015 genes discovered by CTD (y-axis) is plotted against the number of top genes ranked by CTD. We observed that increasing the number of genes in the training set led to an increase in the number of post-2015 genes identified by CTD.

Conclusions and Citations

- CTD/CTD2 are network based approaches that allow users to identify biological signal in complex datasets.
- CTD2 is significantly faster than CTD which allows users to find significantly connected sets within large gene lists
- CTD2 rediscovered known breast cancer genes
- The deployment of CTD2 on the CGC will allow users without a computational background to be able to use this tool.

Funding: NIH grants OT2 OD030547, U54 DA049098, U01 AG072439 to AM.

Citations

[1] Thistlethwaite LR, Petrosyan V, Li X, Miller MJ, Elsea SH, Milosavljevic A. CTD: An information-theoretic algorithm to interpret sets of metabolomic and transcriptomic perturbations in the context of graphical models. PLoS Comput Biol. 2021 Jan 29;17(1):e1008550. doi: 10.1371/journal.pcbi.1008550. Erratum in: PLoS Comput Biol. 2021 Oct 25;17(10):e1009551. PMID: 33513132; PMCID: PMC7875364.

[2] Thistlethwaite LR, Li X, Burrage LC, Riehle K, Hacia JG, Braverman N, Wangler MF, Miller MJ, Elsea SH, Milosavljevic A. Clinical diagnosis of metabolic disorders using untargeted metabolomic profiling and disease-specific networks learned from profiling data. Sci Rep. 2022 Apr 21;12(1):6556. doi: 10.1038/s41598-022-10415-5. PMID: 35449147; PMCID: PMC9023513.

[3] Petrosyan V, Dobrolecki LE, Thistlethwaite L, Lewis AN, Sallas C, Srinivasan RR, Lei JT, Kovacevic V, Obradovic P, Ellis MJ, Osborne CK, Rimawi MF, Pavlick A, Shafaee MN, Dowst H, Jain A, Saltzman AB, Malovannaya A, Marangoni E, Welm AL, Welm BE, Li S, Wulf GM, Sonzogni O, Huang C, Vasaikar S, Hilsenbeck SG, Zhang B, Milosavljevic A, Lewis MT. Identifying biomarkers of differential chemotherapy response in TNBC patient-derived xenografts with a CTD/WGCNA approach. iScience. 2022 Dec 12;26(1):105799. doi: 10.1016/j.isci.2022.105799. PMID: 36619972; PMCID: PMC9813793.

[4] Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, Groves-Kirkby N, Mihajlovic A, DiGiovanna J, Srdic M, Bajcic D, Radenkovic J, Mladenovic V, Krstanovic D, Arsenijevic V, Klisic D, Mitrovic M, Bogicevic I, Kural D, Davis-Dusenbery B; Seven Bridges CGC Team. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. Cancer Res. 2017 Nov 1;77(21):e3-e6. doi: 10.1158/0008-5472.CAN-17-0387. Erratum in: Cancer Res. 2018 Sep 1;78(17):5179. PMID: 29092927; PMCID: PMC5832960.

[5] Agrawal A, Balcı H, Hanspers K, Coort SL, Martens M, Slenter DN, Ehrhart F, Digles D, Waagmeester A, Wassink I, Abbassi-Daloii T, Lopes EN, Iyer A, Acosta JM, Willighagen LG, Nishida K, Riutta A, Basaric H, Evelo CT, Willighagen EL, Kutmon M, Pico AR. WikiPathways 2024: next generation pathway database. Nucleic Acids Res. 2024 Jan 5;52(D1):D679-D689. doi: 10.1093/nar/gkad960. PMID: 37941138; PMCID: PMC10767877.

[6] von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D433-7. doi: 10.1093/nar/gki005. PMID: 15608232; PMCID: PMC539959.

[7]The results <published or shown> here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga."