Open Peer Review on Qeios

Classification of Cancer Response to Antiglycolytic Agents: An Approach to Understanding and Predicting Cancer

Christopher El Hadi[1]

[1] Saint Joseph University of Beirut

Funding: The author(s) received no specific funding for this work.

Potential competing interests: The author(s) declared that no potential competing interests exist.

Abstract

Background: Cancer cells have exquisite metabolic activity in the glycolysis pathway compared to normal cells, which has been shown to provide them with sufficient fuel and molecular building blocks to maintain their proliferation. Inhibition of glycolytic pathways could therefore be an approach to target cancer. In this study, we sought to find the general categories of cancer cell responses to antiglycolytic agents, which could help predict cancer cell responses to drugs.

Methods: Gene expression profiles from 199 experiments were downloaded from the GEO database and transformed into gene fold-changes. The most important genes were selected using the ATC top-value method and the samples were partitioned using "Spherical k-means" clustering, both methods being deemed superior by the authors of the *cola* package. The one-sample chi-square test was used to investigate the predominance of samples in the clusters, followed by the Clopper-Pearson exact test to estimate this proportion of predominance. Signature genes were identified using the F-test for differential analysis; these were grouped using k-means clustering, and each group was functionally enriched with Gene Ontology and Reactome terms to extract biological significance.

Results: Three responses were found. The first showed negative regulation of the mitotic cell cycle and associated intracellular activities and their regulation, as well as negative regulation of the cellular response to stress and DNA damage and antigen processing and presentation via MHC class I. The second response showed positive regulation and the third showed no effect on the above processes. In addition, the tissue samples were not distributed in a way that favored certain classes, an observation that demonstrates that the classification is sensitive to treatment.

Conclusions: We believe that cancer cells function through "molecular states" that are detectable by artificial intelligence and could potentially replace conventional cancer classifications. By studying the response of the cancer state to a given treatment, we can extrapolate this response to all cell lines that are already in that molecular category.

Background

In the 1920s, Otto Warburg and his colleagues observed that tumors took up prodigious amounts of glucose compared to what was observed in the surrounding tissues. Upon investigation, glucose fermentation was detected in these cells, which produced lactate even in the presence of oxygen, hence the inspiration for the term "aerobic glycolysis" [1][2]. The paradoxical abundance of this form of glycolysis in cancer cells is called the "Warburg effect". Knowing that the catabolism of glucose to lactate has an extremely low energy yield, a high rate of glucose consumption is thus required to meet the energetic and anabolic demands of cancer cells [3]. In fact, glycolysis provides numerous metabolic intermediates that can be used for *de novo* synthesis of nucleotides, amino acids, lipids, and NADPH that are essential for promoting rapid cell proliferation. Cancer cells thus have a different metabolism than regular cells, allowing them to maintain a high rate of proliferation and resist apoptosis signals [4].

This said, inhibition of glycolytic pathways could represent a selective approach in cancer research to develop targeted anticancer agents. Chen [5] and Abdel-Wahab [6] and colleagues reviewed the different modalities that exist to forestall glucose utilization by cancer cells. Over the years, many inhibitory drugs have been discovered and tested on cancer cells, followed by gene expression evaluation to find better drugs with a lethal metabolic impact on cancer cells. Several of these glycolytic inhibitors are currently in preclinical and clinical studies that are showing promising results [5][7][8].

In this article, we will be classifying the response of various cancer cells to glycolysis inhibitory drugs in an effort to find a hidden pattern that can summarize the behavior of cancer cells when exposed to anti-glycolytic agents. Myriads of papers have been published describing the discovered biological functions and morbid consequences of these agents on multiple cancers but none, to our knowledge, have sought to find a common denominator linking the treatments' outcomes. This type of classification should help predict the response of malignant cells based on basic cancer characteristics such as tissue of origin or antiglycolytic agent used.

Methods

Data processing

Gene expression profiles were downloaded from the GEO database. The terms entered in the search engine were "Cancer" followed by the name of the agent, or any of the alternative names for the same agent. All the dataset accession numbers and their respective drugs and cell lines can be found in the additional files. The filters used with every search were "Homo sapiens" for organisms, "Expression profiling by array" for the study type, and "Series" under entry type. The inclusion criteria we based our judgment on for choosing the datasets were:

1. Expression data should only be from single-channel microarray chips. Agilent, Affymetrix, and Illumina are the only accepted manufacturers
2. Cells should be treated in an artificial *in vitro* milieu (no mice xenografts etc.)
3. Cells should not have been transfected or mutated to affect the targeted pathway or been made resistant to treatment (only wild cell types under the influence of chemicals)
4. Cells should not have undergone secondary treatments (other than the vehicle and the drug of interest) before RNA extraction
5. Coding RNA, particularly cytoplasmic or total extractions, are preferred
6. The glucose deprivation medium should contain less than 1g/L, i.e., 5.6 mM, of glucose
7. Cells should be quintessentially cancerous, not benign, with no other superimposed disease
8. The data series should contain both controls and treatments.

The raw files downloaded were either in CEL or TXT format. Few TXT files contained processed data, and these were simply used as-is. For the Illumina Beadchips, R's *limma* [9][10] package was used and the function *neqc()* was employed to perform background correction followed by quantile normalization. It should be noted that negative control probes were not always available, therefore the detection p-values of each probe were exploited instead. For probe annotation, the *annotate* package was chosen [11], with *illuminaHumanv3.db* as the annotation data [12], and "gene symbol" as the annotation. For Affymetrix chips, *limma*, *affy* [13], and *oligo* [14] were used to process the CEL files. The *rma()* function normalized the microarray signals and the annotation data packages were adapted to the version of Affymetrix chips available [15][16][17][18][19][20][21][22]. The Agilent chips were read using likewise the *limma* package, annotated according to the chip versions [23][24][25], background-corrected following the normal-exponential method, and quantile normalized. After reading and normalizing the signals, the controls and their respective treatments were exported into excel files. This resulted in a total of 798 microarrays (517 treated assays and 281 controls) comprising 73 cell lines from 17 different tissues treated with 26 different anti-glycolytic agents. An additional 1014 microarrays were also downloaded pertaining to only one study (666 treated assays and 348 controls), which included 59 cell lines from 11 tissues treated with either sorafenib or rapamycin at different drug doses and exposure durations. The log base 2 foldchange (FC) was then calculated for each treated chip, and all the FCs were lastly fitted into two matrices: the first consisted of 199 foldchange retrievals, the second 666. The matrices thus contain expression data in form of FC.

Consensus Clustering

Consensus clustering was performed using the *cola* package (version 1.8.1). This data classification technic was, foremost, done on the 199 experiments found on the GEO, for the sake of finding a common denominator response to all glucose-challenged cells. Data augmentation was performed on the top 6 treatments, each having at least 5% sample prevalence among the experiments at hand, which increased the number of samples to 50 per treatment.Augmentation was achieved by applying the *semiArtifical* package using unsupervised tree set data generation[26]. It confers statistical validation on the repartition of the experiments into their corresponding classes. Lastly, a thirdclustering was performed on specific treatments, namely rapamycin and sorafenib from the 1014-sample accession, to study the effect of sample classification when controlling for the treatment used. Before performing the consensus partitioning, an important step was to clean up the input matrix. *cola* provides the *adjust_matrix()* function that proceeds as follows:

Rows with >25% of treatments having missing values are removed

1. A function is used to impute missing data, using nearest neighbor averaging
2. In every row in the matrix, values larger than the 95th percentile or less than the 5th percentile are replaced by corresponding percentiles
3. Rows with zero variance are removed
4. Rows with variance less than the 5th percentile of all row variances are removed.

After cleaning, Gaussian normalization was performed on each sample to make the sample ranges comparable; this scaling method was used as it demonstrated better results compared to other scaling methods. For clustering, the top n features were first selected by the ATC top-value method (or "Ability To Correlate to other rows"), which was introduced and recommended by *cola*'s authors. Then, the matrix was scaled by the selected rows and randomly sampled; these samples were partitioned by the "Spherical k-means" clustering method, a variant of the standard k-means clustering, which was shown to be superior to other clustering methods by the authors. Finally, the sampling and partitioning process was repeated 50 times to obtain a list of partitions.

The best-fitting number "k" of subgroups was evaluated using the average silhouette score, ambiguous clustering proportion (PAC) score, concordance, and Jaccard index. The PAC score measures the proportion of the ambiguous subgrouping; 1-PAC thus represents the unambiguous, which will be used hereafter. Concordance is defined as the proportion of concordant pairs of samples divided by the total number of possible evaluation pairs; it is a metric to evaluate the predictions made by the clustering algorithm. The silhouette score measures how similar an object is to its class compared to other classes. The Jaccard index is the ratio of the number of sample pairs that are both in the same subgroup in the partitions at k “and” k – 1 and the number of sample pairs that are both in the same subgroup in the partitions at k “or” k – 1. That said, the best number of subgroups is determined based on the majority vote among the highest 1-PAC score, highest mean silhouette, and highest concordance and a PCA plot (Principal Component Analysis plot) showing nonoverlapping and well-delimited groups.

Statistical Inferences

Since the data are categorical, the goodness-of-fit test (i.e., one-sample chi-square test) was used to investigate the null hypothesis of whether the sample frequencies are equally distributed among the partitioned clusters. If not uniformly distributed, we will proceed with calculating the Clopper-Pearson exact test to estimate the population proportions in each cluster. The latter test was used as an alternative to the Wald test knowing that it performs better when it comes to calculating confidence intervals.

Functional Analysis

The “signature genes” were then identified using the F-test for differential analysis. Signatures are simply those rows that show statistically significant specificity in one or more subgroups. These were further grouped by patterns among subgroups using k-means clustering and the appropriate number of signature groups was automatically selected. Functional enrichment was applied to each group of signatures separately. Gene ontology (GO) and Reactome enrichment analysis were performed by applying the R packages *ClusterProfiler* [27] and *ReactomePA* [28], with *cola*’s function *functional_enrichment()*. GO terms were then clustered and visualized as heatmaps by the R package *simplifyEnrichment* [29].

Results

Two matrices were constructed, the first containing 199 experiments and the second 666. The rows amount to 33622, representing the studied genes as expression foldchanges. The *adjust_matrix()* function of the *cola* package was used to clean the matrices. In the following, the ratio of "samples subjected to a treatment in a certain class" to the "total number of samples for that treatment in all classes" will be used instead of the proportion of samples for that treatment to the total samples in the corresponding class. The same applies to the evaluation of tissue frequencies.

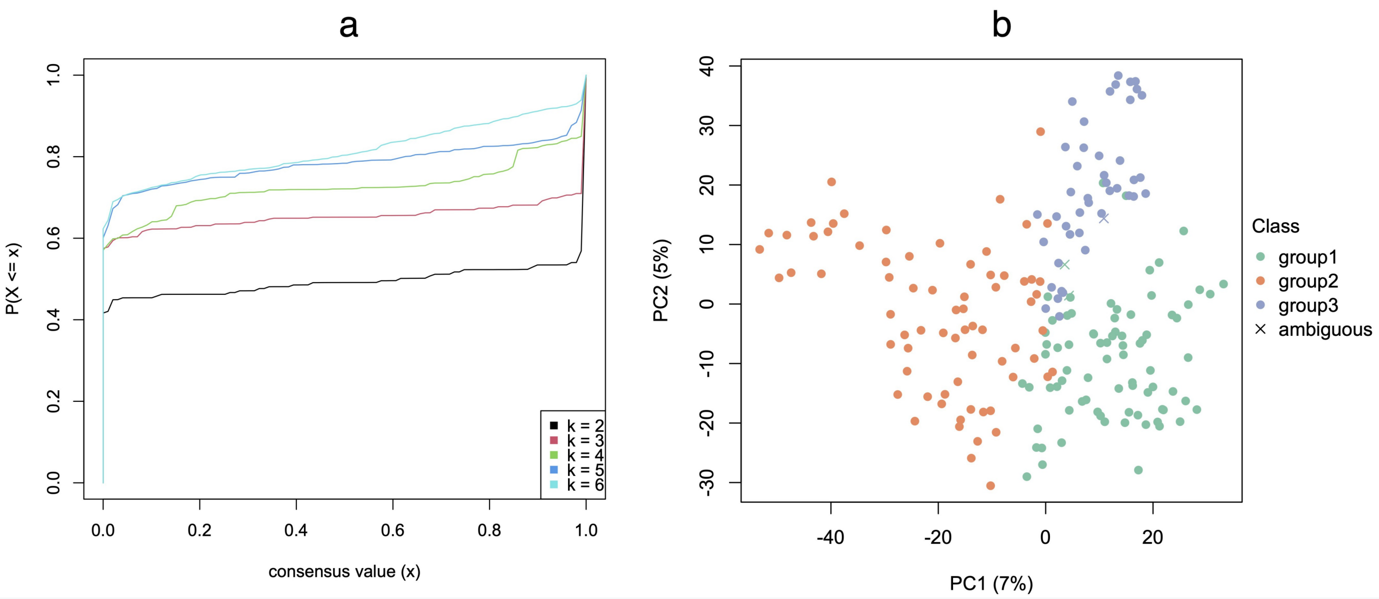

Figure 1. eCDF curve and PCA plot of the consensus clustering. (a) eCDF curve of the consensus matrix from partition by ATC:skmeans, a good k can be selected by aiming at the flatness of the eCDF curve. (b) PCA on 16127 rows with the highest ATC scores, 188 out of 195 confident samples were kept in their classes (silhouette > 0.8)

In the main matrix of 199 columns, 4 treatments were removed because more than 80% of their values were missing, resulting in 195 samples and 16127 significant genes filtered for analysis. Consensus partitioning recommended grouping the samples from the treatments into 5 classes, and alternatively into 2 or 3. Table 1 shows the 1-PAC, mean silhouette, concordance, and Jaccard evaluation indices for all classes of number "k", with a maximum k of 6. Due to the substantial overlap between the different class elements visualized on the PCA graph for k equal to 5 classes, and having realized that the evaluation indices for 2 and 3 classes are almost identical, the choice of 3 classes was preferred. The cumulative distribution function (CDF) curves for a maximum k of 6 and the PCA plot for k equal to 3 are shown in Figure 1A and Figure 1B. Only treatments with silhouette scores ≥0.8 were retained in their corresponding classes, resulting in a new total average silhouette score of 0.98, and a new total of 188 significant treated samples.

Table 1. Clustering evaluation indices for a maximum k of 6

Class	*k*	*1-PAC*	*Mean Silhouette*	*Concordance*	*Jaccard*
2	2	1.00	0.97	0.99	0.51
3	3	0.99	0.96	0.98	0.60
4	4	0.85	0.90	0.94	0.70
5	5	0.91	0.88	0.95	0.72
6	6	0.82	0.72	0.86	0.92

Samples were not uniformly distributed (p = 0.007), with class 1 having the highest number of treated cells (77), followed by class 2 (68) and class 3 (43). Samples treated with genistein, metformin, glucose deprivation (GS), phenformin, BEZ235 and selumetinib accounted for the majority of the GEO samples (65.43%), each with a minimum ratio of >5%. Because the latter drugs have a sufficient number of samples, they are used to analyze the division of treatments into classes and will then be augmented to 50 samples each. The one-sample chi-square test was also applied to the 6 toptreatments and showed that metformin, phenformin, and selumetinib were distributed in far from unequal proportions in the three classes, with no specific distribution pattern for their treated tissues. For the rest, the distribution was not uniform: all 13 BEZ235 samples and 86.67% (CI: 59.54-98.34% and p <0.01) of the GS samples were in class 1, 80% (CI: 64.35-90.95% and p <0.001) of the genistein-treated cells were present in class 2, and all drug combinations (2DG or GS combined with phenformin, metformin, or buformin) were in class 3. The p-values for genistein, GS, and BEZ235 thus prove that they have a prevalence of response in a single class, with ratios significantly >0.5. These results are further illustrated in Table 2.

Table 2.	The proportion of the most relevant treatments in each cluster to their total							
		Class 1		Class 2		Class 3		
Treatment	*n*	*%*	*p*	*%*	*p*	*%*	*p*	*2 p*
Genistein	40	12.50%	<0.001	80.00%	<0.001	7.50%	<0.001	<0.001
Metformin	30	20.00%	<0.005	30.00%	0.043	50.00%	1	0.123
GS	15	86.67%	<0.01	-	-	13.33%	<0.01	<0.005
Phenformin	14	21.43%	0.057	35.71%	0.424	42.86%	0.791	0.607
BEZ235	13	100.00%	<0.001	-	-	-	-	<0.001
Selumetinib	11	45.45%	1	45.45%	1	9.09%	0.01	0.234

*Note. p=t*ests for the null hypothesis, under which we assume that the random variable follows a binomial distribution with size n and probability of success π. 2 *p=* p-value for the 2 one-sample test.

With respect to cell types, all breast, liver, uterus, bone, and colon tissues had a sample prevalence greater than 5%. The one-sample chi-square test resulted in a p-value >0.05 for colon and liver, <0.05 for bone, and <0.005 for breast and uterus which reflects significant ratio differences between classes. This came alongside an ambivalent prevalence regarding classes 1 and 2 (CI: 17.86-44.61% and 39.32-68.19%) for housing the majority of treated breast samples, classes 2 and 3 (CI: 45.72-88.11% and 8.66-49.1%) for uterus, and classes 1 and 3 (CI: 30.76-78.47% and 17.3-64.25%) for bone samples, none of which has a confidence interval excluding 0.5 and all of which instead have overlapping confidence intervals (Table 3). This demonstrates that no tissue has a predominance of samples in a certain class, unlike the treated samples.

Table 3.	The proportion of the most relevant tissues in each cluster to their total							
		Class 1		Class 2		Class 3		
Tissue	*n*	*%*	*p*	*%*	*p*	*%*	*p*	*2 p*
Breast	50	30.00%	0.007	54.00%	0.672	16.00%	<0.001	0.004
Liver	29	24.14%	0.008	44.83%	0.711	31.03%	0.061	0.381
Uterus	20	5.00%	<0.001	70.00%	0.115	25.00%	0.041	0.001
Bone	18	55.56%	0.815	5.56%	<0.001	38.89%	0.481	0.030
Colon	17	47.06%	1.000	35.29%	0.332	17.65%	0.013	0.327

*Note. p=t*ests for the null hypothesis, under which we assume that the random variable follows a binomial distribution with size n and probability of success π. 2 *p=* p-value for the 2 one-sample test.

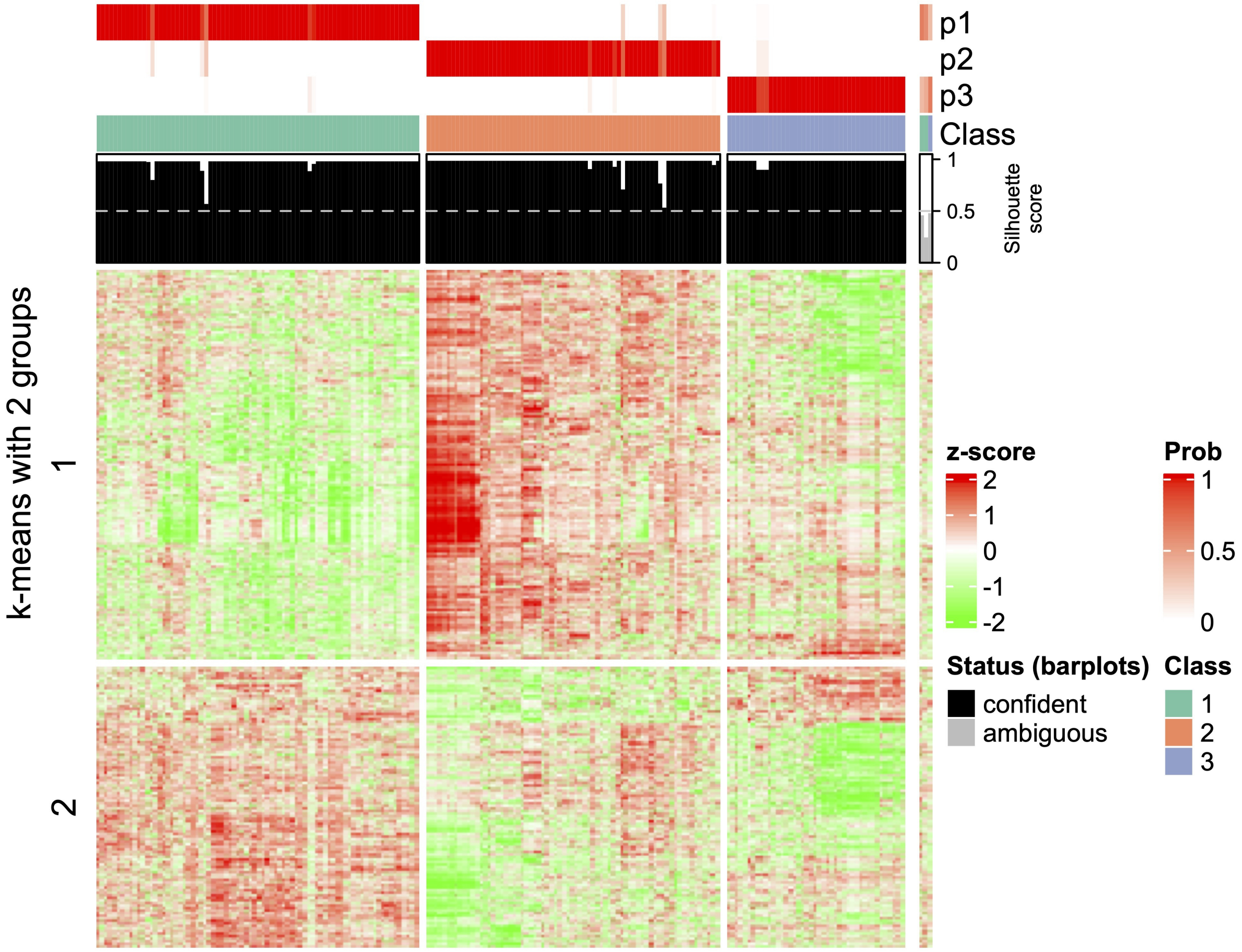

Figure 2. Heatmap of the signature genes and their regulation in each class. 2372 signature genes (14.7% of total genes) chosen for an FDR < 1e-06. These characterize the 3 classes, the first class being downregulated for the first group of signature genes and upregulated for the second group, class 2 has the opposite pattern, and class 3 has no significant regulation for both groups.

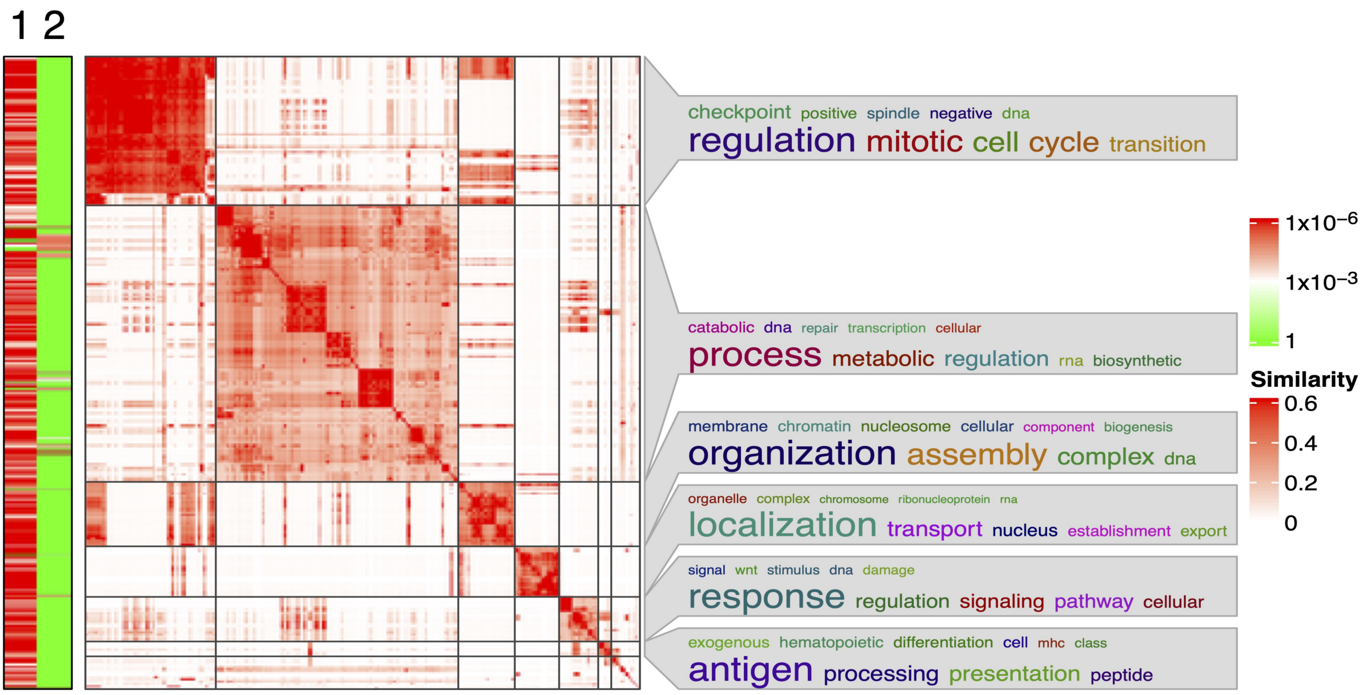

Figure 3. Heatmap of the similarities of significant GO terms from the 195-sample gene list . These resulted from the enrichment (using FDR < 0.01) of the genes in either group. The green-red columns show for which group (1 or 2) the respective GO terms are significant. The word cloud annotation visualizes and highlights the biological functions in each GO group. GO enrichment analysis was performed by a hypergeometric test.

Signature genes were identified using an FDR threshold of $<10^{6}$. K-means clustering on the genes yielded two signature groups, with the first group being more biologically significant than the second. The "signature heat map" showing the classification of the samples and that of the genes is shown in Figure 2. Functional analysis of the first group using Gene Ontology (GO) terms for biological properties resulted in 28 term clusters (Figure 3), with a Benjamini-Hochberg adjusted p-value <0.001. The most relevant ones describe mitotic cell cycle process and regulation; DNA/nucleotide metabolic process; DNA replication and repair; chromosome organization and nuclear division; nucleocytoplasmic RNA transport; cellular response to stress and DNA damage; and antigen processing and exogenous peptide antigen presentation via MHC class I. After clustering of Reactome terms, the pathways involved were related to the cell cycle, mitotic G1 phase and G1/S transition, G2/M transition, DNA damage response and p53-dependent G1 checkpoint, and SUMOylation of RNA binding proteins. These functions and pathways are shown to be downregulated in class 1, upregulated in class 2, and rather unaffected in class 3 (Figure 2). The full results can be found in the additional files.

Consensus partitioning of the augmented samples of the top 6 treatments resulted in two classes, and Pearson's chi-square test was used to compare the two groups of the augmented samples to the three groups obtained from the partitioning of 195 samples. Although the frequency counts of the augmented samples showed no statistical consistency with the distribution of the original samples, it should be noted that the resulting first class was closest to class 2 of the 195-sample clustering (having the highest p-value), and vice versa for the second class. Class 3 was statistically unstable in finding the class that best included it. As an illustration, BEZ235 and GS were seen exclusively in the second class (closest to class 1), most genistein-treated samples were in the first class (closest to class 2), and the same was true for phenformin and selumetinib, while metformin was ambivalent to both classes. It is also worth noting that the regulations are also consistent with those described above; namely the downregulation of mitotic cell cycle processes and chromatin assembly in the second class, and upregulation in the first. The analysis can be reviewed in the additional files.

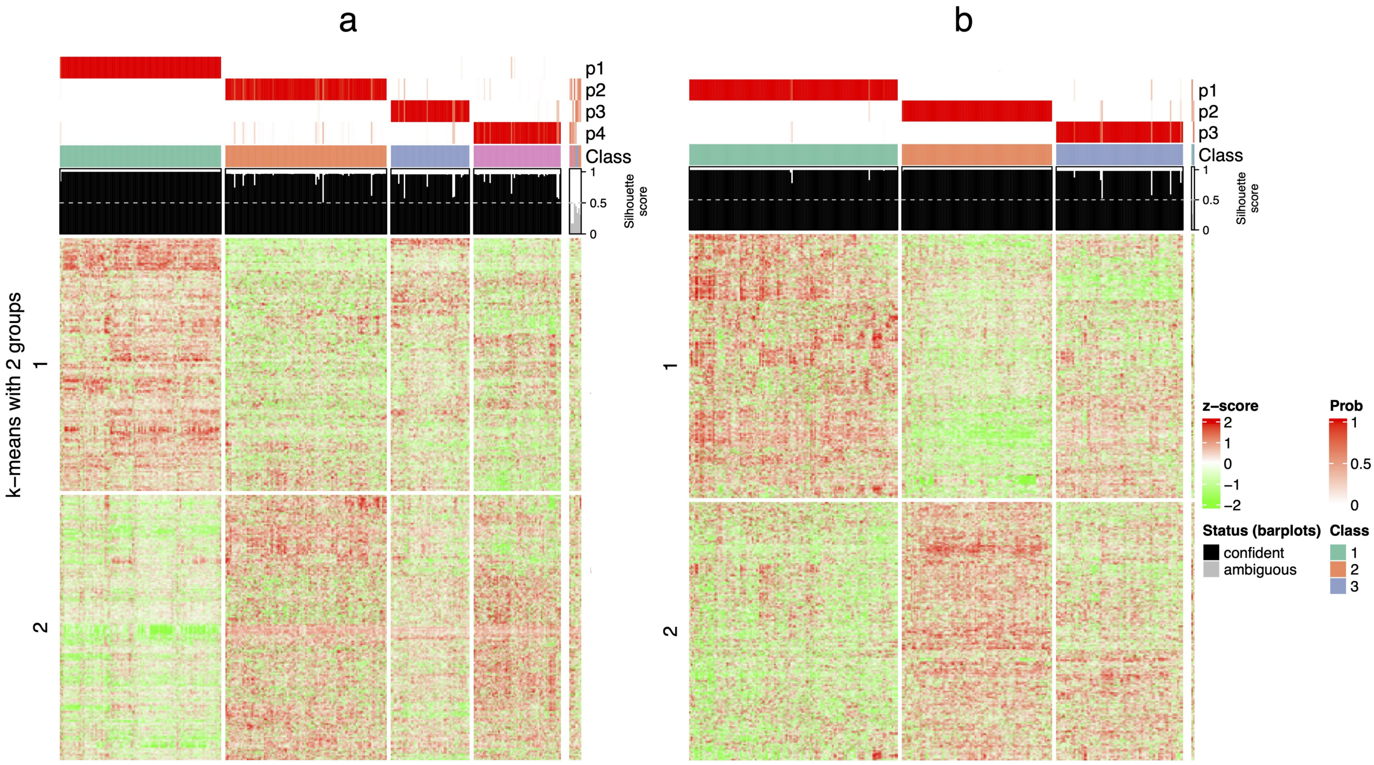

Figure 4. Heatmaps of the top signature genes for sorafenib and rapamycin samples. (a) Sorafenib's heatmap shows ATC:skmeans clustering into 4 classes and (b) 3 classes for rapamycin. Both drugs had 2 distinct signature groups characterizing the classes' biological properties.

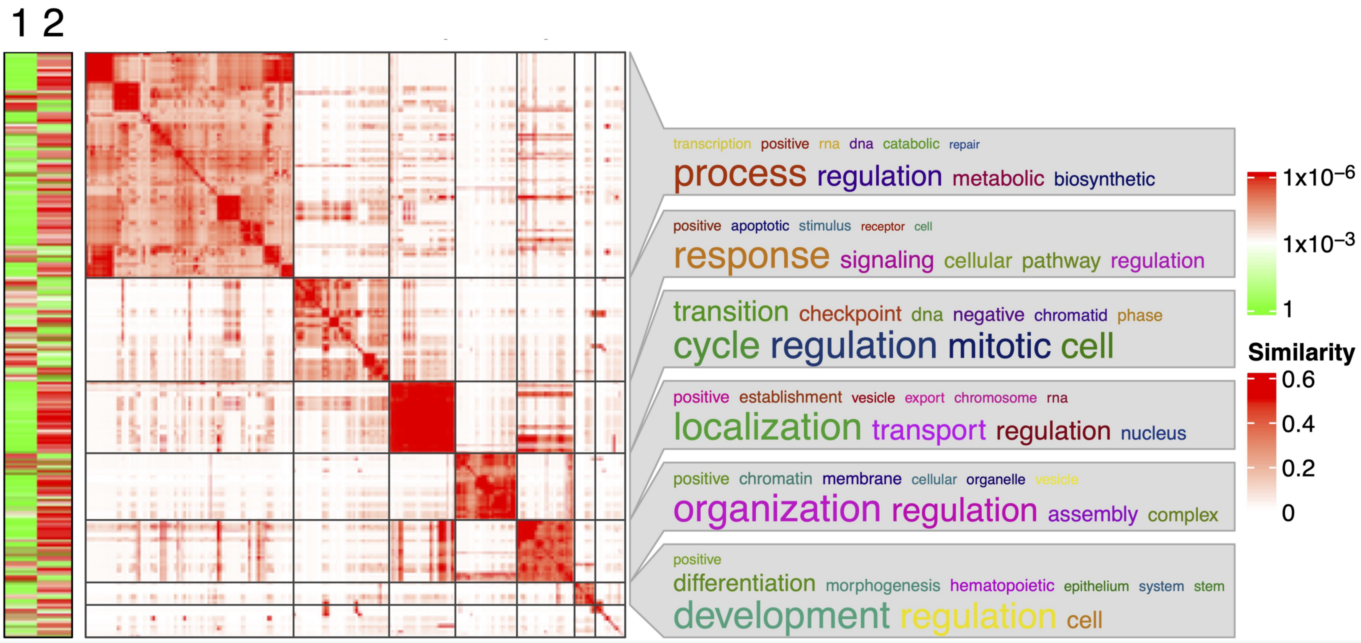

Figure 5. Heatmap of the similarities of significant GO terms (FDR < 0.01) from sorafenib's gene list. These originated from the sorafenib gene list using ATC:skmeans clustering method.

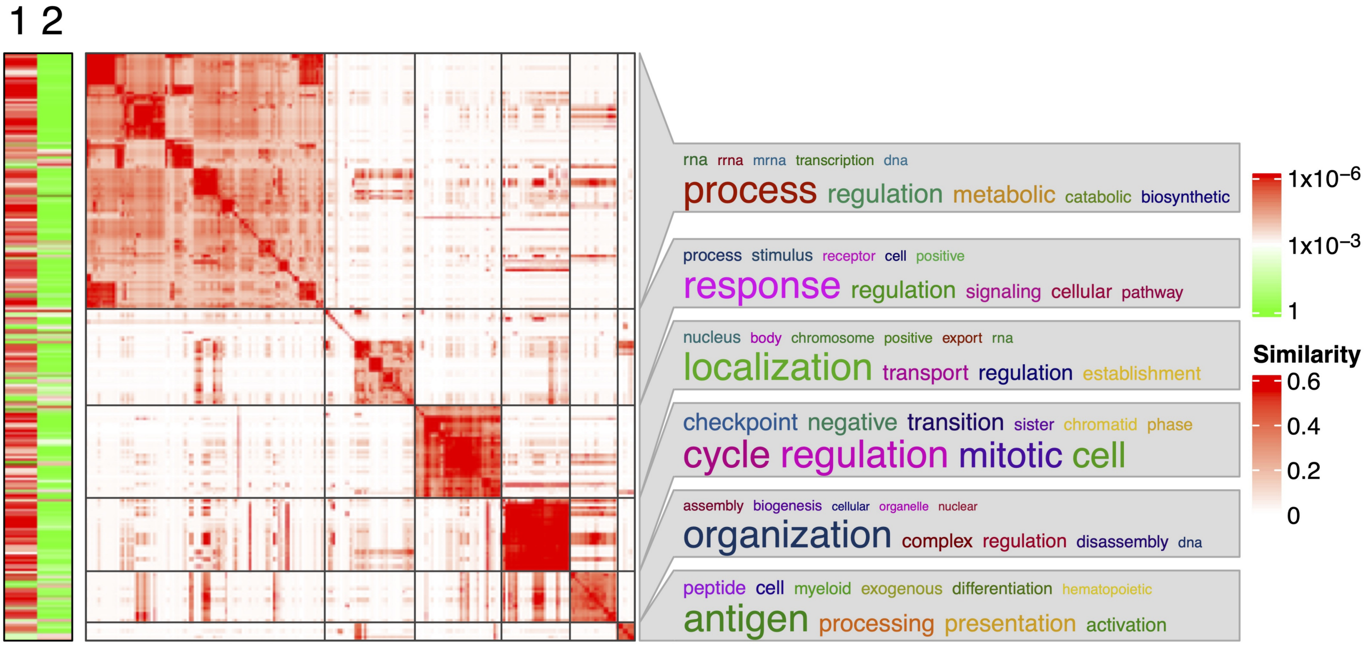

Figure 6. Heatmap of the similarities of significant GO terms (FDR < 0.01) from rapamycin's gene list. These originated from the rapamycin gene list using ATC:skmeans clustering method.

Additional consensus clustering was performed on 341 sorafenib-treated and 325 rapamycin-treated samples (from the accession of 1014 samples), each treatment having mainly 3 available doses given for 3 exposure times. Consensus partitioning of sorafenib samples resulted in 4 classes (1-PAC=0.977, silhouette= 0.946, concordance=0.977), and 3 rapamycin classes (1-PAC=1, silhouette= 0.982, concordance=0.992) (Figure 4). This classification controls for the drug used, leaving room for tissue-, dose-, and duration-based classification and interpretation. The classes obtained for sorafenib simply shared the samples almost uniformly between them, with no preference for any tissue (except breast, lung, and skin, which were more common in classes 1 and 2), treatment duration, and dose; therefore, no hidden trends or categorization of these parameters could be inferred. As with rapamycin, tissues were almost evenly distributed among the three classes (except blood, which was more prevalent in class 1), but sensitivity to dose and duration was salient, i.e., class 1 included reactions lasting only 2 hours, 24 hours in class 2 and 6 hours in class 3. The signature genes for sorafenib and rapamycin were classified into two groups both of which showed significant biological properties for sorafenib (summarized in Figure 5), but the second group has no significant biological translation for rapamycin (Figure 6). The cluster-specific biological regulations of the functional groups are illustrated in Figures 4A and 4B for sorafeniband rapamycin, respectively. A full discussion of the results can be found in the additional files.

It was deemed unnecessary to combine the 666 experiments with the 195 and perform a clustering analysis because of the potential bias that could result from adding 300 samples treated with only 2 agents, to 195 samples treated with a total of 26 agents. In addition, partitioning that maintains constant the tissue variable and allows for a treatment-based study was also deemed impossible because of the limited number of samples per tissue.

Discussion

Consensus clustering was performed on 199 samples comprising 26 different treatment approaches, with each treatment having different doses and time of drug exposure. This range of treatments means that the sole purpose of clustering is to find a typical biological regulatory pattern at each resulting cluster. The clustering resulted in 3 meaningful response classes. The first class, class 1, contains the substantial majority of BEZ235 and GS samples and shows down-regulation of the mitotic cell cycle, its regulation and all associated intracellular activities, as well as down-regulation of the cellular response to stress and DNA damage and of antigen processing and presentation via MHC class I. Class 2 contains almost all genistein samples and shows negative regulation of all the above biological processes. Finally, all drug combinations, i.e., 2DG or GS combined with phenformin, metformin, or buformin, fell into class 3, which includes cells that were not affected by the antiglycolytic agents. Augmentation of the data and then reclassification resulted in a distribution of samples very similar to the main classification. Metformin, phenformin, and selumetinib were not unevenly distributed among the three classes, thus showing no special pattern for their dispersion.

Furthermore, the tissue samples were not distributed in a way that favored certain response classes, an observation that demonstrates that classification is treatment sensitive. This tissue independence was confirmed after classifying 341 sorafenib and 325 rapamycin samples: controlling for treatment resulted in a distribution of tissues that was not disproportionate between classes, thus no tissue-type predominance in any class. Interestingly, the dose and duration of treatment had an impact on the cancer response. All of the studies we reviewed for data interrogation, which examined the response of specific cell lines to a single dose and duration of exposure of an antiglycolytic drug, demonstrated similar biological functions to ours (GEO accession numbers of studies that demonstrated class 1 deregulation: GSE31058, GSE97346, GSE59882, GSE79316, GSE73923, GSE59228, GSE36847, GSE25412, GSE114060, GSE96794, GSE79246, GSE9008, GSE62663, GSE116387, GSE137553, GSE112079; and class 2 upregulation: GSE59704, GSE5200, GSE85257, GSE112079. More information on the 69 studies is available in the additional files).

The goal of our study was to find categories of the response of treated tissues or to the drugs used, which could allow us to predict the response of a cell line to any drug, or the response to a drug independently of the cell line, simply by associating that drug or cell with its response category. In other words, we wanted to know if specific tissues respond biologically in the same way to any antiglycolytic drug. This could be of interest in predicting the response of cancer cells that have not been previously treated with an antiglycolytic agent, based on our knowledge of the biological response of the tissue from which those cells originate. If this latter hypothesis proves to be unfounded, then it would be interesting to know whether each agent has a specific response independent of the tissue treated, which would allow us to predict the response of any cell just by knowing the type of agent administered. That said, we can state from what we have found that antiglycolytic agents appear to have universal biological responses, making it possible to predict the effect of a drug independently of the tissues tested given the statistically significant predominance of drugs in single classes. Furthermore, it is not possible to predict the behavior of a tissue independently of the drug, as it has been shown that tissues can have multiple responses under the influence of a single drug.

The reasons for this difference in drug/tissue distribution are not simple. It is partly due to the diversity of the experiments, which included 79 cell lines, each with nuanced metabolisms, treated with 26 reagents at different doses and exposure times. The mechanism of action of the drugs and the targeted pathways should explain much of this. Simply put, each cell expresses the same biological pathways but at completely different levels of activity, due to endogenous factors, such as developmental and differentiation programs and epigenetic states of the cells of origin, in conjunction with exogenous factors, such as mutagenic exposures, pathogens, and inflammation. The knowledge that each antiglycolytic agent acts on pathways that are ubiquitous in all living human cells, regardless of lineage, should go a long way toward explaining why the classification applies to drugs and not to tissues, and why responses differ greatly between drug classes.

It is reasonable to mention at this point that further *in silico* studies are needed to validate our observations. Asking for more *in vivo* studies in order to have a larger population (more than 199) of experiments for classification and more statistically significant results is theoretically advantageous, but practically useless knowing that we are asking for time-consuming combinations of experiments (combining between thousands of cell types and infinite doses and exposure times of antiglycolytic agents) that take us away from the purpose of this study and what it should verily evaluate and explain. That said, we believe that cancer remains an ambiguous cellular mechanism and that there must be a hidden pattern that links human cancers together, enabling the positioning of cancer cells on a categorizable continuum. A limitation of this study is that we assumed a categorization of cancer by tissue type and expected that there would be tissues with nearly unique and mutually exclusive molecular reactions that should quintessentially respond to anti-glycolytic agents. This response-quintessence of tissues has been disproven in this study. That notwithstanding, the classification of cancer cells based on data on chromosome morphology, DNA epigenetics and somatic variants, mRNA, miRNA, and protein expression levels should be an alternative approach to the study of cancer drug responsiveness. The latter attempts the individualization of the fundamental "molecular activity states" at the origin of malignancy, whichshould uncover hidden biological patterns, not perceptible to the human brain, requiring artificial intelligence and network analyses to detect them. It will also result in classes that can be used as "study units" that should complement the traditional anatomical cancer classification system, or even replace it entirely.

Hoadly *et al.* had this same idea in 2014, hypothesizing that molecular signatures could provide a different taxonomy from the current classification of pathology based on organ and tissue histology [30]. After analyzing 3,527 tumors from 12 different cancer types (dubbed "Pan-Cancer-12") in 2014, they recently presented a new integrative "PanCancer Atlas" analysis including all TCGA tumors, or approximately 33 tumor types (i.e., ~10,000 samples) [31]. They first iteratively clustered (either by hierarchical clustering or unsupervised consensus clustering) chromosome arm-level aneuploidy, DNA methylation data, mRNA and miRNA expression profiles, and protein expression data. They then performed an integrative clustering of all these types of data and were able to identify 28 distinct molecular subtypes, the closest to the idea of "molecular activity states" to date.

Conclusion

Many studies have tested the effects of a single antiglycolytic drug on specific cell lines, but no study to date has sought to find a commonality among all cancer cell responses to antiglycolytic agents used on cancer cells. In this study, we sought to find categories of cancer cell responses to a large variety of antiglycolytic agents. This classification should give a general idea of what to expect from any drug targeting glycolysis and may help predict the effect of drugs on cancer cells. We found that there are 3 possible types of responses for all cancer cells treated with any antiglycolytic agent. This classification seemingly depended on the treatments used but not the type of tissue treated. A better classification approach to study the response of various cancers is to first classify all known cancers, which will help to disentangle the different "molecular activity states" of cancers. These response categories can then help predict the molecular state response of cancer to a given treatment, and thus extrapolate that response to all cell lines that are already in that molecular category.

Abbreviations

2DG: 2 deoxy-d-glucose

mRNA: messenger RNA

miRNA: microRNA

ATC: Ability To Correlate to other rows

BH : Benjamini-Hochberg

CDF : cumulative distribution function

MHC: Major histocompatibility complex

FC: foldchange

FDR: False Discovery Rate

GEO: Gene Expression Omnibus

GO: Gene Ontology

GS: glucose starvation

CI: confidence interval

NADPH: nicotinamide adenine dinucleotide phosphate

PAC: proportion of ambiguous clustering

PCA: Principal Component Analysis

TCGA: The Cancer Genome Atlas

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to the availability of the raw data on the GEO but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

C.H. is the sole author of this manuscript. C.H. undertook the entire analysis process, wrote the main text of the manuscript, and prepared all figures and tables.

Acknowledgments

Not applicable.

Footnotes

Not applicable.

Additional files

Additional file 1 (.xlsx) Samples accessions. A table containing the access number of each downloaded sample, the cell line and corresponding tissue used, the antiglycolytic agent administered as well as the dose and duration of exposure, and the type of microarray chip.

Additional file 2 (.xlsx) 199-sample classification. A table containing the 199-experiments and their respective classes that resulted from the ATC:skmeans consensus clustering. The evaluation of the sample's membership in each class is given by the silhouette score. A statistical analysis of the treatments and tissues is also included.

Additional file 3 (.xlsx) Enrichment analysis of the signature genes in each signature group for the 199-sample list. GO and Reactome terms and term clusters are included on separate sheets.

Additional file 4 (.xlsx) Enrichment analysis of the signature genes in each signature group for the 666-sample list. GO and Reactome terms and term clusters are included on separate sheets.

Additional file 5 (.docx) Discussion of sorafenib and rapamycin's classes and biological regulations.

Additional file 6 (.xlsx) 666-sample classification. A table containing the sorafenib and rapamycin samples and their respective classes that resulted from the ATC:skmeans consensus clustering. The evaluation of the sample's membership in each class is given by the silhouette score. A statistical analysis of the treatments and tissues is also included.

Additional file 7 (.xlsx) Classification of the augmented data. A table containing the samples from the augmentation and their respective classes that resulted from the ATC:skmeans consensus clustering. The evaluation of the sample's membership in each class is given by the silhouette score. A statistical analysis of the treatments and tissues is also included.

Additional file 8 (.xlsx) Enrichment analysis of the signature genes in each signature group for the 199-sample list. GO terms and term clusters are only include

References

1. ^*O. Warburg. (1956). On the origin of cancer cells. Science. 123(3191):309–314. doi:10.1126/science.123.3191.309PubMed PMID: 13298683*
2. ^*O. Warburg. (1925). The Metabolism of Carcinoma Cells. J Cancer Res. 9(1):148–163. doi:10.1158/jcr.1925.148*
3. ^*Alexei Vazquez, Jurre J. Kamphorst, Elke K. Markert, Zachary T. Schug, Saverio Tardito, et al. (2016). Cancer metabolism at a glance. J Cell Sci. 129(18):3367–3373. doi:10.1242/jcs.181016*
4. ^*H. Liu, Y. P. Hu, N. Savaraj, W. Priebe, T. J. Lampidis. (2001). Hypersensitization of Tumor Cells to Glycolytic Inhibitors. Biochemistry. 40(18):5542–5547. doi:10.1021/bi002426w*

5. [a, b] *Xi-sha Chen, Lan-ya Li, Yi-di Guan, Jin-ming Yang, Yan Cheng. (2016). Anticancer strategies based on the metabolic profile of tumor cells: therapeutic targeting of the Warburg effect. Acta Pharmacol Sin. 37(8):1013–1019. doi:10.1038/aps.2016.47*
6. ^ *Ali F. Abdel-Wahab, Waheed Mahmoud, Randa M. Al-Harizy. (2019). Targeting glucose metabolism to suppress cancer progression: prospective of anti-glycolytic cancer therapy. Pharmacol Res. 150:104511. doi:10.1016/j.phrs.2019.104511*
7. ^ *Lynn Jeanette Savic, Julius Chapiro, Gregor Duwe, Jean-François Geschwind. (2016). Targeting glucose metabolism in cancer: a new class of agents for loco-regional and systemic therapy of liver cancer and beyond? Hepatic Oncol.3(1):19–28. doi:10.2217/hep.15.36*
8. ^ *Nicholas S. Akins, Tanner C. Nielson, Hoang V. Le. (2018). Inhibition of Glycolysis and Glutaminolysis: An Emerging Drug Discovery Approach to Combat Cancer. Curr Top Med Chem. 18(6):494–504. doi:10.2174/1568026618666180523111351*
9. ^ *R: The R Project for Statistical Computing. [cited 17 Sep 2021]. Available from: https://www.r-project.org/*
10. ^ *Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43(7):e47–e47. doi:10.1093/nar/gkv007*
11. ^ *R. Gentleman. annotate. Bioconductor 2017. doi:10.18129/B9.BIOC.ANNOTATE*
12. ^ *Andy Lynch Mark Dunning. illuminaHumanv3.db. Bioconductor 2017. doi:10.18129/B9.BIOC.ILLUMINAHUMANV3.DB*
13. ^ *L. Gautier, L. Cope, B. M. Bolstad, R. A. Irizarry. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 20(3):307–315. doi:10.1093/bioinformatics/btg405*
14. ^ *Benilton S. Carvalho, Rafael A. Irizarry. (2010). A framework for oligonucleotide microarray preprocessing. Bioinformatics. 26(19):2363–2367. doi:10.1093/bioinformatics/btq431*
15. ^ *Marc Carlson. hgu133plus2.db. Bioconductor 2017. doi:10.18129/B9.BIOC.HGU133PLUS2.DB*
16. ^ *James W. MacDonald. hugene10sttranscriptcluster.db. Bioconductor 2017. doi:10.18129/B9.BIOC.HUGENE10STTRANSCRIPTCLUSTER.DB*
17. ^ *James W. MacDonald. hugene11sttranscriptcluster.db. Bioconductor 2017. doi:10.18129/B9.BIOC.HUGENE11STTRANSCRIPTCLUSTER.DB*
18. ^ *James W. MacDonald. hugene20sttranscriptcluster.db. Bioconductor 2017. doi:10.18129/B9.BIOC.HUGENE20STTRANSCRIPTCLUSTER.DB*
19. ^ *James W. MacDonald. hugene21sttranscriptcluster.db. Bioconductor 2017. doi:10.18129/B9.BIOC.HUGENE21STTRANSCRIPTCLUSTER.DB*
20. ^ *Bioconductor Core Team. human.db0. Bioconductor 2017. doi:10.18129/B9.BIOC.HUMAN.DB0*
21. ^ *Marc Carlson. hthgu133a.db. Bioconductor 2017. doi:10.18129/B9.BIOC.HTHGU133A.DB*
22. ^ *The Bioconductor Project. hgu219cdf. Bioconductor 2017. doi:10.18129/B9.BIOC.HGU219CDF*
23. ^ *Marc Carlson. RnAgilentDesign028282.db. Bioconductor 2017. doi:10.18129/B9.BIOC.RNAGILENTDESIGN028282.DB*
24. ^ *Marc Carlson. HsAgilentDesign026652.db. Bioconductor 2017.*

doi:10.18129/B9.BIOC.HSAGILENTDESIGN026652.DB

25. ^*Marc Carlson. hgug4112a.db. Bioconductor 2017. doi:10.18129/B9.BIOC.HGUG4112A.DB*
26. ^*Marko Robnik-Sikonja. semiArtificial: Generator of Semi-Artificial Data. 2021. Available from: https://CRAN.R-project.org/package=semiArtificial*
27. ^*Guangchuang Yu, Li-Gen Wang, Yanyan Han, Qing-Yu He. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS J Integr Biol. 16(5):284–287. doi:10.1089/omi.2011.0118PubMed PMID: 22455463; PubMed Central PMCID: PMC3339379*
28. ^*Guangchuang Yu, Qing-Yu He. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst. 12(2):477–479. doi:10.1039/c5mb00663ePubMed PMID: 26661513*
29. ^*Zuguang Gu, Daniel Hübschmann. (2021). simplifyEnrichment: an R/Bioconductor package for Clustering and Visualizing Functional Enrichment Results. bioRxiv. :2020.10.27.312116. doi:10.1101/2020.10.27.312116*
30. ^*Katherine A. Hoadley, Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, et al. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. Cell. 158(4):929–944. doi:10.1016/j.cell.2014.06.049*
31. ^*Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 173(2):291-304.e6. doi:10.1016/j.cell.2018.03.022*