Genobolitics: Extending Metabolitics to Include Gene Expression Data

Ammar Raşid - Abdurrahman Aboudakika

February 2018

1 Introduction

Organisms' bio-systems consist of many sophisticated layers each of which can be studied extensively alone or better in context with other layers. Metabolomics is the field of measuring and analyzing the concentration, abundance or scarcity of metabolite molecules in biological systems, typically the ones smaller than 1.5 kDa. Metabolites interact with each other in a complex metabolic network of groups of closely related biochemical reactions called pathways. Metabolic networks have been studied extensively in the pathway-level. Pathways are regulated through enzymes and inhibitors, thus ultimately those reactions are regulated with genes. Recently, high-throughput omics datasets have been collected which provides the opportunity of integrating multiple layers of omics data (e.g. metabolic data and gene expression data) and perform more holistic analyses. Our goal is to combine metabolic data and gene-expression data to bridge the gap between genome-scale modelling and pathway-level analysis.

2 Literature Review

2.1 Modeling cancer metabolism on a genome scale

Genome-Scale Metabolic Modellings (GSMMs) leverage the availability of high-throughput data omics data by integrating this data into a generic human GSMM. Their goal is to predict the distribution of the network's reactions' flux rates on a genome-scale level. Their approach is a Constraint-Based Modelling (CBM) where phsysico-chemcial constraints are imposed on the space of metabolic behavior possibilities (e.g. mass balance, thermodynamic directionality and maximal flux capacity). Their optimization is done on cellular objective functions such as biomass yield or ATP production. Though GSMMs had been successful in microorganisms, the authors of this paper had to deal with the challenge of modelling more complex species (i.e. human). The challenge is manifested in that those models are not cell- or tissue-specific. Therefore, their solution space contains multiple feasible solutions, which means they need to force more constraints to achieve a better level of cell- and tissue-specificity. Moreover, some objective functions would simply do not apply to some parts of the metabolic network, e.g. the objective function of biomass yield is not appropriate for non-proliferating cells. In order to overcome this challenge they simulate genetic and environmental perturbations by changing the media composition, gene over-expression or genes knockouts. This paper is more of a survey on techniques, used in the past 10 years, of human model reconstructions and their application in various cancer types. Their paper discusses the potential of integrating more omics data, not just the mainly used transcriptomics and proteomics and what challenges are lying ahead in modelling cancer metabolism. Their conclusion, therefore, is that there are relatively sufficient and various omics data than before, and they provide the opportunity of building integrated kinetic and stoichiometric models of cancer metabolism. They discussed what has been being done, collected the power points of those previous works and proposed a generic GSMM. However, had they implemented some of their proposals and shown concrete results, their arguments would have been more cogent. We consider this paper a bootstrap for diving into the problem as it covers the typical ways of integrating and analysing omics data and their application in cancer metabolism.

2.2 Integration Analysis of Three Omics Data Using Penalized Regression Methods: An Application to Bladder Cancer

The goal of the authors is to investigate the genomic mechanisms involved in complex disease through integrating three omics data- genomics, epigenomics and transcriptomics and analyze them using penalized regression methods (i.e. LASSO and ENET), assess the significance of relationships between common genetic variants through statistical permutation-based methods, correct multiple testing with MaxT algorithm and apply their model to DNA methylation and gene expression measured in bladder tumor samples. First, they select single-nucleotide polymorphism (SNPs) and 5'—C—phosphate—G—3' (CpG)s for each gene probe with 1MB window up- and down-stream of the gene. Secondly, they estimate the associatoin of each gene and the selected SNP and CpG using LASSO and ENET. Lastly, they use the permutation-based MaxT algorithm to assess the significance of each model (i.e. SNP, CpG and Global -integrated SNP and CpG). They have identified 48 genes with significant association between their expression and both SNP and CpG. They also applied enrichment analysis through biological interpretation to support their results. It should be said that the results are visualized conveniently in the paper through Venn diagrams for the overlap between significant genes, tables for p-values of gene association with SNP and CpG in both original and validation data and graphs for deviance across the genome when applying LASSO and ENET. Since they claim their model is computationally efficient, they should have included more than just three omics data. It is arguable that 48 is a small number for significantly associated gene expression and SNP/CpG and they replicated only 36 (75%) of those genes in (TCGA) to apply a simulation study and assess the performance of the proposed method. They did not really include This paper is a typical example of integrating omics data for cancer metabolism analysis. However, their objective is to detect significant associations between gene expressions and SNPs/CpGs, while our objective is predicting the reactions' flux distribution.

2.3 Pathway-based personalized analysis of cancer.

Using gene expression data to make inference of the degree of a pathway deregulation, method transforms gene expression information into pathway level information, Aim of the paper was to create a compact, biologically meaningful representation of each sample. Pathifier assigns a score that represent the degree of pathway deregulation **in a context-specific manner**, This is done by first applying principle component analysis (PCA) on collected samples then projecting deviation scores into the principle curve, which according to the authors of this paper makes their approach more robust to perturbations such as removing some of the pathways or samples from the analysis. authors approach consist of the four phases:

1. Pathfier analysis N_p pathways, one at a time and assigns to each sample i and pathway P a score $DP(i)$.

2. To determine pathway deregulation score (PDS), the expression levels of those dP genes that belong to P are used.

3. Each sample i is represented by a point in this dP dimensional space; the entire set of samples forms a cloud of points, and we calculate the (nonlinear) "principal curve" that captures the variation of this cloud.

4. Each sample is projected into the curve obtained previously.

Paper discovered several pathways that are significantly associated with survival of Glioblastoma patients, and two pathways whose score is predictive of survival in colorectal cancer: CXCR3-mediated signaling and oxidative phosphorylation, Authors also identified a subclass of proneural and neural Glioblastoma with significantly better survival, and an EGF receptor-deregulated subclass of colon cancers. Paper made exclusive use of gene expression data, and used the gene level information to make inferences about pathway level perturbations, authors used single source of information to make inference about pathway deregulation not taking into account other types of datasets (eg., metabolomics, copy number variation, gene expression, DNA methylation and epigenetics) Pathifier shares a common objective with us of discovering pathway-level associations in diseased individuals.

2.4 Inference Of Patient-specific Pathway Activities From Multi-dimensional Cancer Genomics Data Using PARADIGM.

The aim of the authors of this paper was to detect the common pathways involved in genetic alterations between patients by integrating multiple omics data (i.e. copy number variation, gene expression, DNA methylation and epigenetics) of tumor samples and cancer cell lines. Their case study involves breast cancer whoe samples they obtained from NCBI's GEO data set, under accessions GPL5737, glioblastoma data from TCGA data portal, and for pathways, they collected a set of curated pathways available from NCI Pathway Interaction Database (PID). Their approach is based on a probabilistic assumption that genetic interactions in pathways are part of the prior for detecting correlations between gene expressions and predicting their changes in cancer-related pathways. They propose a Probabilistic Graphical Model (PGM) framework where a gene is modelled by a factor graph as a set of interconnected variables. Such model allows encoding the expression and known activity of a gene and its products while incorporating many omics data. They combine multiple genome-scale measurements from a patient sample to infer the activities of genes, products and abstract process inputs and outputs for a single pathway. Biological pathways also are modelled by factor graphs to describe the state of entities in a cell, such as mRNA, and use the factors to represent the interactions among those entities. Their approach, PARADIGM (PAthway Recognition Algorithm using Data Integration on Genomic Models), produces a matrix of integrated pathway activities (IPAs) A where A_{ij} represents the inferred activity of entity i in patient sample j. The matrix A can then be used in place of the original constituent data sets to identify associations with clinical outcomes. They infer the parameters of observed factors using an Expectation Maximization (EM). Those inferences refer to the degree of alteration in pathways' activities in patient samples. After inference, they compute a single IPA for gene i based on the log-likelihood L. If the gene is more likely to be activated, the IPA is set to L. Alternatively, if the gene is more likely to be inactivated, the IPA is set to L and 0 otherwise. Since each pathway is analyzed independently, a gene can be associated with inferences of each pathway it appears in, which implies that genes activities differ in different pathway contexts. Finally, they applied Uncentered correlation hierarchical clustering with centroid linkage on glioblastoma patient samples in order to apply Kaplan-Meier (survival rates) analysis. The proposed probabilistic approach outperforms the previous state-of-the-art, SPIA (Signaling Pathway Impact Analysis). They also assessed the significance of their Integrated Pathway Activities (IPA) scores by permuting gene expressions and copy numbers in two different ways, choosing random gene from within the same pathway and choosing a gene from anywhere in the genome. Their Kaplan–Meier analysis showed a significantly better survival profile for patients in the fourth cluster. They were able to relate this finding biological interpretations, which supports the cogency of their model. PARADIGM was also more robust than SPIA in avoiding decoy pathways; each NCI pathway was used to create a decoy pathway which consisted of the same structure but where every gene in the pathway was substituted for a random gene. Their PGM-based factor graph allowed them to incorporate multiple omics data, yet they did not cover transcriptomics, proteomics or epigenomics. Enrichment analysis and biological interpretations of their findings, though present, were scarce and arguably not sufficient. They also should have replicated at least some of the gene-expression data into another independent data set (e.g. EPICURO) for validation. Moreover, although their paper is about inferring genetic activities and interactions, the word "flux" was not mentioned in the paper not even once; they did not cover reactions' flux rates in pathways, which was likely to give even more potential the prior in their probabilistic model. Since we are aiming to integrate multiple omics data to infer patient-specific reactions' flux distribution and this paper is using integrated omics data to infer genetic associations in pathway-level and patient-specific pathway activities, we share partially similar objectives, especially in analyzing the genetic alterations in cancer-related pathways.

2.5 Personalized metabolic analysis of diseases.

The authors of this paper propose an algorithm, *Metabolitics*, that allow system-level analysis of the biochemical alterations of diseased cells. In their case study, Breast Cancer, they focus on reaction-level and pathway-level, but their algorithm is applicable to other system-levels too (e.g. gene expression). The pipeline of their approach consists of 6 phases:

1. Matching metabolites names of the human metabolic model reconstraction (in their case, Recon 2) to the ones in HDMP and CheBI data sets.

2. Dynamically creating a linear programming model; they optimize on an objective function based on a steady-state assumption -the overall reaction flux throughout the entire metabolic network is 0- to obtain the metabolites fold changes.

3. For each reaction they solve the objective function from previous phase once to maximize the flux and another to minimize it. The computed maximal and minimal flux rates are stored as the upper and lower boundaries of that reaction respectively.

4. They compute a *Diff* value for every pathway, which is basically the mean of the differences between the average of the flux upper and lower boundaries of the reference (healthy) sample and the diseased sample computed for each and every reaction in that pathway.

5. They apply ANOVA to assess the significance of the *Diff* values computed for the diseased pathways.

6. They use Machine Learning to classify metabolomics analysis results as whether they belong to healthy or patient individual using *Diff* values as features and Logistic Regression for the classifier model. They also use PCA to eliminate the highly correlated features from the vector representation before feeding those vectors to the classifier.

Metabolitics showed better results in capturing pathways with significant *Diff* values; $P - val < 5E - 5$. They also showed that using *Diff* values as features for the LR classifier accurately stratified breast cancer patients. Not only stratifying the breast cancer samples, but also *Metabolitics* revealed some insights on the changes of pathways activities through different stages of breast cancer. Though their classification accuracy was less than the last state-of-the-art, *Pathifier*, they showed that *Metabolitics* outperforms *Pathifier* when less number of metabolites measurements are used; *Metabolitics* is more resilient to data loss than *Pathifier*. They did not see a need to compare their results to *PARADIGM*, another succussful pathway-based analysis model, as *Pathifier* has been proved to outperform *PARADIGM* before. Though they had insights about the pathway alterations through different cancer stages, they did not perform a Kaplan-Meier analysis and did not mention if they had limitations that hindered them from doing so. Metabolitics has the potential of integrating multiple omics data, which makes their paper just an inception for many extensions (e.g. incorporating gene expression data as well) to come. We are using Metabolitcs algorithm after incorporating gene expression data, to bridge the gap between metabolomics and gene expression data aiming to have a holistic model with genome-scale analysis and pathway-level insights.

3 Methods

3.1 Data

Our data is a mixture of normalized gene expression fold changes in different individuals and genome-scale human metabolism model[15] such as recon2. For gene expression data, we used GDS3952 data set from Gene Expression Omnibus (GEO). In this 88-sample dataset 31 samples are of healthy patients and the rest 57 samples are for breast cancer patients. We use recon2 as our network model to access 7785 reactions and 1675 from the human metabolic network along with the boolean relations among genes associated with each reaction. The genes fold changes from the GEO dataset are used in the objective function as objective coefficients for the reactions. To maintain consistency among all samples we discard any gene that does not have a fold change available for each sample.

3.2 Genobolitics

Genobolitics is essentially an extension of Metabolitics[16], which were performed on metabolomics data. Our objective is to analyze gene-level deviations in diseased samples. Our approach is compared against Pathifier[1] and Paradigm[12] model.

3.2.1 Matching genes names

In order to appeal to the genes ids standard for cobrapy, we had to translate gene symbols in the GEO dataset to HGNC ids. For that we used pyhgnc which provides a query-based mapping from gene symbols to HGNC ids.

3.2.2 Translating boolean relations among genes

In a reaction, a gene with minimum fold change is chosen over another gene if the boolean relation between them is an AND relation. On the other hand, if the relation between two genes is an OR relation, the gene with the maximum fold change is chosen as an objective coefficient for that reaction. Cobrapy provides a string representation for genes reactions rules. We created a data structure for a gene fold change and overwritten the comparator functions for addition and subtraction of the data structure so that addition picks the maximum and subtraction picks the minimum. We then substitute ORs with plus signs and ANDs with minus signs in the gene reaction rule string which we then feed into a literal evaluation function provided by python ast library.

3.2.3 Dynamic Linear Programming

We perform a Flux Balance Analysis (FBA)[10] with a dynamically built linear programming model. Based on the steady state hypothesis, the total production of each metabolite in the metabolic network is equal to the total consumption excluding the external environmental effects where metabolites are traded with external environmental factors. The same hypothesis can arguably be made for genes associated with reactions in the metabolic network. Having calculated the fold change associated with each reaction based on its gene reaction rule, we set that calculated value as the objective coefficient of that reaction. Lest we should have infeasible solution space, we discard any reaction whose gene reaction rule does not have all fold changes available in the GEO dataset. Reactions objective coefficients, therefore, are the wall bricks of the steady state constraints on the optimization problem.

3.2.4 Flux Variability Analysis

Flux Variability Analysis (FVA)[2], in contrast to FBA, calculates both the upper and lower reaction flux boundaries. Having computed the objective coefficient from the gene reaction rule associated with a reaction, we add this reaction objective coefficient to the optimization model as an additional constraint. For every reaction R, the objective function is constructed once to maximize and once to minimize the flux of R and the computed objective function values for that reactions are the upper and lower flux boundaries of that reaction respectively.

3.2.5 Calculating Diff Values at Reaction and Pathway Levels

Reaction flux boundaries reference values are calculated by taking the average of all reference samples boundaries values computed by the objective function. Diff values are then computed for all samples by averaging the difference between the lower boundary of that sample and the reference lower boundary and the upper boundary of that sample and the reference upper boundary. Pathway-level Diff values are computed simply by taking the average of the Diff values of the reactions in the corresponding pathway.

3.2.6 Feature Extraction

Having calculated the Diff value of each pathway, features are represented as a 100 dimensional vector of Diff values where each Diff value corresponds to the deviation of that pathway in the corresponding sample from the average of the reference (healthy) samples. In order to avoid high correlation among features and thus hinder the classification model, we employ Principal Component Analysis for dimensionality reduction.

3.2.7 Machine-Learning-based Classification for disease association

The classifiers are meant to predict whether a sample is healthy or diseased based on the Diff values or reaction flux boundaries computed by the objective function in section 2.4 as features. We are using Random Forest, Multi-Layer Perceptron, Logistic Regression and Stochastic Gradient Descent classifiers separately, in 500-estimator bags or combining them in one hard voting classifier. Our metrics for assessing the classification model are precision, recall, F1-score and accuracy. In order to avoid specious metric scores, we run a stratified K-Folds validation test and take the average of each metric score of all K-folds while maintaining the ratio between the different labels in each training and testing subset in all k-folds.

3.3 Biological Interpretations

3.3.1 Statistical Significance Analysis

In order to assess the statistical significance of disease associations, we run Analysis of Variance (ANOVA) on Diff values computed in section 2.5. Using P-value and F-value as significance metrics, we assess whether Diff values of different pathways in diseased samples are significantly different than healthy samples and thus conclude that Diff values successfully capture relative biological information. Furthermore, ANOVA provides an insight of which pathways are significantly associated with certain phenotypes.

3.3.2 Pathway-level Clustering

We use hierarchical clustering to both identify significant pathways in each disease and in multiple diseases and to construct a disease ontology based on their Diff values similarity that are extracted from their gene expression array.

3.3.3 Robustness Evaluation

In order to assess how robust our model is, we drop chunks of gene expression data gradually and run the model after dropping every chunk. We have run the analysis on breast cancer dataset multiple runs. Each run with different number of samples and ratio of labels. Starting from 13 samples all the way to 597 samples, the classification accuracy ranged from 78% to 90%.

3.3.4 Visualization

We visualize the most k-significantly changing pathways in breast cancer, lung cancer and/or brain cancer based on their Diff values. We use t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction of the 100-dimensional Diff values vectors to visualize pathways and color them with colors corresponding to their labels (i.e. healthy, breast cancer, lung cancer or brain cancer). Trusting the representative power of Diff values as clustering features, each cluster should have a dominant color.

4 Results and Discussion

4.1 Classification

Nested cross-validation (Nested CV) was used to select the classification pipeline, model selection without nested CV uses the same data for model tuning and evaluation, and as a result data may leak into the model (Cawley and Talbot [1]). Nested CV consist of two loops in the inner loop model selecting/tuning is done and in the outer loop, generalization error is estimated by averaging test scores over the different splits, 10 stratified k-folds were used for the inner and outer CV loops, this procedure was repeated 10 times (trials) and resulting test scores was averaged to estimate model generalization error.

Using aforementioned procedure we obtained our classification pipeline, final pipeline consisted of a Principal Component Analysis (PCA) stage followed by a Support Vector Machine classifier (SVM) with a linear kernel, Classification error on underrepresented classes was weighted according to the inverse of class weights to account for the imbalance in the distribution of training dataset during training phase. Classifier was tested

on a lung-cancer datasets obtained from GEO (Gene Expression Omnibus) GDS3257 and a breast-cancer dataset obtained from UCSC Xena (University Of California Santa Cruz datahub) Agilent G4502A , datasets consisted of 107 (49 healthy and 58 unhealthy) and 598 (64 healthy and 534 unhealthy) samples respectively. Classifier is tested with and without diff-scores (raw values) and classification accuracy is compared.

UCSC Xena's Agilent G4502A Breast Cancer dataset. Using raw values an Accuracy, F1 Score, Precision and Recall of 0.898602, 0.945778, 0.907145, 0.988763 respectively was obtained for classification (mean of the 10 trials' scores is used to report the metrics), While an Accuracy, F1, Precision and Recall of 0.907263, 0.947750, 0.952113, 0.944375 was obtained for classification on diff. scores (mean of the 10 trials' scores is used to report the metrics). Anova (Analysis of variance) was then applied on the features(pathways) to obtain the pathways that correlate best with the classification, Multiple testing was applied after correcting p values for multiple testing (using Bonferroni correction with a family-wise error rate of 0.05), Most significant pathways (based on Raw-values) are Nucleotide interconversion, Purine synthesis, Pyrimidine synthesis and Vitamin A metabolism, All of which being in common with the most significant pathways obtained from Metabolitics.

On	Accuracy	F1	Precision	Recall
Raw values	0.898602	0.945778	0.907145	0.988763
Diff. score	0.907263	0.947750	0.952113	0.944375
Base classifier	0.903323	0.4717314	0.892976	1.0

Table 1: Classification accuracy on breast cancer

Reaction	Corrected P Value
RDH3	3.86802e-07
RDH2	3.86802e-07
r2311	1.1918e-06
ACACt2	2.76820e-06
r2308	5.86771e-06
2HBt2	8.1529e-06
r2309	1.07530e-05
RAI1	1.390341e-05
NDPK1m	2.32484e-05
r2310	2.988401e-05

Table 2: Significant reactions for breast cancer dataset

GEO GDS3257 Lung cancer dataset. Using diff. scores an Accuracy, F1 Score, Precision and Recall of 0.827717, 0.830682, 0.884762, 0.805333 respectively was obtained for classification ,Obtained most significant pathways include 10 pathways, Blood group synthesis, Cholesterol metabolism, Eicosanoid metabolism, Fatty acid oxidation, Folate metabolism, Inositol phosphate metabolism, Nucleotide interconversion, Phosphatidylinositol phosphate metabolism, Pyrimidine synthesis, Steroid metabolism, Thiamine metabolism, Urea cycle, all of which are linked to cancer in literature (Itzkowitz, Steven H., et al. 1990) [4], (Luo, Xiangjian, et al 2010) [5], (Wang, Dingzhi, and Raymond N. DuBois 2010)[14], (Monaco, Marie E. 2017)[7], (Ulrich, Cornelia M. 2005) [11], (Vivanco, Igor, and Charles L. Sawyers. 2002) [13], (Mitrunen, Katja, et al. 2000) [6], (Nagamani, Sandesh CS, and Ayelet Erez. 2016)[8]

On	Accuracy	F1	Precision	Recall
Raw values	0.827717	0.830682	0.884762	0.805333
Diff. score	0.827717	0.830682	0.884762	0.805333
Base classifier	0.685897	0.351515	0.542056	1.0

Table 3: Classification accuracy on lung cancer dataset

Pathway	Corrected P Value
Blood group synthesis	2.08690647e-02
Cholesterol metabolism	1.12970569e-04
Eicosanoid metabolism	2.36918749e-05
Fatty acid oxidation	8.29606021e-05
Folate metabolism	8.85016909e-07
Inositol phosphate metabolism	1.33076578e-02
Nucleotide interconversion	1.31230972e-02
Phosphatidylinositol phosphate metabolism	4.05578296e-02
Pyrimidine synthesis	1.80736248e-08
Steroid metabolism	1.03422531e-02
Thiamine metabolism	2.51005162e-03
Urea cycle	1.43497777e-03

Table 4: Significant pathways for lung cancer

4.2 Clustering

The aim of clustering is to both construct a disease ontology and identify significantly similar pathways in groups of diseases. Using various dimensionality reduction algorithms to visualize and computed clustering metrics, we got the results shown in the following table:

Metric	LDA$_2d$	PCA$_2d$	k-best$_2d$	t-SNE$_2d$	LDA$_3d$	PCA$_3d$	k-best$_3d$	t-SNE$_3d$
completeness	0.378431	0.146051	0.165582	0.161552	0.433107	0.138710	0.180055	0.095596
homogeneity	0.514187	0.192522	0.222582	0.222492	0.582785	0.175087	0.232406	0.130141
v$_m$*easure*	0.435986	0.166098	0.189897	0.187187	0.496920	0.154790	0.202908	0.110225

Table 5: Clustering Metrics on Diff Values

Metric	LDA$_2d$	PCA$_2d$	k-best$_2d$	t-SNE$_2d$	LDA$_3d$	PCA$_3d$	k-best$_3d$	t-SNE$_3d$
completeness	0.924294	0.198222	0.332398	0.232236	0.930902	0.138710	0.332398	0.126032
homogeneity	0.991393	0.260313	0.272538	0.316967	0.991870	0.175087	0.272538	0.174126
v$_m$*easure*	0.956669	0.225064	0.299506	0.268065	0.960420	0.154790	0.299506	0.146226

Table 6: Clustering Metrics on Raw Values

In order to construct a disease ontology, hierarchical clustering is used on 19 different diseases. Using Diff values in hierarchical clustering proved essential as the raw values resulted in recursion errors, that is, the clustering algorithm cannot make use of the raw values without processing them into Diff values.

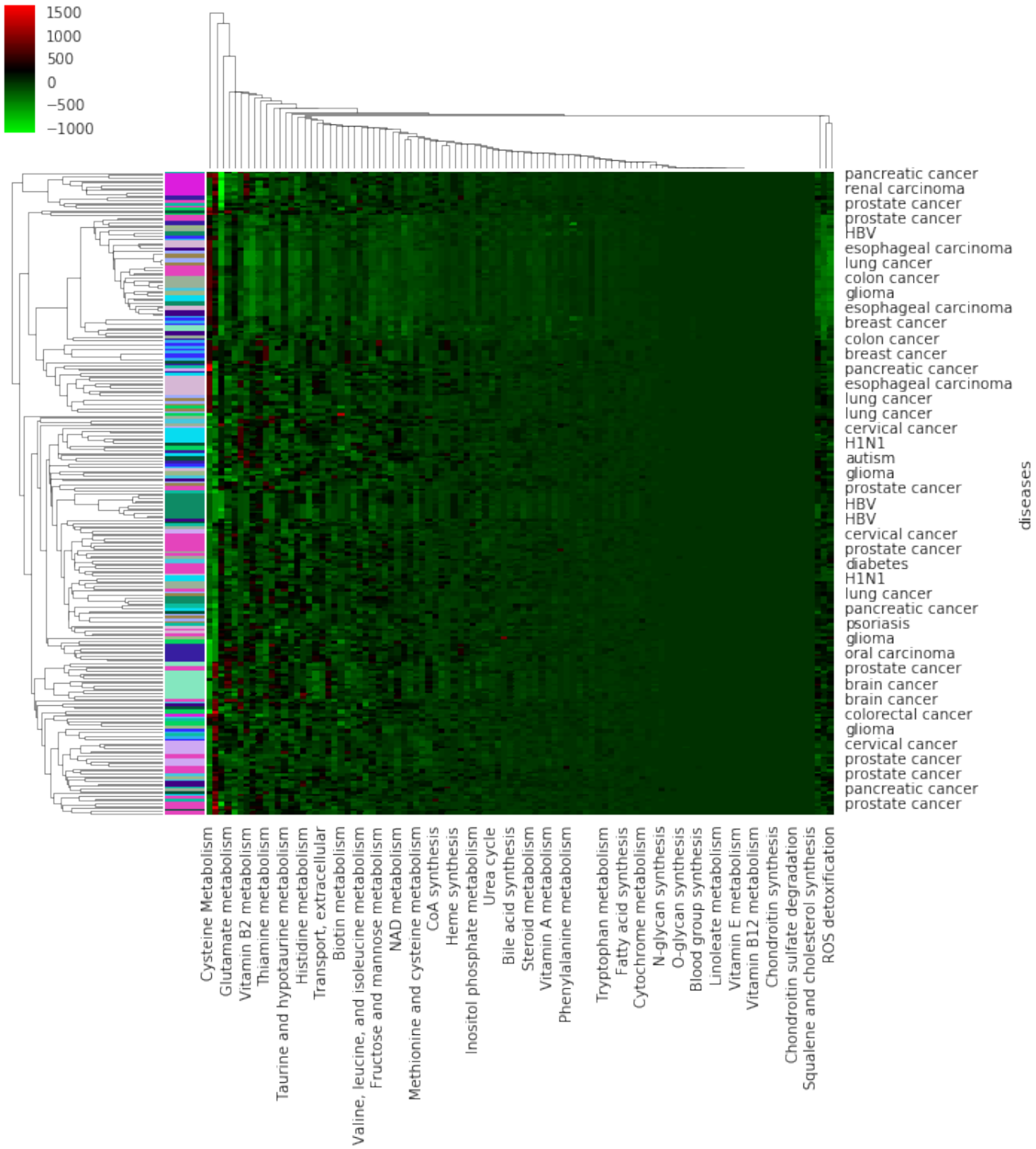

Figure 1: Clustermap of 19 diseases with with the Diff values of 101 Pathways

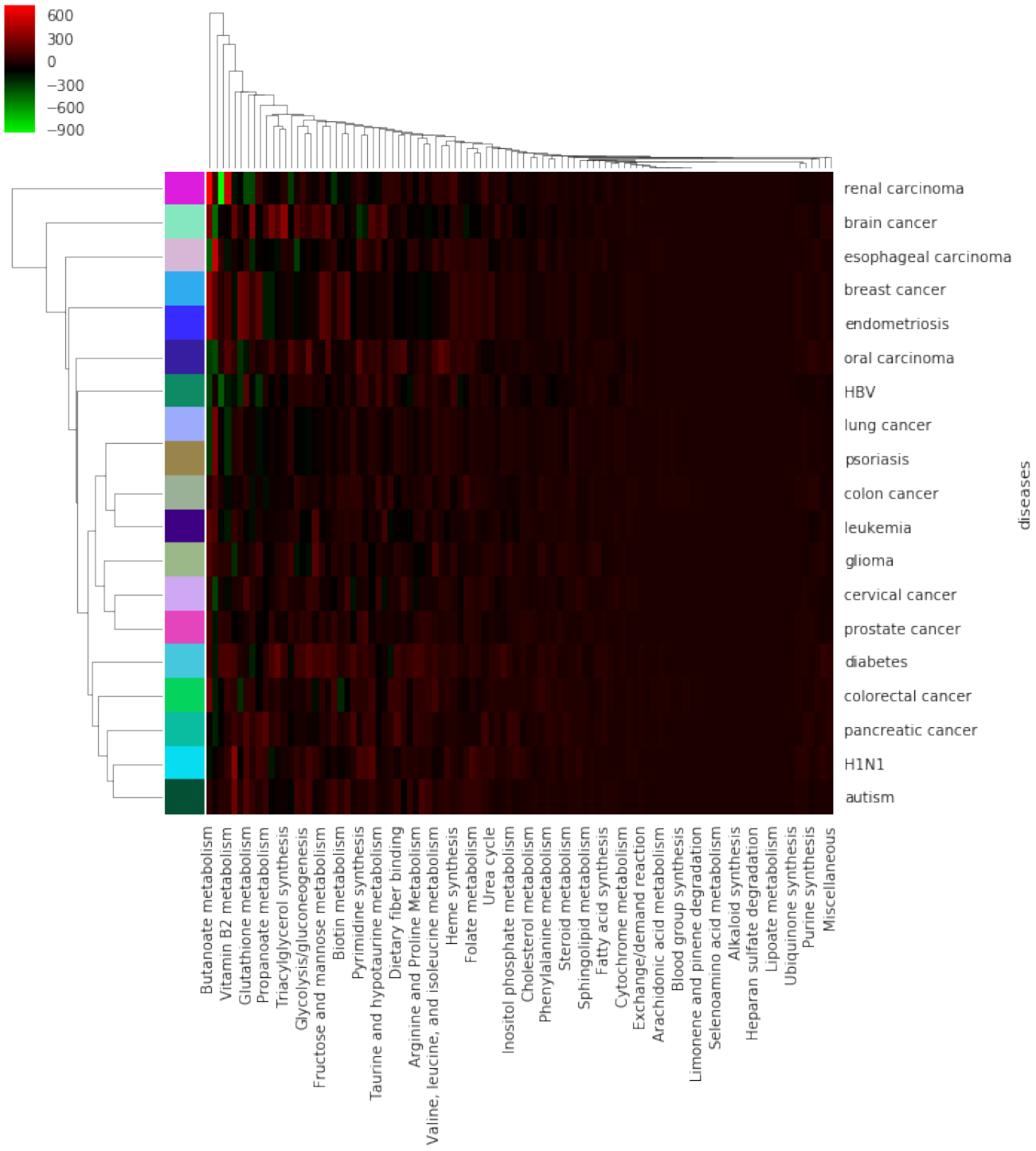

Figure 2: Clustermap - taking the average of all diff values of all samples for each disease

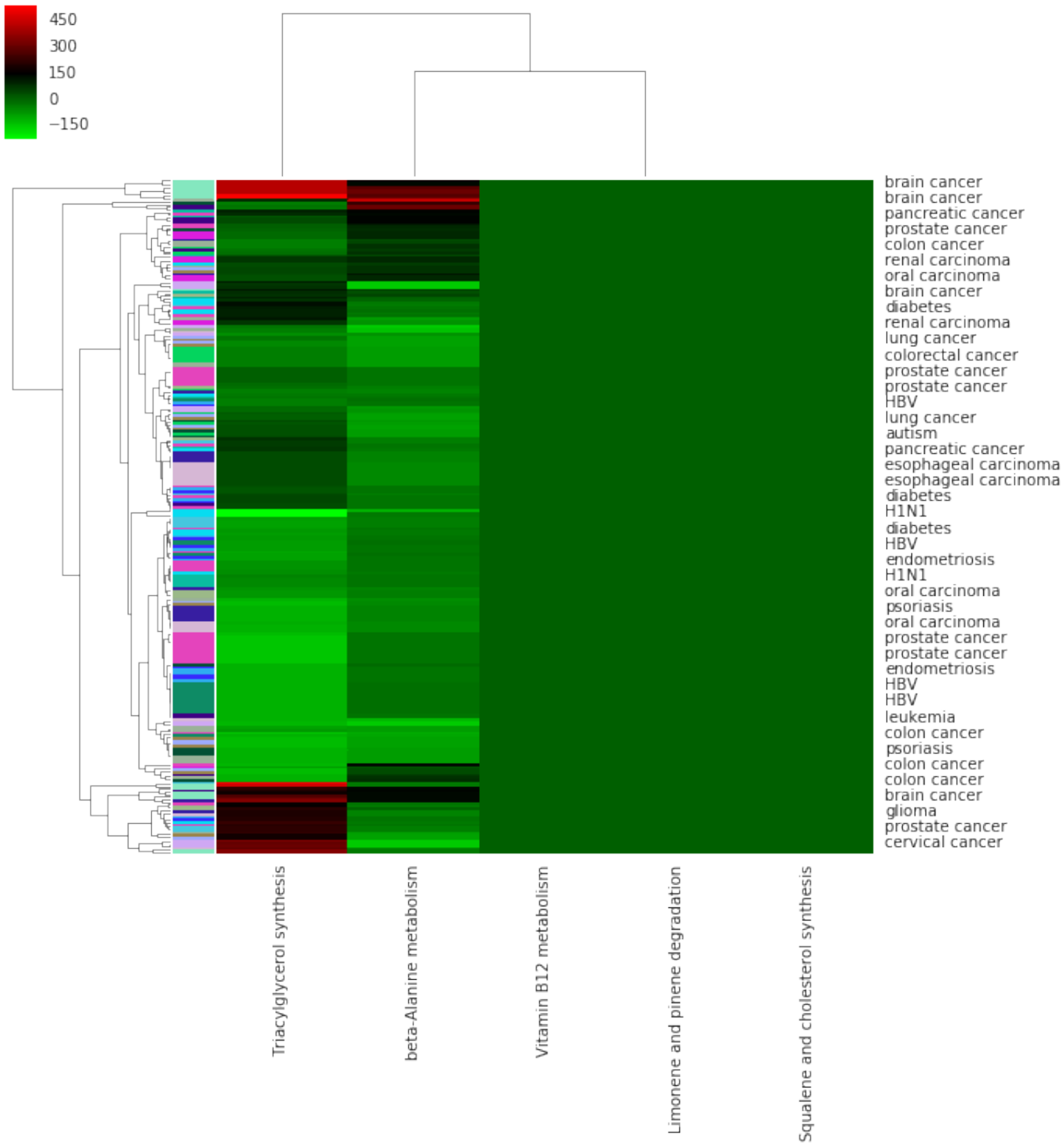

Figure 3: Clustermap using only the most significant 5 pathways

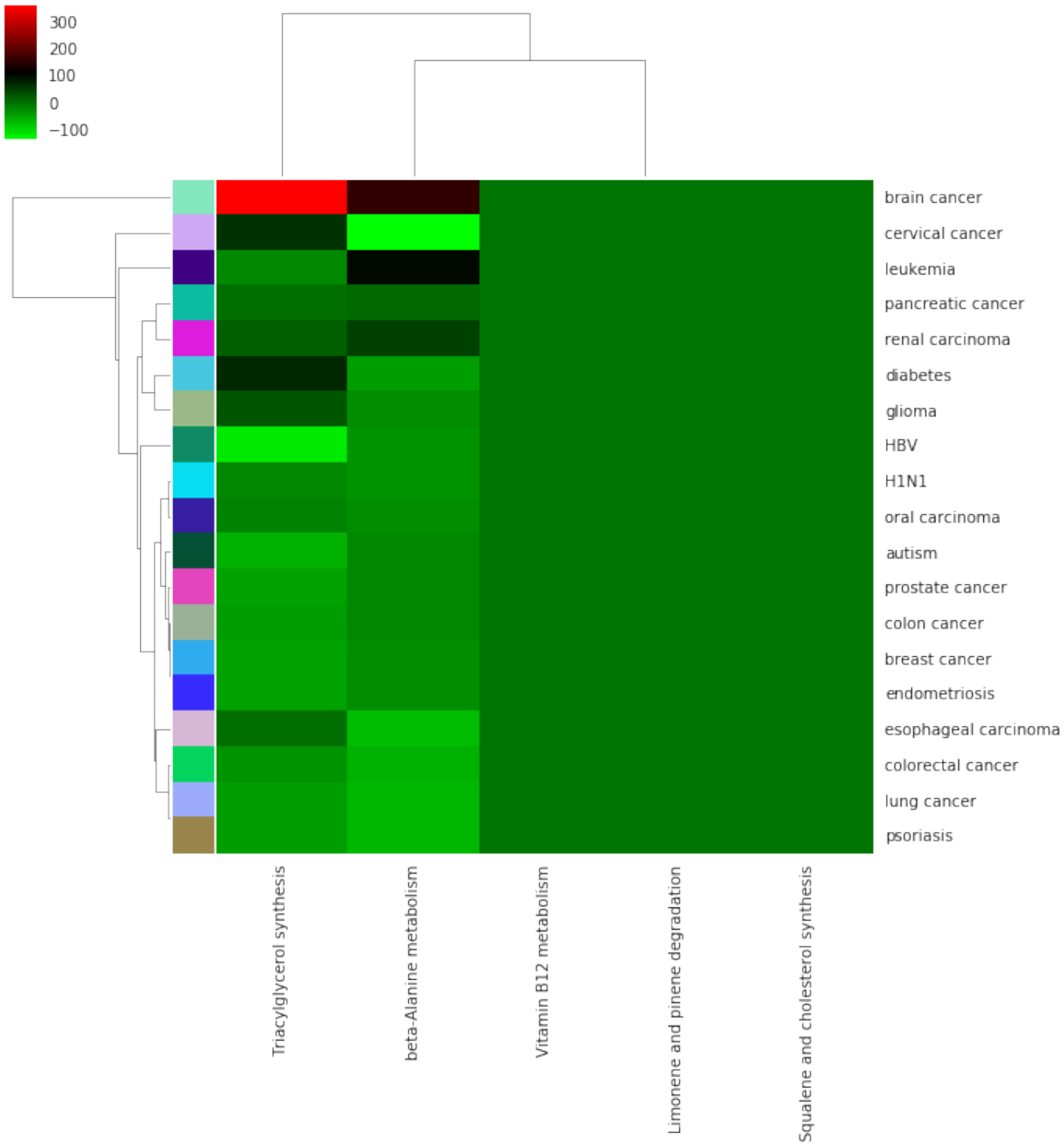

Figure 4: Clustermap - taking the average of all diff values of all samples for each disease, using only the most significant 5 pathways

Taking the average of the Diff values of all the samples for each disease show that on average most pathways on average have less activities in diseased states. Taking the k-most significant pathways across all the 19 diseases show that Triacylglycerol synthesis and beta-Alanine metabolism play a significant role on average for all diseases. Triacylglycerol synthesis pathway, which is in the heat-map show a significant influence in brain cancer, is involved in the metabolism of fatty acids by cultured neuroblastoma and glioma cells [3]. On the other hand cervical cancer in the heat-map is shown to be influenced by a low activity in

beta-Alanine metabolism pathway, which is proved in literature to indeed play a significant role in cervical and renal tumors [9].

5 Conclusion

In this paper we have shown that leveraging the relatively abundant gene expression data to extract relative features. Gene expression array of an individual can identify what disease the individual has, if any, which phase of the disease s/he is going through and which pathways are most significant to address in his or her treatment plan. A disease ontology can be constructed based on the features extracted from gene expression data pertaining to samples of patients of various diseases. Our model identifies the significant pathways in each disease and the similarity among pathways across multiple diseases. The correlation of certain pathways in diseases is captured by gene expression data. Discovering the similarity among pathways has the potential to generalize drugs used for certain pathways to other similar pathways.

6 Future work

Genobolitics has the potential of integrating different omics data such as methylation and transcription factors which could improve the performance of the model and provide a more thorough coverage of the metabolic network.

7 Acknowledgements

We would like to thanks Prof. Ali Çakmak and Muhammed Hasan Çelik for their support and help throughout the research cycle.

8 Code availability

The code used for this project is available at: `https://github.com/AmmarRashed/Genobolitics`

References

[1] Yotam Drier, Michal Sheffer, and Eytan Domany. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, 110(16):6388–6393, 2013.

[2] Steinn Gudmundsson and Ines Thiele. Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11(1):489, Sep 2010.

[3] J. T. R. Clarke H. W. Cook and M. W. Spence. Involvement of triacylglycerol in the metabolism of fatty acids by cultured neuroblastoma and glioma cells. *Journal of Lipid Research*, 23, 1982.

[4] Steven H Itzkowitz, Rajvir Dahiya, James C Byrd, and Young S Kim. Blood group antigen synthesis and degradation in normal and cancerous colonic tissues. *Gastroenterology*, 99(2):431–442, 1990.

[5] Xiangjian Luo, Can Cheng, Zheqiong Tan, Namei Li, Min Tang, Lifang Yang, and Ya Cao. Emerging roles of lipid metabolism in cancer metastasis. *Molecular cancer*, 16(1):76, 2017.

[6] Katja Mitrunen, Nadejda Jourenkova, Vesa Kataja, Matti Eskelinen, Veli-Matti Kosma, Simone Benhamou, Harri Vainio, Matti Uusitupa, and Ari Hirvonen. Steroid metabolism gene cyp17 polymorphism and the development of breast cancer. *Cancer Epidemiology and Prevention Biomarkers*, 9(12):1343–1348, 2000.

[7] Marie E Monaco. Fatty acid metabolism in breast cancer subtypes. *Oncotarget*, 8(17):29487, 2017.

[8] Sandesh CS Nagamani and Ayelet Erez. A metabolic link between the urea cycle and cancer cell proliferation. *Molecular & cellular oncology*, 3(2):e1127314, 2016.

[9] Muthuraman Pandurangan, Gansukh Enkhtaivan, Bhupendra Mistry, Rahul V. Patel, Sohyun Moon, and Doo Hwan Kim. β-alanine intercede metabolic recovery for amelioration of human cervical and renal tumors. *Amino Acids*, 49(8):1373–1380, Aug 2017.

[10] Karthik Raman and Nagasuma Chandra. Flux balance analysis of biological systems: applications and challenges. *Briefings in Bioinformatics*, 10(4):435–449, 2009.

[11] Cornelia M Ulrich. Nutrigenetics in cancer research—folate metabolism and colorectal cancer. *The Journal of Nutrition*, 135(11):2698–2702, 2005.

[12] Charles J. Vaske, Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–i245, 2010.

[13] Igor Vivanco and Charles L Sawyers. The phosphatidylinositol 3-kinase–akt pathway in human cancer. *Nature Reviews Cancer*, 2(7):489, 2002.

[14] Dingzhi Wang and Raymond N DuBois. Eicosanoids and cancer. *Nature Reviews Cancer*, 10(3):181, 2010.

[15] Keren Yizhak, Barbara Chaneton, Eyal Gottlieb, and Eytan Ruppin. Modeling cancer metabolism on a genome scale. *Molecular Systems Biology*, 11(6), 2015.

[16] A. Çakmak and M. H. Çelik. Personalized metabolic analysis of diseases. *Systems Biology*, 2017.