doi:10.20944/preprints202011.0474.v1

Article

Identifying personalized metabolic signatures in breast cancer

Priyanka Baloni[1], Wikum Dinalankara[2], John C. Earls[1], Theo A. Knijnenburg[1], Donald Geman[3], Luigi Marchionni[2,4*], and Nathan D. Price[1*]

[1] Institute for Systems Biology, Seattle, Washington, USA
[2] Department of Oncology, Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
[3] Department of Applied Mathematics and Statistics, Johns Hopkins University, USA
[4] Weill-Cornell Medicine, New York, NY, USA

* Correspondence: nprice@isbscience.org (N.D.P.); lum4003@med.cornell.edu (L.M.)

Abstract: Cancer cells are adept at reprogramming energy metabolism and the precise manifestation of this metabolic reprogramming exhibits heterogeneity across individuals (and from cell to cell). In this study, we analyzed the metabolic differences between interpersonal heterogeneous cancer phenotypes. We used divergence analysis on gene expression data of 1156 breast normal and tumor samples from The Cancer Genome Atlas (TCGA) and integrated this information with a genome-scale reconstruction of human metabolism to generate personalized, context-specific metabolic networks. Using this approach, we classified the samples into four distinct groups based on their metabolic profiles. Enrichment analysis of the subsystems indicated that amino acid metabolism, fatty acid oxidation, citric acid cycle, androgen and estrogen metabolism and ROS detoxification distinguished these four groups. Additionally, we developed a workflow to identify potential drugs that can selectively target genes associated with the reactions of interest. MG-132 (a proteasome inhibitor) and OSU-03012 (a celecoxib derivative) were the top-ranking drugs identified from our analysis and known to have anti-tumor activity. Our approach has the potential to provide mechanistic insights into cancer-specific metabolic dependencies, ultimately enabling the identification of potential drug targets for each patient independently, contributing to a rational personalized medicine approach.

Keywords: Breast cancer, genome-scale metabolic models, constraint-based analysis, divergence analysis, gene expression, metabolism, drug targets, personalized metabolic networks.

1. Introduction

The physiological state of a cell is influenced by underlying metabolic processes which exhibit high degrees of heterogeneity across patients and across cells. Cancer cells reprogram their energy metabolism as is needed to meet the energy demands of proliferation and migration. The mechanisms of invasion and metastasis are complex and mortality is mainly caused by progression of cancer to metastatic state [1]. Alteration of interactions between cancer cells and their microenvironment leads to diverse outcomes in the programmed behavior of the cells. Tumor cells exhibit heterogeneous metabolic profiles, with differential utilization of metabolites such as glucose, lactate, glutamine and glycine [2]. Some of the metabolic and genetic changes that are reported in tumor cells are enhanced glycolysis, differential expression of lactate dehydrogenase A (LDH) which is linked with cancer growth and metastasis, mutations in metabolic enzymes such as isocitrate dehydrogenase 1 (IDH1),

doi:10.20944/preprints202011.0474.v1

succinate dehydrogenase (SDH) and fumarate hydratase (FH) involved in initiating tumors [3]. These findings suggest that metabolism is fundamental in determining the cell fate in cancer and should be explored further. Various omics measurements from diverse cancer cell lines have made it easier to study the physiological changes. Integration of these omics measurements with computational models increases the accuracy of predictions.

Transcriptome analysis provides a genome-wide snapshot of differential gene activity, providing important information about key genes that modulate metabolism at the system level. Transcriptomes are complex data types with a high degree of person-to-person heterogeneity that can obfuscate the underlying biological signal, hindering their use in practice. To partially address this issue, we have recently introduced "divergence analysis" [4], a simplified and personalized data representation that captures the departure of omics profiles from a normal reference baseline. Divergence analysis of breast cancer samples in TCGA [4] has been useful in measuring the degree of divergence for genes and other genomic features in cancer versus the normal baseline phenotype, as well as one cancer phenotype versus another. Divergence is a single sample property (unlike e.g. a differentially expressed gene) and our previous work has shown that divergence encoding largely preserves biological signals and helps removing unwanted noise from the data [4]. It is therefore helpful for data preprocessing before complex system-level analyses, including metabolic network modeling.

Combining biological data and modeling enables us to study complex interactions in a biological system. Integrating transcriptomics data onto a genome-scale metabolic network to perform network-level simulations is a useful step to regularize the data and attempt to infer metabolic states from the combined evidence of the enzymes that are expressed in the transcriptome as a whole. Many computational methods for metabolic modeling have been developed [5,6]. Genome-scale metabolic models (GSMs) provide comprehensive information about known genes, metabolites and reactions in organisms and are useful to infer metabolic differences between conditions [7–9]. These models have been used to predict changing metabolic landscapes in cancers and also predict candidate drug targets and biomarkers of cancer [10–13].

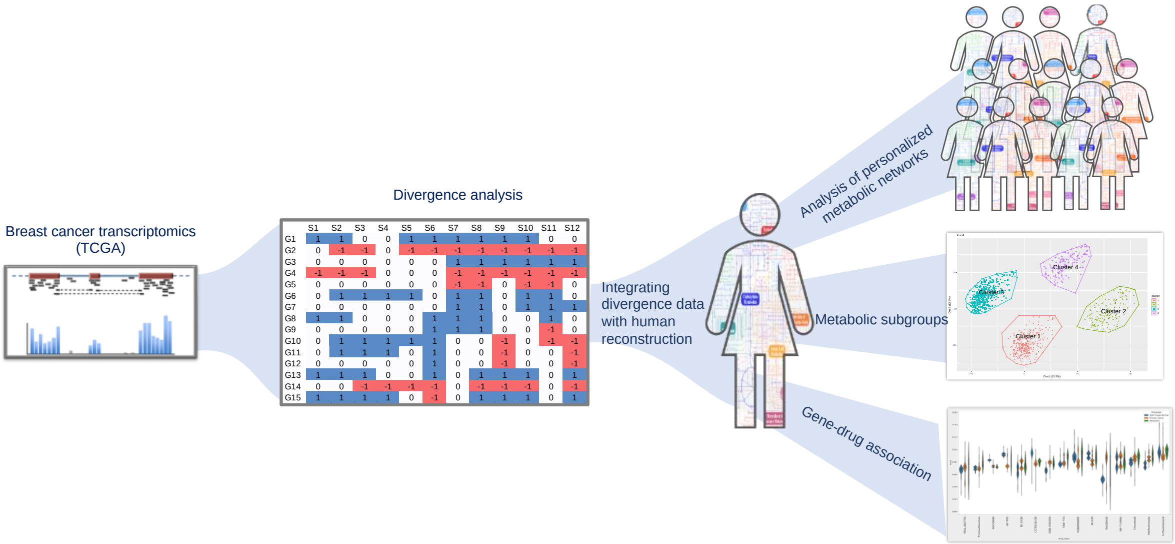
The main contributions of the present work are three-fold: (1) we generate context-specific metabolic networks for 1156 cancer and normal samples by integrating their divergence profiles with a global human metabolic network reconstruction; (2) we develop a framework for identifying key metabolic and regulatory signatures and used it to classify the samples in breast cancer based on their metabolic state; (3) we perform in silico gene knockout in these 1156 context-specific metabolic networks and identify genes that can perturb the system, many of which correspond to known drug targets. Thus, our study provides a novel assessment of metabolic network analysis based on divergence encoding. Herein, we have employed this strategy for breast cancer, but our method can be extended to other cancers and metabolically perturbed diseases to identify key metabolic signatures and potential drug targets.

2. Results

2.1 Understanding metabolic differences in cancer samples using personalized metabolic networks

doi:10.20944/preprints202011.0474.v1

In this study, we used gene expression estimates, encoded as binarized divergence indicators or as TPM values, from 1156 cancer and normal samples from TCGA (https://www.cancer.gov/tcga), and integrated them with a human metabolic model (Recon 3D) [14] to obtain personalized metabolic networks for each sample. This approach allowed us to predict distinct metabolic signatures for each individual sample and classify them according to their metabolic phenotype. In this study, we have referred to personalized metabolic networks generated from divergence and transcriptome analysis as 'divergent networks' and 'normalized networks', respectively. An overview of work done in this study is represented in Figure 1.


Breast cancer transcriptomics (TCGA)
Divergence analysis
Integrating divergence data with human reconstruction
Analysis of personalized metabolic networks
Metabolic subgroups
Cluster 4
Cluster 3
Cluster 2
Cluster 1
Gene-drug association

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
G1	1	1	0	0	1	1	1	1	1	1	0	0
G2	0	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1
G3	0	0	0	0	0	0	1	1	1	1	1	1
G4	-1	-1	-1	0	0	0	-1	-1	-1	-1	-1	-1
G5	0	0	0	0	0	0	-1	-1	0	-1	-1	0
G6	0	1	1	1	1	0	1	1	0	1	1	0
G7	0	0	0	0	0	0	1	1	0	1	1	1
G8	1	1	0	0	0	1	1	1	0	0	1	0
G9	0	0	0	0	0	1	1	1	0	0	-1	0
G10	0	1	1	1	1	1	0	0	-1	0	-1	-1
G11	0	1	1	1	0	1	0	0	-1	0	0	-1
G12	0	0	0	0	0	1	0	0	-1	0	0	-1
G13	1	1	1	0	0	1	0	1	1	1	0	1
G14	0	0	-1	-1	-1	-1	0	-1	-1	-1	0	-1
G15	1	1	1	1	0	-1	0	1	1	1	0	1

Figure 1: Overview of the study design. The breast cancer expression dataset from TCGA was converted to ternary format using divergence analysis (shown in the middle panel) [4]. The divergence values were integrated with human reconstruction and pruned to obtain personalized metabolic networks. The right side of the figure panel shows the identification of metabolic subgroups in the samples using unsupervised clustering. From our analysis we identified important reactions and genes in cancer versus normal and used this information to associate drugs that can target them (bottom panel on right).

2.2 Classifying cancer samples based on their metabolic profile

We used genes present in human reconstruction (Recon 3D) and mapped the divergence values for solid tissue normal, primary tumor and metastatic samples. Principal component analysis of metabolic gene expression in these samples showed two clusters but the normal samples could not be differentiated from cancerous ones (Figure 2a). This suggested that expression profiling is not sufficient to distinguish the samples and classify them. We performed a similar analysis with TPMs and failed to identify a clear clustering of the samples (Supplementary figure 4). To obtain a better understanding of perturbations in the system, we integrated divergence and TPM values with the human metabolic model using iMAT method and generated context-specific metabolic networks for 1156 primary tumor, metastatic and normal tissue samples. We observed distinct clusters for cancer (primary and metastatic) and normal samples, using fluxes measured for reactions in the context-specific networks (Figure 2b). The primary and metastatic samples were mixed in the cancer cluster. This suggested that metabolic networks were able to distinguish various phenotypes and can be used to understand mechanistic changes in the system.

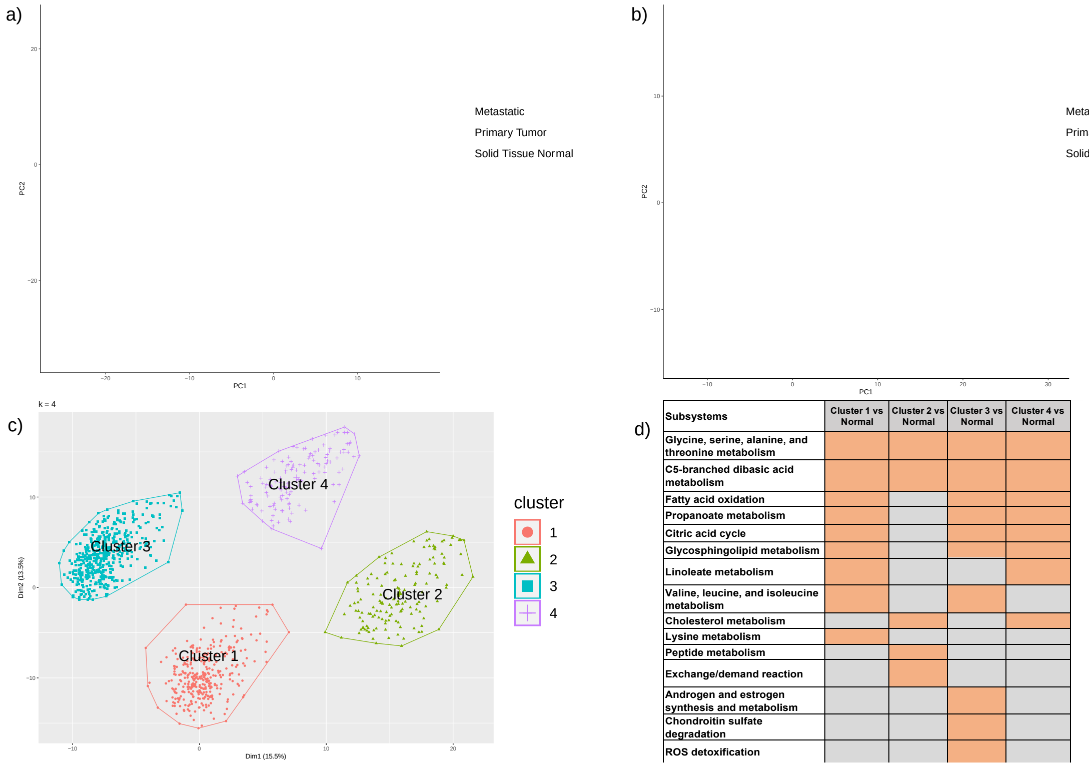
doi:10.20944/preprints202011.0474.v1

(a) *Class comparison*: We compared the reaction fluxes for cancer and normal samples in the dataset and classified reactions in each context-specific network as active or inactive based on their flux measurement (described in methods section). In order to identify active reactions in the context-specific networks, we used the information of reaction fluxes from all 1156 context-specific metabolic networks. If a reaction was present in the network, it was assigned a state of 1, while the remaining reactions were assigned a state of 0 indicating that they were absent in the context-specific metabolic network. Statistical analysis of active reactions in divergent networks identified 471 reactions (p-value < 0.05) that were significantly different in cancer versus normal. These reactions belonged to the following pathways: androgen and estrogen metabolism, bile acid synthesis, cholesterol metabolism, citric acid cycle, drug metabolism, eicosanoid metabolism, exchange reactions, fatty acid oxidation, glutathione metabolism, glycerophospholipid metabolism, glycolysis, steroid metabolism, transport, tyrosine metabolism, urea cycle, and vitamin metabolism. Supplementary Table 1 represents the list of subsystems that were enriched in cancer versus normal.

(b) *Class discovery*: We used an unsupervised machine learning method to classify the cancer samples based on their metabolic state. Using K-means clustering on the simulated reaction fluxes, we obtained four distinct clusters of cancer samples (Figure 2c). The number of clusters was determined by the elbow method; see Supplementary figure 2. The cancer clusters were then labeled from 1-4 and normal tissue samples were assigned as cluster 0. We performed a detailed analysis of the four clusters to identify, if any, associations with standard clinical and pathological tumor characteristics. This analysis showed that the metabolic clusters were significantly associated with PAM50 molecular subtypes and ER status (chi-squared p-value < 0.001), distinguishing the luminal A and B samples from basal-like samples, and also ER positive and negative samples to a greater extent. Specifically, cluster 2 was enriched for luminal subtypes (luminal A and B) and predominantly accounted for ER positive samples, while cluster 3 was enriched in basal-like and ER negative tumors. (Figure 3 and Supplementary file 1). The metabolic clusters of tumor and normal samples were used for identifying important reactions and sub-systems in these clusters.

In addition to identifying differences between cancer and normal phenotypes, we extended our analysis to subsystems that are enriched for each identified cluster. The heatmap of enriched subsystems in cancer versus the normal samples, as shown in Figure 2d, indicated that glycine, serine, alanine and threonine metabolism and C5-branched dibasic acid metabolism were enriched in all the clusters. Fatty acid oxidation, propanoate metabolism, citric acid cycle and glycosphingolipid metabolism were enriched for cluster 1, 3 and 4, whereas cluster 2 showed selective enrichment for peptide metabolism and exchange reactions. Androgen and estrogen metabolism, chondroitin sulphate degradation and ROS detoxification were selectively enriched for cluster 3 samples, indicating that each cluster had a distinct metabolic profile and we can probe their metabolic differences. We compared the reactions in each cluster with respect to those in normal samples and identified 254, 1388, 581, and 324 reactions that were significant in cluster 1-4 respectively (Supplementary figure 3).

doi:10.20944/preprints202011.0474.v1



Subsystems	Cluster 1 vs Normal	Cluster 2 vs Normal	Cluster 3 vs Normal	Cluster 4 vs Normal
Glycine, serine, alanine, and threonine metabolism	🟧	🟧	🟧	🟧
C5-branched dibasic acid metabolism	🟧	🟧	🟧	🟧
Fatty acid oxidation	🟧	⬜	🟧	🟧
Propanoate metabolism	🟧	⬜	🟧	🟧
Citric acid cycle	🟧	⬜	🟧	🟧
Glycosphingolipid metabolism	🟧	⬜	🟧	🟧
Linoleate metabolism	🟧	⬜	⬜	🟧
Valine, leucine, and isoleucine metabolism	🟧	⬜	🟧	⬜
Cholesterol metabolism	⬜	🟧	⬜	🟧
Lysine metabolism	🟧	⬜	⬜	⬜
Peptide metabolism	⬜	🟧	⬜	⬜
Exchange/demand reaction	⬜	🟧	⬜	⬜
Androgen and estrogen synthesis and metabolism	⬜	⬜	🟧	⬜
Chondroitin sulfate degradation	⬜	⬜	🟧	⬜
ROS detoxification	⬜	⬜	🟧	⬜

Figure 2: Cluster analysis of genes and reactions

a) PCA of metabolic genes (divergent values) of 1156 breast cancer samples from TCGA. Samples are colored as brown, green and pink based on normal, primary or metastasis phenotype, respectively.

b) PCA of 1156 samples clustered based on metabolic reaction fluxes and colored with respect to sample type. Samples are colored as brown, green and pink based on normal, primary or metastasis phenotype, respectively.

c) Four clusters of cancer samples indicating distinct metabolic profiles. The clusters have been labelled as 1, 2, 3 and 4.

d) Heatmap representing enriched subsystem for each cluster when compared to normal samples. Orange fields indicate significant subsystems with p-value <0.05 and the gray fields indicate non-significant subsystems with p-value > 0.05.

We extended our analysis to identify which types of samples were enriched in each of the clusters. We mapped information of PAM (Prediction Analysis of Microarray) 50 classifier for breast tumor intrinsic subtyping, known ER status, triple negative status of samples, American Joint Committee on Cancer (AJCC) stage and vital status for samples in the cluster and obtained interesting results and performed chi-squared statistics for these clusters. We found that cluster 2 had a higher proportion of HER2-enriched, luminal A and luminal B samples whereas cluster 4 had higher proportion of basal-like samples (Figure 3a). When we looked at the ER status of the samples, we observed that Cluster 2 had a higher proportion of ER positive samples and cluster 4 had a higher number of samples that were ER negative (Figure 3b). Cluster 2 and 4 had a higher proportion of samples with known cases of triple negative status (Figure 3c). For samples with known AJCC stages, we observed that cluster 2 had a higher proportion of samples that belonged to stage II (Figure 3d).

This suggests that samples belonging to cluster 2 have a distinct metabolic profile and are able to distinguish tissue type and known markers of breast cancer.

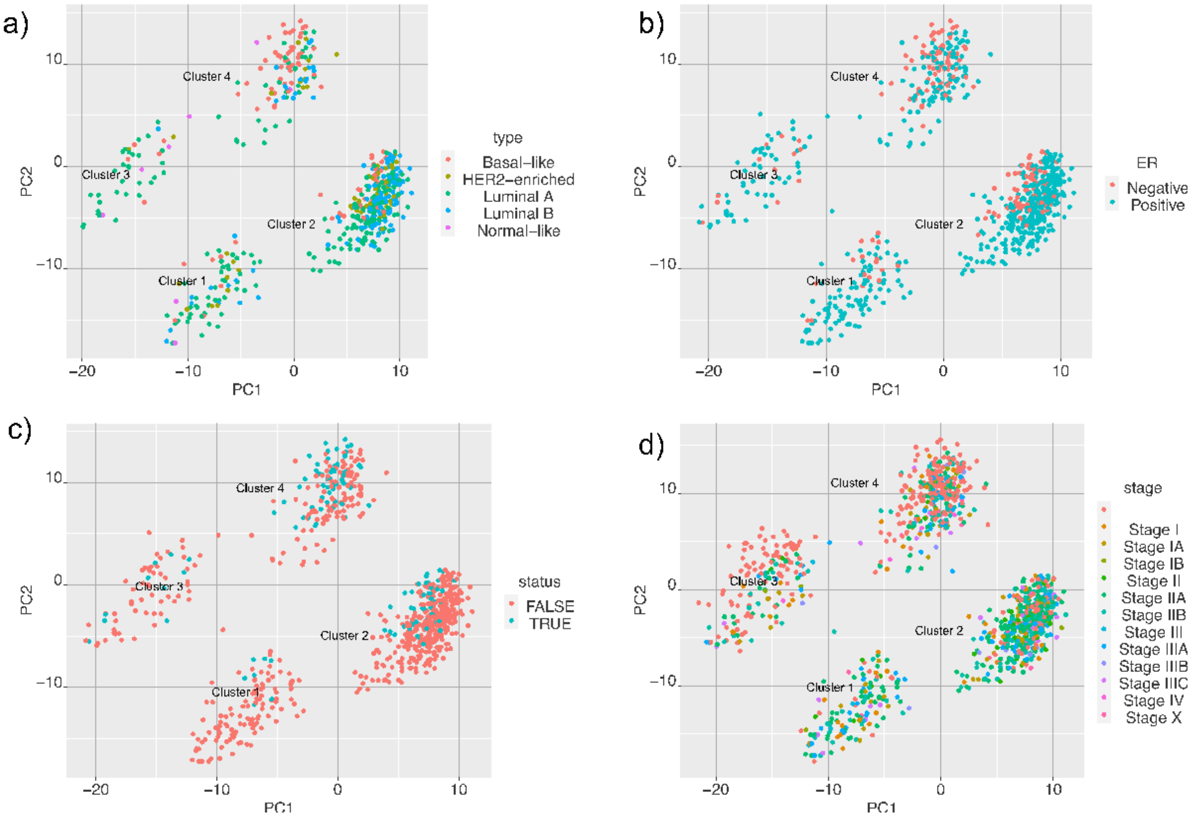


Figure 3: PCA plots of metabolic clusters considering (a) PAM50; (b) ER status; (c) triple negative status and (d) AJCC stage of the samples.

To further analyze these clusters, we measured the recurrence free survival and overall survival (deceased versus living) and observed differences between the 4 clusters. Based on the analysis, samples in cluster 1 and 4 had better survival than cluster 2 and 3. So, we combined clusters 1 and 4, and clusters 2 and 3 to identify differences in survival rate. The plot of Kaplan-Meier estimates in Figure 4 shows differences between the group of clusters. This indicates that the clusters with metabolic differences also have different survival and recurrence rates.

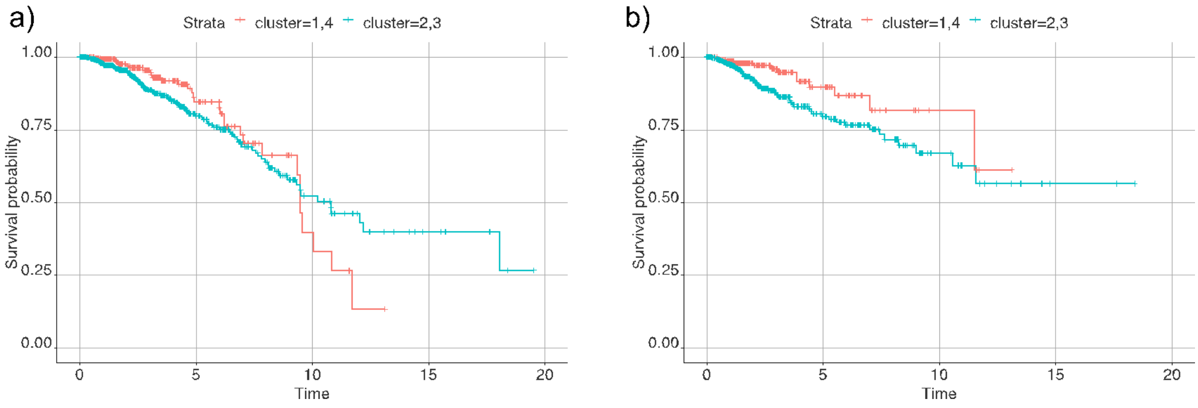


Figure 4: Plots of Kaplan-Meier estimates for (a) overall survival and (b) recurrence of cancer in the individuals. Cluster 1 and 4 are denoted by red line and cluster 2 and 3 in blue line.

Our analyses of personalized metabolic networks showed differences in the metabolic profile of the individuals such that they could be broadly categorized into four clusters and also indicated variations at reactions level, subsystems level and also the survival and recurrence rate. We further identified how metabolic genes contributed to these variations and probed genes that caused perturbations in the system.

2.3 Identifying candidate druggable genes

Deletion of a set of metabolic genes from the models can either have profound effect on the system or no effect at all. In order to predict the genes that cause perturbations in the system, we carried out in silico gene deletion in our personalized metabolic networks. About 53 out of 1884 metabolic genes upon single-gene deletion had a significant effect in the system ($p < 0.05$) upon single gene deletion analysis. Table 1 represents a concise list of genes, the subsystems these genes belong to and drug target information of these genes as reported in the Human Protein Atlas (HPA). The last column has information whether there are known FDA-approved drugs targeting the gene.

Subsystem	**Gene**	**Drug target**
Cholesterol metabolism	SOAT1	FDA approved
Valine, leucine, and isoleucine metabolism	MUT	FDA approved
Citric acid cycle	SDHA, SDHB, SDHC, SDHD	FDA approved (SDHD), Potential drug target
C5-branched dibasic acid metabolism	SUCLA2, SUCLG1, SUCLG2	Potential drug target
Lysine metabolism	DLD, DLST	Potential drug target
Oxidative phosphorylation	ATP5 family, COX family, UQCR family, CYC1, CYTB	Potential drug target
Pyrimidine synthesis	UPRT	
Sphingolipid metabolism	SGMS1	
Transport, mitochondrial	SLC25A10	
Glycerophospholipid metabolism	CEPT1, PCYT2, PDHX	

Table 1: List of genes identified as important from *in silico* gene knockout analysis and mapped to their subsystems and known drug target information from Human Protein Atlas (HPA). Drug target information for the genes is provided in the last column (FDA-approved drugs or potential drug target) using the information from HPA.

Metabolic genes like Sterol O-Acyltransferase 1 (SOAT1), methylmalonyl-CoA mutase (MUT) and isozymes of succinate dehydrogenase (SDHA, SDHB, SDHC, and SDHD) have known FDA approved drugs that can target them. Some of the other genes, identified from our analysis, like uracil

doi:10.20944/preprints202011.0474.v1

phosphoribosyltransferase (UPRT), sphingomyelin synthase 1 (SGMS1), solute carrier protein (SLC25A10), choline/ethanolamine phosphotransferase (CEPT), phosphate cytidylyltransferase 2 (PCYT2) and pyruvate dehydrogenase complex component X (PDHX) did not have drug target information. These genes are involved in diverse metabolic processes as indicated by the subsystems in Table 1. This analysis compared the genes in cancer versus normal samples that alter the system upon deletion and also provided information of the drugs that can target them.

In addition to performing systems-level analysis, we developed a method that can be used for predicting drug effects for each personalized metabolic network or can be used for known phenotypes in the system. In this analysis, we queried a list of genes causing an effect in the system against drug databases and that gave us information of drugs having higher influence in the system. Using the drug response data from the Genomics of Drug Sensitivity in Cancer (GDSC) [15], we also identified drugs that have an influence in the cells when the genes are mutated. Table 2 lists the drugs and their targets based on the number of samples (out of 1156 total samples) that identified the genes reported from our *in silico* gene deletion analysis. These drugs have been tested on 1,001 cancer cell lines including 51 BRCA cell lines. The top-ranking drug, MG-132, is a proteasome inhibitor and blocks the proteolytic activity of the 26S proteasome complex. This drug has been found to be effective in inhibiting the proliferation of BRCA cells. OSU-03012 is a celecoxib and has been shown to have anticancer and antimicrobial activity. The drug in combination with PDE5 inhibitors had shown enhanced anti-tumor activity.

Drug	Brand name	Target	# significant samples (out of 1156)	Cohort
MG-132		Proteasome	599	BRCA
OSU-03012		PDPK1 (PDK1)	474	All cell lines
PAC-1		CASP3 agonist	94	All cell lines
GSK-1904529A		IGF1R	89	All cell lines
PF-562271		FAK	31	All cell lines
QS11		ARFGAP	28	All cell lines
Trametinib	Mekinist	MAP2K1 (MEK1), MAP2K2 (MEK2)	28	All cell lines
XMD11-85h		BRSK2, FLT4, MARK4, PRKCD, RET, SPRK1	23	All cell lines
(5Z)-7-Oxozeaenol		MAP3K7 (TAK1)	14	All cell lines
GSK-650394		SGK3	12	All cell lines
Tipifarnib	Zarnestra,IND58359, R115777	Farnesyl-transferase (FNTA)	12	All cell lines
Vinorelbine	Navelbine	Microtubules	8	All cell lines
5-Fluorouracil		DNA antimetabolite	5	All cell lines

Table 2: Information about drugs ranked based on their influence on genes identified from our *in silico* analysis. The target information and the number of samples in which these genes are observed are also indicated in the table.

From our analysis, it is possible to identify drug combinations that are predicted to have more effect in cancer versus normal samples. Also, we have generated personalized drug profiles for each individual in the study, thus enabling us to predict which drug or drug combination will have a higher drug score in the individual.

3. Discussion

We have generated a personalized metabolic network for each sample in the study using divergence values; classified the samples into different clusters based on their metabolic profile; and identified drug/chemical moieties that can target metabolic genes identified from our analysis. We have applied these steps to breast cancer samples and identified four distinct clusters based on their metabolic profile. From the *in silico* gene deletion analysis we identified metabolic genes that are altered in cancer versus normal conditions. Genes belonging to cholesterol metabolism; valine, leucine, isoleucine metabolism; as well as citric acid cycle had known FDA-approved drugs targeting them. We also carried out N-of-1 analysis and identified drug responses in each sample in our study. We identified proteasome inhibitors (MG-132), COX-2 inhibitor (OSU-03012), CASP3 agonists and inhibitor of IGF-1R (GSK-1904529A) that targeted genes identified from gene deletion analysis of personalized metabolic networks of cancer samples. We have provided evidence that a metabolic analysis is able to provide deeper understanding of the metabolic alterations in cancer. There are three primary findings from this study that are described below.

First, we used individual RNAseq profiles to build personalized metabolic networks to estimate candidate metabolic network states in breast cancer and control samples from TCGA (https://www.cancer.gov/tcga). We first used the divergence approach [4] to identify genes that diverged high or low based on RNAseq data normalized as transcript per millions (TPM). We then integrated this information with a genome-scale human metabolic network [14] to estimate candidate metabolic network states that would be supported by the observed high and low expression of the corresponding enzyme-encoding genes. We used the divergence method for computing values because it has the advantage of removing noise, while keeping important signals in the dataset. Similar results could be obtained using continuous data, but the level of noise in the count is considerable, making it difficult to find anything useful. Whereas transcriptomic data is useful in giving us a snapshot of the extent to which genes are expressed, we need to integrate this information with computational models in order to gain mechanistic insights into the processes that are affected in the system. In this study, we leveraged our knowledge of metabolic networks and integrated divergence data to understand the metabolic landscape in breast cancer. Our workflow also allows us to carry out N-of-1 analysis and generate personalized metabolic networks for each sample in the study.

Second, we identified four distinct metabolic clusters of breast cancer samples from TCGA. Cluster 2 had higher proportion of samples that were HER2-enriched, luminal A and B samples, ER positive samples and triple negative samples compared to cluster 1, 3 and 4. Cluster 4 had a high proportion of samples that were basal-like in origin, ER negative and also triple negative status. Thus, the metabolic clustering analysis gave us information of the metabolic profile of the samples that was not evident from the transcriptome data alone.

Third, from our *in silico* gene deletion analysis, we identified Sterol O-Acyltransferase 1 (SOAT1), methylmalonyl-CoA mutase (MUT) and isozymes of succinate dehydrogenase (SDHA, SDHB, SDHC, and SDHD) as having a significant effect (p-value < 0.05) in cancer as compared to normal samples. These genes had known FDA-approved drug targets that inhibited them. From our drug response analysis, we identified MG-132, a cell-permeable proteasome inhibitor, that has been known to inhibit proliferation of BRCA cells [16,17]. This drug has been known to induce down-regulation of anti-apoptotic proteins Bcl-2 and XIAP and up-regulates expression of pro-apoptotic protein Bax and caspase-3 in glioma cells [18]. We also identified OSU-03012 from our drug response analysis. This drug has been reported to have anti-cancer activity [19] and mediates antitumor effects via the inhibition of PDK1 [20]. The effect of this drug in breast cancer can be tested. PAC-1, identified from our analysis, is an activator of procaspase-3 and induces apoptosis in tumor cells [21]. Our framework provides a list of drugs that can be tested for their effectiveness in breast cancer.

Tumor cells are known to reprogram energy metabolism [22], and metabolic aberrations such as the Warburg effect are considered a hallmark of cancer [23]. Tumor cells exhibit heterogeneous metabolic profiles, with differential utilization of metabolites such as glucose, lactate, glutamine and glycine [2]. Some of the metabolic dysregulation that are reported in tumor cells are enhanced glycolysis, amino acid metabolism, fatty acid metabolism [3,24], that are profoundly dysregulated in cancer and have been linked with mutated genes. Bioavailability of certain metabolites, such as asparagine, has been shown to have an influence on metastatic potential of breast cancer. These studies have shown that metabolism is altered in cancer and it is a fundamental process that needs to be studied in-depth. Tumor cells exhibit variable metabolic profiles making it challenging to decode the heterogeneous metabolic landscape in cancer.

Our framework is generalizable and can be used for generating personalized metabolic networks that will help in categorizing the samples based on their metabolic profile and identifying drug targets that will have an effect on the system.

4. Materials and Methods

Expression data and divergence analysis

We downloaded RNA-Seq data from TCGA breast cancer samples (https://www.cancer.gov/tcga), which consists of 1100 tumor (primary and metastatic) and 56 normal tissue samples. Expression counts summarized at the gene-level were retrieved from the “firehose” data portal. For metabolic model integration with gene expression and to obtain context-specific models for each sample, we used transcript per million (TPM) values that we then simplified into a ternary encoding (up, no change, down) using the divergence method [4].

Divergence analysis is a method for digitizing high dimensional omics data into a binary or ternary representation for simplified analysis. This representation aims to remove inherent population variation in an omics sample to reveal features that are divergent from normal behavior as estimated from a baseline population. In the univariate version of divergence which was utilized here, after transforming the data to the rank space (by replacing the original RNA-Seq counts in each sample profile by their ranks within the profile) and estimating baseline regions, a gene that is differentially expressed above the baseline region is represented by 1 and one that is differentially expressed below the baseline region in represented by -1, with the remaining genes at 0. In this analysis, half of the normal breast samples were used as the reference population to estimate baseline behavior and the divergence coding was computed for each gene for the remaining normal as well as the tumor samples. This step enabled converting the continuous gene expression value to ternary values for genes in the dataset.

Integration of expression data to generate personalized metabolic networks

For our analysis, we used the latest genome-scale reconstruction of known human metabolism, Recon3D, that is a multi-compartment model consisting of 10600 reactions, 5835 metabolites, 2248 metabolic genes as well as 102 subsystems [14]. Gene-protein-reaction (GPR) associations in the genome-scale metabolic models (GEMs) were used for integrating omics information with the models. TPM and divergence values were calculated for RNA-Seq data from TCGA. These values were integrated with the Recon3D model [14] using iMAT [25] (Supplementary figure 1). In this way, we generated 1156 context-specific metabolic networks and predicted a reaction rate ('flux') for each reaction in the network. Reactions related to biomass synthesis and ATP synthase were considered as core reactions and retained for generation of context-specific metabolic networks. We performed flux balance analysis (FBA) using COBRA toolbox v 3.0 [26] and evaluated flux distribution using linear programming (LP) solvers [27], using an objective function that was previously reported for cancer cells [28]. We used Ham's media composition [14] for constraining exchange reactions in the context-specific networks. Using fastFVA [29] the flux values for reactions supporting 90% of biomass production were calculated and used to classify reactions as active or inactive in the context-specific networks. The workflow represented in Figure 1 provides an overview of analyses performed. COBRA toolbox v3.0 was implemented in MATLAB R2018a and academic licenses of Gurobi optimizer v7.5 and IBM CPLEX v12.7.1 were used to solve LP and MILP problems in this study.

Classification of context-specific metabolic networks into metabolic subgroups

We carried out flux variability analysis for all context-specific metabolic networks using fastFVA [29]. Maximum flux values for reactions were used for unsupervised machine learning methods to identify metabolic clusters of cancer samples (Supplementary figure 2). K-means clustering was performed in R using the package cluster and factoextra for cluster and visualization. We computed the distance matrix using Pearson correlation. In order to ascertain the optimal number of clusters, we used the "elbow method" that takes into account the total within-cluster sum of squares (wss). Supplementary figure 2 represents the curve obtained for wss according to the number of clusters k. We distinctly observed four clusters for cancer metabolic networks using K-means clustering [30]. The cancer clusters were then labeled from 1-4 and normal tissue samples were assigned as cluster 0 for our analysis.

Using Fisher's exact test, we identified reactions that were statistically significant in cancer versus normal and also in clusters 1-4 (cancer clusters). The list of active reactions in each cancer cluster was compared with normal tissue samples to determine subsystems that were enriched in each cluster. We also examined the clusters with important phenotypes such as PAM50, ER status of the individual, triple negative status, tissue source site, year of initial pathological diagnosis, and pathological state (Supplementary file 1).

Identifying target genes in the context-specific networks

We performed *in silico* gene deletion analysis using the singleGeneDeletion function in COBRA toolbox [26]. The total number of genes in the Recon3D model was 2248 of which 1883 were unique genes. We deleted genes in the context-specific metabolic networks one at a time and measured the ratio of growth rate of the knockout model versus the wild type model. Genes with growth rate ratio (grRatio) < 0.9 were considered to have impact on the system and were used as input for drug target prediction. A grRatio of 0.9 suggests that the knock-out model was able to attain 90% of its growth compared to the original model. A Wilcoxon rank-sum test was carried out to identify genes that had significant effect on the system upon knock-out in cancer versus normal context-specific networks. Information from the Human Protein Atlas [31] and the Pathology Atlas [32] was used for biological annotation of these genes and identification of these genes as FDA approved drug targets or potential drug targets based on HPA.

Drug target identification for genes shortlisted from metabolic networks

We calculated statistical associations between in vitro drug sensitivity data and the personalized target gene sets, as shown in Supplementary file 2. Specifically, we used the drug response data from Genomics of Drug Sensitivity in Cancer (GDSC) [15] which contains IC50s for 265 anti-cancer drugs across 1,001 cancer cell lines including 51 BRCA cell lines. GDSC also included a genomic and molecular characterization of these 1,001 cell lines. We used the binarized mutation data of more than 19,000 genes, including only protein changing mutations [15,33]. For each of the 1,165 samples, we created a binary vector across the 1,001 GDSC cell lines indicating whether a cell line has at least one mutated gene in the essential gene set of the sample under investigation. A Spearman rank correlation coefficient was computed between the binary vector and the continuous IC50 drug response values for each of the drugs (n=265). We selected drugs for which at least one of the samples the P-value is smaller than 1e-3 (uncorrected). Negative correlation coefficients indicate that mutated cell lines (i.e. those that have mutations in metabolic genes) are more sensitive (low IC50) to a drug.

Statistical analysis

Fisher’s exact test was the statistical method for identifying significant active reactions from the models. For identifying differentially expressed genes in cancer versus normal, we used Wilcoxon rank-sum test. To account for the multiple testing in these analyses we calculated the Benjamini-Hochberg False Discovery Rate correction and a BH-FDR < 0.05 was considered as significant.

Software

The R/Bioconductor package 'divergence' was used for the divergence computation. We used the COBRA toolbox v3.0 [26] in MATLAB 2018a for analyzing the metabolic networks. Academic licenses of the Gurobi optimizer v7.5 and IBM CPLEX v12.7 were used to solve LP and MILP problems. PCA and K-means clustering were done using R 3.5.0 (codename “Joy in Playing”). For K-means clustering we used the package ‘cluster’ and ‘factoextra’ for clustering and visualization.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1,

Supplementary Figure 1: Schematic representation of the iMAT algorithm, operating on reactions (arrows), metabolites (squares) and genes (diamonds). The representative model has reactions labeled with R, metabolites with M and genes with G. The gene expression data is mapped to genes in the model and the user defines a cutoff of gene expression that decides which reactions are retained or eliminated from the model. The genes are colored from green to red denoting higher and lower expression values, respectively.

Supplementary figure 2: Elbow plot to identify optimum number of clusters of cancer samples. Number of clusters k and total within the sum of squares are represented in x- and y-axis respectively.

Supplementary figure 3: Volcano plots for number of significant reactions identified for each cluster after Bonferroni correction and BH-FDR correction

Supplementary figure 4: PCA of normalized TPM values for metabolic genes

Supplementary file 1: Results of cluster-based analysis

Supplementary file 2: Drug sensitivity results for 1156 samples

Author Contributions:
Conceptualization, N.D.P and L.M.; methodology, P.B., W.D., L.M., J.E. and T.A.K.; formal analysis, P.B, W.D., T.A.K.; data curation, W.D. and L.M.; writing—original draft preparation, P.B.; writing—review and editing, N.D.P, L.M., D.G., P.B., J.E., T.A.K. and W.D.; visualization, P.B. and W.D.; supervision, N.D.P and L.M.; funding acquisition, L.M., D.G. and N.D.P.

Funding:
This project has been funded by NIH (R01CA200859).

Acknowledgments: We would like to thank Dr. Ramkumar Hariharan, previous member of the Hood-Price group who was involved in the initial phase of this project.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Chiang, A.C.; Massagué, J. Molecular basis of metastasis. *N. Engl. J. Med.* **2008**, *359*, 2814–2823.
2. Vander Heiden, M.G. Targeting cancer metabolism: a therapeutic window opens. *Nat. Rev. Drug Discov.* **2011**, *10*, 671–684.
3. Wu, W.; Zhao, S. Metabolic changes in cancer: beyond the Warburg effect. *Acta Biochim. Biophys. Sin.* **2013**, *45*, 18–26.
4. Dinalankara, W.; Ke, Q.; Xu, Y.; Ji, L.; Pagane, N.; Lien, A.; Matam, T.; Fertig, E.J.; Price, N.D.; Younes, L.; et al. Digitizing omics profiles by divergence from a baseline. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 4545–4552.
5. Tamura, T.; Lu, W.; Akutsu, T. Computational Methods for Modification of Metabolic Networks. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 376–381.

6. Cazzaniga, P.; Damiani, C.; Besozzi, D.; Colombo, R.; Nobile, M.S.; Gaglio, D.; Pescini, D.; Molinari, S.; Mauri, G.; Alberghina, L.; et al. Computational strategies for a system-level understanding of metabolism. *Metabolites* **2014**, *4*, 1034–1087.
7. Wang, Y.; Eddy, J.A.; Price, N.D. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst. Biol.* **2012**, *6*, 153.
8. Cook, D.J.; Nielsen, J. Genome-scale metabolic models applied to human health and disease. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2017**, *9*, doi:10.1002/wsbm.1393.
9. Raman, K.; Chandra, N. Flux balance analysis of biological systems: applications and challenges. *Brief. Bioinform.* **2009**, *10*, 435–449.
10. Nilsson, A.; Nielsen, J. Genome scale metabolic modeling of cancer. *Metab. Eng.* **2017**, *43*, 103–112.
11. Zhang, C.; Aldrees, M.; Arif, M.; Li, X.; Mardinoglu, A.; Aziz, M.A. Elucidating the Reprograming of Colorectal Cancer Metabolism Using Genome-Scale Metabolic Modeling. *Front. Oncol.* **2019**, *9*, 681.
12. Jerby, L.; Ruppin, E. Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling. *Clin. Cancer Res.* **2012**, *18*, 5572–5584.
13. Yizhak, K.; Chaneton, B.; Gottlieb, E.; Ruppin, E. Modeling cancer metabolism on a genome scale. *Mol. Syst. Biol.* **2015**, *11*, 817.
14. Brunk, E.; Sahoo, S.; Zielinski, D.C.; Altunkaya, A.; Dräger, A.; Mih, N.; Gatto, F.; Nilsson, A.; Preciat Gonzalez, G.A.; Aurich, M.K.; et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol.* **2018**, *36*, 272–281.
15. Iorio, F.; Knijnenburg, T.A.; Vis, D.J.; Bignell, G.R.; Menden, M.P.; Schubert, M.; Aben, N.; Gonçalves, E.; Barthorpe, S.; Lightfoot, H.; et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **2016**, *166*, 740–754.
16. Hunter, C.A.; Roberts, C.W.; Alexander, J. Kinetics of cytokine mRNA production in the brains of mice with progressive toxoplasmic encephalitis. *Eur. J. Immunol.* **1992**, *22*, 2317–2322.
17. Hammond-Martel, I.; Pak, H.; Yu, H.; Rouget, R.; Horwitz, A.A.; Parvin, J.D.; Drobetsky, E.A.; Affar, E.B. PI 3 kinase related kinases-independent proteolysis of BRCA1 regulates Rad51 recruitment during genotoxic stress in human cells. *PLoS One* **2010**, *5*, e14027.
18. Guo, N.; Peng, Z. MG132, a proteasome inhibitor, induces apoptosis in tumor cells. *Asia Pac. J. Clin. Oncol.* **2013**, *9*, 6–11.
19. Weng, S.-C.; Kashida, Y.; Kulp, S.K.; Wang, D.; Brueggemeier, R.W.; Shapiro, C.L.; Chen, C.-S. Sensitizing estrogen receptor-negative breast cancer cells to tamoxifen with OSU-03012, a novel celecoxib-derived phosphoinositide-dependent protein kinase-1/Akt signaling inhibitor. *Mol. Cancer Ther.* **2008**, *7*, 800–808.
20. McCubrey, J.A.; Lahair, M.M.; Franklin, R.A. OSU-03012 in the treatment of glioblastoma. *Mol. Pharmacol.* **2006**, *70*, 437–439.
21. Wang, F.; Wang, L.; Zhao, Y.; Li, Y.; Ping, G.; Xiao, S.; Chen, K.; Zhu, W.; Gong, P.; Yang, J.; et al. A novel small-molecule activator of procaspase-3 induces apoptosis in cancer cells and reduces tumor growth in human breast, liver and gallbladder cancer xenografts. *Mol. Oncol.* **2014**, *8*, 1640–1652.
22. Phan, L.M.; Yeung, S.-C.J.; Lee, M.-H. Cancer metabolic reprogramming: importance, main features, and potentials for precise targeted anti-cancer therapies. *Cancer Biol Med* **2014**, *11*, 1–19.
23. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **2011**, *144*, 646–674.
24. Knott, S.R.V.; Wagenblast, E.; Khan, S.; Kim, S.Y.; Soto, M.; Wagner, M.; Turgeon, M.-O.; Fish, L.; Erard, N.; Gable, A.L.; et al. Asparagine bioavailability governs metastasis in a model of breast cancer. *Nature* **2018**, *554*, 378–381.
25. Zur, H.; Ruppin, E.; Shlomi, T. iMAT: an integrative metabolic analysis tool. *Bioinformatics* **2010**, *26*, 3140–3142.
26. Heirendt, L.; Arreckx, S.; Pfau, T.; Mendoza, S.N.; Richelle, A.; Heinken, A.; Haraldsdóttir, H.S.; Wachowiak, J.; Keating, S.M.; Vlasov, V.; et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* **2019**, *14*, 639–702.
27. Orth, J.D.; Thiele, I.; Palsson, B.Ø. What is flux balance analysis? *Nat. Biotechnol.* **2010**, *28*, 245–248.
28. Agren, R.; Bordel, S.; Mardinoglu, A.; Pornputtapong, N.; Nookaew, I.; Nielsen, J. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput. Biol.* **2012**, *8*, e1002518.
29. Gudmundsson, S.; Thiele, I. Computationally efficient flux variability analysis. *BMC Bioinformatics* **2010**, *11*, 489.
30. MacQueen, J.; Others Some methods for classification and analysis of multivariate observations. In Proceedings of the Proceedings of the fifth Berkeley symposium on mathematical statistics and probability; Oakland, CA, USA, 1967; Vol. 1, pp. 281–297.
31. Uhlén, M.; Fagerberg, L.; Hallström, B.M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.;

Kampf, C.; Sjöstedt, E.; Asplund, A.; et al. Proteomics. Tissue-based map of the human proteome. *Science* **2015**, *347*, 1260419.

32. Uhlen, M.; Zhang, C.; Lee, S.; Sjöstedt, E.; Fagerberg, L.; Bidkhori, G.; Benfeitas, R.; Arif, M.; Liu, Z.; Edfors, F.; et al. A pathology atlas of the human cancer transcriptome. *Science* **2017**, *357*, doi:10.1126/science.aan2507.
33. Drug Download Page - Cancerrxgene - Genomics of Drug Sensitivity in Cancer Available online: https://www.cancerrxgene.org/downloads/bulk_download (accessed on Feb 24, 2020).