Article

A mechanistic model for genome-scale estimations of metabolic activity

Cankut Cubuk [1,2], Carlos Loucera[1,3], Marta R. Hidalgo[4], Alicia Amadoz[5], Jose Carbonell-Caballero[6], María Peña-Chilet[1,3,7], Joaquin Dopazo [1,3,7,8,*]

[1] Clinical Bioinformatics Area, Fundación Progreso y Salud (FPS), CDCA, Hospital Virgen del Rocio, 41013, Sevilla, Spain;
[2] Division of Genetics and Epidemiology, Institute of Cancer Research, London, SW7 3RP, UK
[3] Computational Systems Medicine. Institute of Biomedicine of Seville (IBiS), Sevilla, 41013, Spain
[4] Bioinformatics and Biostatistics Unit, Centro de Investigación Príncipe Felipe (CIPF), 46012, Valencia, Spain
[5] Igenomix S.L. 46980 Valencia, Spain
[6] Centre for Genomic Regulation, 08003, Barcelona, Spain.
[7] Bioinformatics in RareDiseases (BiER), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Sevilla, 41013, Spain
[8] Functional Genomics Node (INB-ELIXIR-es), Sevilla, Spain
* Correspondence: joaquin.dopazo@juntadeandalucia.es;

Abstract: In the last decade, different omic technologies have experiences an exponential technological advancement. However, metabolomics has not followed a similarly vertiginous improving improvement and are far from genomics or transcriptomics in terms of throughput, cost and even accuracy. Therefore, genome-scale in-silico methodologies to estimate metabolic activities from genomic data constitute an active field. The solutions available fall into two extremes: those with few assumptions about the relationships among proteins and metabolites, which are easy-to-use but less accurate and those that account for the complex relationships among molecules and proteins defined in the metabolic pathways, which are more accurate but require mathematical skills. Here, we introduce Metabolica, an algorithm that considers the complex functional relationships among all the molecules and proteins involved in the metabolic pathway analyzed but keeping an easy use that do not require of advanced mathematical skills. Metabolica has been implemented in a freely available software R package. The software inputs transcriptomic data and infers the activities of the reactions that produce the different metabolites in the pathway analyzed. An example shows how detected dysregulated metabolites in several cancers are related to patient survival.

Keywords: metabolic pathways; mechanistic model; Transcriptomic;

1. Introduction

The deviant levels of metabolites are precise indicators of metabolic irregularities behind the neoplastic cell requirements during cancer initiation and progression. These requirements include mainly the biosynthesis of building blocks (nucleotides, lipids, and amino acids) and the fulfilment of the enormous energy dependencies of rapidly proliferating tumor cells. The main elements of metabolism are the metabolites and the biochemical reactions that are catalyzed by enzymes, which participate in a network of complex interactions that are described in detail in different pathway repositories, like Reactome [1], KEGG [2], Wikipathways [3], etc., or other ones more specific, like disease maps [4]. The quantification of these elements is essential to understand their mechanistic interactions with cellular functions (e.g. apoptosis, metastasis, etc.) [5] and to elucidate their cancerous/oncogenic activities [6].

Despite the relevance of metabolism in cancer was known for almost one century [7], with the observation of enhanced aerobic glycolysis (the well-known Warburg effect) [8], the number of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 January 2021 doi:10.20944/preprints202101.0280.v1

studies using differential metabolite profiles is still very limited [9, 10]. Actually, large-scale, consortium-based studies like TCGA [11], TARGET or METABRIC [12] lack of a systematic metabolomics characterization. This is mainly due to the technical difficulties for the correct detection and the quantitative systematic measurement of small molecules. The current challenges in the metabolomics area make of genome-scale metabolic network based *in-silico* methods an active area [13].

While pathway enrichment analysis (EA) methods [14] were initially used for exploratory genome-scale metabolic analysis, other more sophisticated methods, such as Constraint-based flux Balance analysis Methods (CBM) [15], demonstrated to be more accurate for specific tasks, like the characterization of oncometabolites [16, 17]. Several features of metabolism, such as the numerous feedback loops, the reversibility of reactions, the existence of alternate paths (bifurcations) for the production of metabolites or the stoichiometric equilibrium between metabolites [18] cannot be properly captured by EA methods, which focus on sets of genes, even if these are weighted by the proximity in the pathways [19]. Contrarily, CBM consider the features generated by the complex wiring of proteins within metabolic pathways. However, the tradeoff for this analytic advantage is that CBM applications require prior experience in the selection of thresholds for setting the correct constraints to shrink the solution space and need previous skills in the mathematical solvers to obtain an optimal solution space [20-22]. Moreover, setting a correct objective function for the metabolic network modeling of complex animal cells, especially for the cancer cells, is one of the main challenges in CBM [15]. Therefore, it is not surprising that EA algorithms, such as the Reporter Metabolites (RM) [23], are still frequently used to identify metabolites related to significant transcriptional changes. RM aggregates the p-value of the genes that directly influence the production and consumption of metabolites, and assigns a unique p-value to any metabolite. Two recent studies presented the extensions of RM on pathways: the metabolite-centric Reporter Pathway Analysis (RPA) [24], and Metabolic Classifier and Feature generator (MCF) [25].

A recent genome-scale analysis strategy that takes into consideration the topology of the pathways in a simpler algorithmic framework is mechanistic modeling of pathway activity [26]. Mechanistic models have successfully been applied to uncover details of the disease mechanisms behind different cancers [5, 27], including neuroblastoma [28, 29] and glioblastoma [30], mechanisms of action of drugs [31] including drug repurposing [32], gender-specific effects of drugs in cancer [33] and the description of the mechanisms of emergence of drug resistances in cancer at single-cell level [30]. Moreover, mechanistic models have an interesting property: they allow predicting the consequences of perturbations in a given condition [34], which make them excellent tools for drug discovery or personalized therapeutic interventions. A specific version for metabolic analysis that models metabolic modules, defined within KEGG [2] as a comprehensive curated summary of the main aspects of metabolic activity, accounting for the production of the main classes of metabolites (nucleotides, carbohydrates, lipids and amino acids) [35] was recently proposed [36]. Despite metabolic modules provide only a reduced coverage of the whole complexity of the metabolism, this approach was successfully applied to predict gene essentiality in cancer with a high accuracy [5].

Here we present the Metabolica algorithm, an extension of the mechanistic modeling of metabolic modules extended to the whole metabolism, to predict the dysregulated metabolite production activities in cancer using the transcriptomics and/or genomic data. Metabolica considers reversibility of reactions, deals with feedback loops and does not need mathematical solvers. Some examples illustrate how metabolites whose production was dysregulated in cancer were clearly associated to bad prognostic.

2. Results

2.1. Prediction of metabolite production in BRCA and KIRK

To demonstrate Metabolica, the BRCA and KIRC datasets of TCGA project were used. The results obtained were compared with the reporter metabolites algorithm (RM) and real metabolomics

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 January 2021 doi:10.20944/preprints202101.0280.v1

measurements. The benchmark was restricted to 267 metabolites that were contained within metabolomics datasets and had KEGG compound identifiers. The significant differentially regulated metabolites (DRM) are depicted in Figure 1. A total of 74% of DRMs predicted by Metabolica were coincident with the the experimental metabolic profiling. The percentages provided by RM for the same comparison were lower than %1.

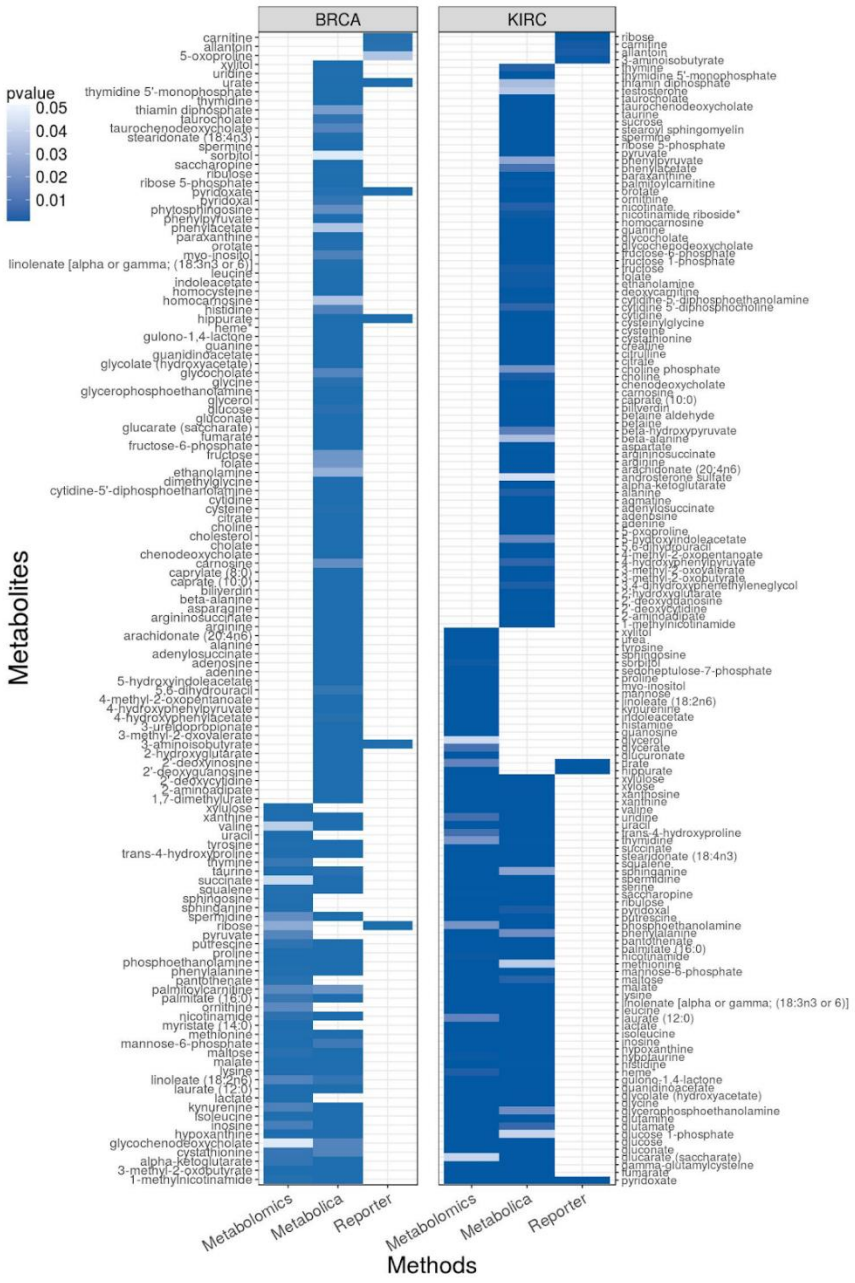

Figure 1. Results of the algorithms compared. DiReMs found by the Metabolica, the reporter metabolites algorithm (RM), and metabolomics dataset (as ground truth) in BRCA and KIRC cancer types.

The total number of DRMs predicted by Metabolica was almost twice the number obtained from the metabolomics datasets. It is important to note that the Metabolica determines DRMs based on the production rate of metabolites but does not necessarily report the actual balance of the metabolite, which is what experimental technologies such as gas chromatography and mass spectrometry detect, given that intermediate metabolites could be further depleted, transformed by other reactions. In order to assess the real ratio of false positives of the algorithm a test was carried by a systematic

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 January 2021 doi:10.20944/preprints202101.0280.v1

comparisons between pairs of groups of individuals sampled from the same condition. Since the individuals in the compared groups belong to the same condition, and are actually identical, any difference reported by Metabolica can be considered a false positive. (see Methods). The mean percentage of false positive results was always lower than % 0.003 when the conventional alpha value of 0.05 set as the threshold of significance (Figure 2). Therefore, biases due to false positives in Metabolica can be discarded.

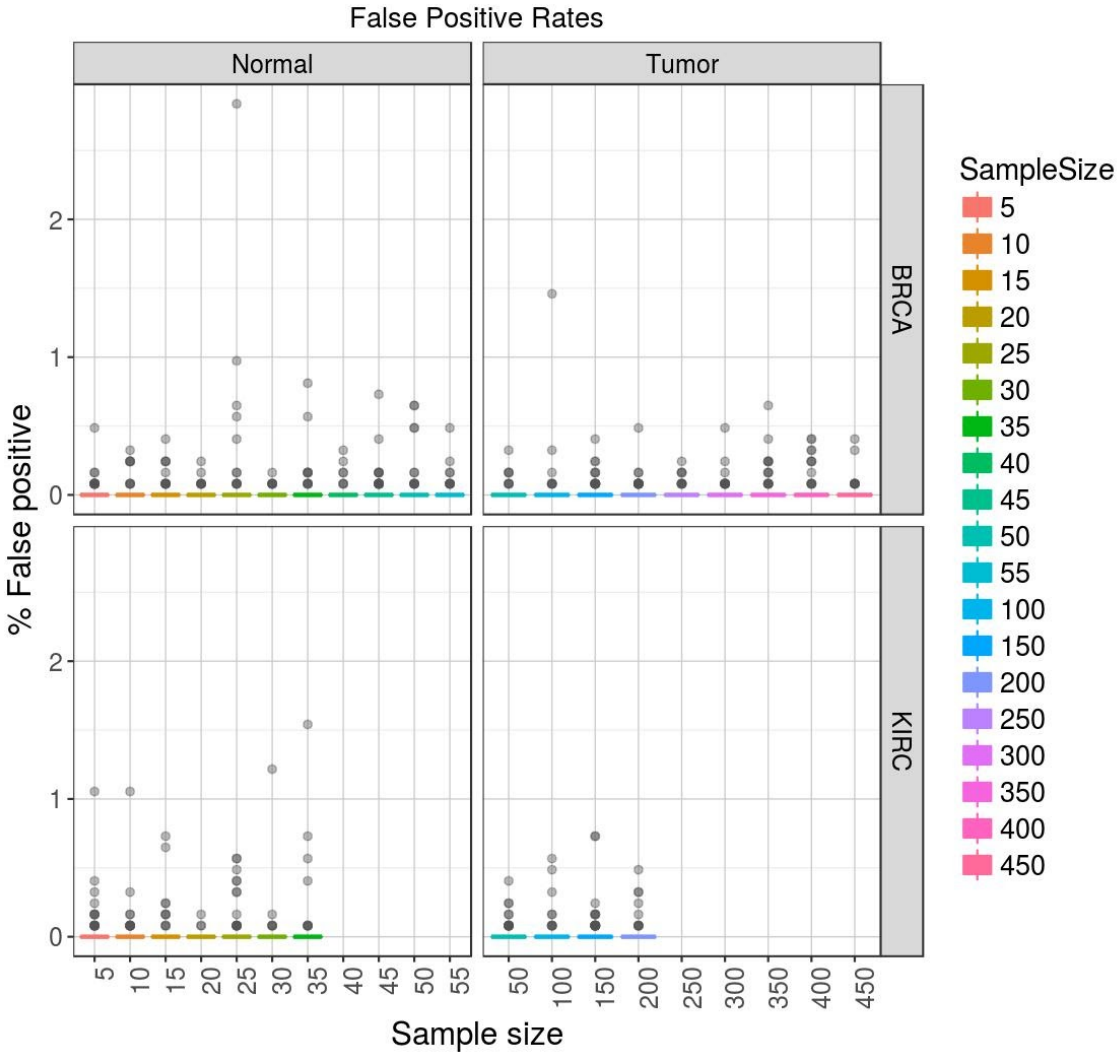

Figure 2: Specificity study. False positive rate of the Metabolica when the different groups of same condition were compared. For each sample size, 1000 different samplings were done and compared. The percentages of significant (FDR-adjusted p<0.05) DRMs that were found in all iterations are given in this figure. The median percentage of false positive results of the Metabolica is given as line and the outliers as dots.

2.2. Impact of metabolites in cancer progression

In order to check whether the estimated metabolite production activities were related with cancer progression and patient prognosis, patient survival and cancer stage analyses were carried out. The number of metabolites significantly associated with patient survival was remarkably different between the two cancer types studied: only 2 metabolites in BRCA and 307 metabolites in KIRC (FDR-adjusted p<0.05 of hazard ratio for Cox model). Out of 307 significant metabolites found in KIRC, 19 were members of arginine and proline metabolism, glutathione metabolism, and primary bile acid biosynthesis pathways. The abundances of these metabolites also showed high correlation with the cancer stages (pearson $|r| > 0.9$). The results of patient survival and cancer stage of three

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 January 2021 doi:10.20944/preprints202101.0280.v1

clinically relevant metabolites from this list, putrescine (C00134), Acyl-CoA (C00040) and Acetyl-CoA (C00024) are given in Figure 3.There is a significant increasing trend for Acetyl-CoA (Correlation coefficient=0.73) and a decreasing trend for putrescine (CC=-0.92) and Acyl-CoA (CC=-0.90).

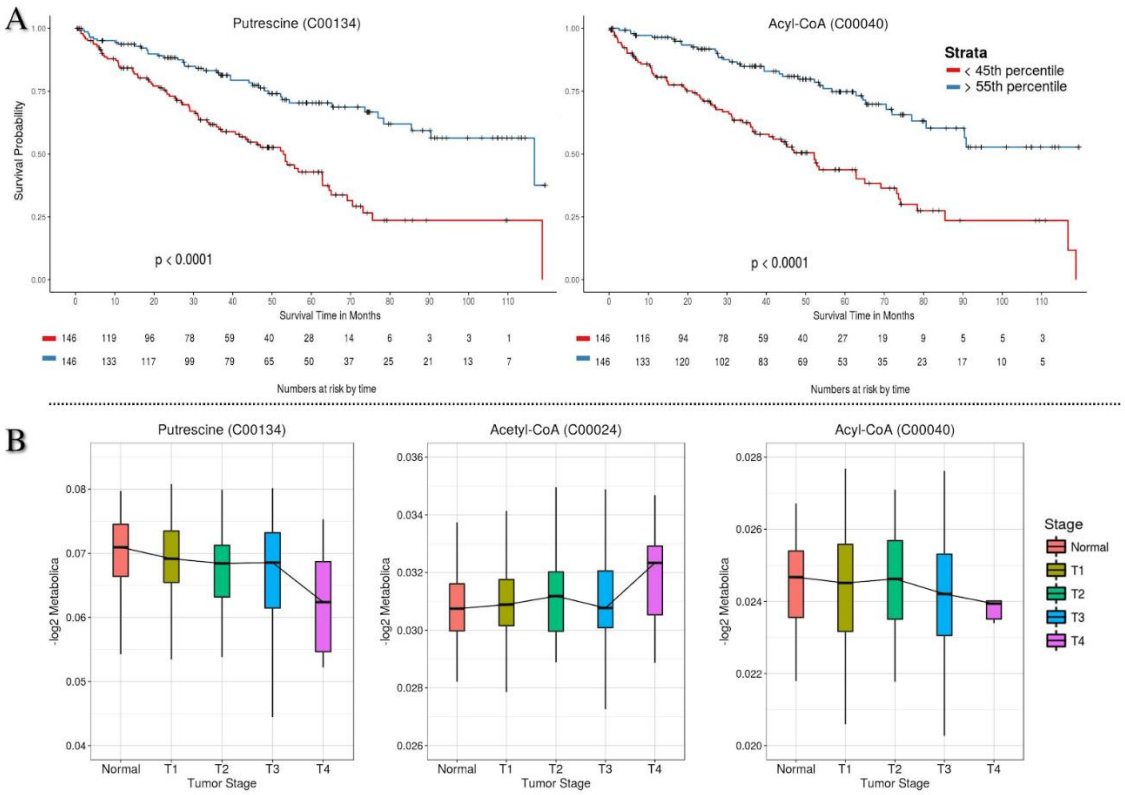

Figure 3: Relationship between metabolite prodiuction and patient survival. (A) The K-M plots of putrescine (C00134) and Acyl-CoA (C00040) are showing the significant relationship between the abundance of these metabolites and the patient survival in KIRC. The 45th and 55th percentiles (red and blue lines) were used to discretize the samples into two groups. The x-axis shows time in months and the number of patients at risk in the high two groups of samples. (B) The box plots of putrescine (C00134), Acetyl-CoA (C00024) and Acyl-CoA (C00040) are showing the relationship between metabolite abundance and cancer stage. The x-axis shows sample groups; normal and tumor stages. The y-axis shows -log2 metabolite abundances by the Metabolica. Survival data are available in Table S1.

3. Discussion

Metabolica is based in a generalized mechanistic model that accounts for the activity of the reactions required to produce any metabolite in the whole metabolic pathway. The results suggest that there is a production of the metabolite, which doesn not mean that the metabolite is detectable given that it can be consumed in another reaction. However, if the metabolite plays a relevant role, this information about its production is relevant.

Figure 1 shows a remarkable difference in the number of predicted DRMs by the Metabolica and RM, suggesting that the use of sub-pathways or circuits in the context of mechanistic models produces a more accurate detection of metabolite production activity [26, 36, 37]. In particular Pyridoxate, also known as vitamin B_6, was identified as a DRM in KIRC by both methods (Figure 1). Pyridoxate and its bioactive form has been identified as a critical factor to alter apoptosis induction properties of chemotherapeutics. Actually, In lung cancer, low expression of pyridoxal kinase gene, which encodes the enzyme that generates the bioactive form of vitamin B_6, was found to be associated with poor prognosis and proposed as a biomarker for risk stratification among lung cancer patients [38]. Glucose dependence is known to be the main metabolic characteristic exhibited by tumor cells

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 January 2021 doi:10.20944/preprints202101.0280.v1

[7]. Interestingly, both Metabolica and real measurements show different levels of differential activity in the production of this metabolite between KIRC and BRCA (Figure 1). This observation may be a consequence of the differences that exist between the glucose metabolization in different cancer cells, that was recently reported by isotope tracing in KIRC [39]. There were metabolites which have been identified as DRMs only by the Metabolica in both cancer types. For example, 5,6-dihydrouracil (Figure 1), which is an intermediate product of breakdown process of uracil into beta-alanine required for epithelial-mesenchymal transition [40]. This catabolic process is also called pyrimidine degradation module and was found to be essential for cancer cell survival in some tumor types and experimentally validated by us [5]. Although a comprehensive description of the results is beyond the scope of this manuscript, it is worth mentioning how the dysregulation of metabolites in BRCA like succinate and kynurenine (Figure 1) that are taking a key role in the initiation of tumorigenesis and its progression were correctly predicted [41, 42].

The therapeutic perspective of the metabolites is an important aspect that cannot be despised. Since metabolites can act as direct regulators of gene expression, they have been used for decades as therapeutic agents, targets, or biomarkers [43]. Thus, one the main motivations here was to demonstrate the clinical relevance of metabolites. To achieve so, patient survival and cancer stage analyses were performed for the metabolites. As previously mentioned, both cancers showed a remarkably different number of metabolites significantly associated with patient survival, which was in agreement with recent observations of pan-cancer analyses of prognostic genes and metabolic modules [5, 44]. On the other hand, the low number of significant associations observed in BRCA has been suggested to be an artifact derived from the short follow-up time of the TCGA samples [45]. Figure 3 shows three clinically relevant metabolites detected by Metabolica from this 10 members of the arginine and proline metabolism, glutathione metabolism, and primary bile acid biosynthesis pathways, putrescine (C00134), Acyl-CoA (C00040) and Acetyl-CoA (C00024). Putrescine is a polycationic alkylamine and the precursor of spermidine and spermine. These polyamines are involved in many fundamental processes, essential for normal cell growth and their depletion may have a cytostatic effect on some tumors. It is known that polyamine metabolism is frequently dysregulated in cancer and polyamine blocking therapies are used to heighten immune responses in cancer [46-48]. Acyl-CoA and Acetyl-CoA are the important metabolites of the fatty acid biosynthesis and elongation processes. They have been identified as critical factors for tumor growth by means of their effect on histone acetylation and gene expression [49, 50]. The high production rate of Acetyl-CoA from acetate is known to be associated with poor prognosis cancer cell survival [51]. Mitochondrial respiration under prolonged hypoxic conditions increases the generation of reactive oxygen species (ROS) that results in the cell death [52]. Therefore, the cancer cells adapt to the hypoxic microenvironment by limiting the conversion of pyruvate to acetyl-CoA that is entering into the tricarboxylic acid (TCA) cycle. For all that, the required acetyl-CoA in the other cellular processes can be generated by alternative routes like using acyl-CoA in β-oxidation [52, 53]. The negative correlation between the tendencies of acyl-CoA and acetyl-CoA from healthy tissue to tumor stage 4 is also confirming this alternative process (Figure 3).

In summary, the Metabolica has been designed to predict metabolite production activity that can be used as potential diagnostic biomarkers or drug targets for complex traits. In particular, it can be used for in-silico targeted enrichment of metabolites to prioritize them for experimental metabolomics studies. The Metabolica tool, the data used in this study and more comprehensive results can be found at https://github.com/babelomics/Metabolica.

4. Materials and Methods

4.1. Breaking down the pathway into elementary sub-pathways

Metabolica requires the definition of sub-pathways in which the activity of the reactions of metabolite synthesis are modeled. Breaking down a pathway into sub-pathways and estimating the activity of a sub-pathway do not depend on a particular pathway repository, but it requires essential information for metabolic reactions (substrate, product, and reversibility descriptions). Here, the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 January 2021 doi:10.20944/preprints202101.0280.v1

canonical pathways presented in KEGG database [54] were used. A total of 78 human metabolic pathways, containing 1901 reactions and 1270 metabolites (Table S2) were downloaded. The KGML files were parsed using the KEGGgraph Bioconductor package [55].

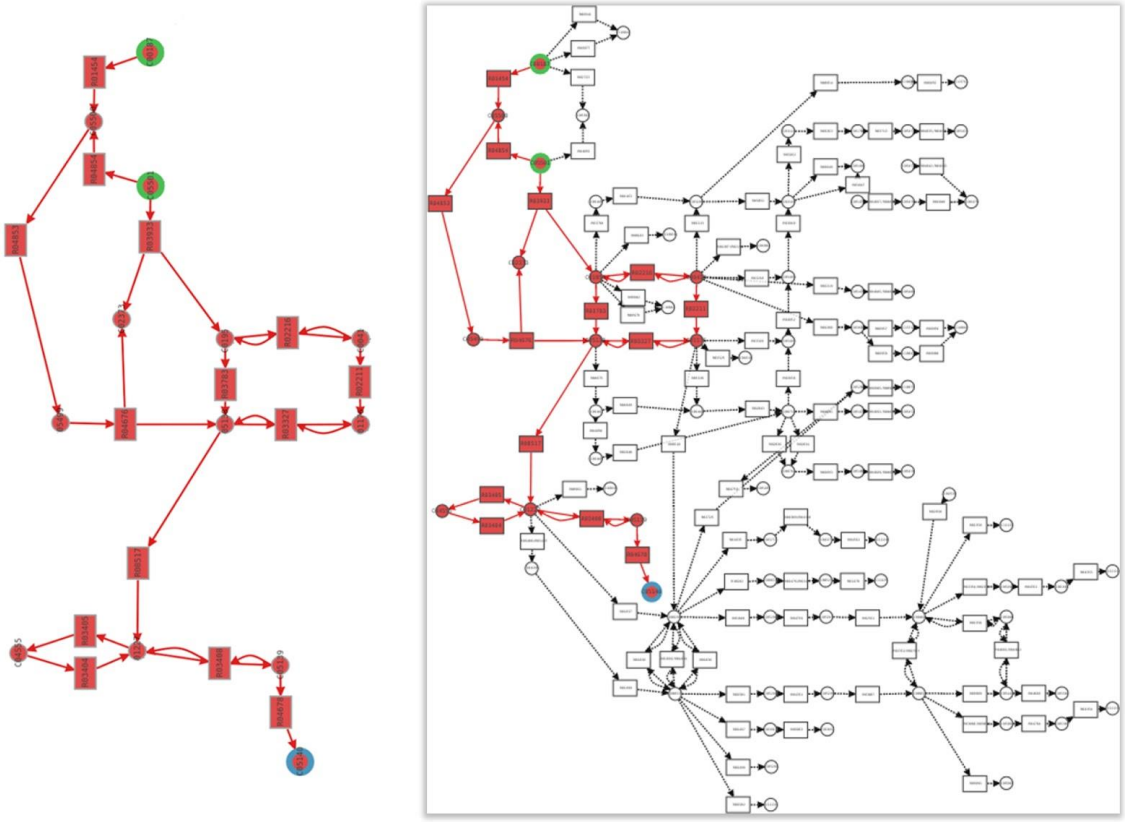

Figure 4: Example of a sub-pathway. The production sub-pathway of 4-Androsten-16alpha-ol-3,17-dione (blue node) starting from cholesterol and 20alpha,22beta-Dihydroxycholesterol (green nodes). This sub-pathway (left) was dissected from the steroid hormone biosynthesis pathway (right). The circles, rectangles and arrows are representing metabolites, metabolic reactions and reaction reversibility, respectively.

The sub-pathway that produces a given metabolite is defined by all the nodes which were visited inside its pathway using breadth-first search algorithm. This process starts from the metabolite produced (so-called product) and continues iteratively on the direction of the edges which are arriving at the product and its connected neighbor nodes. Figure 4 shows a real example of a sub-pathway which is extracted from its pathway as described above.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 January 2021 doi:10.20944/preprints202101.0280.v1

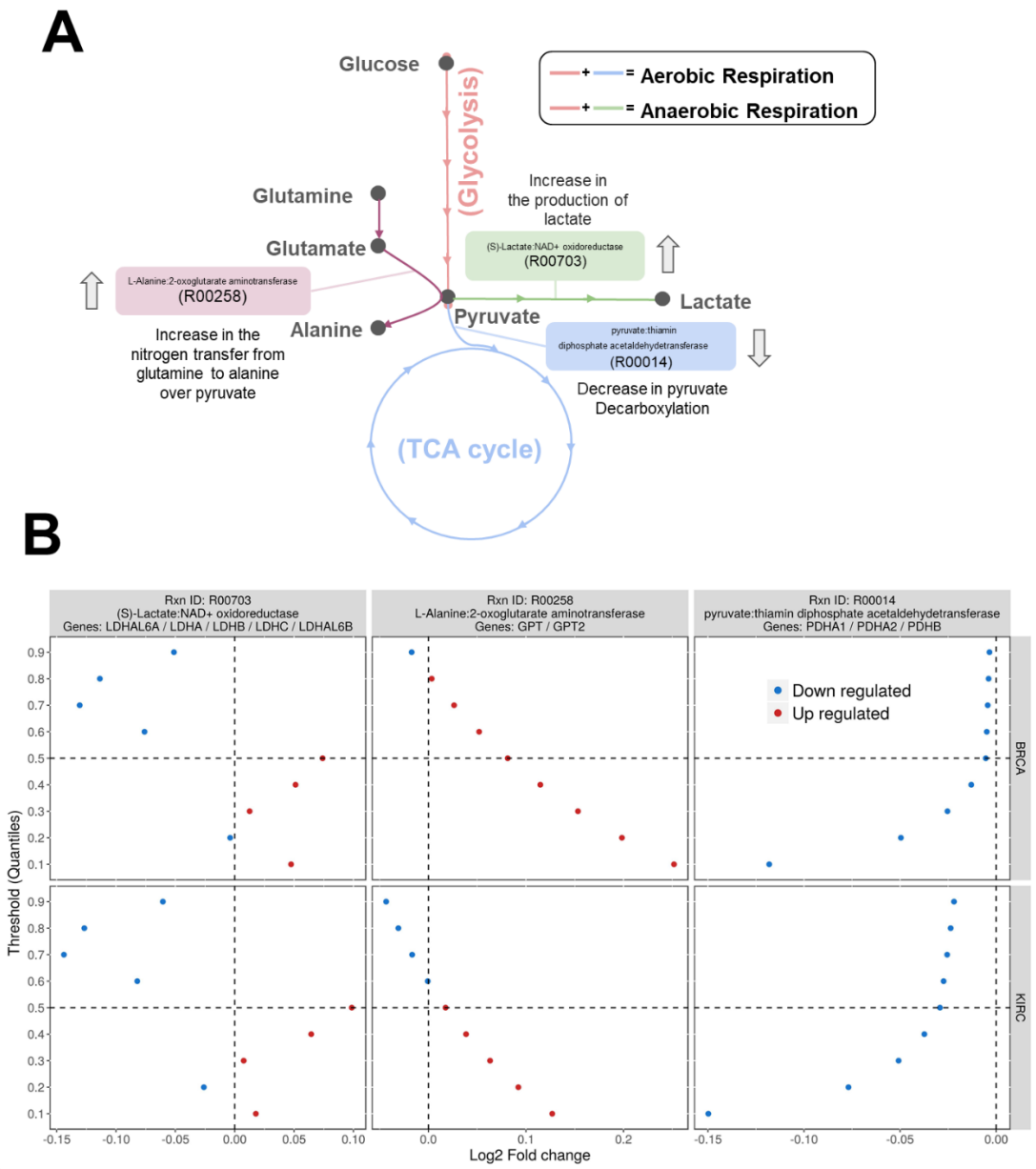

Figure 5: Effect of thresholds on the reactions. A) The selected threshold was fulfilling the following expectations: the highest amount of lactate production (reaction (S)-Lactate + NAD^+ <=> Pyruvate + NADH + H^+; KEGG ID: R00703) [56], decrease in pyruvate decarboxylation (L-Alanine + 2-Oxoglutarate <=> Pyruvate + L-Glutamate, KEGG ID: R00258) [56] and increase in the nitrogen transfer from glutamine to alanine over pyruvate (Pyruvate + Thiamin diphosphate <=> 2-(alpha-Hydroxyethyl)thiamine diphosphate + CO_2, KEGG ID: R00014) [57]. B) The different percentile thresholds applied for transcriptomics data mapping into from reactions nodes and for each threshold the activities of (S)-Lactate + NAD+ <=> Pyruvate + NADH + H+ (KEGG ID: R00703), L-Alanine + 2-Oxoglutarate <=> Pyruvate + L-Glutamate (KEGG ID: R00258) and Pyruvate + Thiamin diphosphate <=> 2-(alpha-Hydroxyethyl)thiamine diphosphate + CO2 (KEGG ID: R00014) are compared between healthy and tumor samples of BRCA and KIRC. The y-axis shows the percentile thresholds tested and the x-axis log2 fold-change of reactions. Down regulated and up regulated reactions in tumor samples are shown with blue and red dots, respectively. Dashed lines are showing zero and 50th percentile fold-changes.

Because of the highly interconnected nature of metabolic pathways and the numerous feedback loops, the convergence of calculations is challenging when the propagation algorithms are applied on metabolic hypergraphs. To deal with this issue, the feedback loops which are not derived from

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 January 2021 doi:10.20944/preprints202101.0280.v1

the product were kept, however, all the feedback loops (outdegree edges) of the product were removed. By this means, we also restrict the consumption of the product by its producing pathway.

Similarly to the signaling [27] or metabolic module [36] implementations, Metabolica requires starting node(s) to initialize the propagation of metabolic flux along sub-pathways. The definition of starting nodes is a 2-steps process: first, the metabolites with indegree of zero and the metabolites at the farthest position in the sub-pathway (products) are selected, and the propagation algorithm (see below) is ran without any objective function. In the second step, the nodes which were not visited in the previous run to the list of starting nodes are included in order to guarantee that all the nodes can be visited in the further runs. All the decomposing steps were done only one time and saved for future analysis of the pathway.

4.2. Estimating reaction node activities

The method proposed in this study uses the expression levels of the genes that encode enzymes as proxies of levels of the corresponding gene product enzymes and, consequently, of their activity levels [58-60]. The transcriptomics data mapping from genes to enzymes (proteins) and reactions were done using the gene-protein-reaction (GPR) relationships [61]. In the case of single gene (protein) - reaction relationships, the normalized gene expression level is used as the activity level of the reaction. For reaction nodes composed of multiple genes (proteins), where the reaction is catalyzed by isozymes or enzyme complexes, a 50th percentile of expression values of the genes is assigned as the activity of the reaction. The percentile threshold (50th) used to summarize gene expression values into a reaction activity was empirically obtained from the observed regulation patterns of 3 well characterized metabolic reactions in the context of the Warburg effect: (S)-Lactate + NAD+ <=> Pyruvate + NADH + H+ (KEGG ID: R00703), L-Alanine + 2-Oxoglutarate <=> Pyruvate + L-Glutamate (KEGG ID: R00258) and Pyruvate + Thiamin diphosphate <=> 2-(alpha-Hydroxyethyl)thiamine diphosphate + CO2 (KEGG ID: R00014), depicted in Figure 5A.

To define the optimal threshold the Warburg Effect, defined as an increase in the rate of glucose uptake and preferential production of lactate (anaerobic), even in the presence of oxygen (aerobic) [56], was used. Therefore, the observation of an increase in the reaction activities of anaerobic respiration compared to aerobic in tumor samples (Figure 5A) was expected. Figure 5B illustrates the upregulation of the main reaction of anaerobic branch (R00703) at 10th, 30th, 40th and 50th percentiles when tumor samples compared to healthy samples. Actually, the highest up regulation was observed at 50th percentile for both cancer types. The first reaction of the aerobic branch, R00014, was down regulated at all percentiles (from 10th to 90th percentiles). Finally, the reaction R00258 was up regulated at 10th, 20th, 30th, 40th and 50th percentiles in both cancer types [57]. Therefore, although the whole range from 10^{th} to 50^{th} percentiles is compatible with the Warburg Effect, the highest amount of lactate production, decrease in pyruvate decarboxylation and increase in the nitrogen transfer from glutamine to alanine over pyruvate is attained at 50^{th} percentile, with the highest lactate production using pyruvate, which is the essential metabolic process in tumorigenesis [57] .

4.3. Computing the sub-pathway activity

For each reaction node r_i of a sub-pathway, the metabolic flux propagation is computed by the given formula according to the following recursive rules:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 January 2021 doi:10.20944/preprints202101.0280.v1

$$w_{m_s,r_i} = m_s \cdot \frac{r_i}{\sum_{r_i \in \mathcal{N}^+(m_s)} r_i} \quad (1)$$

$$w_{r_i,m_p} = \min_{m_s \in \mathcal{N}^-(r_i)} (w_{m_s,r_i}) \cdot r_i \quad (2)$$

$$m_p = \sqrt[n]{m_p \cdot \prod_{r_i \in \mathcal{N}^-(m_p)} w_{r_i,m_p}} \quad (3)$$

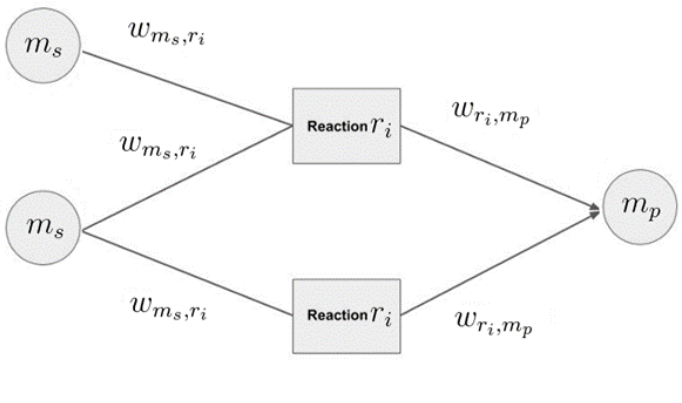

Figure 6. Recursive rules of the propagation algorithm.

Where W_{m_s,r_i} is the amount of substrate (m_s) used by reaction (r_i), $W_{r_i m_p}$ is the amount of product (m_p) produced by the reaction r_i and m_p is the final amount of product which is produced by different reactions. N^+ and N^- denote neighborhood of a node on the direction of its outgoing and incoming edges, respectively. n is the total number of $W_{r_i m_p}$ plus 1 for m_p. Thus, the equation (1) distributes the substrate proportionally with the activities of its consuming reactions. The equation (2) aims to elucidate the reaction rate (limited by the minimum amount of the substrates used). The amount of metabolite produced per unit time depends on the capacity of enzyme (saturation) and the amount of substrate. This is the combination of Michaelis-Menten kinetics and systems-level analysis of mechanisms regulating metabolic fluxes [62]. The equation (3) updates m_p node with the amount of contributing product of the reaction r_i without saturating this node and it can also handle the loops appropriately. The loops in a sub-pathway need a high number of iterations to stabilize the flux propagated. Thus, the Metabolica iterates the flux that is in a loop until it reaches the convergence state. Here, the convergence state is defined as almost-zero flux change between iterations. Therefore, the Metabolica repeats the steps 1, 2 and 3 until the flux initiated in the initial nodes reaches the product in a sub-pathway and while the flux which is propagated in a loop has not reached convergence. Metabolica input values in the [0,1] interval and returns output values in the same interval. Such results are non-dimensional values that, like gene expression values, can be interpreted in the context of a comparison.

4.4. Samples and data processing

RNA-seq counts and simple somatic mutations data for a total of 1550 samples, 1365 corresponding to tumor and 185 to healthy reference tissues, belonging to breast invasive (BRCA) and kidney renal clear cell (KIRC) carcinomas were downloaded from The International Cancer Genome Consortium (ICGC) repository [63]. The trimmed mean of M-values (TMM) normalization method [64] was used for gene expression normalization. Normalized samples were log-transformed and a truncation by quantile 0.99 was applied. The COMBAT method [65] was used for batch effect correction. Finally, the data was re-scaled between 0 and 1.

Annovar tool [66] with its ljb26 database was used for functional annotation of non-synonymous genetic variants. The variants predicted as damaging by at least 3 out of 5 in-silico pathogenicity predictors were considered loss-of-function (LoF) mutations. These in-silico methods are: SIFT [67], Polyphen2 [68], FATHMM [69], MutationTaster [70], MutationAssessor [71]. For each tumor sample, expression value of the genes that were affected by the damaging variants were multiplicated by a decreasing constant: 0.001, to simulate LoF in the enzyme (equivalent to a non-expressed gene).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 January 2021 doi:10.20944/preprints202101.0280.v1

4.5. Comparative performance of Metabolica

The performance of Metabolica and the RM method in detecting DRMs were compared to real mass spectrophotometry (MS) metabolic profiles. An R implementation of RM method, as described in the original paper [23] was used. As it is recommended in the RM article, the aggregated Z scores were corrected for the background distribution by subtracting the mean and dividing by the standard deviation. For each dataset size, random sampling was repeated 10000 times to calculate background mean and standard deviation statistics. Scaled and imputed metabolomics datasets of BRCA and KIRC were downloaded from supplementary materials of their articles [72, 73]. Additionally, quantile normalization was applied to these datasets using preprocessCore Bioconductor package [74].

Here, the samples and sample sizes which were used in RM and the Metabolica calculations belong to TCGA datasets and they are different than the samples of metabolomics datasets analyzed. But all these datasets contain the paired samples of tumor and healthy. Therefore, Student's t-test for paired samples is used to assess the significance of observed changes of metabolites when samples of two conditions are compared.

Finally, significant (FDR-adjusted $p<0.05$) DRMs of 3 methods were compared. The results of metabolomics profiles are taken as gold standards for this comparison.

4.6. False positive rate of Metabolica

In order to estimate the false positive rate (type I error), different testing datasets of samples were generated from the original dataset, and divided them into two equally sized groups that were further compared to each other for finding DRMs with Metabolica. Since the compared groups are composed of the same type of individuals, any significant change found in sub-pathway activities, can be considered a false positive result. To avoid biases derived from sample size or from the type of sample, different group sample sizes were tested both in cancer and in normal individuals. Sample sizes of groups ranged from 5 to 55 for normal samples and from 50 to 450 for tumor samples, which were also proportional to the total number of normal and tumor samples used in this study. For each sample size, 1000 different simulated samplings were carried out and compared. Student's t-test for unpaired samples used to assess the significance of the observations.

4.7. Clinical Relevance of Metabolite Producing Sub-pathway Activities

Since the Metabolica provides individual-level results the clinical relevance of metabolites can be tested. In order to show the potential prognostic value of sub-pathway activities, tumor stage and overall survival time of BRCA and KIRC patients were analyzed. Clinical data were obtained from the cBIOportal [75].

Cox hazards multivariate regression model was used considering the following variables: metabolite abundances, age, tumor stage, sex, race, and tumor purity. Metabolites having significant Cox proportional hazards regression models (p-value of log rank test < 0.05, in which metabolite abundances presented significant contribution with $p<0.05$) were prognostic metabolite candidates. For the visualization of the survival results Kaplan–Meier (K-M) plots were used. The K-M analysis needs a stratification variable with at least two categorical groups [76]. For this reason, we have discretized the metabolite level values. As a common discretization procedure, for each metabolite, lower 45th and upper 55th percentiles of its abundance were used to categorice the samples into two groups. Samples between these two percentiles were not used. Survival analysis was carried out using survival R package [77]. Similarly to the case of two class comparison, multiple testing effects are corrected by FDR [78].

For the tumor stage analysis continuous values of the metabolite abundances were used. The tumor stage codes of American Joint Committee on Cancer, which are divided into four categories (T1, T2, T3 and T4) were used for this analysis. Samples with detailed subdivision codes were pooled under their main designators (e.g. T3′ = |T3| U |T3a| U |T3b| U … |T3n|). To find metabolites with a significant role in tumor progression, a linear correlation between tumor stages and metabolite abundances was used.

5. Conclusions

The Metabolica provides a simple and elegant algorithmic framework for genome-scale metabolic analysis of transcriptomic and genomic data, which accounts for the complexity of the relationships between proteins within metabolic pathways, and delivers estimations of metabolite production activity.

Supplementary Materials: Table S1: Death events for the survival analysis. Table S2: List of human metabolic pathways, reactions and metabolites downloaded from KEGG and used in this study.

Author Contributions: Conceptualization, C.L. M.P.C. and C.C.; methodology, M.R.H. and J.C.C..; software, C.C..; validation, C.C. and A.A.; writing—original draft preparation, C.C.; writing—review and editing, J.D.; supervision, J.D.; funding acquisition, J.D. All authors have read and agreed to the published version of the manuscript.

Funding: Please add: This research was funded by grants SAF2017-88908-R from the Spanish Ministry of Economy and Competitiveness and PT17/0009/0006 from the ISCIII, both co-funded with European Regional Development Funds (ERDF) as well as by European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie (GA 813533) and "ELIXIR-EXCELERATE fast-track ELIXIR implementation and drive early user exploitation across the life sciences" (GA 676559).

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. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B: **The reactome pathway knowledgebase.** *Nucleic acids research* 2017, **46:**D649-D655.
2. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: **KEGG as a reference resource for gene and protein annotation.** *Nucleic acids research* 2015, **44:**D457-D462.
3. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D: **WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research.** *Nucleic acids research* 2017.
4. Ostaszewski M, Gebel S, Kuperstein I, Mazein A, Zinovyev A, Dogrusoz U, Hasenauer J, Fleming RM, Le Novere N, Gawron P: **Community-driven roadmap for integrated disease maps.** *Briefings in bioinformatics* 2019, **20:**659-670.
5. Cubuk C, Hidalgo MR, Amadoz A, Pujana MA, Mateo F, Herranz C, Carbonell-Caballero J, Dopazo J: **Gene expression integration into pathway modules reveals a pan-cancer metabolic landscape.** *Cancer research* 2018, **78:**6059-6072.
6. Nowicki S, Gottlieb E: **Oncometabolites: tailoring our genes.** *Febs j* 2015, **282:**2796-2805.
7. Warburg O: **The metabolism of carcinoma cells.** *J Cancer Res* 1925 **9:**148-163.
8. Carracedo A, Cantley LC, Pandolfi PP: **Cancer metabolism: fatty acid oxidation in the limelight.** *Nat Rev Cancer* 2013, **13:**227-232.
9. Terunuma A, Putluri N, Mishra P, Mathé EA, Dorsey TH, Yi M, Wallace TA, Issaq HJ, Zhou M, Killian JK, et al: **MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis.** *J Clin Invest* 2014, **124:**398-412.
10. Hakimi AA, Reznik E, Lee C-H, Creighton CJ, Brannon AR, Luna A, Aksoy BA, Liu EM, Shen R, Lee W, et al: **An Integrated Metabolic Atlas of Clear Cell Renal Cell Carcinoma.** *Cancer Cell* 2016, **29:**104-116.
11. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, Perry MD, Nahal-Bose HK, Ouellette BFF, Li CH, et al: **Pan-cancer analysis of whole genomes.** *Nature* 2020, **578:**82-93.

12. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature* 2012, **486:**346-352.
13. Eicher T, Kinnebrew G, Patt A, Spencer K, Ying K, Ma Q, Machiraju R: **Metabolomics and Multi-Omics Integration: A Survey of Computational Methods and Resources.** *Metabolites* 2020, **10:**202.
14. Goeman JJ, Buhlmann P: **Analyzing gene expression data in terms of gene sets: methodological issues.** *Bioinformatics* 2007, **23:**980-987.
15. Bordbar A, Monk JM, King ZA, Palsson BO: **Constraint-based models predict metabolic and associated cellular functions.** *Nat Rev Genet* 2014, **15:**107-120.
16. Nam H, Campodonico M, Bordbar A, Hyduke DR, Kim S, Zielinski DC, Palsson BO: **A systems approach to predict oncometabolites via context-specific genome-scale metabolic networks.** *PLoS Comput Biol* 2014, **10:**e1003837.
17. Havas KM, Milchevskaya V, Radic K, Alladin A, Kafkia E, Garcia M, Stolte J, Klaus B, Rotmensz N, Gibson TJ, et al: **Metabolic shifts in residual breast cancer drive tumor recurrence.** *The Journal of clinical investigation* 2017, **127:**2091-2105.
18. Orth JD, Thiele I, Palsson BO: **What is flux balance analysis?** *Nat Biotechnol* 2010, **28:**245-248.
19. Khatri P, Sirota M, Butte AJ: **Ten years of pathway analysis: current approaches and outstanding challenges.** *PLoS Comput Biol* 2012, **8:**e1002375.
20. Opdam S, Richelle A, Kellman B, Li S, Zielinski DC, Lewis NE: **A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models.** *Cell Syst* 2017.
21. Chindelevitch L, Trigg J, Regev A, Berger B: **An exact arithmetic toolbox for a consistent and reproducible structural analysis of metabolic network models.** *Nature communications* 2014, **5:**4893-4893.
22. Ebrahim A, Almaas E, Bauer E, Bordbar A, Burgard AP, Chang RL, Dräger A, Famili I, Feist AM, Fleming RM, et al: **Do genome-scale models need exact solvers or clearer standards?** *Molecular systems biology* 2015, **11:**831-831.
23. Patil KR, Nielsen J: **Uncovering transcriptional regulation of metabolism by using metabolic network topology.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102:**2685-2689.
24. Çakır T: **Reporter pathway analysis from transcriptome data: Metabolite-centric versus Reaction-centric approach.** *Scientific Reports* 2015, **5:**14563.
25. Auslander N, Wagner A, Oberhardt M, Ruppin E: **Data-driven metabolic pathway compositions enhance cancer survival prediction.** *PLoS computational biology* 2016, **12:**e1005125.
26. Amadoz A, Hidalgo MR, Çubuk C, Carbonell-Caballero J, Dopazo J: **A comparison of mechanistic signaling pathway activity analysis methods.** *Briefings in bioinformatics* 2019, **20:**1655–1668.
27. Hidalgo MR, Cubuk C, Amadoz A, Salavert F, Carbonell-Caballero J, Dopazo J: **High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes.** *Oncotarget* 2017, **8:**5160-5178.
28. Fey D, Halasz M, Dreidax D, Kennedy SP, Hastings JF, Rauch N, Munoz AG, Pilkington R, Fischer M, Westermann F, et al: **Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients.** *Sci Signal* 2015, **8:**ra130.
29. Hidalgo MR, Amadoz A, Cubuk C, Carbonell-Caballero J, Dopazo J: **Models of cell signaling uncover molecular mechanisms of high-risk neuroblastoma and predict disease outcome** *Biology Direct* 2018, **13:**16.

30. Falco MM, Peña-Chilet M, Loucera C, Hidalgo MR, Dopazo J: **Mechanistic models of signaling pathways deconvolute the glioblastoma single-cell functional landscape.** *NAR Cancer* 2020, **2**.
31. Amadoz A, Sebastian-Leon P, Vidal E, Salavert F, Dopazo J: **Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity.** *Sci Rep* 2015, **5:**18494.
32. Esteban-Medina M, Peña-Chilet M, Loucera C, Dopazo J: **Exploring the druggable space around the Fanconi anemia pathway using machine learning and mechanistic models.** *BMC Bioinformatics* 2019, **20:**370.
33. Çubuk C, Can FE, Peña-Chilet M, Dopazo J: **Mechanistic Models of Signaling Pathways Reveal the Drug Action Mechanisms behind Gender-Specific Gene Expression for Cancer Treatments.** *Cells* 2020, **9:**1579.
34. Salavert F, Hidalgo MR, Amadoz A, Cubuk C, Medina I, Crespo D, Carbonell-Caballero J, Dopazo J: **Actionable pathways: interactive discovery of therapeutic targets using signaling pathway models.** *Nucleic Acids Res* 2016, **44:**W212-216.
35. Muto A, Kotera M, Tokimatsu T, Nakagawa Z, Goto S, Kanehisa M: **Modular architecture of metabolic pathways revealed by conserved sequences of reactions.** *J Chem Inf Model* 2013, **53:**613-622.
36. Çubuk C, Hidalgo MR, Amadoz A, Rian K, Salavert F, Pujana MA, Mateo F, Herranz C, Carbonell-Caballero J, Dopazo J, Applications: **Differential metabolic activity and discovery of therapeutic targets using summarized metabolic pathway models.** *NPJ Systems Biology* 2019, **5:**7.
37. Li X, Li C, Shang D, Li J, Han J, Miao Y, Wang Y, Wang Q, Li W, Wu C, et al: **The implications of relationships between human diseases and metabolic subpathways.** *PloS one* 2011, **6:**e21131-e21131.
38. Galluzzi L, Vitale I, Senovilla L, Olaussen Ken A, Pinna G, Eisenberg T, Goubar A, Martins I, Michels J, Kratassiouk G, et al: **Prognostic Impact of Vitamin B6 Metabolism in Lung Cancer.** *Cell Reports* 2012, **2:**257-269.
39. Courtney KD, Bezwada D, Mashimo T, Pichumani K, Vemireddy V, Funk AM, Wimberly J, McNeil SS, Kapur P, Lotan Y, et al: **Isotope Tracing of Human Clear Cell Renal Cell Carcinomas Demonstrates Suppressed Glucose Oxidation In Vivo.** *Cell Metabolism* 2018, **28:**793-800.e792.
40. Shaul YD, Freinkman E, Comb WC, Cantor JR, Tam WL, Thiru P, Kim D, Kanarek N, Pacold ME, Chen WW, et al: **Dihydropyrimidine accumulation is required for the epithelial-mesenchymal transition.** *Cell* 2014, **158:**1094-1109.
41. Zhao T, Mu X, You Q: **Succinate: An initiator in tumorigenesis and progression.** *Oncotarget* 2017, **8:**53819-53828.
42. Reznik E, Luna A, Aksoy BA, Liu EM, La K, Ostrovnaya I, Creighton CJ, Hakimi AA, Sander C: **A Landscape of Metabolic Variation across Tumor Types.** *Cell Syst* 2018, **6:**301-313.e303.
43. Arakaki AK, Skolnick J, McDonald JF: **Marker metabolites can be therapeutic targets as well.** *Nature* 2008, **456:**443-443.
44. Anaya J, Reon B, Chen WM, Bekiranov S, Dutta A: **A pan-cancer analysis of prognostic genes.** *PeerJ* 2015, **3:**e1499.
45. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al: **An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.** *Cell* 2018, **173:**400-416.e411.
46. Casero RA, Murray Stewart T, Pegg AE: **Polyamine metabolism and cancer: treatments, challenges and opportunities.** *Nature Reviews Cancer* 2018, **18:**681-695.

47. Hayes CS, Shicora AC, Keough MP, Snook AE, Burns MR, Gilmour SK: **Polyamine-blocking therapy reverses immunosuppression in the tumor microenvironment.** *Cancer immunology research* 2014, **2:**274-285.
48. Mandal S, Mandal A, Johansson HE, Orjalo AV, Park MH: **Depletion of cellular polyamines, spermidine and spermine, causes a total arrest in translation and growth in mammalian cells.** *Proceedings of the National Academy of Sciences* 2013, **110:**2169-2174.
49. Martínez-Reyes I, Chandel NS: **Acetyl-CoA-directed gene transcription in cancer cells.** *Genes & development* 2018, **32:**463-465.
50. Yoshii Y, Furukawa T, Saga T, Fujibayashi Y: **Acetate/acetyl-CoA metabolism associated with cancer fatty acid synthesis: overview and application.** *Cancer letters* 2015, **356:**211-216.
51. Mashima T, Sato S, Okabe S, Miyata S, Matsuura M, Sugimoto Y, Tsuruo T, Seimiya H: **Acyl-CoA synthetase as a cancer survival factor: its inhibition enhances the efficacy of etoposide.** *Cancer Sci* 2009, **100:**1556-1562.
52. Huang D, Li T, Li X, Zhang L, Sun L, He X, Zhong X, Jia D, Song L, Semenza Gregg L, et al: **HIF-1-Mediated Suppression of Acyl-CoA Dehydrogenases and Fatty Acid Oxidation Is Critical for Cancer Progression.** *Cell Reports* 2014, **8:**1930-1942.
53. Pietrocola F, Galluzzi L, Bravo-San Pedro JM, Madeo F, Kroemer G: **Acetyl coenzyme A: a central metabolite and second messenger.** *Cell metabolism* 2015, **21:**805-821.
54. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K: **KEGG: new perspectives on genomes, pathways, diseases and drugs.** *Nucleic acids research* 2016, **45:**D353-D361.
55. Zhang JD, Wiemann S: **KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor.** *Bioinformatics* 2009, **25:**1470-1471.
56. Liberti MV, Locasale JW: **The Warburg effect: how does it benefit cancer cells?** *Trends in biochemical sciences* 2016, **41:**211-218.
57. Smith B, Schafer XL, Ambeskovic A, Spencer CM, Land H, Munger J: **Addiction to Coupling of the Warburg Effect with Glutamine Catabolism in Cancer Cells.** *Cell Rep* 2016, **17:**821-836.
58. Blazier AS, Papin JA: **Integration of expression data in genome-scale metabolic network reconstructions.** *Frontiers in physiology* 2012, **3:**299-299.
59. Kim MK, Lun DS: **Methods for integration of transcriptomic data in genome-scale metabolic models.** *Computational and structural biotechnology journal* 2014, **11:**59-65.
60. Machado D, Herrgård M: **Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism.** *PLoS Comput Biol* 2014, **10:**e1003580.
61. Machado D, Herrgård MJ, Rocha I: **Stoichiometric Representation of Gene-Protein-Reaction Associations Leverages Constraint-Based Analysis from Reaction to Gene-Level Phenotype Prediction.** *PLoS computational biology* 2016, **12:**e1005140-e1005140.
62. Hackett SR, Zanotelli VRT, Xu W, Goya J, Park JO, Perlman DH, Gibney PA, Botstein D, Storey JD, Rabinowitz JD: **Systems-level analysis of mechanisms regulating yeast metabolic flux.** *Science* 2016, **354:**aaf2786.
63. **The International Cancer Genome Consortium (ICGC) repository** [https://dcc.icgc.org/releases/release_26/Projects]
64. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11:**R25.
65. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8:**118-127.

66. Yang H, Wang K: **Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR.** *Nature protocols* 2015, **10:**1556-1566.
67. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31:**3812-3814.
68. Adzhubei I, Jordan DM, Sunyaev SR: **Predicting functional effect of human missense mutations using PolyPhen-2.** *Current protocols in human genetics* 2013, **76:**7.20. 21-27.20. 41.
69. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C: **FATHMM-XF: accurate prediction of pathogenic point mutations via extended features.** *Bioinformatics* 2017, **34:**511-513.
70. Schwarz JM, Cooper DN, Schuelke M, Seelow D: **MutationTaster2: mutation prediction for the deep-sequencing age.** *Nature methods* 2014, **11:**361.
71. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics.** *Nucleic acids research* 2011, **39:**e118.
72. Terunuma A, Putluri N, Mishra P, Mathé EA, Dorsey TH, Yi M, Wallace TA, Issaq HJ, Zhou M, Killian JK: **MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis.** *The Journal of clinical investigation* 2014, **124:**398-412.
73. Hakimi AA, Reznik E, Lee C-H, Creighton CJ, Brannon AR, Luna A, Aksoy BA, Liu EM, Shen R, Lee W: **An integrated metabolic atlas of clear cell renal cell carcinoma.** *Cancer cell* 2016, **29:**104-116.
74. Bolstad BM, Irizarry RA, Åstrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19:**185-193.
75. **cBIOportal** [http://www.cbioportal.org]
76. Kaplan E, Meier P: **Nonparametric estimation from incomplete observations.** *Journal of the American Statistical Association* 1958, **53:**457-481.
77. **A Package for Survival Analysis in R** [https://CRAN.R-project.org/package=survival]
78. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society Series B* 1995, **57:**289-300.