UC San Diego

UC San Diego Previously Published Works

Title

Assessing key decisions for transcriptomic data integration in biochemical networks

Permalink

https://escholarship.org/uc/item/07w948k1

Journal

PLOS Computational Biology, 15(7)

ISSN

1553-734X

Authors

Richelle, Anne
Joshi, Chintan
Lewis, Nathan E

Publication Date

2019

DOI

10.1371/journal.pcbi.1007185

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Assessing key decisions for transcriptomic data integration in biochemical networks

Anne Richelle[1,2], **Chintan Joshi**[1,2], **Nathan E. Lewis**[1,2,3]*

1 Novo Nordisk Foundation Center for Biosustainability at the University of California, San Diego, School of Medicine, La Jolla, California, United States of America, **2** Department of Pediatrics, University of California, San Diego, School of Medicine, La Jolla, California, United States of America, **3** Department of Bioengineering, University of California, San Diego, La Jolla, California, United States of America

* nlewisres@ucsd.edu

Check for updates

OPEN ACCESS

Citation: Richelle A, Joshi C, Lewis NE (2019) Assessing key decisions for transcriptomic data integration in biochemical networks. PLoS Comput Biol 15(7): e1007185. https://doi.org/10.1371/journal.pcbi.1007185

Editor: Vassily Hatzimanikatis, Ecole Polytechnique Fédérale de Lausanne, SWITZERLAND

Received: November 10, 2018

Accepted: June 14, 2019

Published: July 19, 2019

Copyright: © 2019 Richelle et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The Human Protein Atlas transcriptomic dataset (HPA) were acquired from supplementary material of Uhlen et al., 2015. Recon 2.2 model was downloaded from https://github.com/u003f/recon2/tree/master/models.

Funding: This work was supported by a Lilly Innovation Fellowship Award (Elli Lilly Company - https://www.lilly.com) to A.R., a Keck Fellowship (W. M. Keck Foundation - http://www.wmkeck.org) to C.J., and grants to N.E.L. from Novo Nordisk Foundation: Center for Biosustainability at the Technical University of Denmark (grant no.

Abstract

To gain insights into complex biological processes, genome-scale data (e.g., RNA-Seq) are often overlaid on biochemical networks. However, many networks do not have a one-to-one relationship between genes and network edges, due to the existence of isozymes and protein complexes. Therefore, decisions must be made on how to overlay data onto networks. For example, for metabolic networks, these decisions include (1) how to integrate gene expression levels using gene-protein-reaction rules, (2) the approach used for selection of thresholds on expression data to consider the associated gene as "active", and (3) the order in which these steps are imposed. However, the influence of these decisions has not been systematically tested. We compared 20 decision combinations using a transcriptomic dataset across 32 tissues and showed that definition of which reaction may be considered as active (i.e., reactions of the genome-scale metabolic network with a non-zero expression level after overlaying the data) is mainly influenced by thresholding approach used. To determine the most appropriate decisions, we evaluated how these decisions impact the acquisition of tissue-specific active reaction lists that recapitulate organ-system tissue groups. These results will provide guidelines to improve data analyses with biochemical networks and facilitate the construction of context-specific metabolic models.

Author summary

Over the past two decades, systems biology approaches have permeated all fields of biology. Indeed, biological networks are commonly used for the analysis of omics datasets and to test hypotheses through computational simulations. Many of these types of studies require one to overlay omics data to a biological network. In many biochemical networks, such as metabolic networks, there is not a one-to-one relationship between genes and network edges (e.g., metabolic reactions), due to the existence of isozymes and protein complexes. Therefore, decisions must be made on how to overlay data onto networks. For example, these decisions include (1) how to integrate gene expression levels using the Boolean relationships between genes, proteins, enzymes, and reactions, (2) the selection of thresholds on expression data to consider the associated gene as "active" or "inactive",

NNF16CC0021858 - http://novonordiskfonden.dk/en) and from National Institute of General Medical Sciences (grant no. R35 GM119850 - https://www.nigms.nih.gov). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

and (3) the order in which these steps are imposed. While these decisions have been made in many published studies, there has been no systematic evaluation of the impact of these decisions on the biological accuracy of the resulting networks. To this end, we benchmarked the different existing decisions made to integrate the data into the network. The results provide guidelines to improve data analyses with biochemical networks that should translate to many other network types.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Most biological systems can be structured as networks, from cell signaling pathways to cell metabolism. These networks are invaluable for describing and understanding complex biological processes. For example, metabolic network reconstructions can illuminate the molecular basis of phenotypes exhibited by an organism, when used as a platform for analyzing data measuring gene expression, protein expression, enzymatic activity, or metabolite concentrations. For these analyses, the data are overlaid on the biological networks using Boolean rules that describe the relationship between the measured molecules (e.g., mRNAs, metabolites) and the network edges and nodes. These logical rules capture how the molecules influence each other's activity (i.e., activation, inhibition, or cooperation), and allow users to quantify each network edge or define the status of each network component as either "on" or "off". Therefore, a biological process can be described in a given context by adding or removing nodes and/or edges based on genome-scale data.

Genome-scale metabolic networks utilize this Boolean formulation connecting genes to reaction, and therefore have been used extensively as platforms for analyzing mRNA expression data to elucidate how changes in gene expression impacts cell phenotypes [1–8]. These studies have spanned diverse applications from identification of disease mechanisms [9,10] to identification of drug targets [11,12], and the evaluation of cell responses to drugs [13]. Despite the success of the many studies integrating omics data with biochemical networks, there are several challenges in the integration of omics data with networks that are infrequently discussed. These challenges impact the accuracy of context-specific networks, and include experimental and inherent biological noise, differences among experimental platforms, detection bias, and the unclear relationship between gene expression and reaction flux [14]. Furthermore, algorithmic assumptions influence the quality and functionality of resulting models and the physiological accuracy of their predictions [15–19].

While previous work has discussed the impact of various algorithms on obtaining physiologically accurate metabolic networks, the influence of the initial steps of data integration with biological networks has not been clearly evaluated and discussed in the literature. Thus, no universal rules have been established on how to integrate transcriptomic data, referred to here as "preprocessing", leading often to inappropriate decisions for model development. These preprocessing steps include (1) how to account for network elements (e.g., reactions) that do not have a one-to-one relationship with genes and reactions (e.g., isozymes, complexes, and promiscuous enzymes), referred to here as *gene mapping*, and (2) how to define which genes are expressed or not, referred to here as *thresholding*, and (3) the order of gene mapping and thresholding in data integration. Here we evaluate the influence of the transcriptomic

preprocessing steps and their consequences on the biological meaning captured by the data. Specifically, we do this by evaluating 20 different combinations of preprocessing steps, using transcriptomic data from 32 tissues. By evaluating the resulting 640 tissue-specific active reaction lists, we identify which decisions have the largest impact on list content, and which decisions best capture the similarities seen within tissues from the same organ-systems. This study aims to help researcher make proper decision on omics data integration by stress-testing existing methods and proposing improvements on their implementation. This results in guidelines for overlaying transcriptomic data in metabolic networks and the lessons learned should be applicable to the analysis of transcriptomic data in all sorts of biological networks used for systems biology analysis.

Results

Preprocessing decisions for overlaying omics data on biochemical networks

Biochemical and other network types provide valuable platforms for analyzing and interpreting data. In these networks, links between nodes often represent enzyme-catalyzed reactions, and as such there is often not a one-to-one relationship between the genes and reactions. This relationship is represented using logical rules, referred as Gene-Protein-Reaction rules (GPRs, Fig A in S1 Text). When overlaying mRNA abundances on biochemical networks, GPRs are used to define which genes are the main determinants of the enzyme activity catalyzing a reaction. We refer here to this step as *gene mapping*. The most common assumption for multimeric enzyme complexes is that the gene with the minimum expression governs the activity. For isoenzymes, the activity may either depend on the total expression of all isoenzyme genes [20] or the isoenzyme gene with highest expression [21] (Fig 1A).

Furthermore, the absolute mRNA abundance is often considered to represent a gene's potential activity by using a thresholding approach. That is, if the gene is expressed at a level above a threshold, it is often considered to be active. This threshold definition has been implemented in many different ways in the literature, from the use of only a single threshold to more complex rules involving multiple thresholds. For example, one unique threshold value can be applied to all genes (i.e., the global thresholding approach, [22,23]) while others have applied different thresholds to each gene (i.e., the local thresholding approach, [24,25]) (Fig 1B). When using one single threshold in a global context (i.e., global T1), the genes presenting an expression above this value are considered as active (i.e., ON) while the others are inactive (i.e., OFF) (Fig 1C). However, when multiple samples are available, one can compute a gene-specific threshold based on the distribution of the expression levels observed for this gene over all the samples (e.g., a local rule that sets a threshold equal to the mean expression level across all samples). This gene-specific thresholding approach can be implemented in combination with a defined global threshold for genes presenting low expression values among all the samples (e.g., below the usual detection level associated to the measurement method) to prevent their inclusion with the active genes for some samples (i.e., local T1). Therefore, the genes whose expression is below the value defined by this global lower bound will always be considered as inactive (i.e., OFF), while other genes will fall under the local rule for gene-specific threshold definition (i.e., MAYBE ON) (Fig 1C). Another similar extreme case can be encountered when the gene expression level is high in all the samples. Therefore, we propose to also analyze the influence of using one lower and one upper threshold values defined based on the distribution of expression level off all the genes in all the samples (i.e. local T2). Doing so, the local rule for gene-specific threshold definition is actually applied only to the genes whose expression is between the range of values defined by the lower and upper bounds (i.e., MAYBE ON), ensuring that genes presenting low expression values among all the samples are always

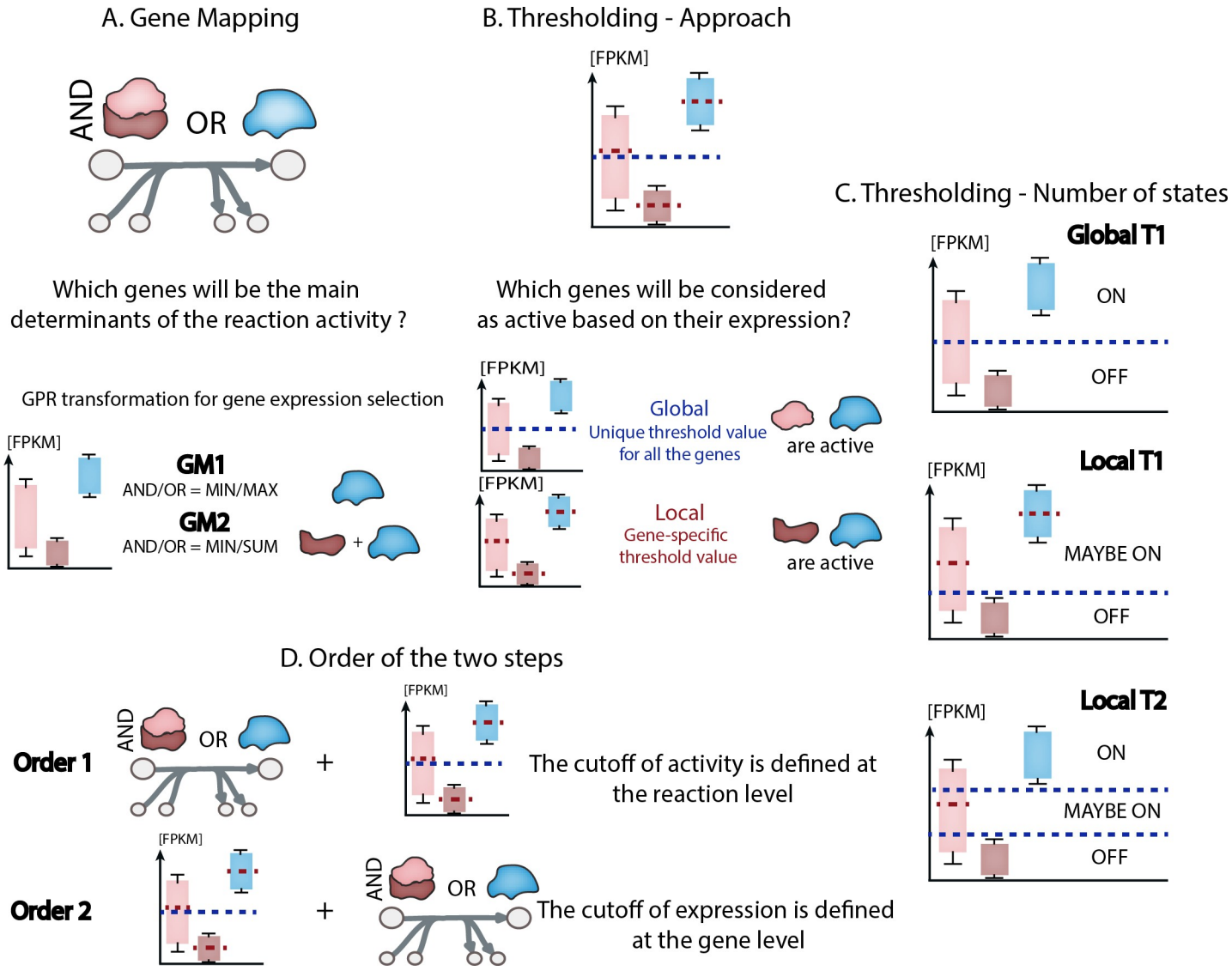

Fig 1. Formulation and implementation of preprocessing decisions. (A) Two types of gene mapping methods (GM1 and GM2) are compared. (B) Two types of thresholding approaches (global and local) are compared. (C) Formulation of three combinations of number of states (Global T1, Local T1, and Local T2) (D) Decisions about the order in which thresholding and gene mapping are performed. For Order 1, gene expression is converted to reaction activity followed by thresholding of reaction activity; for Order 2, thresholding of gene expression is followed by its conversion to reaction activity.

https://doi.org/10.1371/journal.pcbi.1007185.g001

considered as inactive (i.e., OFF) while the ones with very high expression values among all the samples are always considered as active (i.e., ON) (Fig 1C).

Preprocessing of transcriptomic data for their integration into biochemical networks relies mainly on these two decisions: *gene mapping* and *thresholding*, but these can be implemented in different orders, with either gene mapping or thresholding occurring first (Fig 1D). Therefore, multiple combinations of these decisions could be made when overlaying data onto biochemical networks, and these decisions may influence the data integration and the subsequent biological interpretation (see Table 1, Fig 1, and a detailed explanation about the decisions presented in Methods section).

Here, we integrated transcriptomic data from 32 different tissues in the Human Protein Atlas [25] with the Human genome scale model Recon 2.2 [26] using 20 different

Table 1. Decisions involved in transcriptomic data preprocessing.

Decisions	Variables	Existing approaches	Biological meaning
Gene Mapping	GPR transformation for expression selection	AND/OR = MIN/MAX	Isoenzyme reaction activity is given by isoenzymes presenting the maximum activity
		AND/OR = MIN/SUM	Isoenzyme reaction activity is given by the sum of the isoenzyme activities
Thresholding	Approach	local	Gene-specific threshold values
		global	Unique threshold value for all the genes
	Number of states	2 states = 1 global threshold	OFF/ON
		2 states = 1 global threshold and 1 local rule	OFF/MAYBE ON
		3 states = 2 global thresholds and 1 local rule	OFF/MAYBE ON/ ON
Order of the steps	Gene Mapping (GM) Thresholding (T)	GM + T	The cutoff of activity is defined at the reaction level
		T + GM	The cutoff of expression is defined at the gene level

https://doi.org/10.1371/journal.pcbi.1007185.t001

combinations of the 3 main preprocessing decisions (Table 1, Fig 1). This resulted in 640 different tissue-specific profiles of "expression" values for all gene-associated reactions in Recon 2.2. To specifically evaluate the immediate impact of the preprocessing decisions on the resulting networks (i.e., list of reactions of a genome-scale model (GEM) with a non-zero expression level after overlaying the data), we focused our analysis on the content of the networks themselves (i.e., the definition of active biochemical pathways therein) and the biological interpretation of these networks.

Active reaction sets are influenced by preprocessing decisions

Decisions regarding gene mapping, thresholding (i.e., approach and number of states), and order of steps affect the definition of active reaction sets. Specifically, the sets of active reactions (i.e., reactions of the GEM with a non-zero expression level after overlaying the data) varied considerably in size from 358 reactions to 3286 reactions across all tissues, depending on preprocessing decisions and tissue type (Fig 2A). To assess the impact of each decision, we conducted a principal component analysis (PCA) of the reaction sets considered as active, depending on the preprocessing decisions (i.e., a PCA on the matrix of all active reactions vs. all combinations of decisions and tissues; see Methods for details). The first principal component explains >35% of the overall variance in active reaction content (Fig 2B). The thresholding related parameters (global/local and T1/T2) provide the most significant contribution to the variation in the first principal component (38.5%), with the differences between the global and local approaches having the greatest impact (Fig 2C, 2F and Fig B in S1 Text). The effect of thresholding impacted the networks more than the differences across tissues, which only explained 15.8% of the variation in the first principal component. Tissue specific effects did not dominate until the second principal component, where it explained 73% of the variation in the component. The order of the preprocessing steps only provides a small contribution to the explained variation in the first principal component (Fig 2C and 2D). Meanwhile, the type of gene mapping has the least influence on active reaction sets (Fig 2C and 2E). These results indicate that the identification of active reactions is most heavily affected by the thresholding approach (as defined in Fig 1B), followed by the state definition used for thresholding (as defined in Fig 1C) and the order of preprocessing steps (as defined in Fig 1D) while the gene mapping method does not seem to have an influence.

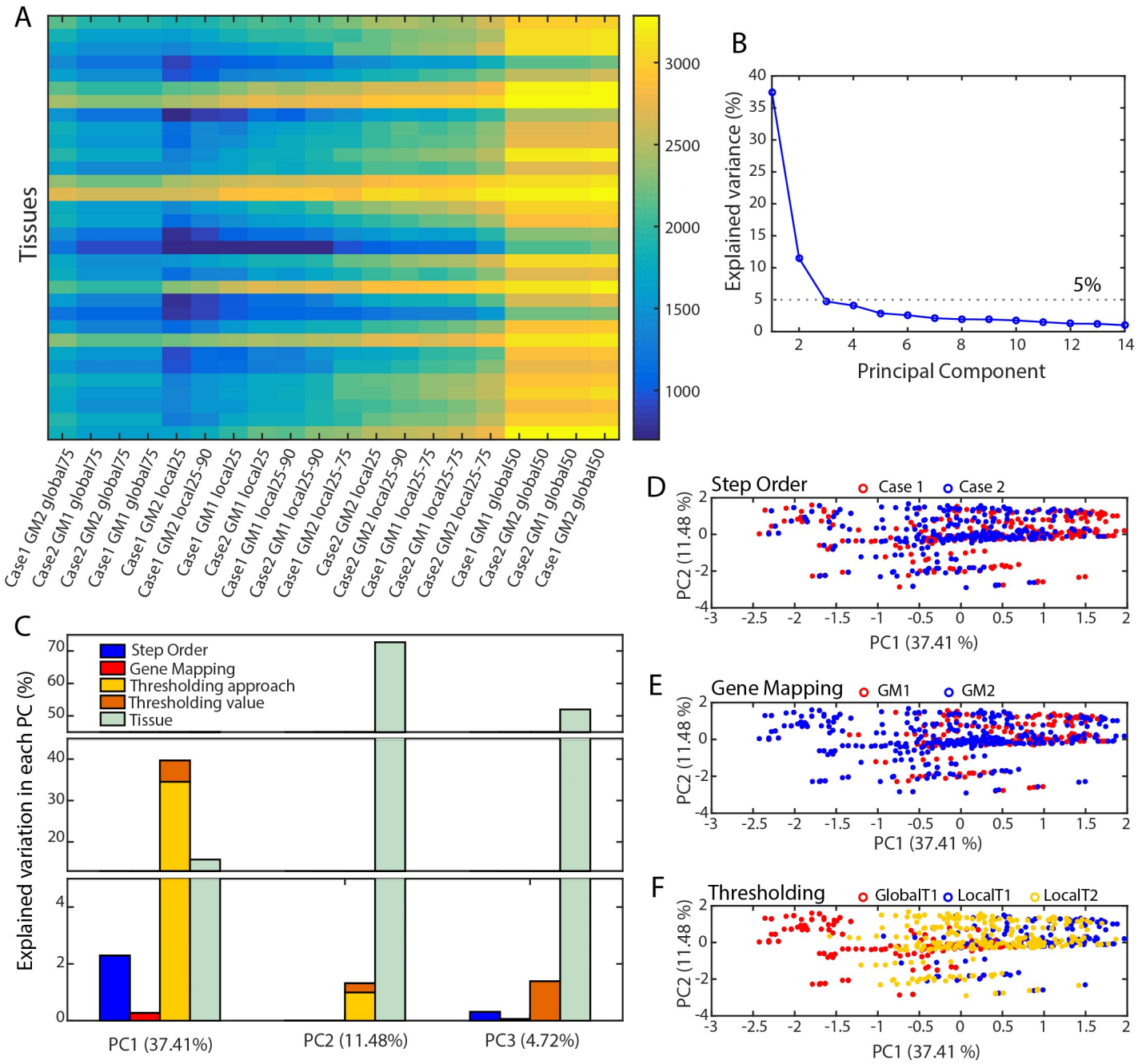

Fig 2. Preprocessing decisions affect the definition of active reactions sets. (A) Twenty different combinations of preprocessing decisions led to a large diversity number of reactions considered as active. (B) The first three principal components (PCs) explain most of the variance in the number of active reactions in a GEM. (C) Thresholding contributes the most to the first PC and more specifically the main contributor is the thresholding approach (i.e. local or global). (D, E and F). The influence of thresholding parameter selection is clear in the first PC (F), while the networks are less influenced by the gene mapping method (E) and the order of preprocessing steps used (D).

https://doi.org/10.1371/journal.pcbi.1007185.g002

Preprocessing decisions influence ability to capture tissue similarities within organ-systems

We assessed the similarities of tissues belonging to the same organ-system, based on the knowledge of the set of active reactions. We assumed that organ-system groups are formed by

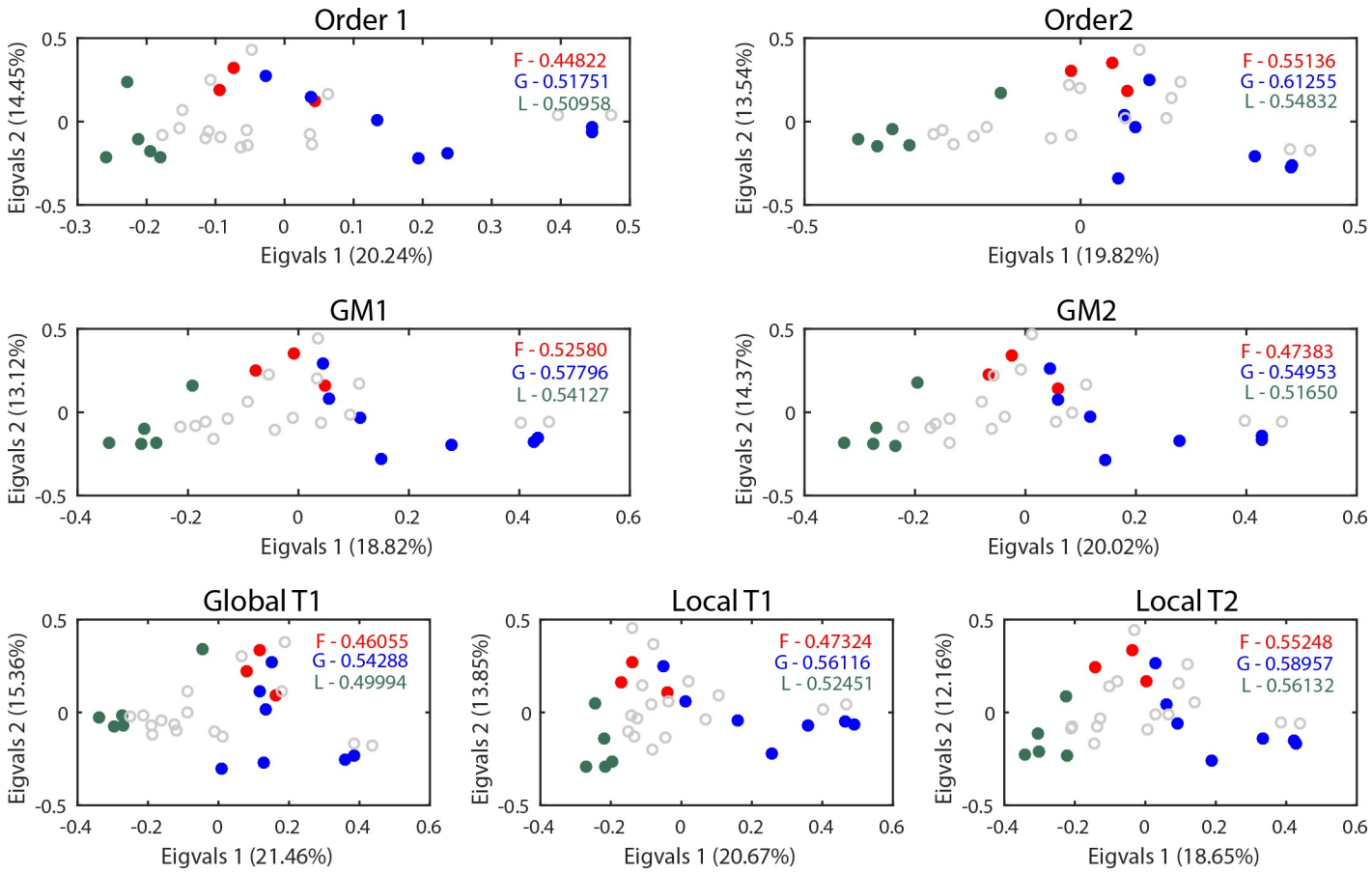

Fig 3. Influence of preprocessing decisions on capturing tissue similarities. Visual representation using a Principal Coordinates Analysis of the similarity between tissues grouped by organ system for each preprocessing decision (numbers in legends are the mean Euclidean distance of the tissues belonging to each group; F–Female reproductive group, G–Gastrointestinal group, and L–Lymphatic group).

https://doi.org/10.1371/journal.pcbi.1007185.g003

tissues working collaboratively to achieve a specific function (e.g., the gastrointestinal system turns food into energy). Therefore, we hypothesized that similarities of tissues within an organ system may lead to a more similar set of active metabolic reactions within the system, in comparison to other systems, as suggested by previous transcriptomic analyses [27,28]. To this end, we calculated Euclidean distances between pairs of tissues belonging to the same organ-system (Fig 3, Fig C in S1 Text, see Methods for more details). Our results highlight the influence of preprocessing decisions on the significance of tissue grouping at the reaction level. Moreover, we observed that some decisions improved the significance of tissue grouping: Order 2 works generally better than Order 1. Local T2 also is better than GlobalT1 and LocalT1. However, there was not a clearly superior approach for gene mapping in our analysis (Fig 4, Fig D in S1 Text).

Some organ classification systems will group dissimilar organs together into a single organ-system, and we wondered if our analysis would still suggest the removal of such tissues from the organ-systems based on metabolic differences. For example, our previous analysis was done without associating the placenta to the *Female reproductive* organ-system group. However, the Human Protein Atlas groups it into the *Female reproductive* organ group (S1 Table). The placenta is functionally and histologically different from the other tissues of this group, being derived from both maternal and fetal tissue. This biological difference was successfully

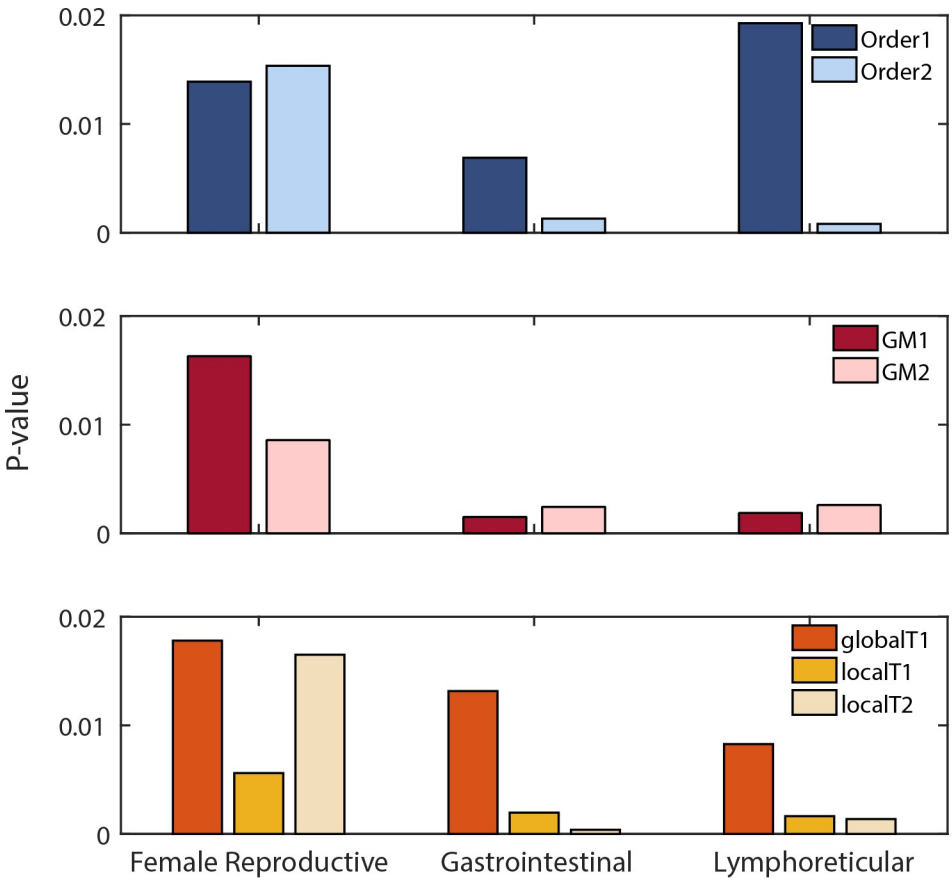

Fig 4. Preprocessing decisions influence the significance of tissue grouping at organ-system level. We compared the mean Euclidean distance observed between tissues belonging to the same organ-system to the mean Euclidean distance for 10000 randomly selected groups with the same number of tissues. The significance of the grouping (P-value) is computed as the proportion of random distances lower than the observed distance for each organ-system.

https://doi.org/10.1371/journal.pcbi.1007185.g004

captured when we compared the tissue similarity analysis with and without the placenta in the *Female reproductive* organ-system group (Fig E in S1 Text).

LocalT2 reduces false negative predictions

Our results above showed that tissues belonging to the same organ system grouped together. This suggests that active reaction sets of tissues contain biological meaning that facilitates grouping tissues belonging to the same organ system. To this end, we identified pathways known to be active in a tissue based on literature. We found 154 pathway-tissue pairs (i.e. a given pathway is known to take place in a given tissue; S2 Table; Fig F1 in S1 Text) and used it to evaluate different thresholding methods (See Methods, Fig F in S1 Text). This resource suggested that pathways could be classified into three categories based on observed ubiquity (i.e., how many tissues wherein the pathways were active): tissue-specific (low ubiquity, pathway is known to occur in 1–2 tissues); group-specific (medium ubiquity, pathway is known to occur in 3–10 tissues belonging to the same organ-system group); and ubiquitous (pathway is known to occur in nearly all tissues) (Fig F2 in S1 Text). The coverage of each of the pathways (as defined in Recon 2.2) in each of the active reaction sets was evaluated using two metrics: (i) ubiquity, and (ii) false negative rate. We evaluate here the false negative rate (i.e., when a pathway known to be present in a tissue is not enriched in this tissue) as opposed to false positive

rate since pathways may also be present in non-canonical tissues (i.e. tissues in which the pathway may be poorly studied).

We first quantified across tissues the ubiquity of each pathway for each thresholding method, resulting in a predicted ubiquity matrix. Pathways grouped together into 5 clusters (Fig G4 in S1 Text). We found that 2 of our 3 ubiquitous pathways matched the pathway cluster P5, 2 out of our 6 group-specific pathways were found in clusters P2 & P3, and 8 of the 20 tissue-specific pathways matched to cluster P4 (Fig G4 in S1 Text). The coverage of pathway types in different pathway clusters suggests that different thresholding methods can be distinguished in their ability to enrich certain types of pathways based on where they are localized across tissues (tissue-specific, group-specific, or ubiquitous). Therefore, we tested the ability of thresholding methods to accurately predict presence of a pathway in the right tissue (hypergeometric test, enriched if $p < 0.05$). For this, we compared the false negative rates for each of the thresholding methods (global50, global75, local25, local25-90, and local25-75) in the 5 clusters. We found that global75 generated the most false negatives among pathways in P2, global50 and local25 generated highest false negatives in P3, and local25 and local25-90 generated the most false negatives in P5 (Figs. H and I in S1 Text).

Interestingly, difference in the number of false negative pathways in P4 were never significant between the different methods. However, nearly 40% of tissue-specific pathways existed in cluster P4, likely because nearly all methods perform equally in making accurate predictions about tissue-specificity of pathways (i.e., pathways with very low ubiquity). The results, here, indicate that nearly all methods perform equally well in predicting highly tissue-specific pathways such as heme synthesis (Fig 5A) and bile acid synthesis (Fig 5B). However, they perform differently when considering group-specific and/or ubiquitous pathways such that global50 and global75 captured fewer accurate tissues for androgen/estrogen metabolism (Fig 5B) and xenobiotic metabolism; and local25 and local25-90 did not capture ubiquity of glycolysis/gluconeogenesis and citric acid cycle (see Methods; Fig 5C). Further, we also found that local25-75 never produced significantly (Fig I in S1 Text) less false negatives in pathway clusters P2, P3, and P5 (Fig H in S1 Text). Therefore, these results together suggest that local25-75 presents most accurate list of active reactions.

Discussion

Several methods have been developed to integrate transcriptomic data in GEMs, thus enabling the comprehensive study of metabolism for different cell types, tissue types, patients, or environmental conditions [8,12,22,23,29,30]. However, while these, and many other studies rely on preprocessing decisions to integrate the transcriptomic data in biochemical networks, each study makes different decisions without reporting the reason for their approach. Indeed, no rigorous comparison of the impacts of such decisions has been previously reported clearly in the literature.

Here, we highlighted how different preprocessing decisions might influence information extracted from tissue specific gene expression data. We evaluated the influence of each preprocessing decision quantitatively by studying the active reaction sets and qualitatively by evaluating tissue grouping at an organ-system level. Our analysis suggested that thresholding related decisions have the strongest influence over the set of active pathways, and more specifically the thresholding approach (i.e., global or local; Fig 1C). This can be explained by the considerable influence of the decision on thresholding on the number of genes selected as expressed (Fig J in S1 Text). We note that threshold value choice for global thresholding was previously found to be the dominant factor influencing cell type-specific model content when context specific extraction methods were benchmarked [18]. When using global thresholds, the number of the

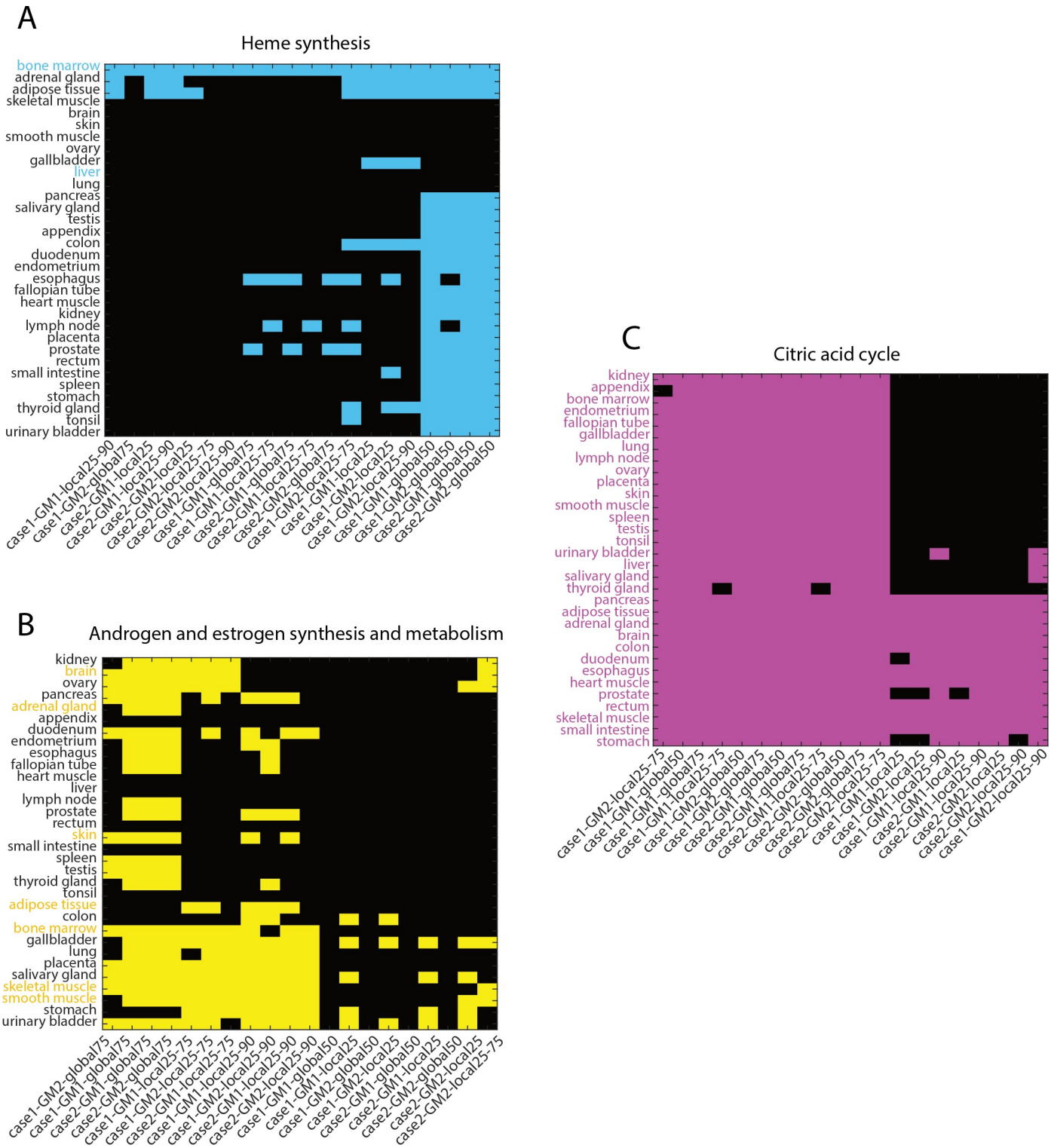

Fig 5. Comparison of active reaction lists obtained using different thresholding methods with manually-curated resource. Tissues where pathways are known to be active are bolded and colored on the y-axis and colored on the binary heatmap. (A) All thresholding methods accurately capture heme synthesis in bone marrow but not in liver. Global50 enriches the pathway in many other tissues but the pathway is known to occur only in bone marrow and liver. (B) Androgen and estrogen synthesis and metabolism is known to occur in brain, adrenal gland, skin, adipose tissue, bone marrow, skeletal muscle, and smooth muscle. Global75 enriches the pathway in higher number of tissues; and global50 and local25 does not enrich it in lower number of tissues compared to other methods. (C) Citric acid cycle is known to occur in all tissues but local25 and local25-90 does not enrich in many of the tissues compared to other methods. Thus, suggesting that local25-75 performs better than other thresholding methods in all pathway types together (Fig F in S1 Text).

https://doi.org/10.1371/journal.pcbi.1007185.g005

genes selected to be active significantly decreases with increasing threshold value. However, the use of local thresholding leads to a smaller variation in the number of genes predicted to be active (Fig K in S1 Text). Furthermore, for similar state and value attribution (e.g., *T1 25th percentile*), the use of the global thresholding approach leads to the selection of a larger number of genes predicted to be active in all tissues than the local approach (Fig J in S1 Text). Therefore, using a global threshold leads to fewer differences between tissues and a higher correlation of active reaction sets across tissues (Fig L in S1 Text), thus losing improved tissue specificity of the networks seen with the local thresholding approaches (Fig 4). This may have an important impact on analyses of tissue specific metabolism. Furthermore, the use of global thresholding is likely to lead to many false-negative reactions (i.e., reactions predicted to be inactive but are active), such as housekeeping genes that might be lowly expressed since they make essential vitamins, prosthetic groups, and micronutrients that are needed in low concentrations. Interestingly, the use of the T2 state definition seems to be less dependent on threshold values attributed than the T1 state definition when using a local approach (Fig K in S1 Text). Therefore, the use of a T2 state definition in combination of a local approach seems to successfully overcome the arbitrary aspect of threshold value selection and its influence on data preprocessing.

The order of preprocessing steps only moderately influences the definition of active reactions sets (Fig 1C). This decision implies two different interpretations of the influence of the RNA transcript levels on the determination of the enzyme abundance and activity associated to a given reaction. Indeed, the *Order 1* suggests that the measured expression levels determine the enzyme abundance available for a reaction while its associated activity will be defined depending on the gene chosen as the main determinant of the reaction behavior. On the other hand, the *Order 2* relies on a comparison of the activities of each gene associated with enzymes that might catalyze a reaction without directly accounting for the absolute transcript abundance. Our analyses suggest that *Order 2* provides more significant grouping for the *Gastrointestinal* and *Lymphoreticular* systems and does not considerably influence the grouping of the *Female reproductive* system. Advances in fluxomic measurement techniques will be invaluable to further investigate this preprocessing decision. Indeed, this would allow the analysis of the correlation between the RNA transcript levels and gene activity (expression data transformed using thresholding) of all the genes contributing to the definition of a reaction activity. Furthermore, this correlation analysis will further help with biological interpretation of this preprocessing decision and further refine guidelines for gene mapping decisions.

In our analysis, both gene mapping methods handle the AND relationships within a GPR rule in the same way but they differ in the treatment of OR relationships by either considering the maximum expression value (GM1) or a sum of expression values (GM2). Therefore, GM1 assumes that a reaction activity is determined by only one enzyme while GM2 accounts for the activity of all potential isoenzymes for a reaction. Surprisingly, while most of the reactions in Recon 2.2 are associated with at least two isoenzymes (Fig M1 in S1 Text), the distributions of these reaction activities do not significantly change between the gene mapping approaches (Fig N in S1 Text). Indeed, even if there is a significant difference in the number of genes mapped to the model depending on the techniques used: an average of 58.3% of the genes present in the model and available in the HPA dataset are mapped to the model reactions using GM1 while 89.5% are mapped using GM2. The expression value of genes that are unmapped using GM1 but mapped with GM2 is often below the 50th percentile of the overall transcriptomic data available (Fig O in S1 Text) and therefore seems to not significantly influence the distribution of the reaction activities obtained. This is why the decisions relating to the gene mapping method do not influence the set of active reactions in the case of the transcriptomic dataset used in this study. However, it may not be the case for all transcriptomic datasets,

especially if more metabolic genes are associated to high gene expression values. In this context, the development of more biologically meaningful gene mapping methods might be the key to capture differences between cell-types or tissues. Current gene mapping methods consider all enzymes as specialists (i.e., one enzyme is associated to one reaction). However, numerous enzymes are actually "generalists" as they exhibit promiscuity [31,32] (Fig M4 in S1 Text). This functional promiscuity of an enzyme may be manifested in the form of competition between reactions catalyzed by this enzyme, and therefore influence the catalytic activity of an enzyme. In this context, future work may benefit from exploring strategies to handle enzyme promiscuity [33].

In conclusion, decisions must be made on how to best handle and incorporate transcriptomic data into biochemical networks. This benchmarking study emphasizes for the first time the importance of carefully evaluating these decisions and associated parameters. Our analysis highlights that the choice of thresholding approach influences the active reaction sets the most, even more than tissue-specific effects. Meanwhile, gene mapping decisions had the lowest influence. We showed that some decisions better capture the functional tissue similarity across different organ systems. Overall, our analysis showed that transcriptomic data preprocessing decisions influence the ability to capture meaningful information about tissues. However, current preprocessing techniques present important limitations and decisions associated to this process should be made very carefully. Indeed, numerous steps and decisions involved in the estimation of enzyme abundance and activity from transcriptomic data rely on biological assumptions that have not yet been leveraged.

With the increasing availability and affordability of omic measurement techniques, studies filling the gap between mRNA expression and enzymatic activity will be of crucial importance. In this context, we hope that the guidelines provided by this study will help researchers develop more robust and biologically meaningful preprocessing techniques, leading to more accurate models, and deeper insights into the tissue-specific behavior of an animal.

Methods

Transcriptomic data

We used the Human Protein Atlas transcriptomic dataset (HPA) which includes RNA-Seq data of 20344 genes across 32 different human tissues [25]. Out of 20344 genes, 1663 can be mapped to the metabolic genes present in Recon 2.2 (99.4% of coverage) [26]. S3 Table presents the 10 genes of Recon 2.2 that are not associated with expression values in the HPA dataset and Fig P in S1 Text presents the distribution of gene expression values in the HPA dataset.

Genome-scale model of human metabolism–Recon2.2

Recon 2.2 [26] includes 1673 genes, 5324 metabolites and 7785 reactions. 3061 reactions do not have GPR associations. The remaining 4724 reactions are associated to 1797 different enzymes and about 20% of these reactions can be catalyzed by multiple isoenzymes. Almost 21% of the enzymes are formed by enzyme complexes (up to 46 subunits—reaction: NADH2_u10m) and about 54% of the enzymes are promiscuous enzymes (Fig M in S1 Text).

Gene mapping

In metabolic networks, the relationship between genes and reactions is represented using logical rules, referred as Gene-Protein-Reaction rules (GPRs). These rules describe the association between the genes responsible for the expression of protein subunits forming the enzyme that catalyzes a reaction (AND for enzyme complexes; OR for isoenzymes). This relationship linking enzymes to reactions may have different types of GPR patterns. Some relationships are

simple, with one gene encoding one enzyme that catalyzes one reaction. However, many are more complicated, in which one enzyme catalyzes multiple reactions (promiscuous), multiple proteins form an enzyme complex that catalyzes one reaction (multimeric), multiple enzymes catalyze one reaction (isoenzymatic), or multiple enzymes could catalyze multiple reactions (isoenzymatic promiscuous) [32] (Fig A in S1 Text). Gene mapping methods (GMMs) require combined use of the GPR rule and gene expression data to determine the enzyme activity associated to a reaction. In this regard, two methods have been used prominently in the field:

i. Selection of the *minimum* expression value among all the genes associated to an enzyme complex (AND rule) and the *maximum* expression value among all genes associated with an isoenzyme (OR rule). We refer to this method as GM1 [21].

ii. Selection of the *minimum* expression value among all the genes associated to an enzyme complex (AND rule) and *sum* of expression values of all the genes associated to an isoenzyme (OR rule). We refer to this method as GM2 [20].

Thresholds

Thresholding Approaches: Thresholding approaches describe the scheme of threshold imposition on the gene expression value for a gene and/or reaction to be considered as "active".

i. *Global approach*: The threshold value is the same for all the genes. The global approach is often applied when only one sample or condition is available and/or no information is available in the literature to define an expression threshold for a single gene. The "global threshold" is most often defined using the distribution of expression values for all the genes, and across all samples if multiple samples are available. This type of thresholding approach has been used, for example, in combination with a model extraction method called Gene Inactivity Moderated by Metabolism and Expression (GIMME) [22].

ii. *Local approach*: The threshold value is different for all the genes. The local approach is often applied when multiple samples are available as it allows a comparison of expression relative to many other samples and conditions. The "local threshold" for a gene is most often defined as the mean expression value of this gene across all the samples, tissues, or conditions [24,25,29].

The definition of thresholding criteria requires one to decide on how to partition the gene expression or reaction activity. In this regard, the ON/OFF state definition is often used in the literature. This type of state definition requires only one value to qualify if a gene/reaction is active. For example, when using one single threshold in a global context (i.e., hereafter referred as global T1), the genes presenting an expression above this value are considered as active (i.e., ON) while the others are inactive (i.e., OFF). However, this type of gene expression partition in a local context (e.g., expression threshold of a gene defined by its mean expression across all samples) presents limitations when facing genes with very low or very high expression values for all the samples. Indeed, when a gene presents always very low expression values, the use of the mean as threshold will lead to the consideration of its expression in some samples. Contrarily, some genes may be associated with very high expression values in all the samples. Doing so, while this gene should be considered as active, the current state partition will lead to considering this gene as non-expressed in all the samples presenting an expression value below the mean.

To overcome this problem, we propose to implement the gene-specific thresholding approach in combination with the definition of a global threshold definition for genes presenting low expression values among all the samples (e.g., below the usual detection level associated to the measurement method) to prevent their definition as an active set for some samples

(i.e., local T1). Therefore, the genes whose expression is below the value defined by this global lower bound will always be considered as inactive (i.e., OFF), while other genes will fall under the local rule for gene-specific threshold definition (i.e., MAYBE ON). The T1 state definition of local thresholding approach can be defined as follows "*the expression threshold for a gene is determined by the mean of expression values observed for that gene among all the tissues BUT the threshold must be higher or equal to a lower percentile bound globally defined*".

Another similar extreme case can be encountered for genes with high expression values in all samples. To this end, we propose to introduce an upper and a lower bound can be introduced to define the expression values for which a gene should always be considered as expressed or non-expressed. This will ensure that genes with very low expression values across all the samples will never be considered as active (i.e., OFF) and genes with very high expression across samples are always considered as active (i.e., ON). Doing so, the local rule for gene-specific threshold definition is applied only to the genes whose expression is in between the range of values defined by the lower and upper bounds (i.e., MAYBE ON). The definition of the local threshold with a T2 state definition can be expressed as follows: "*the expression threshold for a gene is determined by the mean of expression values observed for that gene among all the tissues BUT the threshold:(i) must be higher or equal to a lower percentile bound globally defined and (ii) must be lower or equal to an upper percentile bound globally defined.*"

Threshold values: The threshold values depend on the approach (i.e., local or global) and on the number of states (i.e., T1 or T2) used for thresholding. The global approach can only be associated with T1 state definition as it requires the assignment of only one threshold value. On the other hand, the local thresholding approach can be used in combination with either a T1 or a T2 state definition, as mentioned above. In the context of this study, we have chosen to compare the following combination of threshold value attribution:

i. *Global thresholding values*: The global threshold values chosen in this study are either the 50^{th} or the 75^{th} percentile (named respectively *global50* and *global75*). We also assessed the impact of using the 25^{th} and the 90^{th} percentiles. These thresholds exhibited tissue-specific sets of active genes that were more highly correlated and therefore decreased the ability to differentiate between tissues (*global25*, Fig L in S1 Text) or more uncorrelated with more highly expressed genes removed, leading to a decreased ability to connect similar tissues (*global90*, Fig J in S1 Text).

ii. *Local thresholding values*: we used the 25^{th} percentile of the overall gene expression distribution as lower bound for the local thresholding approach. This combination is referred as *local25* when used alone. Note that, in the case of the HPA dataset, the 25^{th} percentile is equal to 1.2 FPKM and the detection limit of RNA-Seq technique is often considered at 1 FPKM (Fig P in S1 Text). Two different upper bounds have been used for the T2 state definition of the local approach: the 75^{th} (referred as *local25-75*) or the 90^{th} (referred as *local25-90*) percentiles of the overall gene expression distribution. The choice of the 75^{th} percentile as upper bound is based on the distribution of the mean expression value for each gene in the HPA dataset. Indeed, more than half of the genes are associated to a mean higher than the 50^{th} percentile of gene expression distribution (Fig Q in S1 Text). To objectively assess the influence of the choice of this upper bound, we also compared the results obtained by imposing an upper bound of 90^{th} percentile.

Ordering of preprocessing steps

The preprocessing for transcriptomic data could be done in two possible ways as described below:

i. *Order 1*: The gene expression values are associated to reactions using one of the gene mapping methods. These "reaction expressions" can further be used to define the set of active reactions by imposing thresholds of activity.

ii. *Order 2*: The thresholding is used to define the activity of each gene based on gene expression data. These "gene activities" are mapped to the reactions using one of the mapping methods.

For Order 1, thresholding is imposed on these "reaction expressions" and no longer on the gene expression. This leads to the necessity to adapt the local threshold definition in the case of a preprocessing combination using Order 1 with GM2 gene mapping. Indeed, as the GM2 approach map multiple genes to a reaction, the activity of this reaction can no longer be defined by using the gene expression distribution. Therefore, the activity threshold for a reaction is determined by the sum of mean expression values observed for the genes mapped to this reactions) among all the tissues, but the mean expression value of each gene mapped to the reaction must be higher or equal to a lower percentile bound globally defined. Furthermore, it must be lower or equal to upper percentile bound globally defined.

Principal Component Analysis (PCA)

A binary matrix is constructed in which each row represents one of the 20 preprocessing approaches for each tissue (i.e., total of 640 rows) and each column represents a reaction of the GEM (i.e., 7785 columns): a reaction being active (1) or not active (0) in the GEM. The PCA analysis was conducted on this matrix after the removal of reactions being active for all or no preprocessing combinations and having each row centered to have zero mean. The variance explained by the different factors (each preprocessing decision and the tissue origin) within each of the principal components is calculated as follows. Within one factor, the maximum Pearson correlation coefficient (R) of the component scores and categories is calculated across all possible orderings of the categories. Reported is the R^2 scaled to percentages.

Assessment of tissues similarities

The set of active reactions have been used to compute the Euclidean distance between each tissue. We associated each tissue to an organ system using the classification proposed in the Human Protein Atlas (S1 Table) and computed the average Euclidean distance between tissues belonging to the same organ system. Note that, we only considered organ systems presenting more than two tissues within the same group (i.e., Female Reproductive, Lymphatic and Gastrointestinal). To compute the significance of our results, we generated the mean Euclidean distance for 10000 randomly selected group with the same number of tissues and computed the exact p-value (i.e., proportion of random distance lower than the observed distance) associated to each organ system.

Manual identification of pathway-tissue pairs

Pathway-tissue pairs were manually searched using the tissue and pathway names. The search results directed to published articles were then manually selected. The articles that specifically studied the presence of enzymes in the tissues (extracted post-mortem) or culture of human cells were chosen as evidence. We also considered review articles that discussed pathways, their ubiquity among human cell types, or role of a pathway in the tissue (S2 Table).

Calculating enrichment of pathways in active reaction sets

To evaluate the biological differences in core reaction lists obtained from different thresholding methods, we tested if, for a thresholding method, a given pathway was enriched

(hypergeometric p-value < 0.05) in a given tissue based on the number of reactions associated to the pathway (K), the total number of reactions selected (N), the number of reactions selected that associated to the pathway in that tissue (x), and the total number of reactions associated to a pathway. Hypergeometric p-value was calculated using the MATLAB function, *hygecdf*.

For a given thresholding method, our analysis resulted in enrichment matrix (a binary matrix), E. The value of E_{ij} was assigned as 1 if i^{th} pathway was enriched in j^{th} tissue, otherwise the value was assigned as 0 (Fig G3 in S1 Text). The enrichment of a pathway in the active reaction set of a tissue was interpreted as a strong statistical support that the pathway is predicted to be active in the tissue by the thresholding method. We compared this matrix to the known pathway-tissue pairs (Fig G5 in S1 Text) for analysis of false negatives in each thresholding method (Fig G6 in S1 Text).

This matrix was then converted into a vector by summing along the tissue dimension; thus, indicating the number of tissues where a given pathway was enriched. This vector was made for each thresholding method; thus, we ended up with a matrix describing number of tissues a pathway is predicted to occur when a thresholding method is used. Here, we call this matrix predicted ubiquity matrix, U^P. The value of $U^P{}_{ik}$ indicates the number of tissues i^{th} pathway was enriched in by k^{th} thresholding method. The predicted ubiquity matrix was then clustered to identify similarity in ubiquity predictions of thresholding methods among different pathways (Fig G4 in S1 Text). We used complete linkage and Euclidean distance metric for hierarchical clustering. The pathways were then divided into 5 clusters.

Data and software availability

The Human Protein Atlas transcriptomic dataset (HPA) were acquired from supplementary material of [25]. Recon 2.2 model was downloaded from https://github.com/u003f/recon2/tree/master/models. The MATLAB code for applying the different preprocessing combinations is available in COBRA toolbox v3.0, in the papers section [34].

Supporting information

S1 Text. Supplementary materials.
(DOCX)

S1 Table. Organ system grouping for each tissue present in the HPA dataset.
(XLSX)

S2 Table. Pathway-tissue pairs.
(XLSX)

S3 Table. List of genes of Recon 2.2 missing in HPA dataset.
(XLSX)

Author Contributions

Conceptualization: Anne Richelle, Chintan Joshi.

Data curation: Anne Richelle.

Formal analysis: Anne Richelle, Chintan Joshi.

Funding acquisition: Nathan E. Lewis.

Investigation: Anne Richelle, Chintan Joshi.

Methodology: Anne Richelle, Chintan Joshi.

Project administration: Nathan E. Lewis.

Resources: Anne Richelle.

Software: Anne Richelle.

Supervision: Nathan E. Lewis.

Validation: Anne Richelle.

Visualization: Anne Richelle, Chintan Joshi.

Writing – original draft: Anne Richelle.

Writing – review & editing: Anne Richelle, Chintan Joshi, Nathan E. Lewis.

References

1. Åkesson M, Förster J, Nielsen J. Integration of gene expression data into genome-scale metabolic models. Metab Eng. 2004; 6(4):285–93. https://doi.org/10.1016/j.ymben.2003.12.002 PMID: 15491858
2. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. Nature. 2004; 429(6987):92–6. https://doi.org/10.1038/nature02456 PMID: 15129285
3. Gomes de Oliveira Dal'Molin C, Quek L-E, Saa PA, Nielsen LK. A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. Front Plant Sci [Internet]. 2015; 6. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2015.00004/abstract
4. Hyduke DR, Lewis NE, Palsson BØ. Analysis of omics data with genome-scale models of metabolism. Mol BioSyst [Internet]. 2013; 9(2):167–74. Available from: https://doi.org/10.1039/c2mb25453k PMID: 23247105
5. Lewis NE, Cho BK, Knight EM, Palsson BO. Gene expression profiling and the use of genome-scale in silico models of escherichia coli for analysis: Providing context for content. Vol. 191, Journal of Bacteriology. 2009. p. 3437–44. https://doi.org/10.1128/JB.00034-09 PMID: 19363119
6. Lewis NE, Abdel-Haleem AM. The evolution of genome-scale models of cancer metabolism. Vol. 4 SEP, Frontiers in Physiology. 2013.
7. Pfau T, Pacheco MP, Sauter T. Towards improved genome-scale metabolic network reconstructions: Unification, transcript specificity and beyond. Brief Bioinform. 2016; 17(6):1060–9. https://doi.org/10.1093/bib/bbv100 PMID: 26615025
8. Schultz A, Qutub AA. Reconstruction of Tissue-Specific Metabolic Networks Using CORDA. PLoS Comput Biol. 2016; 12(3).
9. Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, Cheng JK, et al. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. Nat Biotechnol. 2010; 28(12):1279–85. https://doi.org/10.1038/nbt.1711 PMID: 21102456
10. Mardinoglu A, Agren R, Kampf C, Asplund A, Uhlen M, Nielsen J. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nat Commun. 2014; 5.
11. Fouladiha H, Marashi S-A. Biomedical applications of cell- and tissue-specific metabolic network models. J Biomed Inform [Internet]. 2017; 68:35–49. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28242343 https://doi.org/10.1016/j.jbi.2017.02.014 PMID: 28242343
12. Jerby L, Ruppin E. Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling. Vol. 18, Clinical Cancer Research. 2012. p. 5572–84. https://doi.org/10.1158/1078-0432.CCR-12-1856 PMID: 23071359
13. Zielinski DC, Filipp F V., Bordbar A, Jensen K, Smith JW, Herrgard MJ, et al. Pharmacogenomic and clinical data link non-pharmacokinetic metabolic dysregulation to drug side effect pathogenesis. Nat Commun. 2015; 6.
14. Zhang W, Li F, Nie L. Integrating multiple "omics" analysis for microbial biology: Application and methodologies. Vol. 156, Microbiology. 2010. p. 287–301. https://doi.org/10.1099/mic.0.034793-0 PMID: 19910409
15. Correia S, Rocha M. A critical evaluation of methods for the reconstruction of tissue-specific models. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015. p. 340–52.

16. Ferreira J, Correia S, Rocha M. Analysing Algorithms and Data Sources for the Tissue-Specific Reconstruction of Liver Healthy and Cancer Cells. Interdiscip Sci Comput Life Sci. 2017; 9(1):36–45.
17. Machado D, Herrgård M. Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. PLoS Comput Biol. 2014; 10(4).
18. 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; 4(3).
19. Pacheco MP, Pfau T, Sauter T. Benchmarking procedures for high-throughput context specific reconstruction algorithms. Front Physiol. 2016; 6(JAN).
20. Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, Kell DB, et al. Improving metabolic flux predictions using absolute gene expression data. BMC Syst Biol. 2012; 6.
21. Jensen PA, Lutz KA, Papin JA. TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. BMC Syst Biol. 2011; 5.
22. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008; 4(5).
23. Zur H, Ruppin E, Shlomi T. iMAT: An integrative metabolic analysis tool. Bioinformatics. 2010; 26 (24):3140–2. https://doi.org/10.1093/bioinformatics/btq602 PMID: 21081510
24. Agren R, Mardinoglu A, Asplund A, Kampf C, Uhlen M, Nielsen J. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Mol Syst Biol. 2014; 10(3).
25. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science (80-) [Internet]. 2015; 347(6220):1260419–1260419. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1260419
26. Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, Hanscho M, et al. Recon 2.2: from reconstruction to model of human metabolism. Metabolomics. 2016; 12(7).
27. Qin Y, Pan J, Cai M, Yao L, Ji Z. Pattern Genes Suggest Functional Connectivity of Organs. Sci Rep. 2016; 6.
28. Gry M, Oksvold P, Pont$\surd$©n F, Uhl$\surd$©n M. Tissue-specific protein expression in human cells, tissues and organs. J Proteomics Bioinforma [Internet]. 2010; 3(10):286–93. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-78249263183&partnerID=40&md5=43dffb6140a208c173419af4fd38143f
29. 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(5).
30. Vlassis N, Pacheco M, Sauter T. Fast Reconstruction of Compact Context-Specific Metabolic Network Models. arXiv Prepr arXiv13047992 [Internet]. 2013;1–22. Available from: http://arxiv.org/abs/1304.7992
31. Khersonsky O, Tawfik DS. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. Annu Rev Biochem [Internet]. 2010; 79(1):471–505. Available from: http://www.annualreviews.org/doi/10.1146/annurev-biochem-030409-143718
32. Nam H, Lewis NE, Lerman JA, Lee DH, Chang RL, Kim D, et al. Network context and selection in the evolution to enzyme specificity. Science (80-). 2012; 337(6098):1101–4.
33. Barker BE, Sadagopan N, Wang Y, Smallbone K, Myers CR, Xi H, et al. A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. Comput Biol Chem. 2015; 59:98–112. https://doi.org/10.1016/j.compbiolchem.2015.08.002 PMID: 26381164
34. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0. ArXiV. 2017;1710.04038.