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, CA 92093, United States.
[2] Department of Pediatrics, University of California, San Diego, School of Medicine, La Jolla, CA 92093, United States.
[3] Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, United States.
*Correspondence: nlewisres@ucsd.edu

Keywords: omic data, systems biology, biochemical pathways

Abstract

Genome-scale models of metabolism (GEMs) describe all metabolic reactions that may occur organism-wide. It is known that each tissue exhibits differential gene expression patterns and enzymatic activities. Therefore, transcriptomic data are commonly used to tailor GEMs and capture tissue-specific behavior. However, since measured gene expression levels span several orders of magnitude, and many reactions in GEMs involve multiple genes, decisions must be made on how to overlay the data onto the network. Referred to here as "preprocessing", as it addresses the steps prior to context-specific model construction, these decisions include how to map gene expression levels to the gene-protein-reaction rules (i.e. gene mapping), the selection of thresholds on expression data to consider the associated gene as "active" (i.e. thresholding), and the order in which these gene mapping and thresholding are imposed. Each of these decisions could impact the resulting expression values associated with each reaction, and therefore model construction and biological interpretation. However, the influence of these decisions has not been systematically tested, nor is it clear which combination of preprocessing decisions will capture the most appropriate biological description of the available data. To this end, we compared 20 different combinations of existing preprocessing decisions, each of which were imposed on transcriptomic dataset across 32 tissues. Our analysis suggested that the thresholding approach has the greatest influence on the definition of which reaction may be considered as active. Finally, we compared tissue-specific active reaction lists based on their capacity to recapitulate groups of tissues at the organ-system level and through this identified optimal preprocessing decisions. These results now provide guidelines that will facilitate the construction of more accurate context-specific metabolic models and analyses with biochemical networks.

Introduction

Metabolic network reconstructions can illuminate the molecular basis of phenotypes exhibited by an organism. Various cellular characteristics such as gene expression, protein expression, and enzymatic activity differ across cell types or tissues. These cellular characteristics and how their variations influence the acquisition of specific phenotypes have often been studied using omics data (Hyduke et al., 2013; Lewis et al., 2010; Gomes de Oliveira Dal'Molin et al., 2015; Lewis and Abdel-Haleem, 2013; Fouladiha and Marashi, 2017; Pfau et al., 2016; Schultz and Qutub, 2016). These studies have spanned wide array of application from identification of molecular mechanisms (Jerby et al., 2010) to identification of drug targets (Fouladiha and Marashi, 2017; Mardinoglu et al., 2014; Jerby and Ruppin, 2012).

Given the ubiquity of transcriptomic data, many studies have integrated mRNA expression data with metabolic network reconstructions to guide the development of biological hypotheses and discoveries (Lewis et al, 2009; Covert et al., 2004; Akesson et al., 2004). To this end, numerous algorithms have been developed to capture the active metabolic pathways in individual tissues or cell types based on transcriptomic data (Blazier and Papin, 2012; Kim and Lun, 2014). Integration of omics data within genome-scale metabolic reconstructions is now a common step when systemically studying context-specific metabolism (Pacheco et al., 2015; Schultz and Qutub, 2016; Zhang and Hua, 2015). However, the use of expression data faces unique challenges such as experimental and inherent biological noise, differences among experimental platforms, detection bias, and the unclear relationship between gene expression and reaction flux (Zhang et al, 2010). Moreover, omics data integration methods rely on assumptions and decisions that influence the quality and functionality of resulting models and the physiological accuracy of their predictions (Opdam et al, 2017, Machado and Herrgard, 2014).

The challenges in accurately capturing active pathways do not only stem from noisiness in the data. In metabolic networks, further challenges arise since there is often not a one to one relationship between genes and reactions. Rather the relationship is represented using logical rules, referred as GPR rules (i.e., Gene-Protein-Reaction rule). 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 GPR patterns. Some relationships are simple, with one gene encoding for one enzyme that catalyzes one reaction. However, many are more complicated, in which one enzyme could catalyze multiple reactions (promiscuous), multiple proteins could form an enzyme complex that catalyzes one reaction (multimeric), multiple enzymes could catalyze one reaction (isoenzymatic), or multiple enzymes could catalyze multiple reactions (isoenzymatic promiscuous) (Nam et al, 2012; Supplementary Figure 1). Transcriptomic data integration methods use these GPR rules to define which genes will be the main determinant of the activity associated to a given reaction. This preprocessing step is referred to as *gene mapping*. In the literature, gene mapping prominently relies on a model's Boolean definition of multimeric enzymes and isoenzymes. The most common assumption for multimeric enzymes is that gene with the minimum expression governs the activity. In case of isoenzymes, the activity may either depend the total expression of all isoenzyme genes (Lee et al., 2012) or the isoenzyme gene with highest expression (Jensen et al., 2011).

Transcriptomic technologies measure the abundance of all RNA transcripts in an organism at a specific moment. This absolute measurement is often considered to represent a gene's activity (i.e. whether a gene is expressed or not) 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 literature; from one unique threshold value for the entire set of genes (i.e. global threshold, Becker et al., 2008; Zur et al., 2010) to thresholds assigned specifically to each gene (i.e. local threshold, Agren et al., 2014; Uhlen et al., 2015). Algorithms also differ in the complexity, using only a single threshold or more complex rules involving multiple thresholds.

Preprocessing of transcriptomic data for their integration in biochemical networks relies mainly on these two steps: *gene mapping* and *thresholding*, but these can be implemented in different orders, with either gene mapping or thresholding occurring first. 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 interpretation (Table 1, Figure 1).

Decision	Variable	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	Number of thresholds states	2 states = 1 threshold	OFF/ON
		3 states = 2 thresholds	OFF/MAYBE ON/ ON
	Threshold approach	local	Gene-specific threshold values
		global	Unique threshold value for all the genes
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

Table 1: Decision involved in transcriptomic data preprocessing

Multiple studies have highlighted that the assumptions and decisions used by omics data integration methods influence the quality and functionality of resulting models and the physiological accuracy of their predictions (Opdam et al., 2017; Machado and Herrgard, 2014; Ferreira et al., 2017; Correia and Rocha, 2015; Pacheco et al., 2015). However, the influence of preprocessing gene expression data is not discussed in literature. Thus, no universal rules have been established to preprocess transcriptomic data. Here we evaluate the influence of the gene expression preprocessing steps and associated decisions on the definition of biochemical pathways activity and its consequence on the biological meaning captured by the data.

Results

We integrated transcriptomic data from 32 different tissues in the Human Protein Atlas (Uhlen et al, 2015) with Human Recon 2.2 (Swainston et al., 2016) using 20 different combinations of the 3 main preprocessing decisions listed in Table 1 (Figure 1, see Methods for details on the definition and implementation of each decision). This resulted in 640 different tissue-specific profiles of "expression" values for all gene-associated reactions in Recon 2.2. We compared these networks to evaluate the influence of each preprocessing decision on the definition of active biochemical pathways. We further analyzed the capacity of these integrated data to capture functional similarities amongst tissues for three different human organ-system groupings (i.e. female reproductive, gastrointestinal, and lymphatic systems).

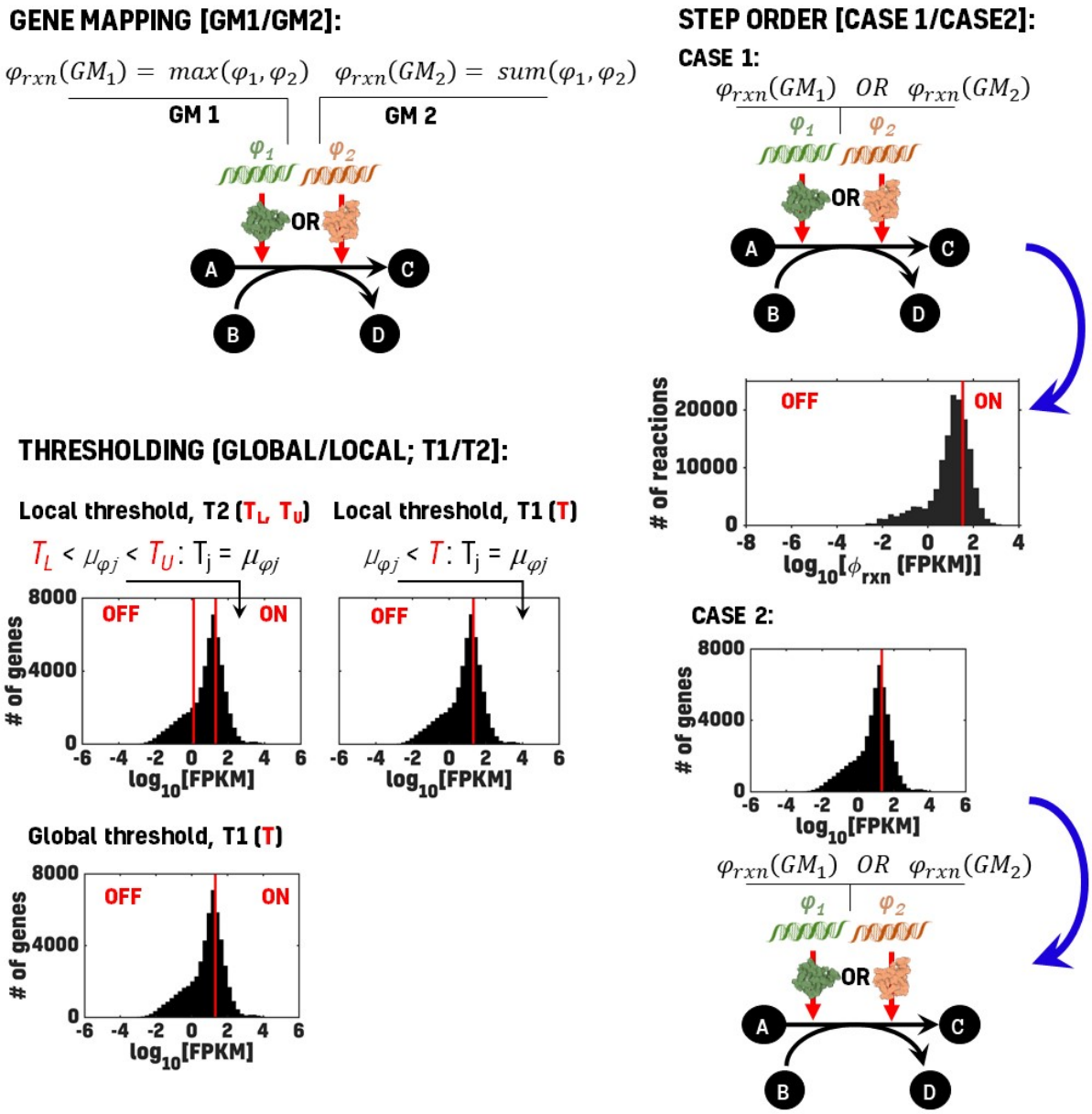

Figure 1 - Formulation and implementation of various preprocessing decisions. Top-left panel, two types of gene mapping methods (GM1 and GM2) used; both of which only differ in their treatment of isoenzymes. ϕ_{rxn} is the

reaction activity; ϕ_1 and ϕ_2 are gene expression values of isoenzymes gene1 and gene2 represented by DNA molecules in green and orange. Bottom-left panel, formulation of three combinations of thresholds (Local Threshold T2, Local Threshold T1, and Global Threshold T1) used. $\mathbf{T_L}$ and $\mathbf{T_U}$ are lower and upper thresholds, respectively, used in Local T2 thresholding; $\mu_{\phi j}$ is the mean of the expression of $\mathbf{j^{th}}$ gene; $\mathbf{T_j}$ is the local threshold of $\mathbf{j^{th}}$ gene; **T** is the global threshold and local T1 threshold governing calculation of gene-specific threshold. Right panel, Decisions about the order in which thresholding and gene mapping are performed. Case 1, gene expression is converted to reaction activity followed by thresholding of reaction activity; Case 2, thresholding of gene expression followed by its conversion to reaction activity.

Active reaction sets are influenced by preprocessing decisions

Decisions regarding gene mapping, thresholds, and step order affect the definition of active reaction sets. Specifically, the sets of active reactions (i.e., reactions 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 (Figure 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 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 (Figure 2B). The thresholding related parameters (global/local and T1/T2) provide the most significant contribution to the variation in all the principal components and more specifically the thresholding approach (global/local) (Figures 2C, 2F, and Supplementary Figure 2). The order of the preprocessing steps only provides a small contribution to the explained variation in the first principal component (Figure 2C, 2D). Meanwhile, the type of gene mapping has the least influence on active reaction sets (Figure 2C, 2E). These results indicate that the selection of active reaction set is most heavily affected by the thresholding approach, followed by thresholding value and the order of preprocessing steps while the gene mapping method does not seems to have an influence.

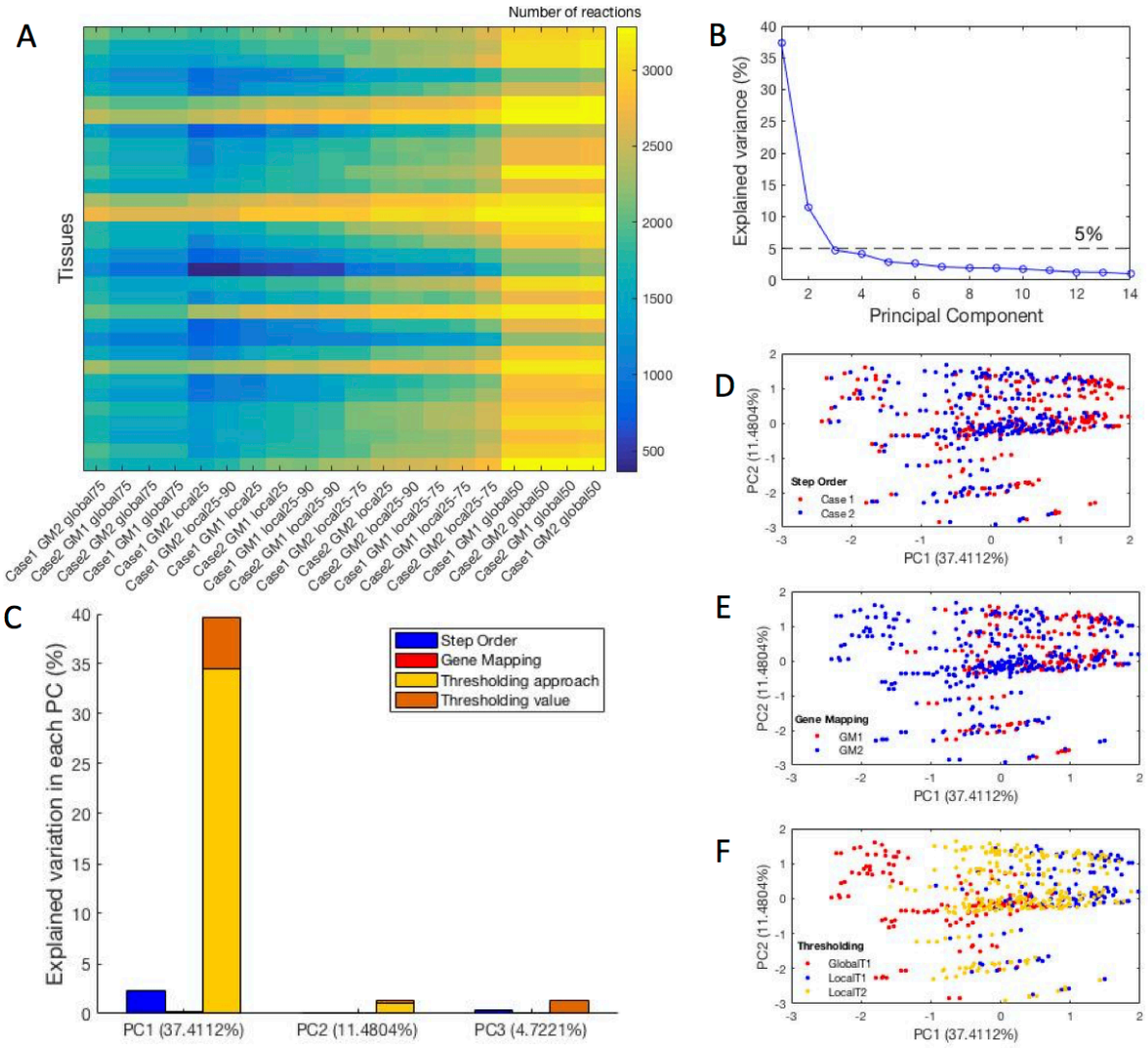

Figure 2 - Preprocessing decisions affect the definition of active reactions sets. (A) Twenty different combinations of preprocessing decisions led to a large diversity number of reactions considered as active. (B) The first three principal components (PCs) explain most of the variance in the number 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 selection is clear in the first PC (F), while this later is less influenced by the gene mapping method (E) and the order of preprocessing steps used (D).

Some preprocessing decisions better 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 tissues working collaboratively to achieve a specific function (e.g., 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. To this end, we calculated Euclidean distances between pairs of tissues belonging to the same organ-system (Figure 3, Supplementary Figure 3, see Methods for more details). Our results highlight the influence of preprocessing decisions on the significance of tissue grouping. Moreover, we observed that some decisions improved the significance of tissue grouping: Case 2

works generally better than Case 1. Local T2 also is better than GlobalT1 and LocalT1 while the influence of the gene mapping approach seems to be more mixed (Figure 4, Supplementary Figure 4).

Note that this analysis has been done without associating the placenta to the *Female reproductive* organ-system group. While this tissue is often associated to this group (i.e. Human Protein Atlas association, Supplementary Table 1), the placenta is actually functionally and histologically different from the other tissues of this group, being derived from both maternal and fetal tissue. This biological difference was successfully captured when we compared the tissue similarity analysis with and without the placenta in the *Female reproductive* organ-system group (Supplementary Figure 5).

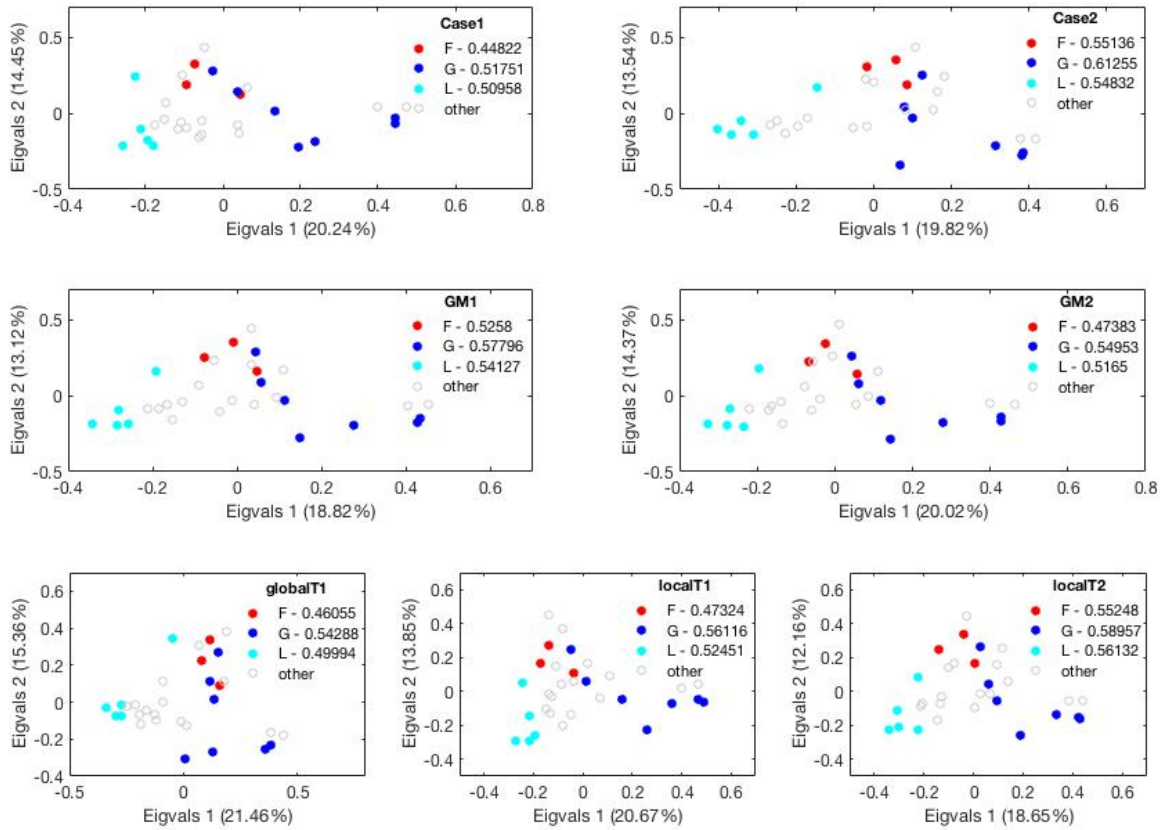

Figure 3 - Influence of preprocessing decisions of the analyze of tissue similarities - Visual representation using a Principal Coordinates Analysis of the similarity between tissues grouped by organ system for each preprocessing decision (number in legend are the mean Euclidean distance of the tissues belonging to each group; F – female reproductive group, G – gastrointestinal group, and L – Lymphatic group.)

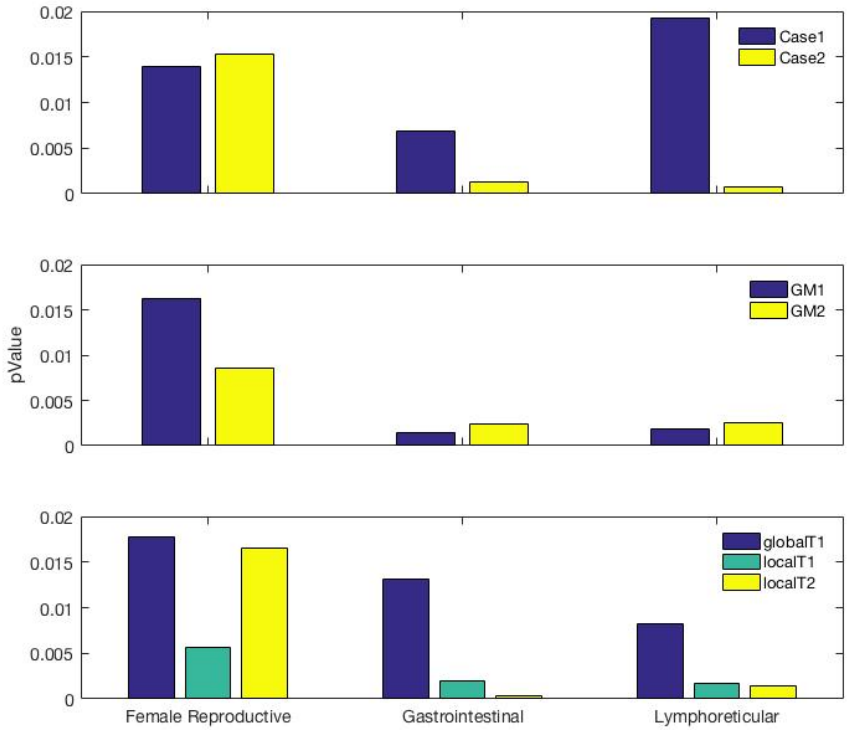

Figure 4 - Preprocessing decisions influence the significance of tissue grouping at organ-system level. pvalues associated to the mean Euclidean distance observed between tissues belonging to the same organ system depending on the preprocessing decision used.

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 (Agren et al., 2012; Schultz and Qutub, 2016; Vlassis et al., 2014; Jerby et al, 2012; Zur et al., 2010; Becker and Palsson, 2008). However, while these, and many other studies rely on preprocessing decisions to integrate the transcriptomic data, each study makes different decisions without reporting the reason for their approach. Indeed, no rigorous comparison of the impacts of such decisions has been reported, to our knowledge.

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 studying the active reaction sets and qualitatively 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; Figure 1C). We note that threshold choice was also the dominant factor influencing model content when context specific extraction methods were recently benchmarked (Opdam et al., 2016). This is expected, since decision on thresholding considerably influences the number of genes selected as expressed (Supplementary Figure 6). When using global thresholds, the number of the genes selected to be active significantly decreases with increasing threshold value. However, local thresholding presents a

smaller variation in the number of genes predicted to be active (Supplementary Figure 7). Furthermore, for similar state and value attribution (e.g. *T1 25th*), 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 (Supplementary Figure 6). Therefore, global thresholds lead to fewer differences between tissues and a higher correlation of active reaction sets across tissues (Supplementary Figure 8), which may be an important issue impacting studies 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 active 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 of threshold values attributed than the T1 state definition when using local approach (Supplementary Figure 7). Therefore, the use of a T2 state definition in combination of a local approach seems to be a good way to overcome the arbitrary aspect of threshold value and its influence on data preprocessing.

The order of preprocessing steps only moderately influences the definition of active reactions sets (Figure 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 *Case 1* order 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 *Case 2* order 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 *Case 2* provides more significant grouping for the *Gastrointestinal* and *Lymphoreticular* systems and does not really influence the grouping of the *Female reproductive* system. It could be interesting to further investigate this preprocessing decision by using fluxomic data. This would allow the analysis of the correlation between the RNA transcripts 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 help leveraging the biological interpretation of this preprocessing decision but also assessing the assumptions used by gene mapping techniques.

Indeed, 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 (Supplementary Figure 9A), the distributions of these reaction activities do not significantly change between the gene mapping approaches (Supplementary Figure 10). Indeed, even if there is a significant difference in the number of genes mapped to the model depending of 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 unmapped genes using GM1 but with GM2 is often below the 50th percentile of the overall transcriptomic data available (Supplementary Figure 11) 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 genes are associated to high gene expression values.

This benchmarking study emphasizes the importance of carefully selecting parameters for integrating transcriptomic data into biochemical networks. With the increasing availability and affordability of omic measurement techniques, studies leveraging the biological assumptions and interpretations underlying the preprocessing of these type of data will be of crucial importance. In this context, the development of more biologically meaningful gene mapping methods might be the key to capturing cell-types or tissues metabolic specificities. 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 (Khersonsky and Tawfik, 2010; Supplementary Figure 9D). 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 (Barker et al., 2015).

Conclusion

Decisions must be made on how to best handle and incorporate transcriptomic data into biochemical networks. Our benchmarking analysis of preprocessing decisions showed that thresholding influences the active reaction sets the most while gene mapping methods influences the least. We were also able to show that some decisions better represent the functional tissue similarity across different organ systems. Overall, our analysis showed that transcriptomic data preprocessing significantly influences 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. In this context, development of more robust and biologically meaningful preprocessing techniques will be the key of the improvement of our understanding of tissue-specific behavior of an animal.

Methods

Transcriptomic data

We used the Human Protein Atlas transcriptomic dataset which includes RNA-seq data of 20344 genes across 32 different human tissues (Uhlen et al., 2015). Out of 20344 genes, 1663 can be mapped to the metabolic genes present in Recon 2.2 (99.4 % of coverage). Supplementary Table 2 presents the 10 genes of Recon 2.2 that are not associated with expression values in HPA dataset and Supplementary Figure 12 presents the distribution of gene expression values in HPA dataset.

Genome-scale model of human metabolism – Recon2.2

Recon 2.2 (Swainston et al, 2016) includes 1673 genes, 5324 metabolites and 7785 reactions. 3061 reactions do not have GPR association. 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 (Supplementary Figure 9).

Gene mapping

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 amongst all the genes associated to an enzyme complex (AND rule) and the *maximum* expression value amongst all the genes associated to an isoenzyme (OR rule). This method will be referred to as GM1 (Jensen et al., 2011).

ii. selection of the *minimum* expression value amongst 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). This method will be referred to as GM2 (Lee et al., 2012).

Thresholds

Approach: Thresholding approach describes the scheme of threshold imposition on expression value for 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 mainly applied when only one sample is available (i.e. sample could be associated to a condition, a cell-type or a tissue) and/or no information is available in the literature to define expression threshold for a single gene. The "global threshold" is most often defined using the distribution of expression value 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) (Becker et al., 2008).

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 having a relative assessment of the activity of a gene across samples. 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 (Agren et al., 2012; Agren et al., 2014; Uhlen et al., 2015).

The definition of thresholding criterion requires making a decision about the partitioning of the gene expression or reaction activity. In this regard, two state definitions are often used in literature:

i. *ON/OFF:* This type of state definition requires only one value to qualify if a gene/reaction is active. Hereafter, it will be referred to as T1.

ii. *ON/MAYBE ON/OFF:* This type of state definition requires the use of two values to qualify if a gene/reaction is associated with high activity, medium activity or low activity. The use of this thresholding criterion is often used in model extraction algorithms as it allows differentiating the genes/reactions that are highly expressed (i.e., high confidence over the inclusion of the gene/reaction) to the ones that will be subject to potential inclusion (i.e., medium confidence over the inclusion of the gene/reaction) if the algorithm parameters permit. Hereafter, it will be referred to as T2.

The concept of states of activity could be of considerable importance in the definition of local thresholds. Actually, the most current practice for local threshold value definition is that a gene will be considered as active in a sample if its expression for this sample is above its mean expression across all samples. However, this approach 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 practice will lead to considering this gene as non-expressed in all the samples presenting an expression value below the mean.

To overcome these problems, local threshold approach may be refined by using the concept of state(s) definition. In the context of this study, we propose to refine the local thresholding approach by using both state definitions described above. The T1 state definition of local thresholding approach will allow to overcome the limitation related to the low expression genes only and 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"*. The T2 state definition of local thresholding approach extends the later to the handling of gene with high expression value. To this end, an upper and a lower bounds 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 and genes with very high expression across samples are always considered as active. Therefore, 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."*

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. Actually, global approach can only be associated with T1 state definition as it requires the assignment of only one threshold value. On the other hand, 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 choosen in this study are either the 50th or the 75th percentile (named respectively *Global T1 50th* and *Global T1 75th*). We also performed some analyses using the 25th and the 90th percentiles, but these thresholds were leading to sets of active genes either too correlated and therefore not allowing differentiation between the differenciation sample tissues (*Global T1 25th*, Supplementary Figure 8) or too small and therefore leading to the non-overlaping of samples tissues (*Global T1 90th*, Supplementary Figure 6).

ii. *Local thresholding values:* we used the 25th percentile of the overall gene expression distribution as lower bound for the local thresholding approach. This combination is referred as *Local T1 25th* when used alone. Note that, in the case of the HPA dataset, the 25th percentile is equal to 1.2 FPKM and the detection limit of RNA-seq technique is often considered at 1 FPKM (Supplementary Figure 12). Two different upper bounds have been used for T2 state definition of the local approach: the 75th (referred as *Local T2 25th & 75th*) or the 90th (referred as *Local T2 25th & 90th*) percentiles of the overall gene expression distribution. The choice of the 75th 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 50th percentile of gene expression distribution (Supplementary Figure 13). To objectively assess the influence of the choice of this upper bound, we proposed to also compare the results obtained by imposing an upper bound of 90th percentile.

Ordering of preprocessing steps

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

i. *Case 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. *Case 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.

Note that for the Case 1 order, thresholding is imposed on these "reaction expressions" and no longer on the gene expression. This lead to the necessity to adapt the local threshold definition in the case of a preprocessing combination using Case 1 order 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 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 (AND 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 and each column represents a variable: 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.

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 (Supplementary Table 1) 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 pvalue (i.e. proportion of random distance lower than the observed distance) associated to each organ system.

References

1. Agren, R. et al. (2012) Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT. *PLoS Comput Biol,* 8(5): e1002518.
2. Agren, R. et al. (2014) Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol Syst Biol,* 10:721.
3. Akesson, M. et al. (2004) Integration of gene expression data into genome-scale metabolic models. *Metab Eng,* 6(4), 285-293.
4. Barker, B.E. et al., (2015) A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. *Comput Biol Chem,* 59 Pt B:98-112.
5. Becker, S.A. and Palsson, B.O. (2008) Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol*, 4, e1000082.
6. Blazier, S.A. and Papin, J.A. (2012) Integration of expression data in genome-scale metabolic network reconstructions. *Front Physiol*, 3, p299.
7. Correia, S. and Rocha, M. (2015) A Critical Evaluation of Methods for the Reconstruction of Tissue-Specific Models. In: Pereira F., Machado P., Costa E., Cardoso A. (eds) Progress in Artificial Intelligence. EPIA 2015. Lecture Notes in Computer Science, vol 9273. Springer, Cham.
8. Covert, M.W. et al. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429, 92–96.
9. Ferreira, J. et al. (2017) Analysing Algorithms and Data Sources for the Tissue-Specific Reconstruction of Liver Healthy and Cancer Cells. *Interdiscip Sci*, (1):36-45.
10. Mardinoglu, A. et al., (2013) Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Mol Syst Biol*, 9: 649.
11. Gomes de Oliveira Dal'Molin, C. et al., (2015) A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. *Front Plant Sci*, 6: 4.
12. Hyduke, D.R. et al. (2013) Analysis of omics data with genome-scale models of metabolism. *Mol Biosyst*, 9, 167–174.
13. Jensen, P.A. et al. (2011) TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. *BMC Syst Biol*, 5:147.
14. Jerby, L. and Ruppin, E., (2012) Predicting Drug Targets and Biomarkers of Cancer via Genome-Scale Metabolic Modeling. *Clin Cancer Res*, 18, 20, 5572-5584.
15. Jerby, L. et al. (2012) Metabolic associations of reduced proliferation and oxidative stress in advanced breast cancer. *Cancer Res*, 72(22):5712-20.
16. Jerby, L. et al. (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol*, 6, 401.
17. Khersonsky, O. and Tawfik, D.S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem*, 79:471-505.
18. Kim, M.K. and Lun, D.S., (2014) Methods for integration of transcriptomic data in genome-scale metabolic models. *Comput Struct Biotechnol J*, 11(18): 59–65.
19. Lee, D. et al. (2012) Improving metabolic flux predictions using absolute gene expression data. *BMC Syst Biol*, 6:73.
20. Lewis, N.E. and Abdel-Haleem, A.M. (2013) The evolution of genome-scale models of cancer metabolism. *Front Physiol*, 4, 237.

21. Lewis, N.E. et al. (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat Biotechnol*, 28, 1279–1285.
22. Lewis, N.E. et al. (2009) Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: providing context for content. *J. Bacteriol,* 191:3437-44.
23. Machado, D. and Herrgard, M., (2014) Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLOS Comput Biol*, 10(10): e1003989.
24. Mardinoglu, A. et al. (2014) Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat Commun*, 5, 3083.
25. Nam, H., et al. (2012) Network Context and Selection in the Evolution to Enzyme Specificity. *Science*, 337(6098) pp. 1101-1104.
26. Opdam, S., et al. (2017) A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. *Cell Syst*, 4, 1–12.
27. Pacheco, M.P. et al. (2015) Benchmarking procedures for high-throughput context specific reconstruction algorithm. *Front Physiol*, 6, 410.
28. Thiele, I. et al, (2013) A community-driven global reconstruction of human metabolism. *Nat Biotechnol*, 31, 419-425.
29. Thomas, P. et al., (2016) Towards improved genome-scale metabolic network reconstructions: unification, transcript specificity and beyond. *Brief Bioinform*, 17(6), Pages 1060–1069.
30. Schultz, A. and Qutub, A.A. (2016) Reconstruction of Tissue-Specific Metabolic Networks Using CORDA. *PLoS Comput Biol*, 12(3): e1004808.
31. Swainston, N. et al. (2016) Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics,* 12, 109.
32. Uhlen, M. et al. (2015) Tissue-based map of the human proteome. *Science*, 347, 1260419.
33. Vlassis, N. et al. (2014) Fast Reconstruction of Compact Context-Specific Metabolic Network Models. *PLoS Comput Biol,* 10(1): e1003424.
34. Zhang, C. and Hua, Q. (2016) Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine. *Front Physiol*, 6:413.
35. Zhang, W. et al. (2010) Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology*, 156, 287–301.
36. Zur, H. et al. (2010) iMAT: an integrative metabolic analysis tool. *Bioinformatics*, 26, 3140–3142.