bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

StanDep: capturing transcriptomic variability improves context-specific metabolic models

Chintan J. Joshi[1], Song-Min Schinn[1], Anne Richelle[1], Isaac Shamie[2,3], Eyleen J. O'Rourke[5], Nathan E. Lewis[1,2,4*]

[1] Department of Pediatrics, University of California, San Diego, School of Medicine, La Jolla, CA 92093

[2] Novo Nordisk Foundation Center for Biosustainability at the University of California, San Diego, School of Medicine, La Jolla, CA 92093

[3] Bioinformatics and Systems Biology Program, University of California, San Diego, United States

[4] Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093

[5] Department of Biology, University of Virginia, Charlottesville, VA 22903

*Corresponding author:

Name: Nathan E. Lewis

Address: 9500 Gilman Drive MC 0760, La Jolla, CA 92093

E-mail: nlewisres@ucsd.edu

Keywords: omics data, systems biology, metabolism, data integration

Abstract

Diverse algorithms can integrate transcriptomics with genome-scale metabolic models (GEMs) to build context-specific metabolic models. These algorithms require identification of a list of high confidence (core) reactions from transcriptomics, but parameters related to identification of core reactions, such as thresholding of expression profiles, can significantly change model content. Importantly, current thresholding approaches are burdened with setting singular arbitrary thresholds for all genes; thus, resulting in removal of enzymes needed in small amounts and even many housekeeping genes. Here, we describe StanDep, a novel heuristic method for using transcriptomics to identify core reactions prior to building context-specific metabolic models. StanDep clusters gene expression data based on their expression pattern across different contexts and determines thresholds for each cluster using data-dependent statistics, specifically standard deviation and mean. To demonstrate the use of StanDep, we

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

built hundreds of models for the NCI-60 cancer cell lines. These models successfully increased the inclusion of housekeeping reactions, which are often lost in models built using standard thresholding approaches. Further, StanDep also provided a transcriptomic explanation for inclusion of lowly expressed reactions that were otherwise only supported by model extraction methods. Our study also provides novel insights into how cells may deal with context-specific and ubiquitous functions. StanDep, as a MATLAB toolbox, is available at https://github.com/LewisLabUCSD/StanDep

Introduction

Integration of omics data with genome-scale metabolic models (GEMs) has facilitated insights into diverse questions, spanning from the elucidation of disease mechanisms [1,2] to the identification of drug targets [3–5]. Furthermore, a recent rise in comprehensively quantified omics data for many tissues and cell types [6–9] presents an opportunity to study context-specific behavior (i.e., tissue, cell type, environmental conditions, or other variations to which cells are exposed) [10–12]. Such studies depend on the omics-integrated models to include context-relevant genes and reactions. Unfortunately, due to over simplified assumptions of which genes are expressed or not, current omics-integration methods may fail to include important genes, leading to less-predictive models [13]. Here we present an improved method to identify context-relevant genes robustly, leading towards models that better describe context-specific metabolism.

The integration of omics data and GEMs is complicated by the complexity of cellular metabolism and enzyme regulation. Metabolic phenotypes are driven by not gene expression alone but also other orthogonal processes, including enzyme assembly, post-translational modifications, localization, and substrate concentration. In other words, gene expression data provide considerable, but partial, insight into metabolic activity. To address this, data integration efforts often infer a ‘core’ set of active reactions from gene expression data. This ‘core reaction set’ is then used to produce a context-specific model via various model-extraction algorithms [14–20], which take into account network topology, model objective function or additional data (Fig. 1, grey). Inference of the ‘core reaction set’ typically involves defining a gene expression level as a threshold parameter – genes expressed above the defined threshold are interpreted to be active and part of the ‘core’. This threshold parameter has a large influence on the data integration process and its resulting model, according to recent systematic benchmarking studies [13,21].

Despite this importance, thresholds have often been poorly defined. Most often a single threshold value is used to evaluate all genes, disregarding complex and pathway-specific regulations over metabolism. Furthermore, such thresholds are often left to be defined by the user with little guidance or standardization, leading to varying and arbitrary model parameterization. Lastly, such single, catch-all thresholds often fail to identify lowly expressed but biologically important genes, including ‘housekeeping’ genes which are constitutively expressed for tissue maintenance functions [22]. A limited number of such housekeeping reactions can be ‘rescued’ by the model-extracting methods. Specifically, housekeeping reactions involved in the central carbon metabolism or tied to the biomass objective function are particularly favored to be ‘rescued’ by these algorithms. Despite this, a sizeable portion of housekeeping genes are seldom included into the resulting model, preventing context-specific models

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from describing important cellular functions [13]. Such challenges could be addressed by a thresholding method that better captures the complex relationship between gene expression and metabolism.

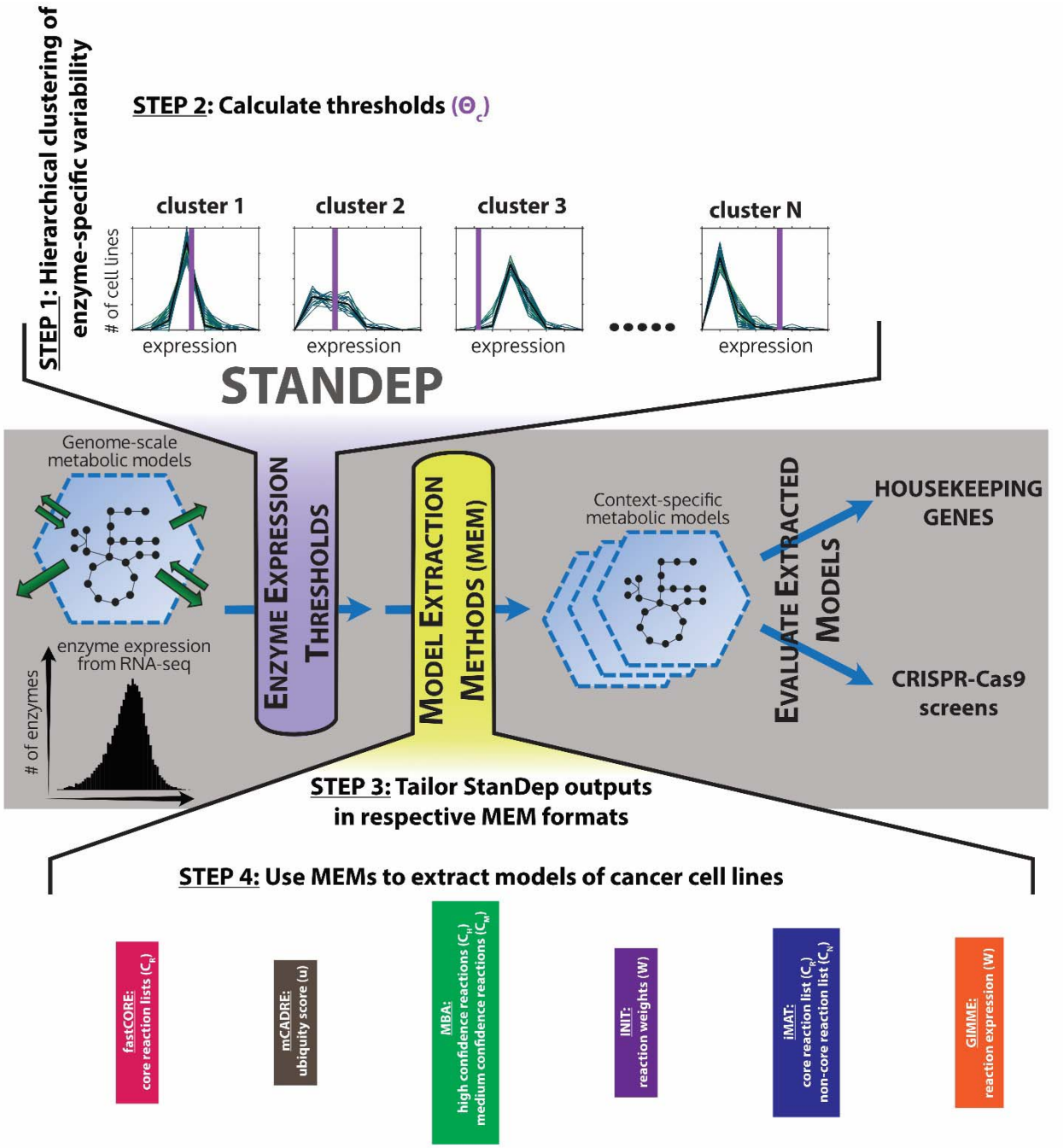


Figure 1: Basic methodology of extracting context-specific metabolic models (CSMMs) using StanDep and model extraction methods (MEMs). In grey panel, metabolic gene expression data is converted to enzyme expression data which are subjected to thresholds using a given approach (for e.g. global, localT1, localT2, or StanDep) and tailored for a given MEM to extract and evaluate CSMMs. In purple extension, StanDep is applied to enzyme expression data by clustering enzyme

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expression values followed by subjecting each cluster to thresholds calculated by eqns (1-4). In green extension, StanDep outputs are tailored (colored equations) to integrate with different MEMs. The extracted CSMMs are then evaluated using housekeeping genes/reactions and gene essentiality screens.

Here we propose a novel thresholding approach called StanDep, which clusters gene expression data based on their expression pattern across conditions (Fig. 1, purple). This enables genes that are expressed similarly to be interpreted together by common thresholds. In other words, StanDep better accounts for regulatory complexity via a group of heuristically derived thresholds rather than a single one-size-fits-all threshold. Using this finer-grained thresholding approach, StanDep captures more housekeeping genes into context-specific models compared to other thresholding methods for a wide variety of model-extraction algorithms. We further validated this method by predicting essential genes in cancer cells [23–26], and analyzing 32 human tissue models and 27 *C. elegans* cell type models. Thus, StanDep provides a novel approach to obtain more complete context-specific models of metabolism from transcriptomics data.

Results

Preprocessing transcriptomics data using StanDep

Established data integration methods have struggled to consistently capture housekeeping features [13], likely because the method rely on only one or few thresholds to interpret thousands of metabolic genes. We hypothesized that interpreting similarly expressed genes together would improve the thresholding process. Accordingly, we developed a novel thresholding method that involves two steps: (1) cluster distribution of individual gene expression, considering multimeric or isozyme relationships, (2) calculate and apply thresholds for each cluster of similarly expressed genes to identify a core reaction set (see Methods for additional detail). These steps (Fig. 1, purple) work in tandem with a variety of model-extracting methods (Fig. 1, yellow), and fit compatibly into the existing general workflow for constructing context-specific metabolic models (Fig. 1, grey).

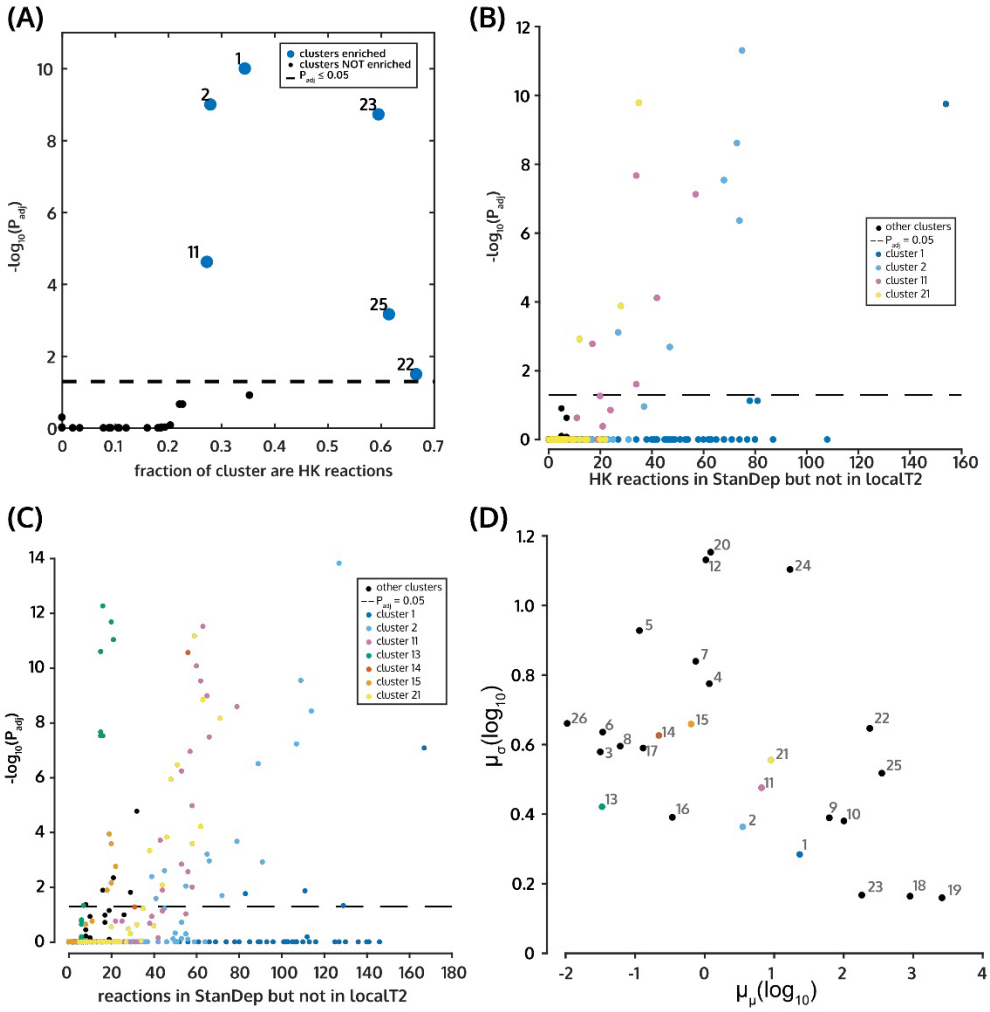
An important parameter in evaluating hierarchical clustering is the number of clusters (N). Here, the number of clusters were selected in two steps. First, we calculated the Jaccard similarity between core reaction list between any pairs of N and N+1, and then, we chose N after which Jaccard similarity between core reaction lists is over 90% (Fig. S13). As we increase the number of clusters, the weaker clusters break into smaller clusters while stronger clusters will remain. Thus, by increasing to sufficiently large number of clusters, we stabilize the selection of core reaction lists.

StanDep seeds core reaction lists with housekeeping reactions

Currently, the common approach to identifying core reactions is to define a single ‘global’ threshold on gene expression [27]. All genes that are expressed above this global threshold are considered to be metabolically active, and their associated reactions constitute the core reactions. Conversely, genes that are expressed below the threshold are unconditionally interpreted as inactive. Alternatively, a recently proposed thresholding method seeks to define ‘local’ thresholds tailored to each gene, which is derived from the gene’s average expression level across tissues [21,28]. Exceptionally high or low expression may still be unconditionally interpreted as active or inactive, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To compare these thresholding methods to StanDep, we applied the methods to a comprehensive transcriptomic profile of human cancer cell-lines [29]. From these gene expression data, core reaction sets were calculated using StanDep and the following three different thresholding methods: (1) Global 75th: genes expressed above the top 25th percentile are considered active, (2) localT2: local thresholds are derived from cross-tissue mean expression; genes expressed above the top 25th percentile and below the



bottom 25th percentile are interpreted unconditionally active and inactive, respectively, and (3) localT1: threshold settings are similar to localT2, but genes are never considered unconditionally inactive, even if they are below the bottom 25th percentile.

Figure 2: StanDep enriches core reaction lists with housekeeping (HK) reactions including those lowly expressed. (A) Housekeeping reactions are enriched in clusters 1, 2, 11, 22, 23, and 25 (big blue dots). (B) Housekeeping reactions which were captured only by StanDep (not in localT2) core reaction lists mainly belong to clusters 2 and 11. (C) Reactions which are captured by only StanDep mainly belong to low but tightly expression clusters. (D) Standard deviation and means for each cluster. The clusters which distinguish StanDep and LocalT2 are shown with colored dots. The colors are the same as that used in (B) and (C).

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Core reactions resulting from these thresholding methods were compared to 929 housekeeping reactions associated with metabolic housekeeping genes [22] in Recon 2.2 [30]. We found that StanDep resulted in the core reaction lists with the largest fraction of metabolism-related housekeeping reactions retained (Table 1, Fig. S1).

Thresholding method	Mean fraction of housekeeping reactions in core reaction lists across 44 NCI-60 cancer cell lines
Global 75th (global) [27]	0.57
Local T2 (25th, 75th) (localT2) [21,28]	0.70
Local T1 25th (localT1)	0.44
StanDep	0.80

Table 1. Core reaction lists from StanDep contain highest fraction of housekeeping reactions compared to those from existing thresholding methods.

We then analyzed the clustering captured housekeeping reactions, and how such captured reactions differed between StanDep and the other methods. Housekeeping reactions were enriched in 6 clusters; out of which, clusters 1, 2, and 11 had the largest number of housekeeping reactions (Fig. S2, Fig. 2A). We found that StanDep selected genes from clusters with moderate expression (10-100 FPKM), such as cluster 1 and low expression (1-10 FPKM), such as clusters 2 and 11 (Fig. 2C&D, S6). Indeed, we also found that housekeeping reactions that were captured by StanDep but not by the localT2 approach mostly belonged to clusters 2 and 11 (Fig. 2B). By contrast, the global approach favored reactions in clusters with fewer housekeeping reactions and reactions with higher expression (>100 FPKM) (Clusters 9 and 10; Fig. S3,S6). Altogether, the results suggest that StanDep seeds core reaction lists with housekeeping reactions that are not captured by existing methods.

StanDep-derived core reaction lists are more self-consistent than localT2

Using StanDep, we built hundreds of models of the NCI-60 cell lines by varying 4 model uptake/secretion constraint types [31] and 6 model extraction methods [14–19]. The resulting models were strongly influenced by model extraction method used (Fig. S11A), but not by constraint type (Fig. S12A). As shown in Table 1, the largest number of housekeeping reactions were captured by localT2 and StanDep. Therefore, we compared StanDep models with localT2 models. Further, given that constraint types did not have strong influence over the model content, we decided to compare only the models that that were built using the exometabolic constraints.

StanDep-based core reaction lists were larger than those from other thresholding methods (Fig. S18). We wondered if StanDep provided better support for inclusion of reactions by MEMs; thus, making them more self-consistent and independent from the extraction methods. We compared the overlap between models and their respective core reaction lists. Indeed, we found that for most MEMs, except MBA and GIMME, StanDep produced core reaction lists that were more self-consistent than localT2 (Fig. S7). Thus, models built using StanDep-derived core reactions had fewer unsupported reactions compared to localT2. Further, we found that reactions that were supported by only the extraction methods but not by localT2 belonged to low expression clusters such as clusters 2 and 11 (Fig. 3E). These are the same clusters that were differentially captured by StanDep (Fig. 2C). Thus, these results indicate that self-

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consistency of StanDep should be interpreted as its ability to provide transcriptomic support for low expression reactions that were otherwise added by the MEMs.

StanDep models accurately capture housekeeping functions

We have shown, so far, that StanDep-derived core reaction lists (Fig. S1) contained significantly more housekeeping reactions than localT2. To see if this is true for the models as well, we compared the coverage of housekeeping genes and housekeeping reactions between StanDep and localT2 models. Overall, we found that StanDep contained more housekeeping reactions than localT2 models (Fig. 3A&B). We then wondered if the housekeeping reactions which are differentially present in StanDep models belong to specific pathways. The housekeeping reactions which were differentially present in the StanDep-derived core reaction list (Fig. S25) and models (Fig. S26,S27) belonged to glycan synthesis and metabolism, metabolism of cofactors and vitamins, and fatty acid oxidation.

Biotin metabolism is part of the metabolism of cofactors and vitamins. We found that the housekeeping reactions belonging to biotin metabolism were differentially present in the StanDep-derived models compared to global or localT2 regardless of extraction method (Fig. S28). There are two genes part of this pathway: HLCS (reaction ID: BTNPL, Gene ID: 3141) and biotinidase (reaction ID: BTND1, Gene ID: 686 (BTD)) (Fig. 3C). Among these BTND1 was identified as the housekeeping reaction. Absence of biotinidase has been identified as an inherited disorder [32] and a lifelong treatment is required. BTND1 is responsible for recycling biotin from biocytin; and low expression has also been identified as a marker for breast cancer [33,34] and thyroid cancer [35]. Out of 4 breast cancer cell lines, global-derived models were least likely to include BTND1 in breast cancer cell lines, followed by localT2-derived breast cancer cell lines. However, StanDep predicted them in over 90% of the models, regardless of extraction method.

In mammalian cells, glycosylation performs essential functions such as protein folding, targeting, stabilization, and adhesion [36]. Further, changes in glycosylation have also been reported to contribute to cancer cell physiology [37,38]. Phosphatidylinositol phosphate metabolism, part of glycan biosynthesis and metabolism, involves glycosylphosphatidylinositol (GPI)-anchor proteins which form into a complex and serves to anchor proteins to the cell surface. An example of such a complex is H8/H7/M4B transamidase (reactions IDs: H7-TAer, H8TAer, and M4BTAer; Gene IDs: 94005 (PIGS), 51604 (PIGT), 128869 (PIGU), 8733 (GPAA1), and 10026 (PIGK)). PIGT and either gene products have been hypothesized to play a role in growth of breast cancer via paxillin phosphorylation [39]. Further, overexpression of PIGT protein has also been observed in several other cancer types [40]. As with BTND1, reactions associated with PIGT had a higher coverage in StanDep-derived models than localT2- or global-derived models (Fig. S31), except in fastCORE (Fig. S29). Further, for global and localT2-derived models, housekeeping reactions from this pathway were included because of the extraction methods (Fig. S32) rather than the core reaction lists.

Given this successful application of StanDep in capturing housekeeping genes and reactions in cancer cell lines, we also applied it to transcriptomics data from Human Protein Atlas (HPA) data [6]. We found that, across 32 human tissues in HPA, the models of human tissues extracted using StanDep resulted in 92% of housekeeping genes in fastCORE and 88% of housekeeping genes in mCADRE. The models of human

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tissues have been discussed in Supplementary Text. Thus, StanDep can be easily extrapolated across datasets.

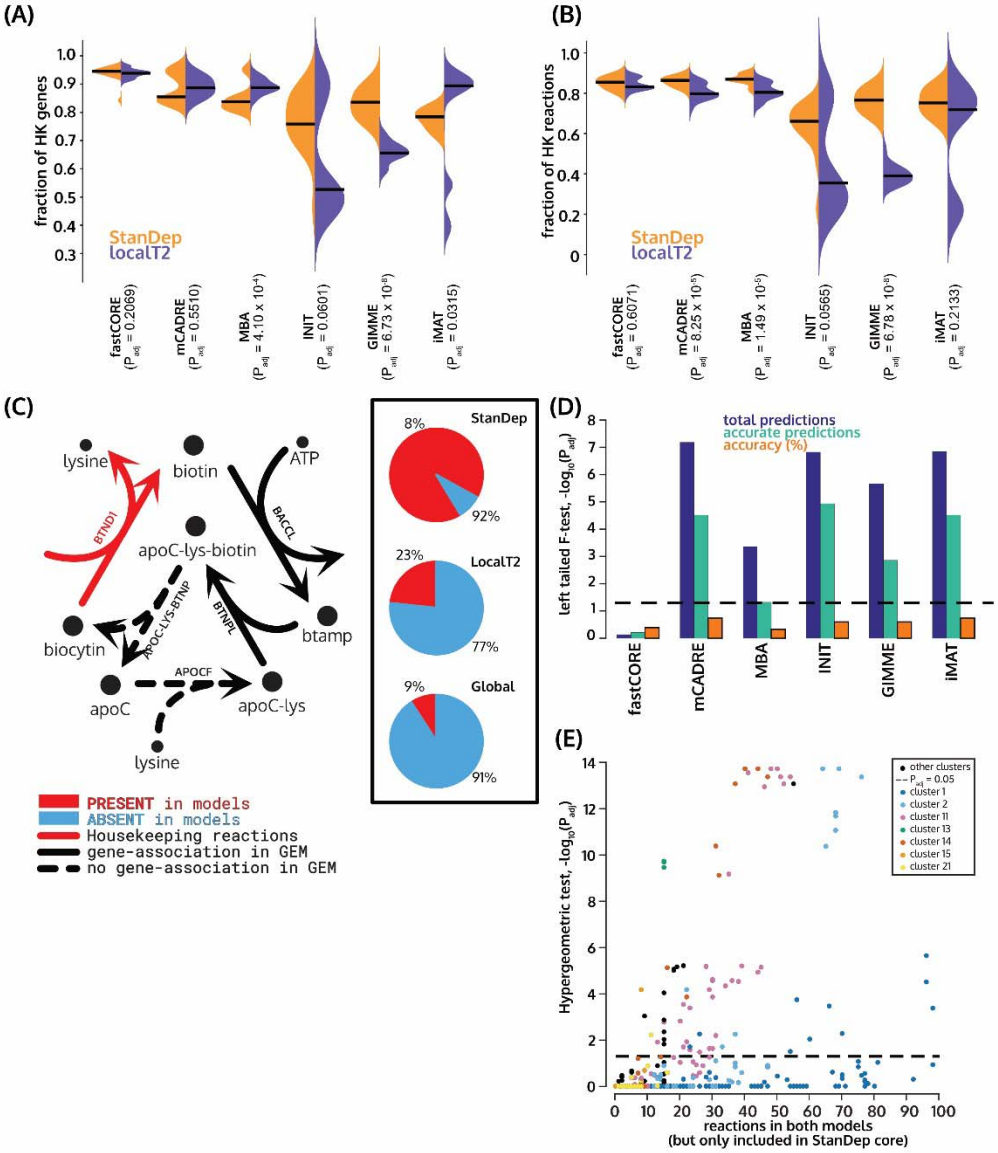


Figure 3: Performance of StanDep models (purple) compared to localT2 models (orange) using a list of housekeeping (HK) genes/reactions, CRISPR-Cas9 essentiality screens. Coverage of (A) housekeeping genes and (B) housekeeping reactions in StanDep and LocalT2 are shown. BHFDR correction for Wilcoxon rank-sum test p-values are reported on the x-axis. (C) Biotinidase (shown in red arrows, BTND1), a housekeeping reaction in Biotin metabolism, is differentially present in the StanDep-derived models. StanDep-derived models have improved coverage of this reaction compared to localT2- or global-derived models (pie plots in inset), regardless of extraction method (Fig. S28). Housekeeping reactions are given in red, gene-associated reactions are in black, and reactions with no gene associated are given in dashed arrows. The pie plots show percentage of models (cell line-extraction method = 120 models in each pie plot) that contained (red) or did not contain (blue) the reaction. (D) StanDep models are more predictable than localT2 models. BHFDR correction for left-tailed F-test for comparison of gene essentiality predictions using CRISPR-Cas9 essentiality screens are plotted. (E) StanDep provides transcriptomic

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

explanation (besides network connectivity evidence) for inclusion of lowly expressed reactions by MEMs. Each point on the plot represents the p-value of reactions belonging to a given cluster which were part of both StanDep and localT2 models but were supported by only StanDep core reaction lists.

StanDep-extracted models accurately capture essential genes

In addition, we analyzed the content of the extracted models and evaluated their capacity to predict gene essentiality, as experimentally measured by CRISPR-Cas9-mediated loss-of-function screens. [23,24,26]. In these predictions, StanDep-derived models performed comparably to models from other methods. In fact, prediction accuracies from StanDep had lower variance (Fig. 3D; Fig. S17), suggesting that StanDep-derived models may be more reliable in such predictions.

Lastly, we compared the list of essential genes in the two sets of fastCORE models, and then analyzed accurate predictions of matches and mismatches among the two thresholding methods. Among accurate matching predictions, we found that 113 genes were accurately predicted in at least one cell line by both models. Among these, 22.12% were ubiquitously essential while 26.54% were cell line-specific (black, Fig. S5). Among 58 genes that were accurately predicted by StanDep-derived models only, 22.41% were essential in at least ten cell lines, while 39.66% were essential in exactly one cell line (blue dotted, Fig. S5). Interestingly, among 46 genes that were accurately predicted by localT2-derived models only, none were essential in more than seven cell lines while 58.7% were essential in only one cell line (orange dash, Fig. S5).

We also extended gene essentiality analysis for models of 27 *C. elegans* cell types using previously published animal-level RNAi screens [25], transcriptomics data [7] and metabolic model [41]. However, due to the lack of systemically identified housekeeping genes for *C. elegans* or cell type-specific essential genes, we could compare only the genes which ubiquitously essential. Here, too, we found that StanDep models contained information not only about animal-level essential genes but also cell type-specific pathways were enriched in respective tissues; for e.g. peroxisomal fatty acid beta oxidation in neurons, hypodermis and intestine. The models of C. elegans cell types have been discussed in Supplementary Text. These results suggest that StanDep-derived models are not only able to predict ubiquitously essential genes but also cell line-specific essential genes.

Discussion

Several methods exist for building context-specific models by integrating transcriptomic data into genome scale models [14–19]. Previous work identified thresholding as the most influential parameter impacting the resulting model content and quality [13,21]. Despite this, established thresholding methods have struggled to reflect biological complexity. Here, we present StanDep, a novel heuristic approach for determining thresholds. We made hundreds of models using StanDep and evaluated them against models constructed using an existing thresholding methods [28].

Existing thresholding methods assume that all genes have the same expression patterns, regulation, and stabilities. This, however, is not true; in particular, for some genes, their metabolic products are needed in lower quantities, or the proteins may be more stable, so the necessary mRNA levels may be low; e.g. glycosylation related genes have low transcript abundance [42]. A thresholding method dependent on a

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

single arbitrary expression threshold may therefore exclude such genes. One example of such class of genes are lowly expressed housekeeping genes, consistently expressed across all cellular contexts as their gene products are required for cellular maintenance or the production of essential enzyme prosthetic groups. Indeed, classification of housekeeping genes is an important part of this study. In the literature, we found only one study that explicitly provided a list of housekeeping genes [22]. Recently a novel tool, GeneGini, has been shown to be an effective way of identifying housekeeping genes [43,44]. Housekeeping genes and reactions identified using this tool (Fig. S33) reproduced the results presented in this study (Fig. S34-35).

We were able to show that StanDep core reaction lists (Fig. S25) and models (Fig. S26,S27) better captured housekeeping reactions in processes such as glycan and cofactor metabolism. Our study identified cluster 2 and cluster 11 (mean expression less than 10 FPKM, Fig. S6A; standard deviation less than 2 FPKM, Fig. S6B) as containing enzymes catalyzing over 300 housekeeping reactions (Fig. S2). Indeed, the success of StanDep in capturing more housekeeping reactions is primarily attributed to considering: (i) patterns of variability in gene expression, and (ii) standard deviation as a measure of biological variability in the formulation of cluster-specific thresholds. Thus, we showed that gene-specific variability is an important determinant when preprocessing transcriptomics data.

Apart from the MEMs, there can be two potential reasons for reactions to be included in the extracted model: (i) presence of reactions in core reaction list calculated using a thresholding method, or (ii) demanded by the exometabolomic constraints applied. In cases such as that of biotin metabolism for which exometabolomic constraints were not available, the reactions had to come from the selection by StanDep. Further, as shown for StanDep, increased inclusion of reactions due to core reaction list will result into a method that is more self-consistent. In cases such as that of phosphatidylinositol phosphate metabolism, the inclusion was attributed to the MEMs in localT2 and global models; and to core reaction list in StanDep models (Fig. S32). It serves as an example for reactions which make StanDep core reaction lists more self-consistent. Thus, model extraction methods are not entirely reliable in recapitulating critical cellular functions [13], highlighting the importance of accurately identifying core reactions during data integration.

Housekeeping reactions of both these pathways showed high coverage in not only MBA-like methods but also GIMME-like methods [45] (Fig. S28, S29). Inclusion of such reactions also explains why the Jaccard similarity for StanDep-derived models is higher across extraction methods than localT2- or global-derived models (Fig. S30). Thus, inclusion of such housekeeping reactions is an important criterion that the models of human cells must satisfy as they not only capture the biology of human cells but also provide better agreement regardless of extraction method. Thus, it is an important observation that the StanDep-derived core reaction lists captured >100-200 additional housekeeping reactions per cell line compared to those derived by global or localT2 approaches. Using global thresholding of top 25th percentile, the selection was favored from clusters where enzyme expression values were higher (Fig. S3). However, without direct experimental evidence (which would be context dependent) there is no way to determine an exact threshold above which a gene and the associated enzyme can be classified as active. StanDep avoids using a single threshold by calculating and applying cluster-specific thresholds. Of course, our rationale for using cluster-specific thresholds was that some clusters are enriched in housekeeping reactions. Nevertheless, applying cluster-specific thresholds was possible because of diversity of the gene

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expression profiles across cancer cell lines and human tissues used in this study and availability of datasets with larger samples. However, since StanDep requires larger datasets with higher diversity of gene expression, a global approach may still be an appropriate choice when integrating highly homogenous or small transcriptomics data sets.

We also benchmarked StanDep using six of the existing MEMs to build models for the NCI-60 cancer cell lines. This showed that StanDep works best with MBA-like extraction methods. Besides using housekeeping genes/reactions and CRISPR-Cas9 gene essentiality screens for validation and comparison with localT2-derived models. In light of StanDep-derived models being at least comparable to localT2-derived models in accuracy (Fig. 3D), we can say that StanDep provided a transcriptomic explanation for why a reaction needs to be included within the model making self-consistency an important quality metric.

The level of diversity in gene expression across cancer cell lines and human tissues helped in the identification of housekeeping genes. The models of cancer cell lines and *C. elegans* cell types were also able to predict essential genes. Furthermore, we showed that StanDep can capture relevant aspects of cell type-specific metabolism, such as the presence of peroxisomal fatty acid β-oxidation in neurons, intestine, and hypodermis cell types of *C. elegans* and NCI-60 cell line-specific essential genes. In conclusion, StanDep demonstrates that in addition to considering the expression level of a gene, the use of its variability across tissues and cell types can help to better define context-specific cellular function.

Acknowledgements

This work was supported by the NIGMS (grant no. R35 GM119850), a Lilly Innovation Fellows Award to CJ and AR, and funding from the Keck Foundation.

Methods

Datasets used

For this study, we used gene expression data for NCI60 cancer cell lines [29], HPA gene expression data for tissues [6], and gene expression data for *C. elegans* cell types [7]. For validating our models, we used a list of housekeeping genes [22], CRISPR data for 20 NCI60 cancer cell lines [23,24,26], and RNAi phenotypic data for *C. elegans* [25]. For further details on data extraction, please see supplementary methods.

Data processing

We selected genes from the human metabolic reconstruction, Recon 2.2 [30]. This included for NCI60 data, 1416 genes; for HPA data, 1661 genes out of 1673 genes in Recon 2.2; and for *C. elegans* cell type data, 1248 genes were part of the global expression datasets and *C. elegans* GEM [41]. We then converted gene expression values into enzyme expression values using gene mapping. Gene mapping involved extracting gene-protein-reaction (GPR) relationships from the model and calculating enzyme expression. The extraction of GPR was done using the COBRA function, *GPRparser.m*. For enzymes that have only

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

one subunit, the value of enzyme expression is same as the value of gene expression. For multimeric enzymes, these relationships share an "AND" relationship; thus, the minimum value amongst genes part of the enzymes were set as enzyme expression value. The assumption for multimeric enzymes was that gene with lowest expression will govern the amount of functional enzyme expressed. It should be noted that we did not resolve OR relationships representing isoenzymes and allowed all functional enzymes to be represented in the enzyme expression dataset. The enzyme expression data spanned 1325 enzymes (4133 reactions) for NCI60 data, 1792 enzymes for HPA data, and 2533 enzymes for *C. elegans* data.

Hierarchical Clustering

Clustering distribution patterns of gene expression

We $\log_{10}$-transformed the calculated enzyme expression dataset and counted the number of samples expressed with each bin width. Bin width were set based on the $\log_{10}$-transformed minimum and maximum enzyme expression values. This resulted in a matrix with rows representing each enzyme, columns representing bins, the value within the matrix representing number of samples from the dataset which were expressed within each bin range. We then performed hierarchical clustering with *Euclidean* distance metric and *complete* linkage metric to cluster genes based on distribution pattern of gene expression. We also show the comparison between using other distance (Fig. S15) and linkage methods (Fig. S16, Supplementary Results).

Deciding number of clusters

Clustering in our work is used as a tool to divide genes into categories based on distribution patterns of their expression across different conditions. These clusters are then responsible for generating their own threshold. Therefore, number of clusters were determined such that all pathway is enriched in at least one cluster. The pathways were extracted from the GEMs, Recon 2.2 (for NCI60 and human tissues) and iCEL1273 (for *C. elegans* cell types). Only pathways which contained at least one gene-associated reaction were considered. For the NCI60 Klijn et al. dataset, we used 26 clusters; for HPA dataset, we used 19 clusters; and for *C. elegans*, we used 14 clusters for enzyme expression and gene expression data respectively. We also show the comparison of choosing different number of clusters (Fig. S13; Supplementary Results).

Clustering core reaction sets or models

For analysis of models, we calculated Jaccard similarity of reaction content across different models which were part of any given analysis. We then performed hierarchical clustering to see how tissues are grouped. Hierarchical clustering was performed with the *Euclidean* distance metric and *complete* linkage metric. The interpretation of clustering Jaccard similarity is that models that are most similar to each other are likely to be equally far from other models.

Pathway enrichment

Pathway enrichment was performed by calculating hypergeometric p-value (p-value < 0.05) for the number of enzymes belonging to a given pathway present within a given cluster. Pathway association of an enzyme was calculated based on pathway association of the reactions being catalyzed by an enzyme.

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Identification of Core Reactions

StanDep

StanDep applies thresholds specific to each cluster of genes (Fig. 1). In the StanDep threshold formulation, we included two terms: (i) standard deviation, and (ii) mean term. Fine-tuned expression level of genes is represented as the Standard deviation term; and is dependent on the difference between standard deviation of the cluster and the dataset. Lower standard deviation favors the selection of enzymes in all contexts while higher standard deviation term reflects context-specificity of the enzymes. The mean term, interpretation of second assumption, is dependent on the magnitude of the expression of enzymes in that cluster. In both cases, we used the difference between cluster and overall data to address inconsequential variations that maybe occurring in expression. The standard deviation is always positive but logarithmic mean may be negative and sometime be even quite large. Therefore, we introduced normalization to make the standard deviation term and mean term at par. The threshold for each cluster is given by the following equations:

$$\Theta_c = (\theta_c - min(\theta_c)) * 100/max(\theta_c - min(\theta_c));\ \Theta_c \in [0,100] \quad (1)$$

$$\theta_c = f(\sigma_c) + g(\mu_c); \quad (2)$$

$$f(\sigma_c) = (\sigma_c - \Delta)/max(\sigma_c - \Delta); \quad (3)$$

$$g(\mu_c) = -(\mu_c - \boldsymbol{M}); \quad (4)$$

In the above set of equations, $\boldsymbol{\Theta_c}$ is the processed threshold value for a given cluster *c*; $\boldsymbol{\theta_c}$ is the raw value of threshold for cluster *c*; $\boldsymbol{\sigma_c}$ is the standard deviation of the cluster *c*; $\boldsymbol{\Delta}$ is the standard deviation of the dataset; $\boldsymbol{\mu_c}$ is the mean of the cluster *c*; and $\boldsymbol{M}$ is the mean of the dataset. The equation is derived by penalizing cluster-specific thresholds based on: (i) how low the cluster mean is compared to the mean of the dataset; (ii) how far the standard deviation of the cluster is from the standard deviation of the dataset. The final normalization was done to ensure that the clusters-specific thresholds are between 0 and 100. The $\boldsymbol{\Theta}$ is the top percentile value of the cluster-specific data above which an enzyme in that cluster in a given context is qualified active. If the value of $\boldsymbol{\Theta_c}$ is 100, we set the threshold value of the cluster as the mean of the data.

The current published literature on the below thresholding methods does not address how the threshold values should be derived. Therefore, we used some of the most commonly used percentile values in previously published studies [13,21,27,28].

Other thresholding methods

We used three of the existing thresholding methods: (i) global [13,21,27], (ii) localT1 [21], and (iii) localT2 [21,28]. The implementation for each of them was same as in a previous study [21,28]. However, they have also been described in detail in supplementary methods.

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Constraining Pre-extraction Models and Model reduction

Exometabolomic constraints

Exometabolomic data of the NCI60 cell line were obtained from previous work [31] and further processed as previously described [13]. After processing, we added 23 new demand reactions, wherein each reaction is secreting a different metabolite. These were added to reflect the experimental observations by Jain et al. The biomass reaction was changed to one that contains precursor molecules from the one that contains macromolecules like DNA, RNA, protein, lipids, carbohydrate, and others. The replacement of the biomass reaction was done to all the models. The global lower and upper bounds for all reactions except biomass and ATP demand were set to -1000 and 1000 respectively. The lower bounds of the biomass reaction and ATP demand were constrained to relatively small values of the order of 1e-2 and 1.833 mmol gDW^{-1} h^{-1} [46] respectively. The cell line specific constraints on 78 demand and exchange reactions were applied on the modified Recon 2.2, followed by making flux consistent constrained genome-scale models for each of the cell lines. This was done by identifying and removing flux-inconsistent reactions using *fastcc.m* in COBRA Toolbox. The flux tolerance was always set to 1e-8.

No constraints

To make unconstrained models, we did not apply exometabolomic constraints but only applied constraints on lower bounds of biomass and ATP demand reaction as described above. The global lower and upper bounds were set to -1000 and 1000 respectively. This was followed by identifying and removing flux inconsistent reactions. The flux tolerance was always set to 1e-8.

Semi constrained

To make semi-constrained models, we applied directional constraints on demand and exchange reactions of each cell line, applied constraints on lower bounds of biomass and ATP demand as described above. The global lower and upper bounds were set to -1000 and 1000 respectively. This was followed by identifying and removing flux-inconsistent reactions. The flux tolerance was always set to 1e-8.

Relaxed constraints

To make relaxed models, we constrained the direction of flow to 10 mmol gDW^{-1} h^{-1} on demand and exchange reactions as suggested by exometabolomic data. The order of magnitude of original constraints on these reactions was 1e-3 to 1e-6. The global lower and upper bounds were set to -1000 and 1000 respectively. This was followed by identifying and removing flux-inconsistent reactions. The flux tolerance was always set to 1e-8.

Implementation with model extraction methods (MEMs)

In this study, we compared the models derived using localT2 and StanDep. This section describes the extraction of StanDep-derived models by tailoring each of the MEMs. Models derived using localT2 were not constructed in this study, rather we extracted those models from a previous study [28]. Therefore, for implementation of each of the extraction methods with these thresholding methods, please see the methods for that study.

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To construct models using 6 of the extraction methods these inputs were common to all: (i) a flux-consistent Recon 2.2 genome-scale model was used, and (ii) epsilon, a.k.a. flux tolerance, was set to 1e-8. Inputs specific to a given MEM are described below.

FASTCORE

To construct models using FASTCORE [19], we used *fastcore.m* in the COBRA toolbox. Other inputs needed for the algorithm are requires core reaction lists. Please see above on how we identified them. The biomass reaction was manually added to the core reaction list.

iMAT

To construct models using iMAT [18], we used *iMAT.m* in the COBRA toolbox. Other inputs needed for the algorithm are: (i) core reactions (i.e., list of reactions identified to be active, including the biomass reaction) and (ii) non-core reactions, which are not part of core reactions (reactions not associated to a gene were not included in non-core reactions).

MBA

To construct models using MBA [14], we used *MBA.m* in the COBRA toolbox. Other inputs needed for the algorithm are: (i) high expression set, list of reactions which are highly expressed and (ii) medium expression set, list of reactions which are moderately expressed. We generated 10% interval around threshold for each cluster. We defined high expression set as the list of reactions catalyzed by enzymes which are above 110% of the threshold value, and medium expression set as the list of reactions catalyzed by enzymes which are between 90% and 110% of the threshold value. For instances where a reaction was present in both high and medium expression set, we interpreted it as at least enzyme associated to the reaction being able to express at high levels. Thus, we put these reactions in high expression sets. The biomass reaction was given the highest value.

mCADRE

To construct models using mCADRE [15], we used *mCADRE.m* in the COBRA toolbox. Other inputs needed for the algorithm are: (i) ubiquity score (i.e., how often a reaction is expressed across samples of the same context); (ii) confidence scores quantifying level of evidence for a reaction to be present in the model; (iii) protected reactions; and (iv) since we did not protect any reactions, we set the functionality check to 0. To calculate ubiquity score ($\boldsymbol{U_{c,i}}$), we calculated threshold distances ($\boldsymbol{D_{c,i}}$), here defined as distance of a given enzyme expression ($\boldsymbol{x_{i,c}}$) in the context *i* from the threshold ($\boldsymbol{\Theta_c}$) of the cluster *c* where the enzyme belongs. The threshold distances and ubiquity scores were calculated using the eqs (5-7).

$$D_{c,i} = x_{i,c} - \Theta_c \tag{5}$$

$$if\ D_{c,i} > 0;\ U_{c,i} = 1 \tag{6}$$

$$if\ D_{c,i} < 0;\ U_{c,i} = 1 - (D_{c,i}/min(D_{c,i})) \tag{7}$$

We used the ubiquity score to quantify how often an enzyme is expressed in samples of the same context. For isoenzymatic reactions, the reaction ubiquity score was set to the enzyme with maximum ubiquity

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

score. For reactions which do not have an associated gene, the ubiquity score was set to -1. Since, we did not have confidence scores, we assigned a confidence of 0 to all reactions, as suggested in COBRA toolbox tutorial for mCADRE. However, we also tried using our list of core reactions as a binary vector specifying whether a reaction is in the core set and if it did not have any effect of the final model. The biomass reaction was manually assigned a ubiquity score of 1. The confidence score of 1 is associated with transcriptomics evidence and our metric ubiquity score already has this information.

INIT

To construct models using INIT [16], we used *INIT.m* in the COBRA toolbox. Other inputs needed for the algorithm are reaction weights, varying between -1 and 1. To calculate enzyme weights, we calculated the threshold distance for each enzyme as described previously, without normalizing. Weights for all reactions catalyzed by an enzyme were same as the enzyme weight. Here, we used a different normalizing scheme. We scaled our threshold distances to a maximum threshold distance for any of the enzymes within the data. For isoenzymatic reactions, the weights of each enzyme were added. We set the weights for non-gene associated reactions to 0. The biomass reaction was manually assigned a weight of 1.

GIMME

To construct models using GIMME [17], used *GIMME.m* in the COBRA toolbox. Other inputs needed for the algorithm are: (i) a reaction expression vector representing gene expression values associated with the reactions; and (ii) a threshold determining whether reaction expression is considered active. We calculated the reaction expression vector in the same way as we calculated enzyme weights for INIT. The thresholds were set to 1. The biomass reaction was given a value of 1.

Gene essentiality in NCI60

NCI60 data

To test our essentiality predictions of NCI60 models with CRISPR screen data, we downloaded pooled CRISPR knockout screen data from DepMap.org [23,24,26] for 20 NCI-60 cell lines. Essential genes were identified based on the CRISPR score. The CRISPR score was calculated as the ratio of abundance of single guide RNA (sgRNA) of a knock out after and before growth selection. A negative CRISPR score suggests a higher probability that the gene is essential. The accuracy was estimated using the percentage of predicted essential genes that have a negative score [47]. We then used 1-tailed Wilcoxan rank sum test to identify if the CRISPR scores for genes predicted to be essential in the metabolic model and CRISPR scores of genes predicted to be non-essential are coming from the same populations.

RNAi phenotypic data

To get the list of essential genes in *C. elegans*, we extracted genes that presented a *Nonv* or *Gro* RNAi phenotype. As described by the authors [25], *Nonv* phenotype refers to all phenotypic classes that result in lethality or sterility (1170 essential genes); and *Gro* refers to phenotypic classes that result in growth defects, slow post-embryonic growth or larval arrest (276 essential genes). Out of these, the iCEL1273 [41] model contained 187 genes. Similarly, we found 900 non-essential genes in iCEL1273.

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. 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: 1279–1285. doi:10.1038/nbt.1711

2. 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: 3083. doi:10.1038/ncomms4083

3. Lewis NE, Abdel-Haleem AM. The evolution of genome-scale models of cancer metabolism. Front Physiol. 2013;4: 237. doi:10.3389/fphys.2013.00237

4. Shen Y, Liu J, Estiu G, Isin B, Ahn Y-Y, Lee D-S, et al. Blueprint for antimicrobial hit discovery targeting metabolic networks. Proc Natl Acad Sci U S A. 2010;107: 1082–7. doi:10.1073/pnas.0909181107

5. Kim MK, Lun DS. Methods for integration of transcriptomic data in genome-scale metabolic models. Comput Struct Biotechnol J. 2014;11: 59–65. doi:10.1016/j.csbj.2014.08.009

6. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science (80-). 2015;347: 1260419–1260419. doi:10.1126/science.1260419

7. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017;357: 661–667. doi:10.1126/science.aam8940

8. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, et al. Mapping the Mouse Cell Atlas by Microwell-Seq. Cell. 2018;172: 1091-1107.e17. doi:10.1016/J.CELL.2018.02.001

9. Creecy JP, Conway T. Quantitative bacterial transcriptomics with RNA-seq. Curr Opin Microbiol. 2015;23: 133–40. doi:10.1016/j.mib.2014.11.011

10. Hyduke DR, Lewis NE, Palsson BØ. Analysis of omics data with genome-scale models of metabolism. Mol Biosyst. 2013;9: 167–74. doi:10.1039/c2mb25453k

11. Cimini D, Patil KR, Schiraldi C, Nielsen J. Global transcriptional response of Saccharomyces cerevisiae to the deletion of SDH3. BMC Syst Biol. 2009;3: 17. doi:10.1186/1752-0509-3-17

12. Machado D, Herrgård M. Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. Maranas CD, editor. PLoS Comput Biol. 2014;10: e1003580. doi:10.1371/journal.pcbi.1003580

13. 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: 318-329.e6. doi:10.1016/j.cels.2017.01.010

14. Jerby L, Shlomi T, Ruppin E. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Mol Syst Biol. 2010;6: 401. doi:10.1038/msb.2010.56

15. Wang Y, Eddy JA, Price ND. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Syst Biol. 2012;6: 153. doi:10.1186/1752-0509-6-153

16. 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

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INIT. Maranas CD, editor. PLoS Comput Biol. 2012;8: e1002518. doi:10.1371/journal.pcbi.1002518

17. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008;4: e1000082. doi:10.1371/journal.pcbi.1000082

18. Zur H, Ruppin E, Shlomi T. iMAT: an integrative metabolic analysis tool. Bioinformatics. 2010;26: 3140–3142. doi:10.1093/bioinformatics/btq602

19. Vlassis N, Pacheco MP, Sauter T. Fast Reconstruction of Compact Context-Specific Metabolic Network Models. Ouzounis CA, editor. PLoS Comput Biol. 2014;10: e1003424. doi:10.1371/journal.pcbi.1003424

20. Schultz A, Qutub AA. Reconstruction of Tissue-Specific Metabolic Networks Using CORDA. Maranas CD, editor. PLOS Comput Biol. 2016;12: e1004808. doi:10.1371/journal.pcbi.1004808

21. Richelle A, Joshi C, Lewis NE. Assessing key decisions for transcriptomic data integration in biochemical networks. Hatzimanikatis V, editor. PLOS Comput Biol. 2019;15: e1007185. doi:10.1371/journal.pcbi.1007185

22. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013;29: 569–574. doi:10.1016/j.tig.2013.05.010

23. Aguirre AJ, Meyers RM, Weir BA, Vazquez F, Zhang C-Z, Ben-David U, et al. Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. Cancer Discov. 2016;6: 914–29. doi:10.1158/2159-8290.CD-16-0154

24. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49: 1779–1784. doi:10.1038/ng.3984

25. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature. 2003;421: 231–237. doi:10.1038/nature01278

26. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016;34: 184–191. doi:10.1038/nbt.3437

27. Pacheco MP, Pfau T, Sauter T. Benchmarking Procedures for High-Throughput Context Specific Reconstruction Algorithms. Front Physiol. 2016;6: 410. doi:10.3389/fphys.2015.00410

28. Richelle A, Chiang AWT, Kuo C-C, Lewis NE. Increasing consensus of context-specific metabolic models by integrating data-inferred cell functions. Ouzounis CA, editor. PLOS Comput Biol. 2019;15: e1006867. doi:10.1371/journal.pcbi.1006867

29. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, et al. A comprehensive transcriptional portrait of human cancer cell lines. Nat Biotechnol. 2015;33: 306–312. doi:10.1038/nbt.3080

30. 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: 109. doi:10.1007/s11306-016-1051-4

31. Jain M, Nilsson R, Sharma S, Madhusudhan N, Kitami T, Souza AL, et al. Metabolite Profiling Identifies a Key Role for Glycine in Rapid Cancer Cell Proliferation. Science (80-). 2012;336:

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1040–1044. doi:10.1126/science.1218595

32. Wolf B. Biotinidase deficiency: "if you have to have an inherited metabolic disease, this is the one to have." Genet Med. 2012;14: 565–575. doi:10.1038/gim.2011.6

33. Faith M, Abraham P. An update on diagnostic value of biotinidase: From liver damage tocancer: Minireview. Biomed Res. 24. Available: http://www.alliedacademies.org/articles/an-update-on-diagnostic-value-of-biotinidase-from-liver-damage-tocancer-minireview.html

34. Kang U-B, Ahn Y, Lee JW, Kim Y-H, Kim J, Yu M-H, et al. Differential profiling of breast cancer plasma proteome by isotope-coded affinity tagging method reveals biotinidase as a breast cancer biomarker. BMC Cancer. 2010;10: 114. doi:10.1186/1471-2407-10-114

35. So AK-C, Kaur J, Kak I, Assi J, MacMillan C, Ralhan R, et al. Biotinidase is a Novel Marker for Papillary Thyroid Cancer Aggressiveness. Mehta K, editor. PLoS One. 2012;7: e40956. doi:10.1371/journal.pone.0040956

36. Dwek RA. Biological importance of glycosylation. Dev Biol Stand. 1998;96: 43–7. Available: http://www.ncbi.nlm.nih.gov/pubmed/9890515

37. Taniguchi N, Kizuka Y. Glycans and Cancer. Advances in cancer research. 2015. pp. 11–51. doi:10.1016/bs.acr.2014.11.001

38. Chan C-P, Mak T-Y, Chin K-T, Ng IO-L, Jin D-Y. N-linked glycosylation is required for optimal proteolytic activation of membrane-bound transcription factor CREB-H. J Cell Sci. 2010;123: 1438–48. doi:10.1242/jcs.067819

39. Wu G, Guo Z, Chatterjee A, Huang X, Rubin E, Wu F, et al. Overexpression of Glycosylphosphatidylinositol (GPI) Transamidase Subunits Phosphatidylinositol Glycan Class T and/or GPI Anchor Attachment 1 Induces Tumorigenesis and Contributes to Invasion in Human Breast Cancer. Cancer Res. 2006;66: 9829–9836. doi:10.1158/0008-5472.CAN-06-0506

40. Nagpal JK, Dasgupta S, Jadallah S, Chae YK, Ratovitski EA, Toubaji A, et al. Profiling the expression pattern of GPI transamidase complex subunits in human cancer. Mod Pathol. 2008;21: 979–91. doi:10.1038/modpathol.2008.76

41. Yilmaz LS, Walhout AJM. A Caenorhabditis elegans Genome-Scale Metabolic Network Model. Cell Syst. 2016;2: 297–311. doi:10.1016/j.cels.2016.04.012

42. Nairn A V, York WS, Harris K, Hall EM, Pierce JM, Moremen KW. Regulation of glycan structures in animal tissues: transcript profiling of glycan-related genes. J Biol Chem. 2008;283: 17298–313. doi:10.1074/jbc.M801964200

43. O'Hagan S, Wright Muelas M, Day PJ, Lundberg E, Kell DB. GeneGini: Assessment via the Gini Coefficient of Reference "Housekeeping" Genes and Diverse Human Transporter Expression Profiles. Cell Syst. 2018;6: 230-244.e1. doi:10.1016/J.CELS.2018.01.003

44. Muelas MW, Mughal F, O'Hagan S, Day PJ, Kell DB. The role and robustness of the Gini coefficient as an unbiased tool for the selection of Gini genes for normalising expression profiling data. bioRxiv. 2019; 718007. doi:10.1101/718007

45. Robaina Estevez S, Nikoloski Z. Generalized framework for context-specific metabolic model extraction methods. Front Plant Sci. 2014;5: 491. doi:10.3389/fpls.2014.00491

46. Kilburn DG, Lilly MD, Webb FC. The Energetics of Mammalian Cell Growth. J Cell Sci. 1969;4.

47. Tobalina L, Pey J, Rezola A, Planes FJ. Assessment of FBA Based Gene Essentiality Analysis in

bioRxiv preprint doi: https://doi.org/10.1101/594861; this version posted September 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cancer with a Fast Context-Specific Network Reconstruction Method. PLoS One. 2016;11: e0154583. doi:10.1371/journal.pone.0154583