UC San Diego
UC San Diego Previously Published Works

Title

Increasing consensus of context-specific metabolic models by integrating data-inferred cell functions

Permalink

https://escholarship.org/uc/item/1t69c8ft

Journal

PLOS Computational Biology, 15(4)

ISSN

1553-734X

Authors

Richelle, Anne

Chiang, Austin WT

Kuo, Chih-Chung

et al.

Publication Date

2019

DOI

10.1371/journal.pcbi.1006867

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

RESEARCH ARTICLE

Increasing consensus of context-specific metabolic models by integrating data-inferred cell functions

Anne Richelle[1,2], Austin W. T. Chiang[1,2], Chih-Chung Kuo[1,3], 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, United States of America, **2** Department of Pediatrics, University of California, San Diego, School of Medicine, La Jolla, CA, United States of America, **3** Department of Bioengineering, University of California, San Diego, La Jolla, CA, United States of America

* nlewisres@ucsd.edu

Abstract

Genome-scale metabolic models provide a valuable context for analyzing data from diverse high-throughput experimental techniques. Models can quantify the activities of diverse pathways and cellular functions. Since some metabolic reactions are only catalyzed in specific environments, several algorithms exist that build context-specific models. However, these methods make differing assumptions that influence the content and associated predictive capacity of resulting models, such that model content varies more due to methods used than cell types. Here we overcome this problem with a novel framework for inferring the metabolic functions of a cell before model construction. For this, we curated a list of metabolic tasks and developed a framework to infer the activity of these functionalities from transcriptomic data. We protected the data-inferred tasks during the implementation of diverse context-specific model extraction algorithms for 44 cancer cell lines. We show that the protection of data-inferred metabolic tasks decreases the variability of models across extraction methods. Furthermore, resulting models better capture the actual biological variability across cell lines. This study highlights the potential of using biological knowledge, inferred from omics data, to obtain a better consensus between existing extraction algorithms. It further provides guidelines for the development of the next-generation of data contextualization methods.

Author summary

Genome-scale models of human metabolism have facilitated numerous exciting discoveries regarding human physiology and therapeutics. The accuracy of results from such studies requires that models capture the tissue or cell-type specific metabolism. In hopes to obtain accurate models, several algorithms have been developed to extract cell- or tissue-specific metabolic models. Each algorithm has provided useful insights into the metabolism of specific cell and tissue types. However, since each of these methods use different assumptions to guide reaction inclusion and removal, they result in considerable

differences in size, functionality, accuracy, and ultimate biological interpretation, even when using the same data set. To overcome this, the enclosed research proposes an approach to infer the functionalities of a cell or tissue from omics data, and then protect these functions to guide the construction of a context-specific model. Through this study, we highlight the value of using experimental data to help infer the set of metabolic functions that should be included in a model, in an effort to obtain greater consensus across existing extraction algorithms. This study further provides guidelines for the development of the next-generation of data contextualization methods.

Introduction

Genome-scale metabolic models (GeMs) have been widely used for model-guided analysis of large omics datasets, since they provide cellular context to these data by establishing a mechanistic link from genotype to phenotype. GeMs include all reactions in an organism. Since not all enzymes are active in each cell type or culture condition, algorithms have been developed to build context-specific models using omics data to recapitulate the metabolism of specific cell types under specific conditions [1,2]. These algorithms have provided useful insights in the metabolism of specific cell and tissue types [1,3–10]. However, since each method uses different assumptions to guide reaction inclusion and removal, they result in considerable differences in size, functionality, accuracy, and ultimate biological interpretation, even when using the same data set [1,2,11].

The poor consensus in generated models requires increased caution in the interpretation of model-derived hypotheses of how metabolism is used under specific environments. Indeed, most generated models, upon construction, will be missing known metabolic functions and this varies considerably for models built using different approaches [1]. To gain confidence in model predictions and reconcile the differences across approaches, users can enforce the inclusion of known metabolic capabilities in the model. In this regard, the tINIT extraction algorithm introduced the possibility to enforce the capacity of context-specific models to represent some cellular functionalities by using a list of metabolic tasks known to occur in all cell types [12]. However, this protectionist approach requires one to know and predefine the functionalities of a specific cell line, tissue, or context.

To overcome this, we propose an approach to infer the functionalities of a cell or tissue from omics data, and then protect these functions to guide the construction of a context-specific model. To this end, we curated and standardized published lists of metabolic tasks [13,14], resulting in a collection of 210 tasks covering 7 major metabolic activities of a cell (energy generation, nucleotide, carbohydrates, amino acid, lipid, vitamin & cofactor and glycan metabolism). We also developed a framework to directly predict the activity of these functionalities from transcriptomic data and subsequently use these for a protectionist approach to several existing extraction algorithms. Models resulting from this approach should more comprehensively capture the unique metabolic functions of a given cell type. We evaluated the validity and variation across models built with this approach, coupled to existing context-specific extraction methods. Specifically, we constructed hundreds of models for 44 cancer cell lines in which we built the models using standard approaches or protected a list of metabolic functions that have been inferred from the original transcriptomic data of each cell line. We also varied the reference human reconstruction and algorithms employed for the generation of cell line specific models, using two different reference models (iHsa [13] and Recon 2.2 [15]) and 6 different algorithms (mCADRE [16], fastCORE [5], GIMME [6], INIT [7], iMAT [4],

and MBA [8]). We compared the sets of extracted models at the level of reaction content, metabolic functions, and capacity to predict essential genes identified in CRISPR-Cas9 loss-of-function screens. Through this study, we highlight the value of using experimental data to help infer the set of metabolic tasks that should be included in a model, in an effort to obtain greater consensus across existing extraction algorithms.

Results

Context-specific extraction methods yield more variation in model content than omics data of different cell lines

We built models from Recon 2.2 [17] and iHsa [13] using six model extraction methods (MEMs: mCADRE, fastCORE, GIMME, INIT, iMAT, MBA) for 44 different cell lines from the NCI-60 panel (S1 Table; 15 cell lines were not used due to the absence RNA-Seq data in [18] for these cell lines). Uptake and secretion rates of the input GeMs were quantitatively constrained using a list of experimentally measured metabolites (S2 Table)[19,20]. Furthermore, a biomass function, consisting of 56 metabolites required for growth, was added and constrained to the experimentally measured growth rate of the cell lines (S3 Table). The biomass function and constraints from exometabolomic data introduced in the GeMs were implemented as described in [1]. The extraction process of cell line specific models was done based on RNA--Seq data [18] to specify active genes in each cell line. Details on the implementation of MEMs tested and the preprocessing of gene expression data for the definition of gene activity are provided in the Methods section.

To assess the relative impact of algorithm and data source on model content, we conducted a principal component analysis (PCA) of the reactions in all models for each reference GeM. As observed previously [1,11], the decisions regarding algorithm choice significantly impact the content of our cell line-specific models. The first principal (PC1) component explains 38% of the overall variance in model reaction content, with >60% of the variation in PC1 explained by the choice of model extraction method (Fig 1A and 1B). Indeed, the different algorithms yielded cell line-specific models that varied considerably in size, with few reactions common to all models extracted from either Recon 2.2 or iHsa (Fig 1C, Fig A in S1 Text). Even among models extracted using the same algorithm, there is non-negligible variability in model reaction content (Fig 1D). This leads to the generation of models that are substantially different with respect to the cell line considered, while the transcriptomic data used to tailor the GeMs shows high consistency across most cell lines (Fig 1E).

Metabolic tasks as a tool for model benchmarking and model extraction

Model reaction content is often evaluated to compare context-specific algorithms. Recently, approaches to benchmark models with their functionalities have been proposed [1,12,21]. Current approaches use repositories of known cellular tasks to assess the capacity of models to achieve specific modeling goals or to enable the representation of specific metabolic functions. This idea of assessing the quality of a metabolic network reconstruction using biological knowledge was introduced in Recon 1 through the characterization of the "human metabolic knowledge landscape" [22]. However, the concept of "metabolic tasks" (Fig 2A) was clearly defined in 2013 by Thiele and coworkers [14] to benchmark the improvements of Recon 2 compared to Recon 1, wherein they stated that "a metabolic task is defined as a nonzero flux through a reaction or through a pathway leading to the production of a metabolite B from a metabolite A". Since then, additional lists of tasks have been published. To standardize these and develop a framework for their easy use with GeMs, we curated the existing lists of
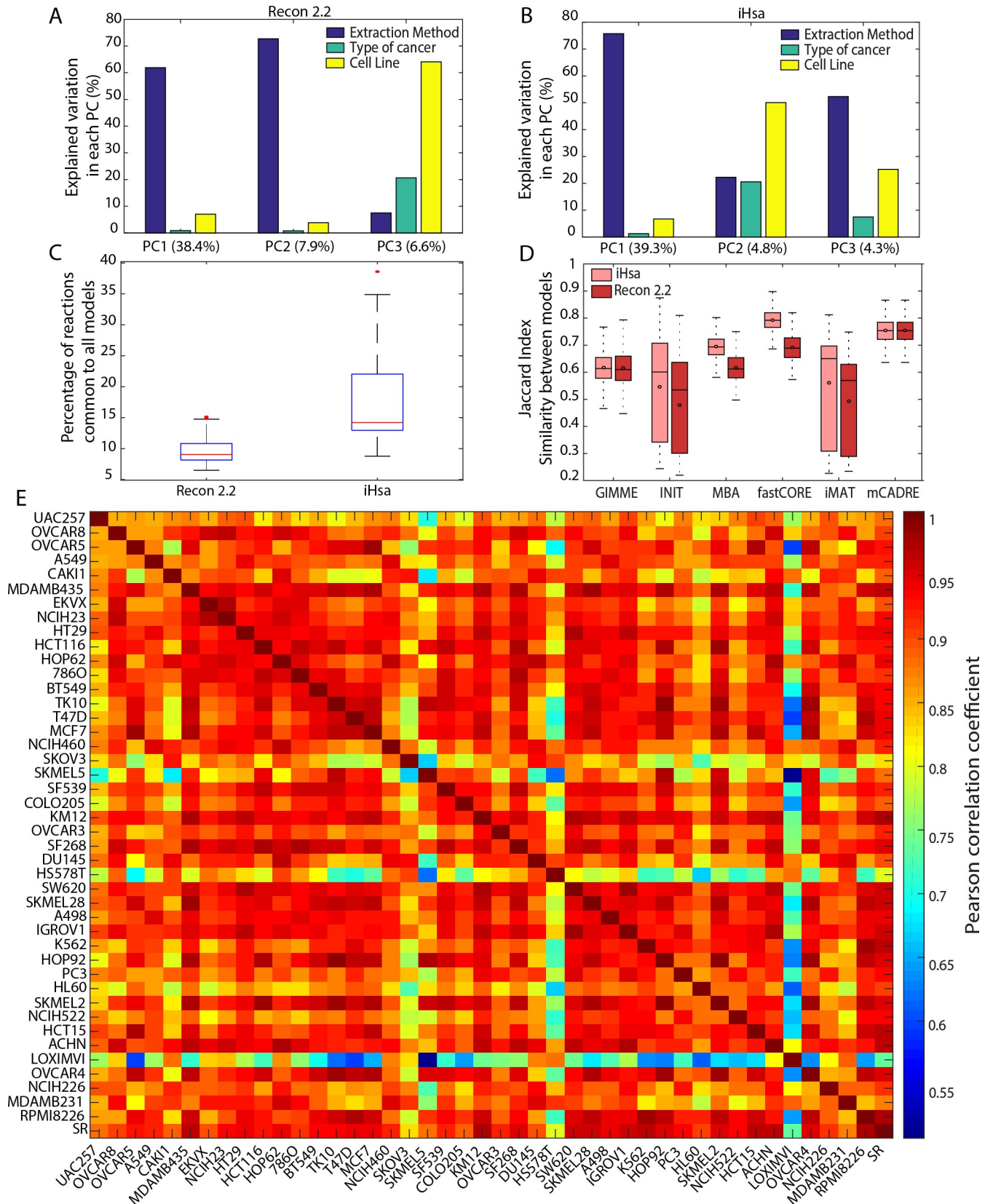
Fig 1. Choice of extraction method is mainly responsible for the variability in the reaction content across models. The extraction method used contributes the most to the first PC for models built using (A) Recon 2.2 or (B) iHsa as a reference GeM. (C) Only a small percentage of reactions are shared in all the models extracted from both Recon 2.2 (shared reactions = 218) and iHsa (shared reactions = 513). (D) For each method, the similarity of models of different cell lines (computed using a Jaccard index based on the reaction content) varies substantially, while (E) the transcriptomic data used to generate these models present a much higher correlation between cell lines.

metabolic tasks (13,14) and obtained a collection of 210 tasks covering 7 major metabolic activities of a cell (energy generation, nucleotide, carbohydrates, amino acid, lipid, vitamin & cofactor and glycan metabolism) (Fig 2B and 2C, S4 Table). We evaluated the task collection using genome-scale metabolic models for human [13,14,17,22,23], CHO cells [9], rat [13] and mouse [24] (Fig 2D, S5 Table). Details on our proposed formalism of the metabolic tasks and the associated computational framework for their use are presented in the Methods.

Metabolic tasks can be used to compare the performance of models extracted from different reference GeMs. As observed at the level of the reaction content, the extraction method strongly influences the model functions (explaining >50% of the overall variance in the first PC; Fig 3A). However, the reference model is the most prominent factor in the second PC underlying a non-negligible influence of this variable in the extraction process. This is mainly due to differences in gene, protein, reaction association (GPR) annotations and reaction content between Recon 2.2 and iHsa. Interestingly, Recon 2.2 captures more metabolic functions with fewer reactions (Fig 3B). However, the number of successful tasks increases proportionally with the number of reactions in a model. Furthermore, as the extraction method used
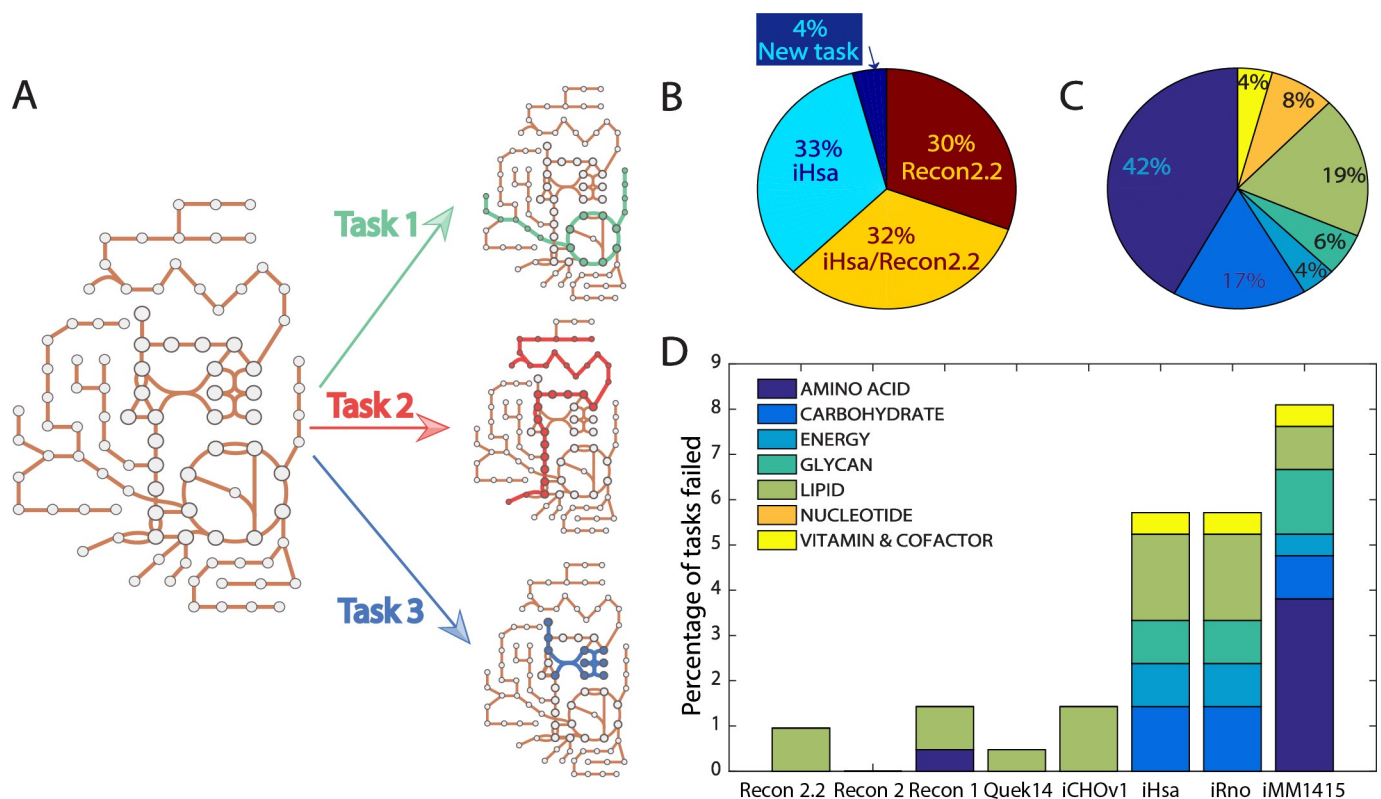
Fig 2. Generation of a collection of 210 tasks representing known metabolic features. (A) A metabolic task can be defined as the set of reactions needed to transform input metabolites into defined products. (B) Original sources of our collection of 210 curated tasks. (C) The curated list of tasks covers 7 main metabolic systems. (D) Several existing genome-scale models were used to evaluate the collection of curated tasks. A small number of tasks were non-functional on specific GeMs, for reasons detailed in S5 Table.
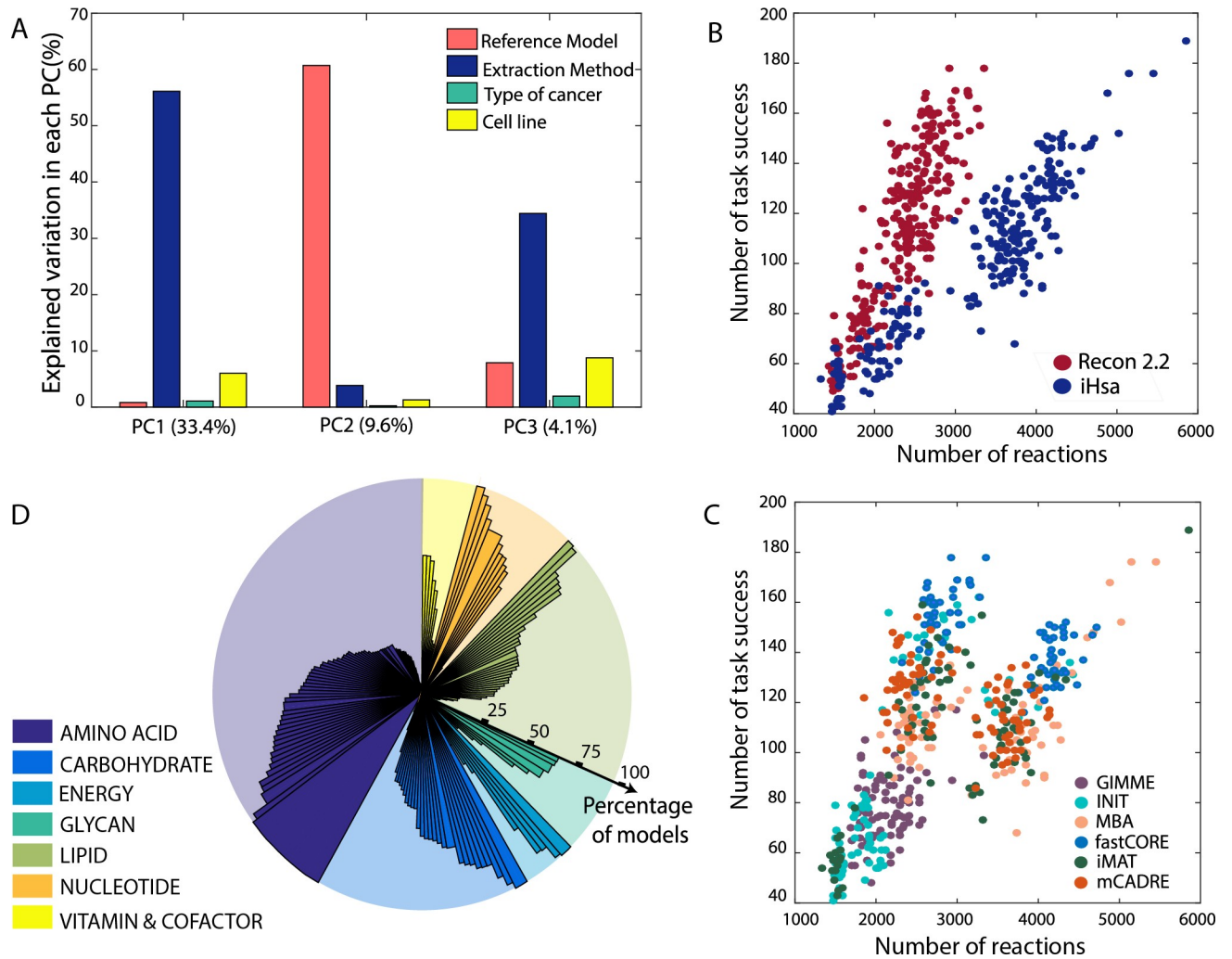
Fig 3. Metabolic tasks can be used to benchmark models. (A) The choice of extraction method contributes the most to the first PC of functional metabolic tasks, and the reference model explains the most variance within the second PC. (B) Recon 2.2 captures more metabolic functions with fewer reactions. (C) Some extraction methods are associated with a higher capacity to conserve more metabolic functions, as they typically retain more reactions. (D) The percentage of tasks present in all extracted models is low, and is predominantly associated with amino acid metabolism. Furthermore, some metabolic functionalities are not retained in any extracted models.

influences the number of reactions removed, distinct patterns are seen from the ratio of the number of metabolic tasks to the number of reactions introduced by the different algorithms (Fig 3C). As for the reaction content, the number of tasks retained in each model varies substantially, depending on the cell lines considered. Surprisingly, only 8% of the tasks are present in all models (Fig 3D, Fig B in S1 Text), thus highlighting the large variation in metabolic functions a model will have, depending on algorithm choice.

Protecting inferred metabolic tasks reduces the variability of model content from different algorithms

We inferred active metabolic tasks directly from transcriptomic data using the whole genome-scale model. To this end, we computed the list of reactions associated with each task and used the GPR rules to determine the gene expression levels associated with each of these reactions.
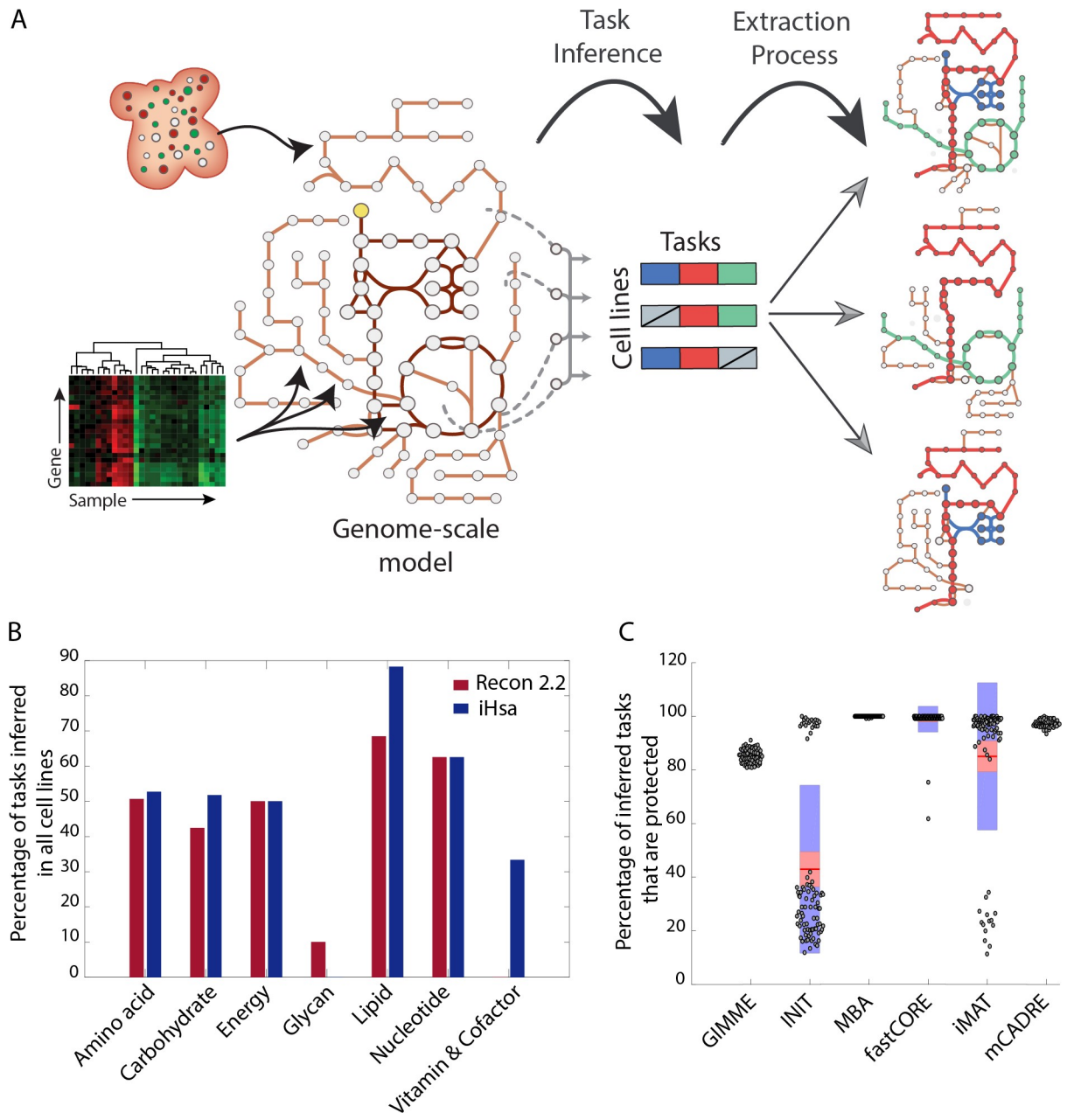
Fig 4. Metabolic tasks can be inferred from omics data to determine which tasks should be protected during the model extraction process. (A) Metabolic functions are inferred from transcriptomic data using the genome-scale model and then protected during the implementation of the extraction algorithms. (B) The functional analysis of transcriptomic data highlights that more than half of the tasks should be conserved across all the cell lines. (C) The computational framework of some extraction algorithms does not allow a complete protection of the inferred metabolic tasks, but protected tasks were almost completely retained for MBA-like methods (i.e. MBA, fastCORE and mCADRE). The red area represents 1.96 SEM (standard error of the mean—95% confidence interval) and the blue 1 SD (standard deviation).

A metabolic score is attributed to each task by using the mean activity level of each reaction (Fig 4A; See Methods). We found that more than the half of the tasks should be conserved across all cell lines (Fig 4B), which is far more than those active using the algorithms in their standard format (i.e., without protecting tasks). Therefore, we generated a new set of models,

wherein we also enforced the inclusion of reactions associated with tasks inferred for each of the 44 different cell lines (S7 Table). We focused on MBA-like algorithms (i.e., MBA, fastCORE and mCADRE), since they are directly amenable to use the protectionist approach with minor modifications to the algorithms (Fig 4C). Indeed, other algorithms do not ensure the inclusion of a reaction even if it is enforced. For example, iMAT relies on the definition of a core set of high-confidence reactions, but core reactions can be removed if it depends on many non-expressed non-core reactions (Fig 4C). See Methods for a detailed description of the implementation of the protectionist approach for each algorithm.

For equivalent extraction setups (i.e., same reference model, extraction method, and cell line), the number of reactions included in the extracted model was not considerably influenced by the protection of the metabolic task, while the number of active tasks clearly increases (Fig C in S1 Text). We performed PCA of the reaction content and the metabolic functions of the models with protected tasks. We observed that the protection of metabolic tasks inferred from data significantly decreased the influence of the extraction method on the final model content. However, the use of this approach remains sensitive to the choice of the reference model (Fig 5A and 5B). The reduction of the variability of model with respect to the extracted methods used can be explained by the increased number of shared tasks across cell lines, all supported by the transcriptomic data. This is seen in particular for tasks involved in amino acid and lipid metabolism (Fig D in S1 Text). Furthermore, we observed that the variation in model content was better explained by the cell lines (Fig E in S1 Text). Actually, the task protection increases the similarities between context-specific models with respect to the cell line (Fig F in S1 Text) but also with respect to the transcriptomic data (Fig G in S1 Text). Finally, all the models now share more than 64% of the metabolic tasks (Fig 5C).

Beyond model content, we evaluated how the protectionist approach influenced model predictions. Thus, we analyzed the influence of protecting inferred tasks on gene-essentiality predictions (i.e. prediction of the genes whose knockdown leads to a growth impairment). We systematically deleted each gene in all generated models, and then used flux balance analysis to test models for normal or impaired growth. Gene deletions associated with impaired growth are considered as essential. We observe that task protection reduces the number of genes predicted to be essential for all thresholds considered (i.e., percentage of the maximum wild type growth rate) for the various extraction methods used on both reference models (Fig 5D; Fig H in S1 Text).

We further evaluated the accuracy of essentiality predictions by comparing these to CRISPR-Cas9 loss-of-function screens for 20 cell lines [25–27]. In these screens, essential genes are identified based on gene scores attributed using single guide RNA (sgRNA) abundance for each knockout before and after growth selection. Gene scores that are more negative have a higher probability of being essential. Therefore, the agreement between model predictions and the CRISPR screen data can be quantified as the percentage of predicted essential genes that have a negative gene score [28]. Furthermore, the significance of the improvement gained from protecting data-inferred metabolic tasks can be computed using a 1-tailed Wilcoxon test. Consistent with previous reports [1,29], we found that the models, without protecting metabolic tasks, correctly predicted many essential genes. However, overall, the protectionist approach provided a small but significant improvement to gene-essentiality predictions (up to 5% improvement; Fig 5E; Fig I in S1 Text). However, the task protection reduced the number of predicted essential genes, which increased the proportion of true positives and reduced the number of false positives.

To further assess the identity of these true positives provided by task protection, we compared our model simulations to a collection of known anti-cancer drug targets (S11 Table) [30]. In this analysis, we found the protectionist approach better captured the gene essentiality
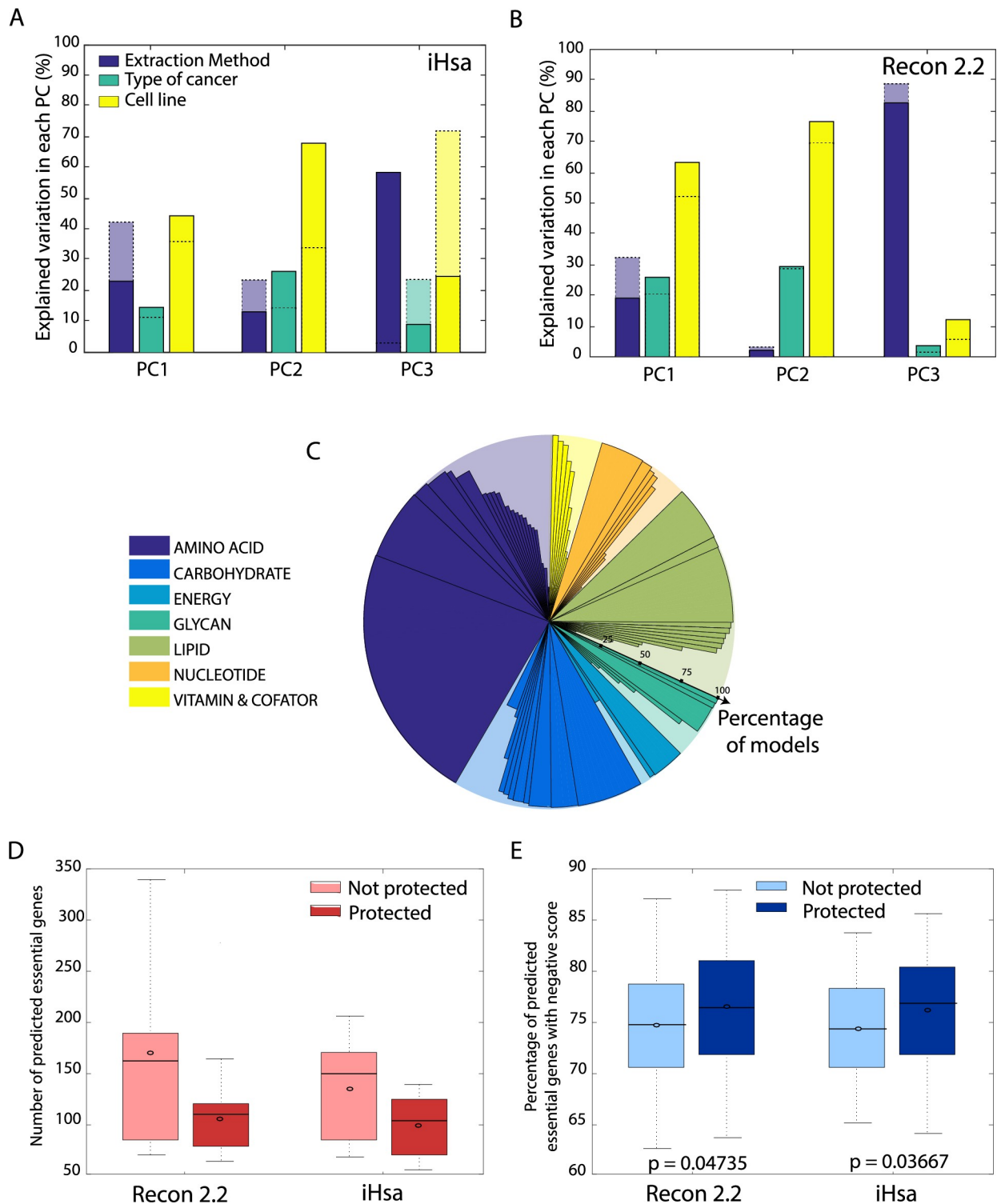
Fig 5. Protection of data-inferred tasks reduces the influence of extraction methods on model content. (A) and (B) After protection of tasks, the influence of the extraction method on the model content at the reaction level is decreased in the 3 first PCs for both reference models. These PCA analyses evaluate the models generated using only the algorithms allowing a significant protection of the inferred metabolic tasks (i.e., fastCORE, mCADRE and MBA). (C) The implementation of the protectionist approach considerably increases the percentage of tasks present in all extracted models. (D) The protection of data-inferred metabolic tasks reduces the number of predicted essential genes (threshold of 90% WT growth) for both reference models. (E) However, task protection increases the percentage of predicted essential genes found in genome-wide CRISPR knock out screens (i.e., genes with a negative score). Significance of increase in percentage was evaluated using the 1-sided Wilcoxon rank-sum test, with the expectation that the protection should improve the prediction.

related to known drug targets (Fig J in S1 Text). We further tested if the models contain cancer hallmark genes [31] and that models built using the protectionist approach increased the proportion of genes in the models that are associated with cancer hallmarks (Fig K in S1 Text; S12 Table).

Discussion

Here we generated hundreds of models for 44 cell lines from the NCI-60 panel using multiple MEMs and two reference GeMs (Recon 2.2 and iHsa) using standard approaches or by protecting metabolic tasks that have been directly inferred from transcriptomic data. We presented a comparative analysis of these two sets of models. As previously observed, the analysis of the first set of extracted models (i.e., models generated without protecting metabolic functions) indicated that the choice of model extraction algorithm significantly influenced the model content at the reaction level [1,2,11]. This leads to considerable variability in context-specific model content, which dwarfed the biological variability across cell lines, otherwise seen in their transcriptomes.

We provided here a curated list of 210 tasks that were used to compare the functionalities of the extracted models. The evaluation of metabolic tasks has emerged as a valuable practice in metabolic modeling studies [12–14,22,32–35]. Such an approach allows one to evaluate the capacity of models to achieve specific modeling goals by capturing known metabolic features. Here we also demonstrated that the approach allows one to objectively compare models that may not share the same structure, such as different reference network reconstructions or models that have been extracted using different methods or parameters. We demonstrated that the selection of a reference model can significantly impact the resulting metabolic functions captured by extracted models, thus possibly impacting the results and interpretations from modeling studies. Indeed, the comparison of the functions of models extracted from both Recon 2.2 and iHsa demonstrated the non-negligible influence of these reference models. We found this is principally due to differences in the GPR annotations in both GeMs. However, these differences in GPR annotations do not considerably influence the inference of metabolic tasks from transcriptomic data. The functional similarity across cell lines captured using data-inferred metabolic tasks is highly consistent between both reference models (Fig L in S1 Text). While community initiatives to standardize the formal representation of GeMs will facilitate cross-comparison between diverse existing GeMs [36], these results highlight the potential of using the inference of functionalities directly from the transcriptome as a way to increase the consensus between extraction methods and reference models.

One challenge in the evaluation of metabolic models is the difficulty of comprehensively defining metabolic functions from a manual search of the literature. Thus, another strength of our approach is that it decreases the need for *a priori* knowledge or assumptions of the metabolic functions that should be included when building a cell or tissue specific model. Therefore, this list of metabolic tasks provides a framework for modelers to develop more physiologically accurate models by inferring the activity of metabolic tasks directly from omics data. Thus, key reactions that need to be included in a model can be protected, without requiring one to know what the cell does. However, the resulting models should still be curated to evaluate expected functionalities, such as for example auxotrophies.

Our protectionist approach can be implemented with diverse model extraction algorithms since it only requires the algorithms to prevent the removal of active metabolic tasks during the extraction process. However, some algorithms will require modifications to ensure the protection of all reactions related to a task. Current implementations of the GIMME-like and iMAT-like families do not favor this type of protection. By minimizing flux through reactions

associated with low gene expression, GIMME-like extraction methods may remove low expression reactions one would want to retain for a validated metabolic task if there are high expression reactions that allow for growth. The iMAT-like methods are similar as they rely on finding an optimal trade-off between removing reactions associated with low gene expression, and keeping reactions whose genes/enzymes are highly expressed. Thus, modified implementations of these algorithms will be needed to allow the protection of reactions based on experimental observations. Finally, this approach can also be extended to any type of network complexity reduction that have been developed in the metabolic modeling field, such as the MILP-based approaches developed to tailor models based on exometabolomic data [37,38].

In our work, we also demonstrated that the models built with the protectionist approach are able to better capture cell-type specific metabolism and accurately predict many essential metabolic genes. Thus, these models may be invaluable for drug development strategies. The emergence of experimental techniques to assess the genetic vulnerabilities of a cell (e.g., CRISPR-Cas9, RNAi) allows researchers to identify sets of genes that should be essential for growth maintenance. These essential genes can further be used to evaluate the capacity of models to represent the interdependence between down-regulation of a gene and the concomitant impairment of growth. Thus, models can be used for interpreting the mechanisms underlying metabolic vulnerabilities that may be invaluable for new drug discoveries. Furthermore, many of the metabolic changes occurring in certain diseases, such as cancer, can be captured by the current list of tasks. Since many of the metabolic tasks are shorter paths, cases where the metabolic flux is redirected due to disease-related metabolic perturbation might be captured by the specific collection of tasks computed using our method. Finally, for the rare cases where a mutation to a specific enzyme leads to a change in the metabolic reaction catalysed by the enzyme change, such as mutations in isocitrate dehydrogenase leading to the production of oncometabolites [39–41]. With such knowledge, researchers are able to define such changes as new metabolic tasks associated with the mutations and incorporate them into their models.

Finally, the list of tasks presented in this study was constructed based on existing repositories. However, a community effort could be undertaken to extend the scope and the definition of these metabolic functions, including the development of tasks seen in plants and microbes and tasks associated with secondary metabolism and microbial gene clusters [42,43]. Furthermore enzymatic mutations leading to new metabolic functions [39] can be systematically defined and added, as currently efforts in the constraint-based modeling community do so on a model by model basis [41]. As these tasks are connected to their associated gene products, this repository of curated tasks would facilitate the description of genome-scale metabolic reconstructions as more than a network of reactions but rather as interconnected maps of cellular functions for diverse organisms. This would be invaluable for the development of algorithms using more relevant biological information and facilitate more comprehensive and accurate descriptions of metabolic adaptations that occur in cells facing a change of context.

Conclusively, context-specific extraction methods are powerful approaches that provide insights in the metabolic state of a cell in specific environments. However, the underlying assumptions used to tailor the GeM based on omics data vary across algorithms, with the consequence that drastically different models can be obtained based on the same data. The poor consensus in generated models may limit the use of context-specific methods for data-driven hypotheses. The definition of metabolic tasks can help with these concerns. Our curated list of tasks and computational framework will allow users to infer metabolic functions directly from transcriptomic data using the whole genome-scale model, and drive the development of improved context specific models. Such models will pave the way toward a better consensus between existing context-specific extraction algorithms, and facilitate the application of models for novel biomedical and engineering applications.

Methods

Preprocessing of gene expression data

RNA-Seq data for the 44 cell lines from the NCI-60 panel were downloaded from [18]. We processed the gene expression data to attribute a gene activity score for each gene and define which genes are active in each cell line. A gene is defined as active in a sample if its expression value is above a threshold defined for this gene within the dataset considered. The threshold of a gene is defined by the mean value of its expression over all the samples coming from the same dataset with exceptions that the threshold needs to be higher or equal the 25th percentile of the overall gene expression value distribution and lower or equal to the 75th percentile. The gene score is computed as follows:

$$Gene\ Score = 5 \cdot \log\left(1 + \frac{Expression\ level}{Threshold}\right)$$

These gene scores are mapped to the models by parsing the GPR rules associated with each reaction. The gene score for each reaction is selected by taking 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 isozyme (OR rule) [44]. Note that we have recently benchmarked the influence of preprocessing methods on the definition of the set of active genes and observed that this parameter combination presented the best performance [45].

Implementation of the MEMs

Model extraction methods (MEMs) employ diverse algorithms to extract cell line- or tissue-specific models from a GeM. The MEMs we have considered can be categorized into three families: "GIMME-like" (i.e., GIMME), "iMAT-like" (i.e., iMAT and INIT) and "MBA-like" (i.e., MBA, FASTCORE, and mCADRE), as proposed previously [2]. The GIMME-like family minimizes flux through reactions associated with low gene expression. The iMAT-like family finds an optimal trade-off between removing reactions associated with low gene expression, and keeping reactions whose genes/enzymes are highly expressed. In the MBA-like family, the algorithms use sets of core reactions that should be retained and active, while removing other reactions if possible. All the algorithms used in this study have been implemented using the function *createTissueSpecificModel* available in the COBRA Toolbox 3.0 [46]. We describe below the list of required parameters needed to run the different methods, all optional parameters have been kept to their default setting.

FASTCORE [5]—The core reactions set (*options.core*) is determined by all the reactions associated to a gene score superior to 5log(2). Note that the biomass reaction was added to the core reactions sets.

GIMME [6]—The implementation of GIMME requires two parameters: the gene scores (*options.expressionRxns*) and a threshold value, the reactions associated with a gene score value below this threshold will be minimized (*options.threshold* = 5log(2)). Note that we manually attributed a gene score of 10log(2) to the biomass reaction to ensure its inclusion.

iMAT [4,47]—Three parameters need to be provided to run iMAT: the gene scores (*options.expressionRxns*), a lower threshold value (reactions with gene score below this value are considered as "non-expressed") and a upper threshold value (reactions with gene score above this value are considered as "expressed"). To simplify the comparison across algorithms, we set both thresholds to the same value: *options.threshold_lb* = *options.threshold_ub* = 5log(2), as done in a previous benchmarking study (1). Note that we manually attributed a gene score of 10log(2) to the biomass reaction to ensure its inclusion.

INIT [7]—The implementation of INIT requires attributing positive weights (*options. weights*) to each reaction with high expression and negative weights for the ones with low expression. All the reactions associated with a gene score below 5log(2) have been assigned a weight of -8 while the weights of remaining reactions were defined as the ratio between the gene score for each reaction and 5log(2). The weight associated with the biomass reaction was put to the maximum of obtained reaction weights.

MBA [8]—The implementation of MBA requires the definition of two set of reactions: high confidence (*options.high_set*) due to their expression and others with medium confidence (*options.medium_set*). The set of reactions with high confidence is defined as reactions with a gene score above the 75th percentile of the distribution of all gene scores and the medium confidence set by all the reactions presenting score above 5log(2) and below the 75th percentile of the distribution of all gene scores. Note that the biomass reaction has been manually added to the high confidence set of reactions.

mCADRE [16]—The implementation of mCADRE requires a score quantifying how often a gene is expressed across samples (*options.ubiquityScore*) and a literature-based evidence score (*options.confidenceScores*). Since the confidence score identification used in the original paper is difficult to transpose in this study, we did not define the confidence score as preformed in the tutorial presenting the implementation of mCADRE in COBRA Toolbox 3.0 (46). Furthermore, as the gene scores are computed based on the knowledge of the gene expression of a gene across all samples, we used the gene scores as ubiquity scores.

Curation of metabolic tasks

The curation has been done by first taking the union of previously published lists of metabolic tasks [13,14]. We removed duplicated tasks and lumped tasks that rely on the description of similar metabolic functions. Each remaining task without strong biological evidence was removed. We also created 9 new tasks that were essential for the acquisition of already described metabolic functions (i.e., intermediate biosynthetic steps for the acquisition of other tasks). Doing so, we obtained a collection of 210 tasks associated with 7 systems (energy, nucleotide, carbohydrates, amino acid, lipid, vitamin & cofactor and glycan metabolism). For each task, we provided its original source (Recon and/or iHsa) and comments on the biological evidence of this metabolic function (S4 Table).

A unified framework for computing metabolic tasks as a model benchmarking tool

In its original version, Thiele and coworkers (2013) [14] define a

"*metabolic task as a nonzero flux through a reaction or through a pathway leading to the production of a metabolite B from a metabolite A. The metabolic capacity of the network was demonstrated by testing nonzero flux values for these metabolic tasks. For each of the simulations, a steady-state flux distribution was calculated. Each metabolic task was optimized individually by choosing the corresponding reaction in the model, if present, as objective function and maximized the flux through the reaction*".

In parallel, Agren and coworkers presented an alternative framework to compute the metabolic tasks present in a model within their RAVEN toolbox [48]. They defined a metabolic task through a list of inputs and outputs for which the pseudo-stationary assumption will be relaxed following a magnitude imposed by the user and assumed that a task successfully passes if the variation imposed to the inputs leads to the imposed variation of the outputs.

We also propose to define a metabolic task as the capacity of producing a defined list of output products when only a defined list of input substrates is available. However, we modified the way to implement it from the RAVEN toolbox. Instead of relying on the relaxation of the steady-state assumption, we take an approach more similar to that proposed by [14] by imposing constraints only at the flux level. Therefore, a model successfully passes a task if the associated LP problem is still solvable when the sole exchange reactions allowed carrying flux in the model are temporary sink reactions associated with each of the inputs and outputs listed in the task. This framework allows the use of known stoichiometry to fix the ratio between the fluxes of the sink reactions associated with each input and output of the task. We implemented the code to compute the tasks in Matlab, and the code, *checkMetabolicTasks*, has been contributed to the COBRA Toolbox3.0 [46].

Validation on existing animal genome-scale models

We tested the list of tasks using published genome-scale models of human [13,14,17,22,23], Chinese hamster [9], rat [13] and mouse [24] cells (Fig 2D, S5 Table). All models successfully pass more than 90% of the tasks. For each failed task, we provided a reason of the failure (i.e. definition of the missing reaction to successfully pass the task) (S5 Table). As the definition of the metabolic tasks depends on the provision of the exact name of the metabolites in each model, we also provide a table of nomenclature compatibility between the different genome-scale models tested (S6 Table).

Inference of metabolic tasks from transcriptomic data

We developed a computational framework for attributing a score to each metabolic task in order to extend the application of the concept beyond the model benchmarking scope. If a task successfully passes in a model, one can compute the list of reactions associated with this task and, in doing so, access the list of genes that may contribute to the acquisition of this metabolic function based on the GPR rules. To this end, we used the parsimonious Flux Balance Analysis (pFBA) algorithm to define the set of reactions and associated genes required to pass a task within a specified model [49]. Thanks to the availability of this information, metabolic functions can now be directly assessed from transcriptomic data. The proposed computation of a metabolic score relies first on the preprocessing of the available transcriptomic data and the attribution of a gene activity score for each gene (see associated Methods section). We further used the GPR rules associated with each reaction required for a task to decide which gene will be the main determinant of the enzyme abundance associated with this reaction and attribute the corresponding gene activity level (i.e., selection of the *minimum* expression value among all the genes associated to an enzyme complex (AND rule) and the *maximum* expression value among all genes associated with an isoenzyme (OR rule)). Therefore, each reaction involved in a task is associated with a reaction activity level (RAL) that corresponds to the preprocessed gene expression value of the gene selected as the main determinant for this reaction. Finally, the metabolic score can be computed as the mean of the activity level of each reaction:

$$\text{MT score} = \text{sum(RAL)}/\text{number of reactions involved in the task}$$

Doing so, a metabolic task will be considered as active if its MT score has a value greater than 5log(2). The list of active metabolic tasks for each of the 44 cell lines from the NCI-60 panel is available in S7 Table.

Protection of data-inferred task during extraction process

We used the list of active metabolic tasks (S7 Table) to determine the set of reactions that should be protected during the extraction process for each of the 44 cell lines. The protectionist

approach has been implemented for each extraction method by using the same set of parameters as previously described with the following modification:

FASTCORE—The set of reactions associated with the metabolic tasks defined as active based on the transcriptomic data has been manually added to the core reactions set (*options. core*).

GIMME & iMAT—A gene score of 10log(2) (*options.expressionRxns*) has been attributed to all the reactions associated to the metabolic tasks defined as active based on the transcriptomic data.

INIT—The weights (*options.weights*) for all reactions associated with the metabolic tasks defined as active based on the transcriptomic data were put to the maximum of obtained reaction weights.

MBA—The reactions associated with the metabolic tasks defined as active based on the transcriptomic data have been manually added to the high confidence set of reactions.

mCADRE—A ubiquity score (*options.ubiquityScore*) of 1 has been attributed to all the reactions associated to the metabolic tasks defined as active based on the transcriptomic data.

Principal component analysis

For the reaction PCAs, a binary matrix is constructed in which each row represents an extracted model and each column represents a reaction, with each element representing the presence (1) or absence (0) of a reaction in a model. Reactions in all or no models were removed from the matrix. Similarly for the metabolic function PCA, the matrix had each row as an extracted model and each column as a metabolic task, with each element in the matrix representing if the task is present (1) or absent (0) in a model. For the PCAs, the matrix was centered to have zero mean within each row. PCA was done on this matrix. The variance explained by the different factors (MEM, cancer type and cell line) within each of the principal components is calculated as follows. Within one factor, the maximum Pearson correlation coefficient (R) of the component scores and categories is calculated across all possible orderings of the categories. Reported is the R^2 scaled to percentages. The same procedure was used to perform the PCA on the model functionalities except that the binary matrix of reactions was replaced by the binary matrix representing the list of metabolic tasks that are successfully passed in each extracted model. The attributes of all extracted models (number of reactions and metabolites, number of successfully passed tasks and predicted growth rate) are available in S8 Table and the results of the extracted model benchmarking using the list of metabolic tasks is available in S9 Table.

Predictions of gene-essentiality

To predict gene-essentiality, FBA was used to optimize biomass production following the removal of each reaction in the cell line-specific models that would be affected by gene removal based on the GPRs. The function used to perform this deletion analysis is available in COBRA Toolbox 3.0, *singleGeneDeletion.m* [46]. To test these essentiality predictions of the models against experimental data, we downloaded CRISPR-Cas9 loss-of-function screens data for 20 NCI-60 cell lines from depmap.org [25–27]. In these screens, essential genes are identified based on genes scores attributed using single guide RNA (sgRNA) abundance for each knock-out before and after growth selection. A more negative gene score suggests a higher probability that the gene is essential. Therefore, the agreement between prediction and data can be analyzed by using the percentage of predicted essential genes that have a negative gene score [28]. A 1-tailed Wilcoxon rank sum test was used to test whether the percentage of predicted essential genes of the model extracted using the protectionist approach were significantly higher

than the ones without protection. The results of the gene deletion study and prediction against CRISPR-Cas9 loss-of-function screens are available in S10 Table.

Supporting information

S1 Text. Supplementary materials.
(DOCX)

S1 Table. Description of the 44 different cell lines from the NCI-60 panel.
(XLSX)

S2 Table. Uptake and secretion rates constraints used in input GeMs.
(XLSX)

S3 Table. Biomass function and experimentally measured growth rate of the cell lines.
(XLSX)

S4 Table. Description of the list of metabolic tasks.
(XLSX)

S5 Table. Validation of the metabolic tasks list on existing animal genome-scale models.
(XLSX)

S6 Table. Table of nomenclature compatibility between the different GeMs tested.
(XLSX)

S7 Table. List of active metabolic tasks for each of the 44 cell lines from the NCI-60 panel.
(XLSX)

S8 Table. Attributes of all extracted cell line-specific models.
(XLSX)

S9 Table. Benchmarking of extracted cell line-specific models using the list of metabolic tasks.
(XLSX)

S10 Table. Results of the gene deletion study and prediction against CRISPR-Cas9 loss-of-function screens.
(XLSX)

S11 Table. Collection of known anti-cancer drug targets.
(XLSX)

S12 Table. Cancer hall mark gene sets.
(XLSX)

Author Contributions

Conceptualization: Anne Richelle, Nathan E. Lewis.

Data curation: Anne Richelle, Austin W. T. Chiang, Chih-Chung Kuo, Nathan E. Lewis.

Formal analysis: Anne Richelle, Nathan E. Lewis.

Funding acquisition: Nathan E. Lewis.

Investigation: Anne Richelle, Nathan E. Lewis.

Methodology: Anne Richelle, Nathan E. Lewis.

Project administration: Anne Richelle, Nathan E. Lewis.

Resources: Anne Richelle, Nathan E. Lewis.

Software: Anne Richelle, Nathan E. Lewis.

Supervision: Anne Richelle, Nathan E. Lewis.

Validation: Anne Richelle, Nathan E. Lewis.

Visualization: Anne Richelle, Nathan E. Lewis.

Writing – original draft: Anne Richelle, Nathan E. Lewis.

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

References

1. Opdam S, Richelle A, Kellman B, Li S, Zielinski DC, Lewis NE. A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. Cell Syst. 2017; 4(3).

2. Robaina Estévez S, Nikoloski Z. Generalized framework for context-specific metabolic model extraction methods. Front Plant Sci. 2014; 5(September):1–11.

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

4. Zur H, Ruppin E, Shlomi T. iMAT: An integrative metabolic analysis tool. Bioinformatics. 2010; 26 (24):3140–2. https://doi.org/10.1093/bioinformatics/btq602 PMID: 21081510

5. Vlassis N, Pacheco M, Sauter T. Fast Reconstruction of Compact Context-Specific Metabolic Network Models. arXiv Prepr arXiv13047992 [Internet]. 2013;1–22. Available from: http://arxiv.org/abs/1304.7992

6. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008; 4(5).

7. Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS Comput Biol. 2012; 8(5).

8. Jerby L, Shlomi T, Ruppin E. Computational reconstruction of tissue-specific metabolic models: Application to human liver metabolism. Mol Syst Biol. 2010; 6.

9. Hefzi H, Ang KS, Hanscho M, Bordbar A, Ruckerbauer D, Lakshmanan M, et al. A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. Cell Syst. 2016; 3(5):434–443.e8. https://doi.org/10.1016/j.cels.2016.10.020 PMID: 27883890

10. Kumar A, Harrelson T, Lewis NE, Gallagher EJ, LeRoith D, Shiloach J, et al. Multi-tissue computational modeling analyzes pathophysiology of type 2 diabetes in MKR mice. PLoS One. 2014; 9(7).

11. Ferreira J, Correia S, Rocha M. Analysing Algorithms and Data Sources for the Tissue-Specific Reconstruction of Liver Healthy and Cancer Cells. Interdiscip Sci Comput Life Sci. 2017; 9(1):36–45.

12. Agren R, Mardinoglu A, Asplund A, Kampf C, Uhlen M, Nielsen J. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Mol Syst Biol. 2014; 10(3).

13. Blais EM, Rawls KD, Dougherty B V., Li ZI, Kolling GL, Ye P, et al. Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions. Nat Commun. 2017; 8.

14. Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 2013; 31(5):419–25. https://doi.org/10.1038/nbt.2488 PMID: 23455439

15. Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, Hanscho M, et al. Recon 2.2: from reconstruction to model of human metabolism. Metabolomics. 2016; 12(7).

16. Wang Y, Eddy J a, Price ND. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Syst Biol. 2012; 6:153. https://doi.org/10.1186/1752-0509-6-153 PMID: 23234303

17. 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 Jul; 12(7):109.

18. 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(3):306–12. https://doi.org/10.1038/nbt.3080 PMID: 25485619

19. 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. 2012 May; 336(6084):1040–4. https://doi.org/10.1126/science.1218595 PMID: 22628656

20. Zielinski DC, Jamshidi N, Corbett AJ, Bordbar A, Thomas A, Palsson BO. Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism. Sci Rep. 2017; 7.

21. Pacheco MP, Pfau T, Sauter T. Benchmarking procedures for high-throughput context specific reconstruction algorithms. Front Physiol. 2016; 6(JAN).

22. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A. 2007 Feb; 104 (6):1777–82. https://doi.org/10.1073/pnas.0610772104 PMID: 17267599

23. Quek LE, Dietmair S, Hanscho M, Martínez VS, Borth N, Nielsen LK. Reducing Recon 2 for steady-state flux analysis of HEK cell culture. J Biotechnol. 2014; 184:172–8. https://doi.org/10.1016/j.jbiotec.2014.05.021 PMID: 24907410

24. Sigurdsson MI, Jamshidi N, Steingrimsson E, Thiele I, Palsson BT. A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. BMC Syst Biol. 2010; 4.

25. 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(2):184–91. https://doi.org/10.1038/nbt.3437 PMID: 26780180

26. 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(12):1779–84. https://doi.org/10.1038/ng.3984 PMID: 29083409

27. Aguirre AJ, Meyers RM, Weir BA, Vazquez F, Zhang CZ, Ben-David U, et al. Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. Cancer Discov. 2016; 6(8):914–29. https://doi.org/10.1158/2159-8290.CD-16-0154 PMID: 27260156

28. Tobalina L, Pey J, Rezola A, Planes FJ. Assessment of FBA Based Gene Essentiality Analysis in Cancer with a Fast Context-Specific Network Reconstruction Method. PLoS One. 2016; 11(5):e0154583. https://doi.org/10.1371/journal.pone.0154583 PMID: 27145226

29. Gatto F, Miess H, Schulze A, Nielsen J. Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. Sci Rep. 2015; 5.

30. Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, Shlomi T. Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol. 2011; 7.

31. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 2015; 1(6):417–25. https://doi.org/10.1016/j.cels.2015.12.004 PMID: 26771021

32. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science (80-) [Internet]. 2015; 347(6220):1260419–1260419. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1260419

33. Chubukov V, Gerosa L, Kochanowski K, Sauer U. Coordination of microbial metabolism. Vol. 12, Nature Reviews Microbiology. 2014. p. 327–40. https://doi.org/10.1038/nrmicro3238 PMID: 24658329

34. Bordbar A, Mo ML, Nakayasu ES, Schrimpe-Rutledge AC, Kim YM, Metz TO, et al. Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation. Mol Syst Biol. 2012; 8.

35. Correia S, Rocha M. A critical evaluation of methods for the reconstruction of tissue-specific models. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015. p. 340–52.

36. Lieven C, Beber ME, Olivier BG, Bergmann FT, Babaei P, Bartell JA, et al. Memote: A community-driven effort towards a standardized genome-scale metabolic model test suite. bioRxiv [Internet]. 2018 Jan 1; Available from: http://biorxiv.org/content/early/2018/06/21/350991.abstract

37. Röhl A, Bockmayr A. A mixed-integer linear programming approach to the reduction of genome-scale metabolic networks. BMC Bioinformatics. 2017; 18(1).

38. Erdrich P, Steuer R, Klamt S. An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. BMC Syst Biol. 2015; 9(1).

39. Dang L, White DW, Gross S, Bennett BD, Bittinger MA, Driggers EM, et al. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature. 2009; 462(7274):739–44. https://doi.org/10.1038/nature08617 PMID: 19935646

40. Intlekofer AM, Wang B, Liu H, Shah H, Carmona-Fontaine C, Rustenburg AS, et al. L-2-Hydroxygluta-rate production arises from noncanonical enzyme function at acidic pH. Nat Chem Biol. 2017; 13 (5):494–500. https://doi.org/10.1038/nchembio.2307 PMID: 28263965

41. Nam H, Campodonico M, Bordbar A, Hyduke DR, Kim S, Zielinski DC, et al. A Systems Approach to Predict Oncometabolites via Context-Specific Genome-Scale Metabolic Networks. PLoS Comput Biol. 2014; 10(9).

42. Chu HY, Wegel E, Osbourn A. From hormones to secondary metabolism: The emergence of metabolic gene clusters in plants. Plant J. 2011; 66(1):66–79. https://doi.org/10.1111/j.1365-313X.2011.04503.x PMID: 21443624

43. Craney A, Ahmed S, Nodwell J. Towards a new science of secondary metabolism. Vol. 66, Journal of Antibiotics. 2013. p. 387–400. https://doi.org/10.1038/ja.2013.25 PMID: 23612726

44. Jensen PA, Lutz KA, Papin JA. TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. BMC Syst Biol. 2011;5. https://doi.org/10.1186/1752-0509-5-5

45. Richelle A, Joshi C, Lewis NE. Assessing key decisions for transcriptomic data integration in biochemical networks. bioRxiv [Internet]. 2018;301945. Available from: https://www.biorxiv.org/content/early/2018/04/16/301945

46. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0. ArXiV. 2017;1710.04038.

47. Shlomi T, Cabili MN, Herrgård MJ, Palsson B, Ruppin E. Network-based prediction of human tissue-specific metabolism. Vol. 26, Nature Biotechnology. 2008. p. 1003–10. https://doi.org/10.1038/nbt.1487 PMID: 18711341

48. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. PLoS Comput Biol. 2013 Jan; 9(3):e1002980. https://doi.org/10.1371/journal.pcbi.1002980 PMID: 23555215

49. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010; 6.