Network-based prediction of human tissue-specific metabolism

Tomer Shlomi[1,4], Moran N Cabili[1,4], Markus J Herrgård[2], Bernhard Ø Palsson[2] & Eytan Ruppin[1,3]

Direct *in vivo* investigation of mammalian metabolism is complicated by the distinct metabolic functions of different tissues. We present a computational method that successfully describes the tissue specificity of human metabolism on a large scale. By integrating tissue-specific gene- and protein-expression data with an existing comprehensive reconstruction of the global human metabolic network, we predict tissue-specific metabolic activity in ten human tissues. This reveals a central role for post-transcriptional regulation in shaping tissue-specific metabolic activity profiles. The predicted tissue specificity of genes responsible for metabolic diseases and tissue-specific differences in metabolite exchange with biofluids extend markedly beyond tissue-specific differences manifest in enzyme-expression data, and are validated by large-scale mining of tissue-specificity data. Our results establish a computational basis for the genome-wide study of normal and abnormal human metabolism in a tissue-specific manner.

Metabolic network modeling of biological systems involves the analysis and prediction of metabolic flux distributions under diverse physiological and genetic conditions. Traditional modeling techniques are based on mathematical approaches that require detailed information on kinetics and on enzyme and metabolite concentrations[1,2]. However, a lack of accurate information of kinetic constants and enzyme and metabolite intracellular concentrations limits the current applicability of such methods to small-scale systems. Constraint-based modeling bypasses this hurdle by analyzing the function of large-scale metabolic networks through relying solely on simple physical-chemical constraints[3]. In recent years, constraint-based modeling has been frequently used to successfully predict various phenotypes of microorganisms, such as their growth rates, rates of nutrient uptake, by-product secretion and the lethality of gene knockouts (see ref. 4 for review).

Despite this progress in applying constraint-based modeling to studying the metabolism of microorganisms, large-scale modeling of human metabolism is still in its infancy. Nonetheless, the emergence of metabolic diseases such as diabetes and obesity as major sources of morbidity and mortality[5,6] has stimulated research into human metabolism and its regulation. Metabolic enzymes and their regulators are increasingly considered viable drug targets for these and other conditions[7,8]. However, in reconstructing human metabolic networks, most of the previous work has focused on characterizing distinct metabolic pathways[9,10]. Until recently, reconstructions of large-scale human metabolic networks had been performed only for specific cell types and organelles[11–13]. Although fundamental steps forward, reconstructions of the global human metabolic network based on an extensive evaluation of genomic and bibliomic data (that is, comprehensive assessment of the literature)[14,15] are not tissue specific. In adapting constraint-based modeling methods from the realm of microorganisms to that of multicellular organisms, one encounters two main hurdles. The first is that different tissues have different metabolic objectives that are not well characterized and remain largely unknown. This is in contrast to modeling microorganisms where a simple objective function (such as maximizing the biomass production rate) can be used together with flux balance analysis[4] to predict biologically plausible flux distributions. The second major obstacle is the lack of information on tissue-specific metabolite uptake and secretion, which is essential for employing flux balance analysis.

We present a new constraint-based computational method for systematically predicting human tissue-specific metabolic behavior by integrating a genome-scale metabolic network with tissue-specific gene- and protein-expression data. Changes in gene- and protein-expression levels play a major role in controlling tissue-specific metabolic functions[16–18], and a strong correlation between gene expression and measured[19,20] and predicted[21–24] metabolic fluxes is reported for microorganisms. To account for metabolic flux activity that is not reflected in the expression data (that is, post-transcriptional regulatory effects), we treat tissue-specific variations in enzyme-expression levels not as the final determinants of enzyme activity, but as cues for the likelihood that the enzyme in question supports metabolic flux in its associated reaction(s). Network integration is then used to accumulate these cues into a global, consistent metabolic behavior, which reflects the outcome of putative post-transcriptional regulatory effects. Our method's reliance on enzyme-expression data to infer tissue-specific metabolic flux eliminates the need for a priori knowledge of tissue-specific objective functions and metabolites exchanged by the tissue with biofluids. Instead, the method provides predictions regarding tissue-specific metabolite uptake and secretion.

To examine our method's ability to correctly predict metabolic behavior based on gene-expression data, we first apply it to predicting the metabolic state of the yeast *Saccharomyces cerevisiae* under

[1]School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. [2]Department of Bioengineering, University of California, San Diego, La Jolla, California 92093-0412, USA. [3]School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel. [4]These authors contributed equally to this work. Correspondence should be addressed to T.S. (shlomito@post.tau.ac.il) or E.R. (ruppin@post.tau.ac.il).

Published online 17 August 2008; doi:10.1038/nbt.1487

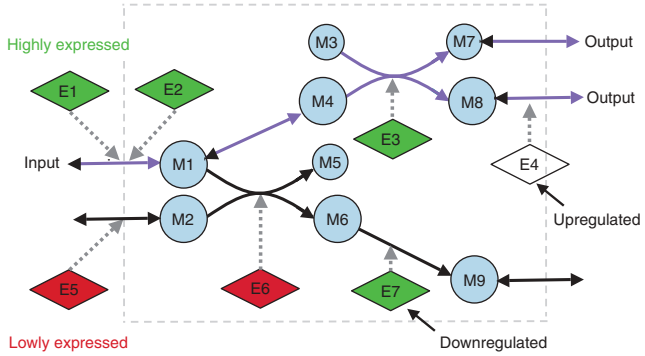


Figure 1 An example of predicting flux-activity states of genes based on a metabolic network model and gene-expression measurements. Circular nodes represent metabolites, whereas diamond nodes represent enzymes. White, red and green represent normal, significantly low and significantly high expression of the enzyme-encoding genes, respectively. Solid edges represent metabolic reactions. Broken edges associate enzymes with the reactions they catalyze. The predicted steady-state flux distribution, involving the activation of reactions, is shown as purple arrows. Enzyme E4 is predicted to be post-transcriptionally upregulated and E7 is predicted to be post-transcriptionally downregulated.

conditions for which reliable data are readily available for validation. We then apply it to a genome-scale human metabolic network model[14] that we have integrated with tissue-specific enzyme-expression data to predict tissue-specific metabolic behavior of ten human tissues. Specifically, for each tissue, we obtain a unique view of metabolic activity that includes information on the predicted exchange of metabolites with surrounding biofluids. The model's tissue-specific activity predictions are validated based on a comprehensive comparison to known large-scale information on tissue specificity of genes, reactions and metabolites obtained from various databases, and by studying the tissue specificity of genes that cause metabolic disease. The predicted tissue-specific metabolic behaviors, made available through our website (http://www.cs.tau.ac.il/~shlomito/tissue-net), should provide valuable reference sources for studying human metabolism under normal and disrupted physiological conditions.

RESULTS

Network-based prediction of metabolic behavior

To account for levels of post-translational regulation that are not reflected in the gene- and protein-expression data we use, we treat the expression levels of enzymes merely as cues for the likelihood that their associated reactions carry metabolic flux. We use network integration to accumulate these cues into a global, consistent prediction of metabolic behavior. To this end, we employ a discrete representation of significantly high or low enzyme-expression levels across tissues, following previous usage of such data in metabolic modeling for other purposes[25,26]. Network integration is then done by solving a constraint-based modeling optimization problem to find a steady-state metabolic flux distribution (that is, an assignment of fluxes to all the reactions in the network) that, first, satisfies the stoichiometric and thermodynamic constraints embedded in the model and, second, maximizes the number of enzymes whose predicted flux activity is consistent with their measured expression level. In other words, the method aims to obtain a flux distribution where the number of flux-carrying reactions associated with highly expressed enzymes is maximized, and the number of flux-carrying reactions associated with lowly expressed genes is minimized. The resulting predicted flux distribution is used to assign flux activity states to the genes, reflecting the presence and/or absence of nonzero flux through the enzymatic reactions they are associated with. For some of the genes, their flux activity state can be uniquely determined to be active or inactive, with associated confidence estimations. For others, their activity state cannot be uniquely determined because of potential alternative flux distributions with the same overall consistency with the expression data (mostly owing to isozymes or alternative pathways; **Supplementary Results**, section 1, and **Supplementary Fig. 1** online)[22,26,27]. Because expression levels are not enforced as exclusive determinants of metabolic flux, the flux activity states of genes may deviate from their expression states. Genes are considered to be post-transcriptionally up- or downregulated based on a difference between their measured expression level and their predicted flux activity state in a given tissue.

An example of flux activity–state predictions obtained with the above method is shown in **Figure 1**. The method predicts a flux distribution that is consistent with the expression state of five of the six genes expressed at significantly high or low levels relative to some reference state. Based on the flux predictions, two enzymes are predicted to be post-transcriptionally regulated. As the highly expressed enzyme E7 is predicted not to support metabolic flux, it is hence considered to be downregulated. As the moderately expressed membrane transporter E4 is predicted to support metabolic flux, it is hence considered to be upregulated. Predicted fluxes through specific exchange reactions that cross the system boundaries represent the uptake and secretion of metabolites from the tissue. Of the five metabolites that can be exchanged with the tissue's surroundings (M1-2, M7-9), the method predicts the uptake of one metabolite (M1) and the secretion of two others (M7 and M8). Notably, the high-expression level of the membrane transporter of M1 indicates that it may be active, but it does not provide information regarding whether M1 is taken up or secreted from the tissue. In contrast, the integrated approach can determine the direction of flux for many reactions by propagating the known thermodynamic constraints on reaction reversibility and directionality throughout the network. In the current example, the direction of the activated pathway is inferred based on the irreversibility of enzymes E3 and E4.

As a basic validation of our method, we applied it to predict the metabolic behavior of the yeast *S. cerevisiae* based on gene-expression data measured in various growth media[20] (**Supplementary Results**, section 2, and **Supplementary Fig. 2** online). Comparing the predicted metabolic activity to measured metabolic fluxes in the central carbon metabolism of *S. cerevisiae*[20] revealed a significant correlation (hypergeometric P-value = 0.01) with a precision of 0.71 and recall of 0.89 (**Supplementary Fig. 3** online). As a further larger-scale validation, the predicted flux activity pattern was found to significantly correlate with flux balance analysis predictions (based on explicit yeast biomass maximization) with a mean precision of 0.89 and recall of 0.79 across growth media (mean hypergeometric P-value = 5.6×10^{-8}; **Supplementary Fig. 4** online). For comparison, the predicted flux activity obtained solely from expression data (that is, only for the subset of the reactions associated with genes in the model) has a markedly lower accuracy, with a precision of 0.83 and recall of 0.61. Moreover, the predicted flux activity pattern correctly captures the directionality of metabolite exchange with the growth media, and the production of essential biomass precursors, in agreement with predictions made using flux balance analysis (**Supplementary Figs. 5** and **6** online). This is notable because our method does not use information on biomass composition or metabolite exchange.

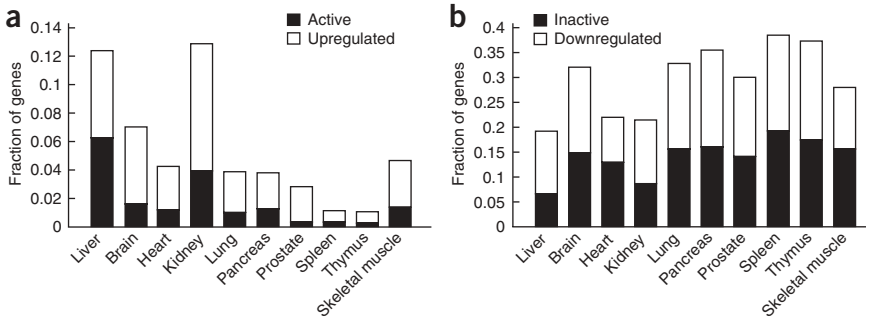


Figure 2 The fraction of all metabolic genes in the model predicted to be active and inactive in ten different tissues. (**a**) The fraction of highly expressed and metabolically active genes is shown in black and the fraction of post-transcriptionally activated genes (that is, genes that are not highly expressed but predicted to be active) in white. (**b**) The fraction of lowly expressed and metabolically inactive genes is shown in black, and the fraction of post-transcriptionally downregulated genes (that is, genes that are not lowly expressed but are predicted to be inactive) in white.

Predicting human tissue-specific metabolism

We computed tissue-specific behavior based on the metabolic network model of Duarte *et al.*[14] and sets of gene-[28] and protein-[29]expression measurements in ten tissues: brain, heart, kidney, liver, lung, pancreas, prostate, spleen, skeletal muscle and thymus (**Supplementary Data Set 1** online). The activity states of 644 of the model genes were uniquely determined (i.e., scored as either active or inactive) in at least one tissue, with an average of 408 genes with a determined activity state per tissue (**Supplementary Data Set 2** online). The activity states of the remaining genes in the model remained undetermined either because of errors in the model (in the form of reactions that reside on dead-end pathways) or because of the existence of alternative flux distributions (mostly resulting from isozymes and alternative pathways). Many of the genes predicted to be active in a certain tissue are not highly expressed there, and, conversely, many of the genes predicted to be inactive are not lowly expressed; this shows the considerable amount of additional information obtained by integrating expression data with the metabolic network to infer metabolic gene activity (**Fig. 2a,b**). Our approach predicts an average of 42 genes (that is, 3.6% of the analyzed genes) to be post-transcriptionally upregulated and 180 (15.4%) genes to be post-transcriptionally downregulated in each tissue. This points to an interesting asymmetric effect of post-transcriptional regulation on metabolic flux activity. The predicted activity states were found to be robust to noise in the expression data, with noise in the expression state of 15% of the genes causing up to only 8% of the predicted activity states to be in error (**Supplementary Table 1** online). To assess the accuracy of the model predictions, we performed a fivefold cross-validation test in which the gene activities for a set of 20% of the genes were predicted, while using the expression of the remaining 80% to constrain the model gene activity. The overlap between the genes predicted as active and the highly expressed genes in the held-out data was significant for all tissues (**Supplementary Fig. 7** online).

To systematically validate the predicted tissue-specific metabolic behavior, we collected data on tissue specificity of metabolites from the Human Metabolome Database (HMDB)[30] and biochemical reactions from the Braunschweig Enzyme Database (BRENDA)[31] (**Table 1** and **Supplementary Data Set 3** online). We also obtained data on tissue-specific gene activity by searching the worldwide web for co-occurrences of genes and tissues in the titles of research papers (**Table 1** and **Supplementary Data Set 3**). The predicted tissue-specificity of genes, reactions and metabolites was significantly correlated with all three data sources (**Table 1**, **Supplementary Table 2** and **Supplementary Results**, section 3 online). Focusing on tissue specificity findings of genes, reactions and metabolites that are inferred by post-transcriptional upregulation (that is, tissue-specificity findings that cannot be inferred by using the original expression data without the metabolic network) provides a significant high number of matches with the known tissue-specific associations in all cases (**Table 1**). Analogously, in the inverse direction, tissue specificity inferred by post-transcriptional downregulation entails a significantly low number of matches with the known tissue-associations (**Table 1**), as one would expect.

As further validation, we classified the reactions to metabolic subsystems (as defined in the original metabolic network model[14], primarily by using the Kyoto Encyclopedia of Genes and Genomes LIGAND database) and surveyed predicted subsystem-tissue associations arising from post-transcriptional regulation of enzymes (**Supplementary Fig. 8** online). Many such associations are consistent with known tissue functions and co-occur on the web in a statistically significant manner (hypergeometric P-value $= 3.6 \times 10^{-4}$). For

Table 1 Accuracy of tissue-specificity predictions of genes, reactions and metabolites

		Global accuracy			Upregulation accuracy			Downregulation accuracy		
Category	Validation data source	Pre.	Rec.	P-value	Pre.	Rec.	P-value	Pre.	Rec.	P-value
All genes	WEB	0.37	0.37	$2.6{\cdot}10^{-9}$	0.33	0.18	$1.6{\cdot}10^{-8}$	0.82	0.24	$2.2{\cdot}10^{-3}$
Disease genes	OMIM[37]	0.49	0.55	$<10^{-300}$	0.47	0.50	$<10^{-300}$	0.85	0.22	$3.6{\cdot}10^{-4}$
Transporter Genes	HMTD[34], TCDB[35]	0.57	0.41	$6.1{\cdot}10^{-7}$	0.64	0.21	0.06	0.8	0.25	0.03
Enzymatic reactions	BRENDA[31]	0.7	0.42	$4.{\cdot}10^{-12}$	0.55	0.21	$2.6{\cdot}10^{-12}$	0.68	0.25	$3.7{\cdot}10^{-25}$
All metabolites	HMDB[30]	0.36	0.47	$<10^{-300}$	0.32	0.32	$7.4{\cdot}10^{-8}$	0.81	0.21	$4.2{\cdot}10^{-7}$
Exchange metabolites	HMDB[30]	0.36	0.38	$3.2{\cdot}10^{-3}$	0.33	0.25	0.06	0.8	0.2	0.01

The accuracy P-value reflects the overlap between the predicted tissue-associations of genes, reactions and metabolites (the different categories, each composing a separate row in the Table) and known tissue-associations derived from various data sources (assuming a hypergeometric distribution of random tissue-associations as the background model). The prediction accuracy is computed over all genes that are predicted to be active in at least a single tissue (see **Supplementary Results**, section 3, for details regarding the choice of validation gene set). The global accuracy column refers to the set of all tissue associations (both up- and downregulated) predicted by the model. The prediction accuracy for the subsets of elements that are inferred based on post-transcriptional up- or downregulation is displayed in the next columns, for each category (row). The complete sets of known and predicted tissue-specific activities of genes, reactions and metabolites are available in **Supplementary Data Sets 2** and **3**. Pre., precision; Rec., recall.

example, the model predicted the post-transcriptional upregulation of the genes *HSD17B4* (hydroxysteroid (17-beta) dehydrogenase 4) and *SCP2* (peroxisomal thiolase 2, a sterol carrier protein) in the liver, in accordance with their known involvement in bile-acid biosynthesis[9]. In another large-scale validation of our predictions, we draw on previous studies that have shown that genes expressed in a small number of tissues tend to have higher evolutionary rates[32,33]. Repeating a similar analysis, we found that genes predicted to be active by the integrated model in a small number of tissues indeed have significantly higher evolutionary rates (Wilcoxon test P-value $= 6 \times 10^{-4}$). In contrast, the subset of metabolic genes whose tissue specificity is determined solely by expression does not manifest significantly higher evolutionary rates for genes expressed in only a small number of tissues than those expressed in a large number of tissues.

Tissue-specific uptake and secretion of metabolites

Tissue-specific metabolic behavior can be further characterized by identifying the metabolites that a particular tissue exchanges with biofluids. Relying solely on gene or protein expression for this task is problematic as only 46% (115/249) of the metabolites that are known to be secreted or taken up by human tissues are associated with known membrane transporters[14]. Moreover, even for these metabolites, expression data do not reflect the direction of transport and they do not account for potential post-transcriptional regulation of the transporter. Based on the model's flux activity predictions, we derived a global map of secretion and uptake of 249 metabolites across different tissues (**Supplementary Fig. 9** and **Supplementary Data Set 2** online). As expected, the major metabolic organs, the liver and kidney, exchange the highest number of metabolites with biofluids (81 and 71, respectively).

The predicted tissue-specific metabolite exchange that depends on membrane transporters was validated based on data on tissue specificity of transporters, obtained from the Human Membrane Transporter Database[34] (HMTD) and from the Transport Classification Database[35] (TCDB) (**Table 1** and **Supplementary Data Set 3**). In many of these cases, such as uptake of norepinephrine by the heart[36],

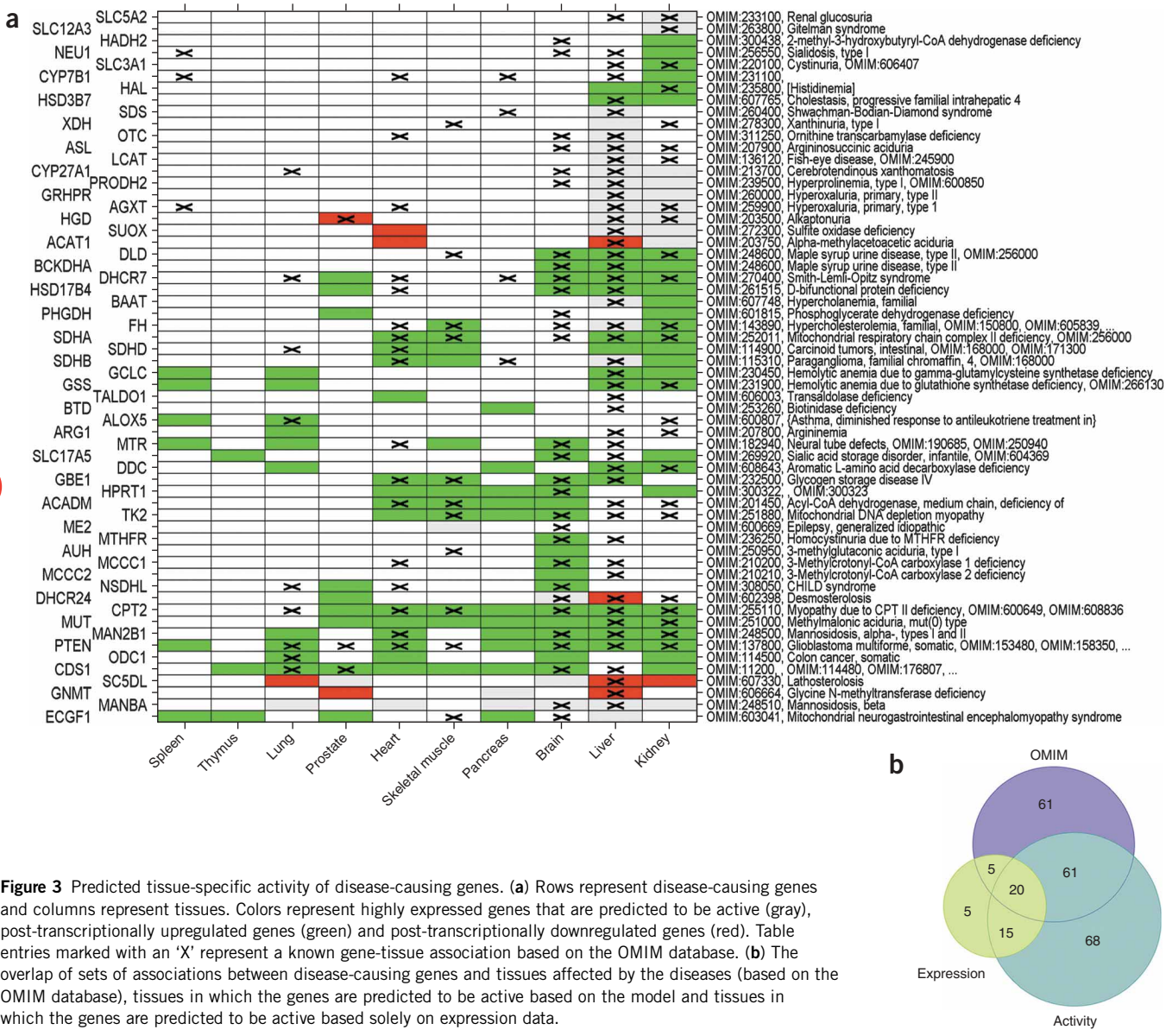


Figure 3 Predicted tissue-specific activity of disease-causing genes. (**a**) Rows represent disease-causing genes and columns represent tissues. Colors represent highly expressed genes that are predicted to be active (gray), post-transcriptionally upregulated genes (green) and post-transcriptionally downregulated genes (red). Table entries marked with an 'X' represent a known gene-tissue association based on the OMIM database. (**b**) The overlap of sets of associations between disease-causing genes and tissues affected by the diseases (based on the OMIM database), tissues in which the genes are predicted to be active based on the model and tissues in which the genes are predicted to be active based solely on expression data.

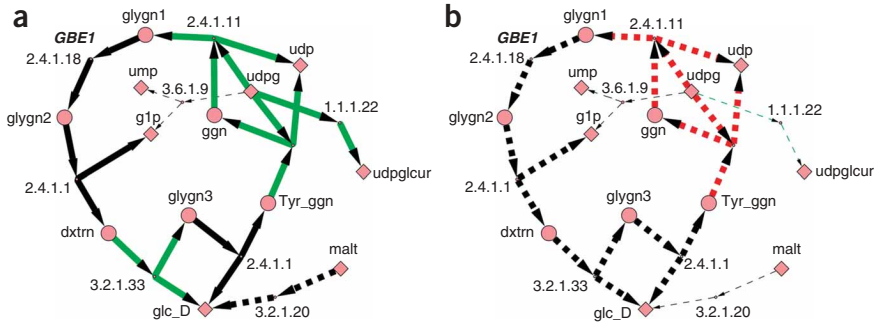


Figure 4 An example of a sub-network of glycogen metabolism. (**a**,**b**) The predicted tissue-specific activity of *GBE1* (1,4-alpha-glucan branching enzyme) in the liver (**a**) and its inactivity in the spleen (**b**) are illustrated. Nodes (circles and diamonds) represent metabolites and edges (arrows) represent biochemical reactions. The presented sub-network is connected to the remaining network via currency metabolites (highly connected metabolites in the network) that are represented by diamond-shaped nodes (**Supplementary Results**, section 5). For simplicity, only abbreviations of metabolite names and enzyme (Enzyme Commission) numbers are specified, with the full information provided in the tissue-specific network visualizations on our website (http://www.cs.tau.ac.il/~shlomito/tissue-net). Gene-expression data are represented by edge colors, with reactions associated with highly, lowly or moderately expressed genes colored in green, red, or black, respectively. An optimal flux distribution (that is, one that is the most similar to the expression data) is shown. Solid edges represent reactions that carry metabolic flux, whereas broken edges represent reactions with zero flux. Reactions whose flux activity state is uniquely determined to be active or inactive (across the space of alternative optimal flux distributions) are marked with thick edges. As evident, *GBE1* is predicted to have an active flux activity state in the liver and an inactive state in the spleen, although it is not highly expressed in both tissues.

the correct metabolite exchange was not reflected in the transporter's expression state. As further validation of these predictions in cases where a membrane transporter is not known, we show that predicted tissue-specificity of the exchanged metabolites matched their known presence in various tissues, based on HMDB (**Table 1** and **Supplementary Data Set 3**). For example, the secretion of the bile acid chenodeoxyglycocholate in the liver[9] is correctly inferred based on the predicted flux activity, even though no membrane transporter is associated with it. Notably, for 33% of the metabolites that are exchanged with biofluids, no tissue-specificity data were found in either of the above data sources, making this the first prediction of their tissue-specific uptake and secretion.

Tissue-specific analysis of metabolic disease genes

We focused on 60 known disease-causing genes in the OMIM database[37] that are predicted to be active in one or more tissues and obtained a set of 164 predicted gene-tissue associations (**Fig. 3a**). The accuracy of these predictions was assessed by comparing them with the actual tissue-associations of the corresponding diseases, obtained by mining the OMIM database (**Supplementary Data Set 3**). The prediction accuracy of disease-related gene-tissue associations is statistically significant (**Table 1**), with a precision of 0.49 and a recall of 0.55 (**Fig. 3b**). When we focused on the 129 gene-tissue associations that are inferred solely based on predicted post-transcriptional regulation effects (that is, excluding those that can be inferred by the expression data directly without the integrated model), we also obtained a significant correlation with the known tissue-associations (**Table 1**). This further shows that additional information is gained by applying the present method, which extends the set of gene-tissue associations threefold compared with using the expression data only. For example, the gene *GBE1* (1,4-alpha-glucan branching enzyme), which causes the glycogen storage disease type IV (Andersen disease; OMIM:232500), is predicted to be post-transcriptionally upregulated specifically in all the tissues whose function is affected by this disease (liver, heart, skeletal muscle and brain[38,39]), whereas it is not highly expressed in any one of them (**Fig. 3a**). A visualization of the glycogen metabolism sub-network showing the predicted activity of *GBE1* in the liver based on the activity of related, highly expressed genes is shown in **Figure 4**. Similarly, the genes *DLD* (dihydrolipoamide dehydrogenase) and *BCKDHA* (branched chain keto acid dehydrogenase E1) that cause the maple syrup urine disease (OMIM:248600) are predicted to be post-transcriptionally upregulated specifically in the brain and liver, in accordance with the information in OMIM. The prediction regarding BCKDHA is further evidence that the activity of branched chain keto acid dehydrogenase is post-transcriptionally regulated via phosphorylation, with the corresponding kinase allosterically regulated by branched-chain keto acids[40]. Finally, disease-causing genes have been shown to be more likely to be expressed in a tissue-specific manner than genes not associated with pathology[32]. Repeating the same analysis for the metabolic genes considered here reveals a moderate association between disease-causing genes and the extent of tissue specificity determined solely based on expression data (hypergeometric P-value = 0.016), but a markedly higher association for tissue-specificity level determined based on the predicted tissue-specific metabolic activities (hypergeometric P-value = 9×10^{-8}). These results further support the central role of post-transcriptional regulation in determining the tissue-specific activity of metabolic disease-causing genes.

DISCUSSION

This study presents a generic approach for systematically predicting human tissue-specific metabolic behavior by integrating a genome-scale metabolic network model with tissue-specific gene-expression and protein-abundance data. The method predicts the metabolic behavior of ten human tissues, suggesting that on average, 18% of the metabolic genes they express are post-transcriptionally regulated. That is, the predicted flux activities of these genes differs from their expression levels. Post-transcriptional regulation of metabolic activities can further be dissected into hierarchical regulation, which affects the maximum activity of enzymes (that is, v_{max}) by controlling protein translation and degradation rates and their phosphorylation, and metabolic regulation, which denotes the effect of metabolite concentrations on actual enzyme activity through allosteric and mass action effects[41]. Notably, the dissection of the predicted post-transcriptional regulation to either hierarchical or metabolic regulation is beyond the scope of our method and would require comprehensive experimental data sets on tissue-specific enzyme activities and metabolic fluxes. The substantial effect that post-transcriptional regulation is predicted to have on human tissue-specific metabolism is consistent with its critical role in determining metabolic fluxes in microorganisms[41,42].

To validate the predicted tissue-specific metabolic behavior, we relied on various data sources for tissue specificity of genes, reactions and metabolites. In all cases, the predicted tissue-specificity was significantly correlated with these data sets, with the precision and recall

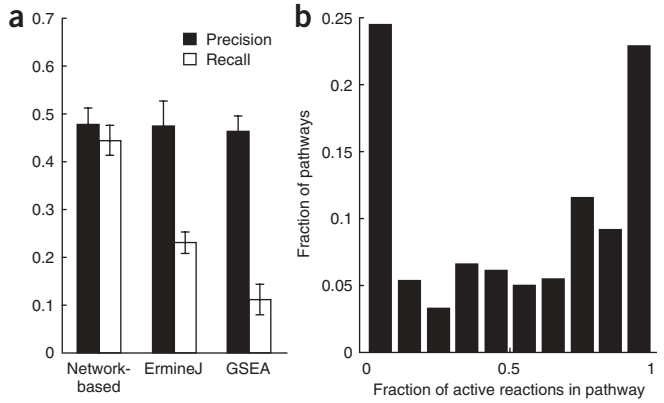


Figure 5 Comparison of network- and pathway-based prediction. (**a**) Comparison of gene tissue-specificity prediction obtained by our method and with two functional enrichment-based methods, ErmineJ and GSEA. Average precision and recall is computed over the various gene tissue-specificity data sources referred to in **Table 1**, focusing on the set of genes that is predicted by our method to be active in at least a single tissue (**Supplementary Results**, sections 3 and 4). (**b**) Coherency of predicted pathway activity based on pathway definitions given in ref. 14, measured by the fraction of the reactions that are predicted to be active in each pathway (across all tissues). As shown, a high fraction of the pathways are only partially activated. For example, as many as 36% of the pathways have between 0.2 and 0.8 of their reactions activated.

varying between 0.36–0.7 and 0.37–0.55, respectively (**Table 1**, **Supplementary Table 2** and **Supplementary Results**, section 3). It should be noted, though, that one cannot expect to get an optimal correlation with the relevant tissue-specificity data as the data are noisy and are inconsistent between data sources (**Supplementary Fig. 10** online). Some of the false predictions are expected to result from incompleteness of the metabolic network model, as well as from several simplifying assumptions incorporated into our computational method to enable the large-scale analysis of a network with thousands of reactions. Specifically, the steady-state assumption (required because of the lack of global enzyme kinetic data) under which the metabolic state is described as a constant flow of metabolic flux does not take into account time-dependent changes in tissue metabolism. Such changes may be reflected in alterations in gene-expression levels (putatively caused by transcriptional regulation or changes in mRNA degradation rates) and hence may potentially be predicted by our method given the relevant time-dependent expression data.

The accuracy of our tissue-specificity gene predictions can be further assessed by a comparison with previous methods that rely on enrichment analysis of highly expressed genes within sets of functionally related genes (e.g., metabolic pathways). Applying two such state-of-the-art methods, ErmineJ[43] and GSEA[44], to predict tissue specificity of metabolic genes and reactions based on tissue-specific gene-expression data[28] testifies to the moderately higher prediction accuracies obtained by our method (**Fig. 5a**, **Supplementary Results**, sections 3,4, and **Supplementary Figs. 11–14** online). **Figure 5a** compares the prediction performance of all methods on a set of genes predicted to be active in at least a single tissue. The recall level of the gene activity predictions obviously decreases if we take into account additional metabolic genes whose activity state cannot be determined by our method (or for that matter, by any constraint-based approach using the same metabolic model) owing either to errors in the model or the existence of multiple flux distributions (**Supplementary Results**, section 3, **Supplementary Figs. 12–14** and **Supplementary Table 2**). Considering these additional genes, but excluding isozymes, the mean recall level of our method is still higher than the other approaches (0.25 for our method compared to 0.22 and 0.08 for ErmineJ and GSEA; **Supplementary Fig. 12**). When we also include isozymes (for which our method is incapable of predicting a unique activity state), ErmineJ presents the highest recall (0.22 compared with 0.14 for our method) but with a considerably lower precision (0.38 compared with 0.47 for our method; **Supplementary Fig. 13**). Predicting isozymes' activity based on a simplifying assumption that they are either coherently activated or inactivated in each tissue considerably raises the recall level of our predictions above those of ErmineJ to a level of 0.33 but lowers the precision slightly below ErmineJ to a level of 0.35 (**Supplementary Fig. 14**).

On a conceptual level, in contrast with constraint-based methods that involve global stoichiometric computations over the whole network, gene set enrichment–based methods rely on pathways defined a priori and do not use information about the activity of neighboring pathways to infer the activity of a given pathway. The latter methods assume that a gene set is either entirely activated or inactivated under a certain condition, whereas in practice, different subsets of a pathway may be activated because of the large fraction of alternative pathways and cross-links among pathways. Indeed, inspecting the coherency of metabolic pathway activity predicted by our approach (based on the pathway annotation in ref. 14), a significant fraction (36%) of classically defined pathways are only partially activated across the different tissues (**Fig. 5b**), with the fraction of activated reactions per pathway varying between 0.2 and 0.8. These results highlight the need to consider the entire metabolic network in a global fashion without a priori dividing the network into pathways that are, in fact, not independent of each other.

To facilitate easy access to the predicted tissue-specific metabolic behavior, we generated network visualizations in which the expression and predicted flux data are projected over the global human network (**Supplementary Results**, section 5). These network visualizations are accessible through our website (http://www.cs.tau.ac.il/~shlomito/tissue-net) and can be explored interactively using the freely available Cytoscape software[45]. To allow for easy browsing through this huge network, alternative views that dissect the network to either cellular compartments or metabolic subsystems are available. An illustrative example of a small sub-network with both tissue-specific expression and predicted metabolic flux (as extracted from Cytoscape) is shown in **Figure 4**.

The basic approach presented here opens the way for future computational investigations of metabolic disorders given the relevant expression data. A first step toward this endeavor by Duarte *et al.*[14] mapped changes in gene expression measured following gastric bypass surgery onto the metabolic network to visualize and interpret genome-scale changes in metabolic behavior. The computational method described here can markedly advance this line of study by predicting a flux distribution that is consistent with disease-state expression data and that concomitantly allows predicting tissue-specific, post-transcriptional regulatory effects. One such important application is the classification of tissue-specific gene-expression measurements from either healthy or sick individuals, based on the predicted metabolic behavior that they induce. Another compelling application would be the prediction of disease and tissue-specific biomarkers that could be identified using biofluid metabolomics[46].

Our approach can be used to predict the metabolic behavior of many additional tissues under different physiological conditions, using readily available tissue-specific expression measurements[47,48]. More

refined tissue-specific models could potentially be generated by identifying tissue-specific objective functions or measuring tissue-specific metabolite exchange patterns that could further limit the space of possible functional states[49]. Future tissue-specific metabolic models may also integrate additional data on microRNA regulation that are known to take a central role in cell-type differentiation[50], as well as incorporate organelle-specific proteomic data[51]. Overall, the method presented here lays the foundation for the rapid development of human tissue-specific metabolic models and is likely to advance the computational study of human metabolic disorders.

METHODS

Tissue-specific modeling of metabolism. The genome-scale human metabolic network model by Duarte *et al.* accounts for 1,496 ORFs, 2,004 proteins, 2,766 metabolites and 3,311 reactions[14]. Data on tissue specificity of genes based on gene-expression data and protein abundance were obtained from GeneNotes[18,28] and HPRD[29], respectively. We consider a gene to be highly or lowly expressed in a certain tissue if it is marked uniformly as expressed or nonexpressed in all GeneNote patterns and in HPRD, respectively, or else it is considered to be moderately expressed. A detailed Boolean gene-to-reaction mapping (part of the metabolic network model of Duarte *et al.*) was employed to identify a tissue-specific expression state for each reaction, reflecting whether its enzyme-encoding genes are classified as expressed in the tissue. Specifically, this was done by modifying the Boolean mapping to account for tri-valued expression states, assigning highly, lowly and moderately expressed genes, values of 1, –1 and 0, respectively, and replacing the logical 'and' and 'or' operators with 'max' and 'min' expressions, respectively (following ref. 22). This analysis resulted in a subset of the reactions in the model (denoted R_H) that is defined to be highly expressed and another subset (denoted R_L) defined as lowly expressed.

For each tissue, we then formulated the following mixed integer linear programming (MILP) problem to find a steady-state flux distribution satisfying stoichiometric and thermodynamic constraints, while maximizing the number of reactions whose activity is consistent with their expression state:

$$\max_{v,y^+,y^-} \left(\sum_{i\in R_H} (y_i^+ + y_i^-) + \sum_{i\in R_L} y_i^+ \right)$$

$$s.t$$

$$S \cdot v = 0 \quad (1)$$

$$v_{\min} \le v \le v_{\max} \quad (2)$$

$$v_i + y_i^+ \left(v_{\min,i} - \varepsilon\right) \ge v_{\min,i}, i \in R_H \quad (3)$$

$$v_i + y_i^- \left(v_{\max,i} + \varepsilon\right) \le v_{\max,i}; i \in R_H \quad (4)$$

$$v_{\min,i}(1 - y_i^+) \le v_i \le v_{\max,i}(1 - y_i^+), i \in R_L \quad (5)$$

$$v \in R^m$$

$$y_i^+,\ y_i^- \in [0, 1]$$

where v is the flux vector and S is a n x m stoichiometric matrix, in which n is the number of metabolites and m is the number of reactions. The mass balance constraint is enforced in equation (1). Thermodynamic constraints that restrict flow direction are imposed by setting $v_{\min}$ and $v_{\max}$ as lower and upper bounds on flux values in equation (2), respectively. For each expressed reaction, the Boolean variables y^+ and y^- represent whether the reaction is active (in either direction) or not. Specifically, a highly expressed reaction is considered to be active if it carries a significant positive flux that is greater than a positive threshold ε (equation (3)) or a significant negative flux $< -\varepsilon$ (equation (4) for reversible reactions). We chose a threshold of $\varepsilon = 1$ to determine reactions' flux activity for highly expressed reactions, though various other choices provide qualitatively similar results. For each lowly expressed reaction, the Boolean variable y^+ represents whether the reaction is inactive (equation (5)). Specifically, lowly expressed reactions are considered to be inactive if they carry zero metabolic flux, though changing equation (5) to enable these reactions to carry a low metabolite flux (that is, with an upper bound lower than ε) and still be considered inactive provides qualitatively similar results (**Supplementary Table 3** online). The optimization maximizes the number of highly expressed reactions (R_H) that are active and the number of lowly expressed reactions (R_L) that are inactive. The commercial CPLEX solver was used for solving MILP problems on a Pentium-4 machine running Linux in a few dozens of seconds per problem.

A solution found by the MILP solver is guaranteed to be an optimal one in terms of the objective function maximized, but the solution identified may not be unique as a space of alternative optimal solutions may exist. In our case, the space of optimal solutions represents alternative steady-state flux distributions obtaining the same similarity with the expression data. To account for these alternative solutions, we employed a variant of Flux Variability Analysis[27] that was used in a previous study on alternative metabolic-regulatory solutions[26] (**Supplementary Results**, section 1). Our method computes for each metabolic reaction whether it is predicted to be always active (or, in the opposite case, always inactive) in a certain tissue across the entire solution space. This is performed by solving two MILP problems (each similar to the one described above) for each reaction to find the maximal attainable similarity with the expression data when the reaction is forced to be activated (denoting this similarity x) and when it is forced to be inactivated (denoting this similarity y). A reaction is then considered to be active in this tissue if $x > y$ (that is, a higher similarity with the expression data is achieved when the reaction is active than when it is inactive) with a confidence level of x–y. Conversely, it is considered to be inactive if $x < y$, with a confidence of $y - x$. In case $x = y$ (that is, the same similarity with the expression data can be achieved both when the reaction is forced to be active or inactive), the activity state is considered to be undetermined. To assign a flux-activity state for a gene, which may be associated with multiple reactions in the model (via one or more enzymes), a similar method is employed. In this case, x denotes the maximal possible similarity with the expression data when at least a single reaction associated with the gene is activated, and y denotes the maximal possible correlation with the expression data when all reactions associated with the gene are inactivated.

Notably, previous studies have used MILP in the context of applying constraint-based methods to studying metabolism in microbial organisms[26,52,53]. A previous investigation has integrated a microorganism's expression data with its metabolic network, showing improved prediction performance of various phenotypes[24]. However, this method analyzes gene-expression data as a preprocessing step that removes inactive genes before employing a standard flux balance analysis procedure, which requires a definition of a cellular objective function and additional a priori data on metabolite uptake rates. As described above, our method does not require these data (which are unavailable for human tissues) as it integrates the expression data as part of an optimization method, which maximizes the consistency between gene expression and the corresponding enzyme activity in a soft-constrained manner.

Web searches for gene-tissue pairs. Web searches were done with the Google search engine, using an automated scripting utility called Query Google (http://www.linguistics.ucla.edu/people/hayes/QueryGoogle/). The searches were restricted to web pages consisting of both the gene and the tissue names in the title, using the "allintitle:" Google search command.

Data acquisition of disease-tissue associations. A list of diseases along with their causing genes was obtained from the OMIM database[37]. The tissues afflicted in each disease were found by parsing the disease description field in the OMIM database in search for the tissue names.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We are grateful to Shiri Freilich and Ben Sandbank for helpful comments and suggestions. We wish to thank the reviewers for their constructive remarks that helped improve this manuscript considerably. T.S. is supported by an Eshkol Fellowship from the Israeli Ministry of Science. M.C. is a fellow of the Edmond J. Safra Program in Tel-Aviv University. M.J.H. was supported by National Institutes of Health grant no. GM071808. This research was supported by grants from the Israeli Science Foundation, the German-Israeli Foundation and the Tauber fund to E.R.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Fell, D.A. *Understanding the Control of Metabolism* (Portland Press, London, 1996).
2. Domach, M.M., Leung, S.K., Cahn, R.E., Cocks, G.G. & Shuler, M.L. Computer model for glucose-limited growth of a single cell of *Escherichia coli* B/r-A. Reprinted from *Biotechnology and Bioengineering* **26**, 203–216 (1984). *Biotechnol. Bioeng.* **67**, 827–840 (2000).
3. Price, N.D., Papin, J.A., Schilling, C.H. & Palsson, B.O. Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol.* **21**, 162–169 (2003).
4. Price, N.D., Reed, J.L. & Palsson, B.O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**, 886–897 (2004).
5. Lanpher, B., Brunetti-Pierri, N. & Lee, B. Inborn errors of metabolism: the flux from Mendelian to complex diseases. *Nat. Rev. Genet.* **7**, 449–460 (2006).
6. Muoio, D.M. & Newgard, C.B. Obesity-related derangements in metabolic regulation. *Annu. Rev. Biochem.* **75**, 367–401 (2006).
7. Altucci, L., Leibowitz, M.D., Ogilvie, K.M., de Lera, A.R. & Gronemeyer, H. RAR and RXR modulation in cancer and metabolic disease. *Nat. Rev. Drug Discov.* **6**, 793–810 (2007).
8. Shi, Y. & Burn, P. Lipid metabolic enzymes: emerging drug targets for the treatment of obesity. *Nat. Rev. Drug Discov.* **3**, 695–710 (2004).
9. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
10. Romero, P. *et al.* Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* **6**, R2 (2005).
11. Wiback, S.J. & Palsson, B.O. Extreme pathway analysis of human red blood cell metabolism. *Biophys. J.* **83**, 808–818 (2002).
12. Vo, T.D., Greenberg, H.J. & Palsson, B.O. Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J. Biol. Chem.* **279**, 39532–39540 (2004).
13. Chatziioannou, A., Palaiologos, G. & Kolisis, F.N. Metabolic flux analysis as a tool for the elucidation of the metabolism of neurotransmitter glutamate. *Metab. Eng.* **5**, 201–210 (2003).
14. Duarte, N.C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. USA* **104**, 1777–1782 (2007).
15. Ma, H. *et al.* The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.* **3**, 135 (2007).
16. Levine, D.M. *et al.* Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. *Genome Biol.* **7**, R93 (2006).
17. Son, C.G. *et al.* Database of mRNA gene expression profiles of multiple human organs. *Genome Res.* **15**, 443–450 (2005).
18. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
19. Fong, S.S. & Palsson, B.O. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat. Genet.* **36**, 1056–1058 (2004).
20. Daran-Lapujade, P. *et al.* Role of transcriptional regulation in controlling fluxes in central carbon metabolism of *Saccharomyces cerevisiae*. A chemostat culture study. *J. Biol. Chem.* **279**, 9125–9138 (2004).
21. Schuster, S., Klamt, S., Weckwerth, S., Moldenhauer, F. & Pfeiffer, T. Use of network analysis of metabolic systems in bioengineering. *Bioprocess Biosyst. Eng.* **24**, 363–372 (2002).
22. Bilu, Y., Shlomi, T., Barkai, N. & Ruppin, E. Conservation of expression and sequence of metabolic genes is reflected by activity across metabolic states. *PLOS Comput. Biol.* **2**, e106 (2006).
23. Famili, I., Forster, J., Nielsen, J. & Palsson, B.O. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci. USA* **100**, 13134–13139 (2003).
24. Akesson, M., Forster, J. & Nielsen, J. Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.* **6**, 285–293 (2004).
25. Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. & Palsson, B.O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96 (2004).
26. Shlomi, T., Eisenberg, Y., Sharan, R. & Ruppin, E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol. Syst. Biol.* **3**, 101 (2007).
27. Mahadevan, R. & Schilling, C.H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264–276 (2003).
28. Shmueli, O. *et al.* GeneNote: whole genome expression profiles in normal human tissues. *C. R. Biol.* **326**, 1067–1072 (2003).
29. Mishra, G.R. *et al.* Human protein reference database–2006 update. *Nucleic Acids Res.* **34**, D411–D414 (2006).
30. Wishart, D.S. *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res.* **35**, D521–D526 (2007).
31. Schomburg, I. *et al.* BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* **32**, D431–D433 (2004).
32. Winter, E.E., Goodstadt, L. & Ponting, C.P. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.* **14**, 54–61 (2004).
33. Hubbard, T.J. *et al.* Ensembl 2007. *Nucleic Acids Res.* **35**, D610–D617 (2007).
34. Yan, Q. & Sadee, W. Human membrane transporter database: a Web-accessible relational database for drug transport studies and pharmacogenomics. *AAPS PharmSci* **2**, E20 (2000).
35. Saier, M.H., Jr., Tran, C.V. & Barabote, R.D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.* **34**, D181–D186 (2006).
36. Bohm, M., La Rosee, K., Schwinger, R.H. & Erdmann, E. Evidence for reduction of norepinephrine uptake sites in the failing human heart. *J. Am. Coll. Cardiol.* **25**, 146–153 (1995).
37. McKusick, V.A. Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
38. Greene, H.L., Brown, B.I., McClenathan, D.T., Agostini, R.M., Jr & Taylor, S.R. A new variant of type IV glycogenosis: deficiency of branching enzyme activity without apparent progressive liver disease. *Hepatology* **8**, 302–306 (1988).
39. Tay, S.K.H. *et al.* Fatal infantile neuromuscular presentation of glycogen storage disease type IV. *Neuromuscul. Disord.* **14**, 253–260 (2004).
40. Brosnan, J.T. & Brosnan, M.E. Branched-chain amino acids: enzyme and substrate regulation. *J. Nutr.* **136**, 207S–211S (2006).
41. Rossell, S. *et al.* Unraveling the complexity of flux regulation: a new method demonstrated for nutrient starvation in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* **103**, 2166–2171 (2006).
42. Daran-Lapujade, P. *et al.* The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels. *Proc. Natl. Acad. Sci. USA* **104**, 15753–15758 (2007).
43. Lee, H.K., Braynen, W., Keshav, K. & Pavlidis, P. ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics* **6**, 269 (2005).
44. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
45. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
46. Kell, D.B. Metabolomic biomarkers: search, discovery and validation. *Expert Rev. Mol. Diagn.* **7**, 329–333 (2007).
47. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
48. Shyamsundar, R. *et al.* A DNA microarray survey of gene expression in normal human tissues. *Genome Biol.* **6**, R22 (2005).
49. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* **3**, 119 (2007).
50. Zhao, Y. & Srivastava, D. A developmental view of microRNA function. *Trends Biochem. Sci.* **32**, 189–197 (2007).
51. Andersen, J.S. *et al.* Nucleolar proteome dynamics. *Nature* **433**, 77–83 (2005).
52. Burgard, A.P., Pharkya, P. & Maranas, C.D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647–657 (2003).
53. Shlomi, T., Berkman, O. & Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. USA* **102**, 7695–7700 (2005).

Network-based Prediction of Human Tissue-specific Metabolism: Supplementary Results

Tomer Shlomi, Moran N. Cabili, Markus J. Herrgård, Bernhard Ø. Palsson, Eytan Ruppin

1. Predicting metabolic flux activity and directionality in the presence of alternative possible flux distributions

Our method predicts metabolic flux activity based on identifying a global flux distribution that is optimal in terms of similarity with expression data. An illustrative example of how our method handles the existence of alternative optimal flux distributions is shown in Supplementary Figure 1. The figure describes an example of a metabolic network, in which metabolic flux, that begins with the uptake of metabolite M1, can go through the lower pathway (i.e., through M3; Supplementary Fig. 1b) or through the upper pathway (i.e., through M2; Supplementary Fig. 1c), or though a combination of both pathways. In this example, every feasible metabolic behavior in which M1 is taken up from the environment has an optimal similarity of 1 with the expression data.

The activity state for each reaction is computed by solving two MILP problems (each similar to the one described in the main text) to find the maximal attainable similarity with the expression data when the reaction is forced to be activated (denoting this similarity x), and when it is forced to be inactivated (denoting this similarity y). Examining reaction R1, the maximal similarity with the expression when this reaction is active is 2 (and hence x=2), and the maximal similarity with the expression when it is inactive is 1 (and hence y=1) and hence this reaction is predicted to be active with a confidence of 1, based on the absolute value of the difference between x and y. Based on this method, reactions R2-R4 are predicted to have an undetermined activity state, as the similarity with the expression data remains the same whether each of them is activated or not (Supplementary Fig. 1d). Reaction R5 is predicted to be active, as it must carry non-zero flux if R1 is active, regardless of which of the alternative pathways are activated (Supplementary Fig. 1d).

For reactions that are specified as reversible in the model (i.e., which may carry metabolic flux in two opposite directions), the method may be able to predict the direction of metabolic flow considering the space of alternative solutions. Specifically, this method was used to predict directionality of exchange reactions, which provides interesting information on the uptake and secretion of metabolites from surrounding biofluids. To do that, we use a similar method to the one described above, and compute the maximal correlation with the expression data when a reaction is activated in one direction (denoted *x1*) and when it is activated in the inverse direction (denoted *x2*). A difference between *x1* and *x2* is used to predict the direction of the reaction. In some cases, a reaction can be predicted to be active, although the direction cannot be determined, considering the space of possible solutions.

2. Predicting metabolic flux activity in Yeast

To validate our method for predicting metabolic flux activity based on gene expression data, we applied it to predict the metabolic behavior of the yeast *S. cerevisiae* under various growth media. The predicted metabolic behavior is validated on a large-scale via a corresponding Flux Balance Analysis (FBA) model, and via experimental flux measurements available for the yeast's central metabolism. The predictions rely on a large-scale metabolic network model of *S. cerevisiae* by Duarte et al[1] and on micro-array data for *S. cerevisiae* under various growth media by Daran et al[2]. In addition to gene expression data, Daran et al. provides experimental measurements of metabolic fluxes in the central carbon metabolism of *S. cerevisiae* that are used for validation.

The setup of this validation study is illustrated in Supplementary Figure 2. Our method employs the metabolic network mode without requiring the biomass synthesis reaction and without prior knowledge of a growth media composition (Fig. 2a). It was used to predict metabolic flux activity under four growth conditions containing glucose, ethanol, acetate or maltose. The micro-array data was discretized by defining lower and upper thresholds on the measured Affymetrix expression levels. Genes with expression level lower than 50 were considered to be lowly-expressed. The analysis was repeated with

several upper thresholds of 200, 400, 600, and 800. Overall, for each growth media and each choice of threshold, each reaction is predicted by employing our method to be active, inactive, or undetermined. For comparison, we employed a Flux Balance Analysis (FBA) model to predict metabolic flux under the same conditions. FBA predicts a flux distribution that provides maximal growth rate by relying on *a priori* data on the composition of the growth media as well as yeast biomass composition. To account for alternative possible FBA solutions, we employed Flux Variability Analysis (FVA), to predict the set of active reactions under each condition, assuming optimal growth rate[3]. Using this method, each reaction is determined to be either: active, non-active, or possibly active (i.e., undetermined - in the latter case it can be either active or in-active in an optimal FBA solution).

Comparing the predicted metabolic activity obtained by our method to measured metabolic fluxes in the central carbon metabolism system of *S. cerevisiae* revealed a significant correlation (hyper geometric p-value = 0.01) with precision of 0.71 and recall of 0.89 (Supplementary Fig. 3).

To compare our predictions with FBA, we considered the reactions which FBA determines to be either active or in-active. The correlation between our method predictions and FBA across the four growth media and parameter choices are highly significant, with a mean precision of 0.89 and recall of 0.79 (hyper-geometric p-value = $5.6 \cdot 10^{-8}$; Supplementary Fig. 4a). Next, we compared the method's predictions to those obtained when relying solely on the expression data (focusing only on reactions that are associated with genes in the model). We found that the while the precision of both methods (compared to FBA predictions) is quite similar, the recall of our method is significantly higher (0.83 for our method and 0.61 for the expression-based predictions; Supplementary Fig. 4b). Finally, we examined the precision and recall of our method for the remaining subset of reactions – those that are not associated with genes in the model and hence their activity state cannot be predicted solely based on expression data. We found that also for these reactions, our predictions achieve high precision of 0.78 and recall of 0.65 (p-value of $6.6 \cdot 10^{-7}$, Supplementary Fig. 5c). In addition to predicting the

reactions' activity, our method predicts the reactions' directionality for the 43% reversible reactions in the model (obviously, the directionality of flux cannot be predicted based on the expression data alone). The correlation between these predictions and FBA predictions of directionality is highly significant (p-value = $1.02 \cdot 10^{-8}$) with a mean precision of 1.0 and recall of 0.67.

A specific interesting characterization of the predicted metabolic behavior is the identity of the metabolites that are exchanged with the surrounding growth media. A similar characterization was presented in the main text for the tissue-specific exchange of metabolites with surrounding biofluids in human tissues. The correlation between our method's predictions of metabolite exchange with those obtained with FBA in the yeast's case are statistically significant (*p*-value = 0.003) with a mean precision of 0.68 with recall of 0.42 (Supplementary Fig. 6a). An additional validation of our approach is its ability to correctly predict the synthesis of yeast biomass precursors, without an explicit constraint of that kind embedded in the model. We find that on average 82% of the yeast biomass constituents (as defined in the growth reaction in the FBA model) are predicted to be synthesized by our approach, across the different media and discretization thresholds (Supplementary Fig. 6b).

3. Validating tissue-specificity predictions

The aim of limiting the set of analyzed genes to those that are predicted to be activated in at least a single tissue has been to exclude genes whose activity cannot be predicted by our method and the model in *any tissue* (i.e. regardless of specific gene expression data). This may result either from errors in the model in the form of dead-end pathways, or due to the presence of isozymes (which account for 56% of the genes in the model), for which the method cannot be expected to pinpoint which of the coding enzymes is activated under different tissues. Additionally, this may result from more complex cases that involve alternative pathways which are harder to identify a-priori (this refers to cases where one of the alternative pathways involves reactions which are not coded by existing ORFs in the model, e.g., spontaneous reactions). As shown, the prediction accuracy is statistically significant in all cases (Table 1 in the main text).

An alternative approach for identifying some of the genes that cannot be predicted to be active by the method that do not depend on the expression data involves the direct identification of genes that reside on dead-end pathways and isozymes. To identify genes that are on dead-end pathways we employed Flux Variability Analysis (FVA)[3] to identify reactions that cannot be activated in the model in all feasible flux distributions. Genes that are associated only with reactions that cannot be activated were then excluded from further consideration. To identify genes whose activity is completely backed-up by isozymes, we employed the Boolean gene-to-reaction mapping in the model of Duarte et al. to compute for each gene the set of reactions affected by its knockout. Genes whose knockout do not affect any reaction were excluded (notably, all genes excluded cannot be predicted by our MILP to be active under any tissue). The prediction accuracy obtained when excluding only genes that reside on dead-end pathways (reflecting errors in the model) is still highly statistically significant across all validation datasets (Supplementary Table 2).

Though constraint-based modeling cannot predict which of several genes that code for the same isozyme is activated in a certain condition, a prediction regarding the activity of the corresponding reaction suggests that at least a one of the coding genes is activated. Based on this observation, we computed also predictions regarding the activity of isozymes, where all genes that code for the same activated reaction are considered as active. The resulting prediction accuracy, computed over all genes in the model (excluding those on dead-end pathways), is highly significant in all cases. Comparing the resulting prediction accuracy with other pathway-based methods (see next section) shows a markedly higher recall by our method, with no clear advantage to any of the methods in terms of precision level (Supplementary Fig. 14).

4. Functional enrichment-based tissue-specificity predictions

To further assess the accuracy of gene tissue-specificity predictions, we compared it with extant methods that identify significant enrichment of highly expressed genes within groups of functionally related genes. Such enrichment tests can be used to infer patterns

of tissue-specific gene activity by defining genes to be active in a tissue if they belong to a gene set that is enriched with highly expressed genes in the tissue. Specifically, we applied two state-of-the-art methods of that kind - ErmineJ[4] and GESA[5]. The tissue-specific genes expression used with both methods was taken from Shmueli et al[6] (GEO accession index GSE803), as was used for computational method. The ErmineJ software, uses gene sets based on GO annotation, was taken from Lee et al[4]. GSEA (Gene Set Enrichment Analysis) was computed using a software from Subramanian et al[5], using all gene sets that represent pathways from the MSigDB.

Comparing the accuracy of the tissue-specific predictions obtained with our method with that obtained with ErmineJ and GSEA shows that our method outperforms both other methods. For all three data sources of genes tissue-specificity (Web, transporter databases, OMIM genes; see validation data sources in Table 1), our method yields a significantly higher recall with a similar (or higher) precision to that obtained by the other methods (Supplementary Fig. 11). The marked advantage of our approach is evident across the different gene sets discussed in the previous section (Supplementary Fig. 12-14).

5. Tissue-specific network visualization

To facilitate easy access to the predicted tissue-specific metabolic behavior, we generated network visualizations in which the expression and predicted flux data are projected over the global human network. These network visualizations are accessible through the supplemental website: http://www.cs.tau.ac.il/~shlomito/tissue-net.hmtl, using the publicly available Cytoscape software [7]. Since many high degree nodes exist in the network, special layouts are required to produce network visualizations that are readily interpretable. To this end we produced network visualizations in which hub nodes are repeated multiple times and hence layouts with a small number of edge crossings can be generated. The following "currency metabolites" hub nodes which had the highest degree are not shown: adp, atp, co2, o2, h2o, h2o2, h, k, na1, nad, nadh, nadp, nadph, nh4, pi, ppi, so4. Using this method, the following layouts were created for each tissue:

1. Compartmental layout - a partition of the global network into 7 sub-networks that represent different cellular-compartments.
2. Pathways layout – a partition of the global network to model pathways, based on an annotation provided in the network model of Duarte et al.

The legend of the network visualization is explained below:

1. Metabolites are represented as large circular and diamond-shaped nodes. Metabolites marked as circles appear only once in the network and metabolites marked as diamonds represent hubs which may appear in multiple places in the network. Metabolite names end with a [#] where # represents a cellular compartment. The same metabolite may appear in multiple compartments. The following compartments are considered:

C	'cytoplasm'
E	'extracellular'
L	'lysosome'
M	'mitochondrion'
N	'nucleus'
R	'endoplasmic
X	'peroxisome'

   Each metabolite node contains an abbreviation, with the full name of the metabolite available in the node attribute - *node_name_long*.

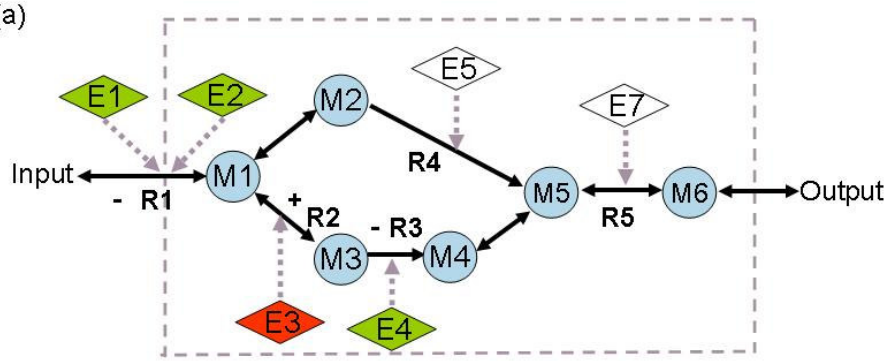
2. Reactions are represented as small circular nodes with one or more substrate and product metabolites connected to them. Reaction directionality is represented with arrows. The following node attributes characterize reactions:
   1. *node_name_long* – full name of the enzymatic reaction.
   2. *node_ec* – E.C (Enzyme Classification) number.
   3. *node_reaction* – a textual description of the reaction's substrates and products.
   4. *node_subsystem* – pathway annotation of the reaction.
   5. *node_gene_exp* – a Boolean equation that shows the expression state of tghe enzyme driving the reaction, depending on the expression state of its

corresponding genes. Values of 1 and -1 represent highly and lowly expressed genes, respectively.

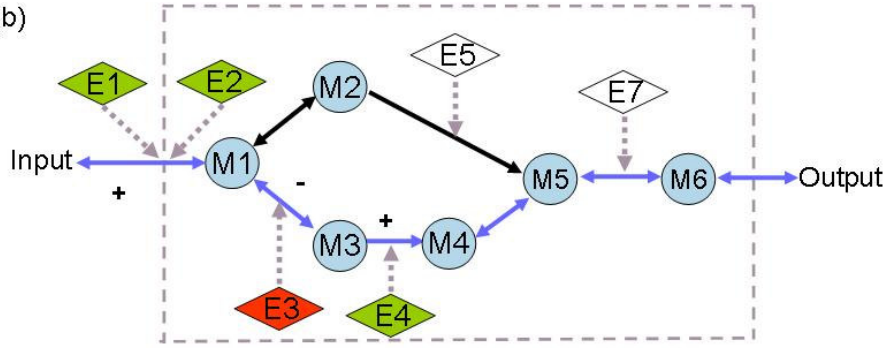
3. Gene expression data is represented by edge colors. Reactions associated with significantly highly or lowly expressed genes are colored in green or red, respectively. Other reactions are colored black.

4. An optimal (in the sense of similarity with the expression data), feasible flux distribution is represented by edge type and direction. Solid edges represent reactions that carry metabolic flux, while dashed edges represent reactions with zero flux. Reactions with a determined flux activity state – i.e., reactions that are predicted to be active or inactive across the entire solution space, are marked with thick edges. For example, a reaction that is determined to have an inactive state is represented as a thick, dashed edge. Alternatively, a reaction that is determined to have an active state is represented as a thick, solid edge.

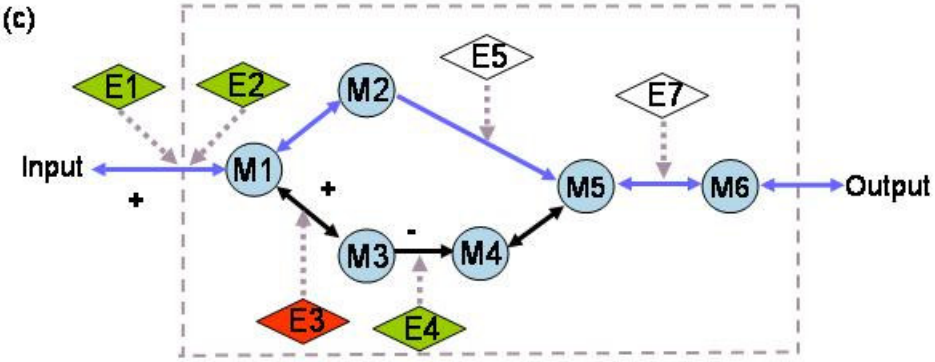
(a)



(b)



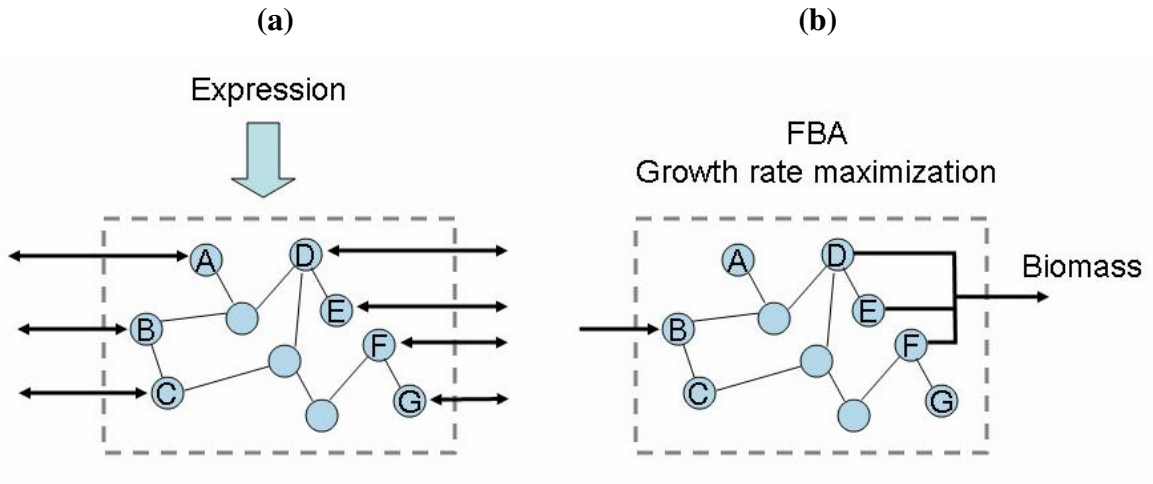
(c)



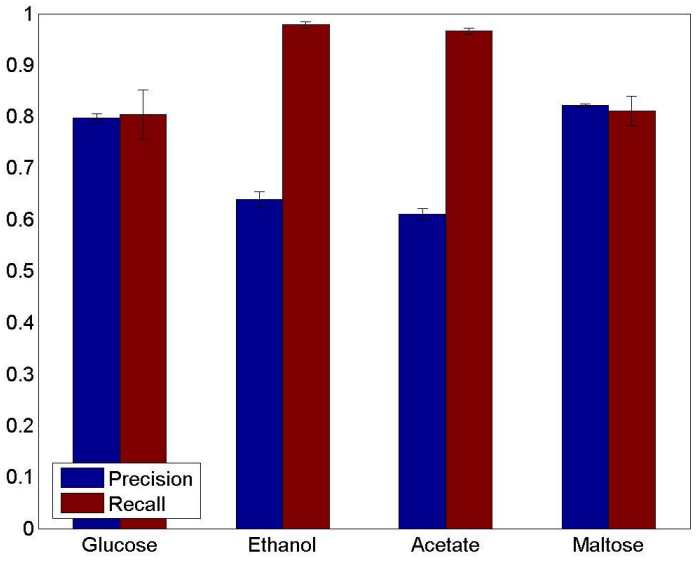
(d)

	Active score (x)	Inactive score (y)	Activity state	Confidence abs (x-y)
R1	2	1	Active	1
R2	2	2	Undetermined	0
R3	2	2	Undetermined	0
R4	2	2	Undetermined	0
R5	2	1	Active	1

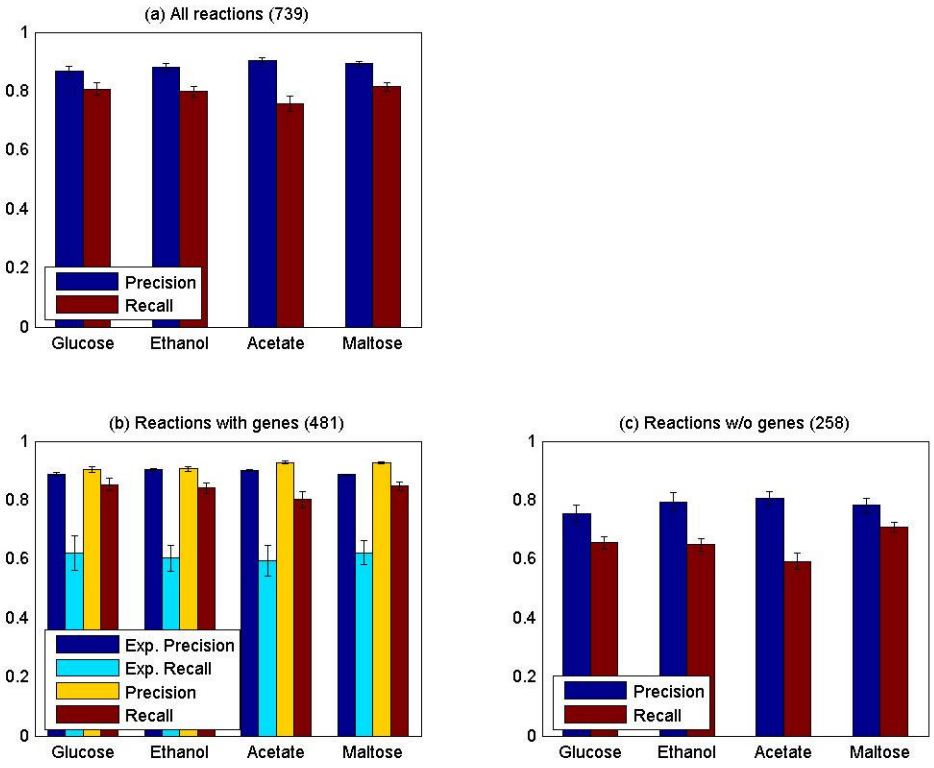
Supplementary figure 1: (a) An illustrative example of prediction of reaction activity state while considering alternative flux distributions. Circular nodes represent metabolites, whereas diamond nodes represent enzymes. Red, green, and white colors represent significantly low, significantly high, and normal expression of the enzyme-coding genes, respectively. Solid edges represent metabolic reactions. Dashed edges associates enzymes with the reactions they catalyze (as in Fig. 1 in the main text). Panels b and c show alternative flux distributions that obtain maximal match with the expression data (and are hence equally plausible). The predicted steady-state flux distribution in each potential flux distribution is colored in purple. (d) Plus and minus signs represent the similarity between the reaction's expression and flux states. The table illustrates the computation of reaction's activity state of several reactions, based on the confidence level that they are active or inactive across all possible flux distributions (as described in the Methods section).



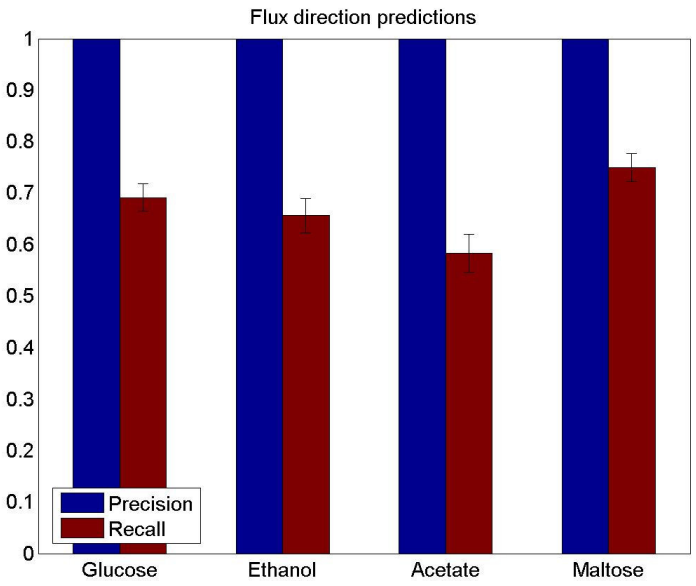
Supplementary Figure 2: A schematic representation of our flux activity prediction method (a) and flux activity prediction via Flux Balance Analysis (FBA) (b), in the yeast validation study. (a) Our method relies on expression data as cues for the likely activity of some of the enzymes in the network (arrows point to reactions for which such information exist). It requires no prior knowledge about the composition of the growth media, or about the biomass composition. (b) In contrast, FBA requires explicit knowledge of the growth media and biomass composition (the left arrow denotes the existence of a certain metabolite in the media, and the right arrows denote the biomass function).



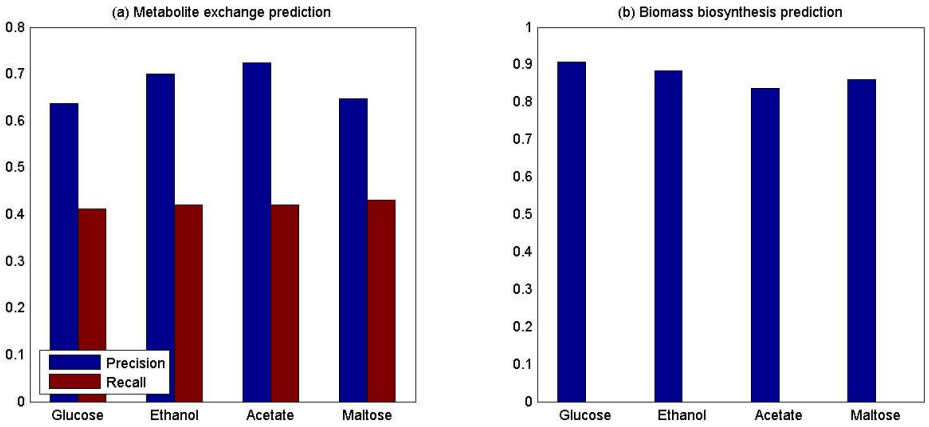
Supplementary Figure 3: Accuracy of predicted metabolic flux activities in the central carbon metabolism of yeast compared to the experimental flux measurements, across four different growth media.



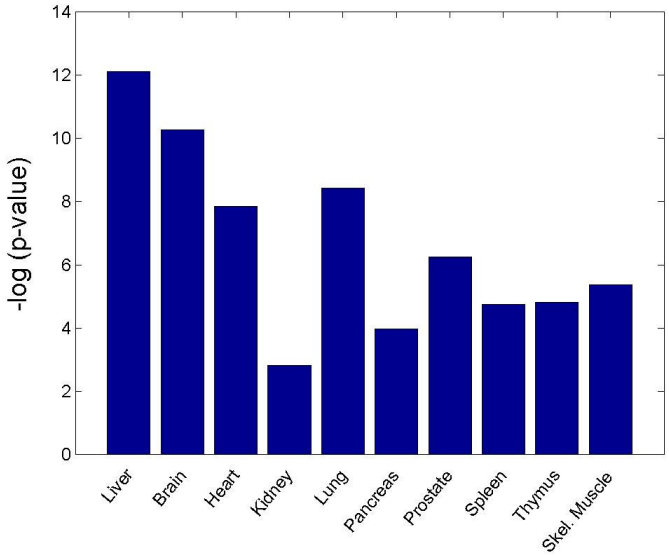
Supplementary Figure 4: Accuracy of predicted metabolic fluxes compared with FBA predictions. (a) Prediction accuracy over all reactions using our method. (b) Prediction accuracy of our method and of predictions that rely solely on expression data (i.e., without the network information), for the set of reactions that are associated with genes in the model. (c) Prediction accuracy of our method for reactions that are not associated with genes in the model (and hence their activity cannot be predicted solely based on expression data).



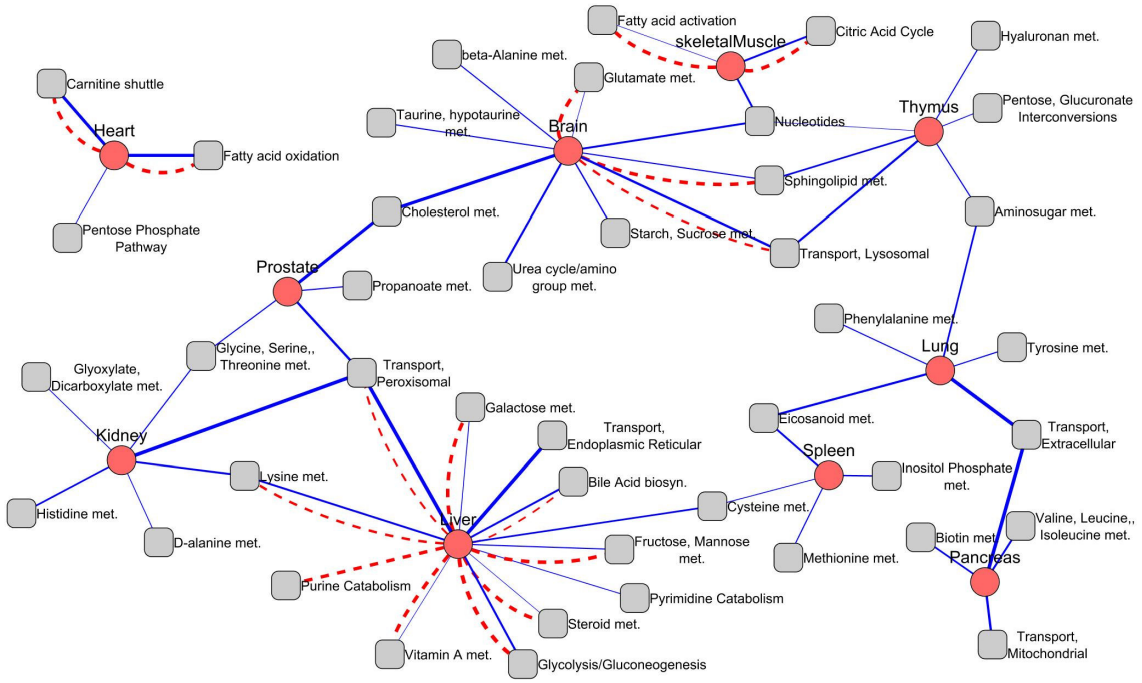
Supplementary Figure 5: Accuracy of predicted metabolic flux direction compared to FBA.



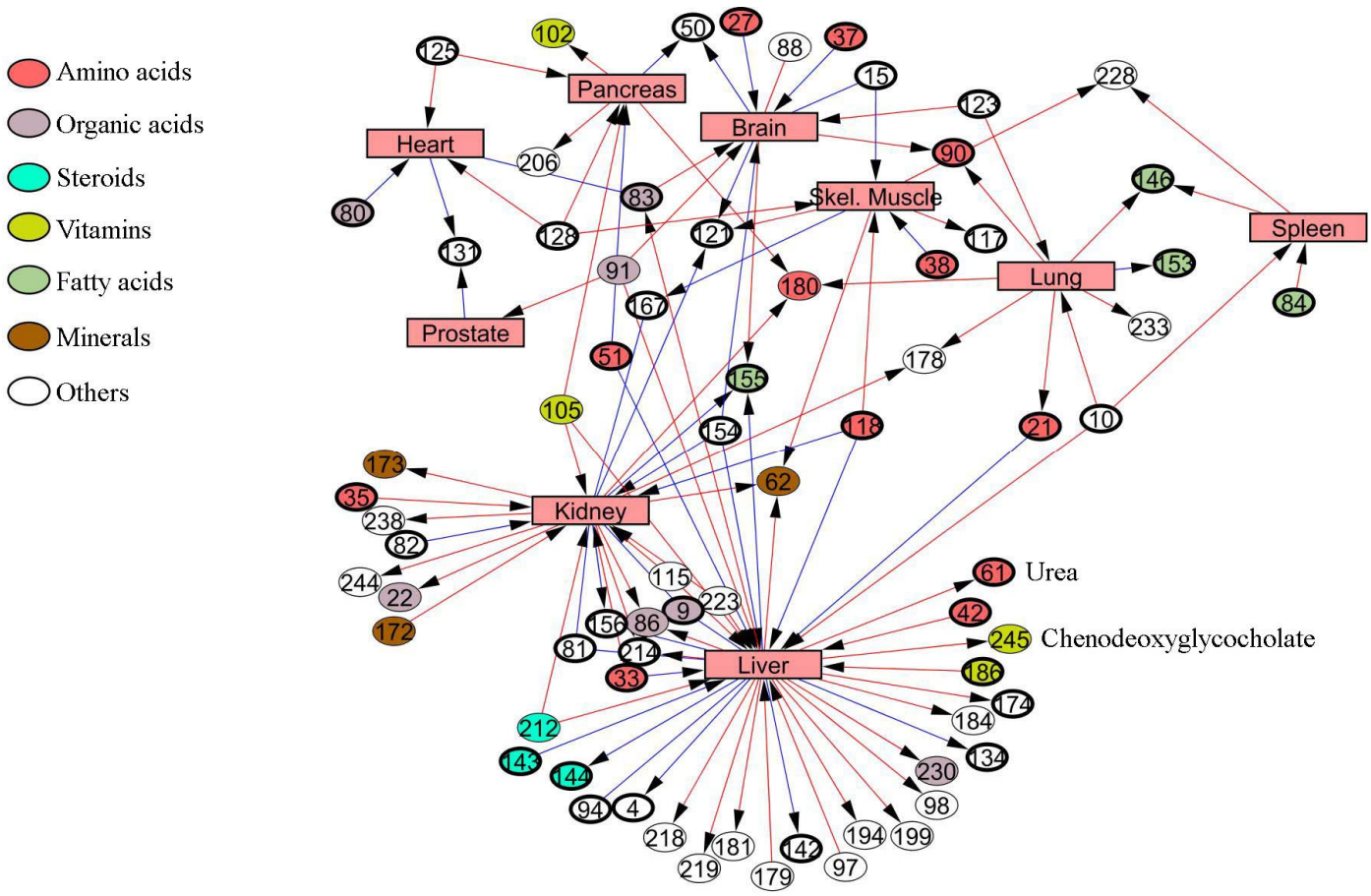
Supplementary Figure 6: (a) Accuracy of predicted metabolite exchange compared to FBA. (b) Fraction of biomass metabolites (as specified in FBA's growth reaction) that are predicted to be synthesized (without the explicit usage of a growth reaction).



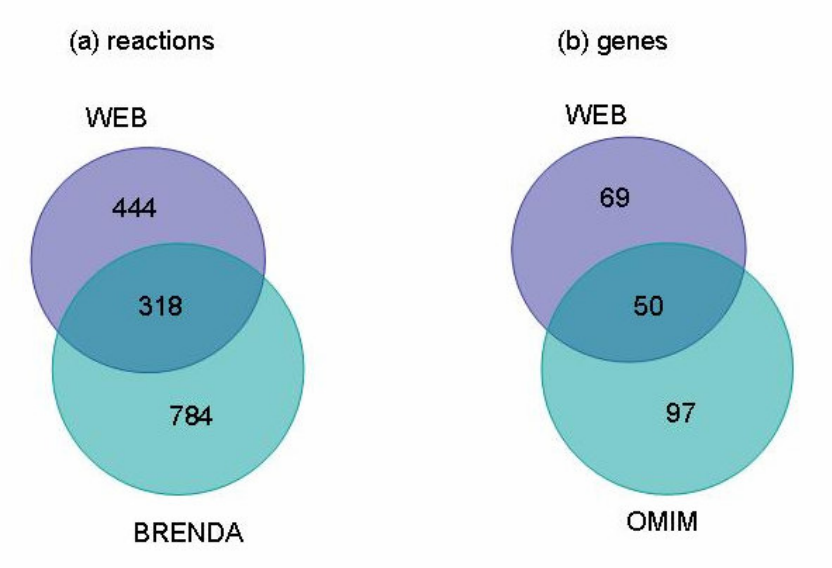
Supplementary Figure 7: The overlap (Hyper-geometric *p*-value) between the set of genes whose protein products are predicted to be metabolically active and the set of highly expressed genes in each tissue, based on the cross-validation test described in the main text.



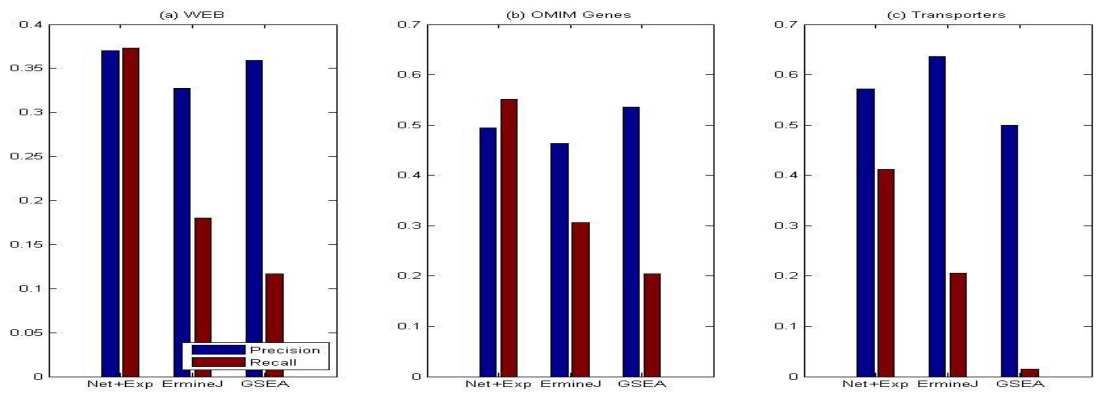
Supplementary Figure 8: A network representation of the metabolic subsystems with post-transcriptionally up-regulated genes in each tissue. Red nodes represent tissues and gray nodes represent metabolic subsystems. Blue edges represent model predictions, with the edge width proportional to the number of subsystem genes predicted to be post-transcriptionally up-regulated. Red edges represent significant tissue-subsystem associations derived from the web queries analysis, with edge width proportional to the number of co-occurrences on the web. The considerable match between the predictions and the tissue-subsystem associations is evident. Among the different tissues, the liver displays a high fraction of sub-systems whose activity is determined via post-transcriptional regulation.



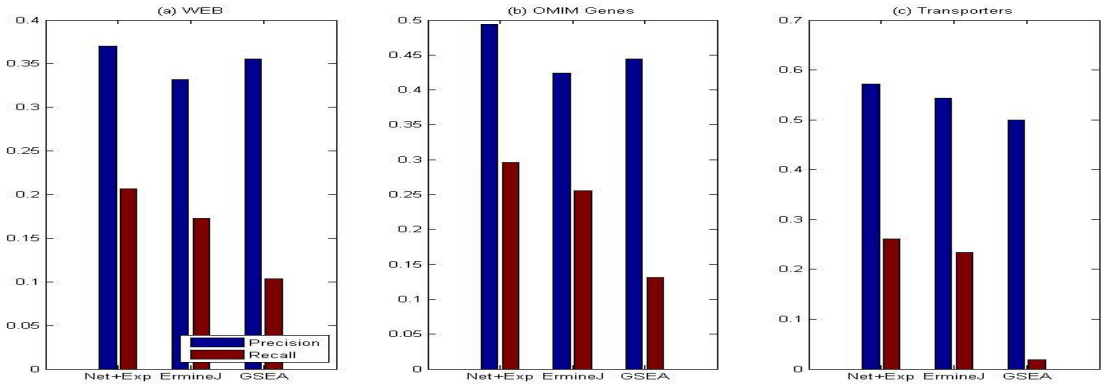
Supplementary Figure 9: A global map of tissue-specific metabolites exchange. Rectangular nodes represent tissues and circular nodes represent metabolites. Metabolites marked with a thick border are associated with known membrane transporters. Metabolites names (matching the indices in the circular nodes) are given in Supplementary Dataset 2. Edges represent a predicted metabolite exchange in a certain tissue. Directed edges represent specific predictions regarding metabolite uptake or secretion, while non-directed edges denote metabolites that could be either taken up or secreted. Blue edges represent metabolite exchange included in the expression data, and red edges represent model predictions that are not reflected in the expression data and require the integrated model. For clarity of exposition, the figure shows only metabolites associated with no more than three tissues.



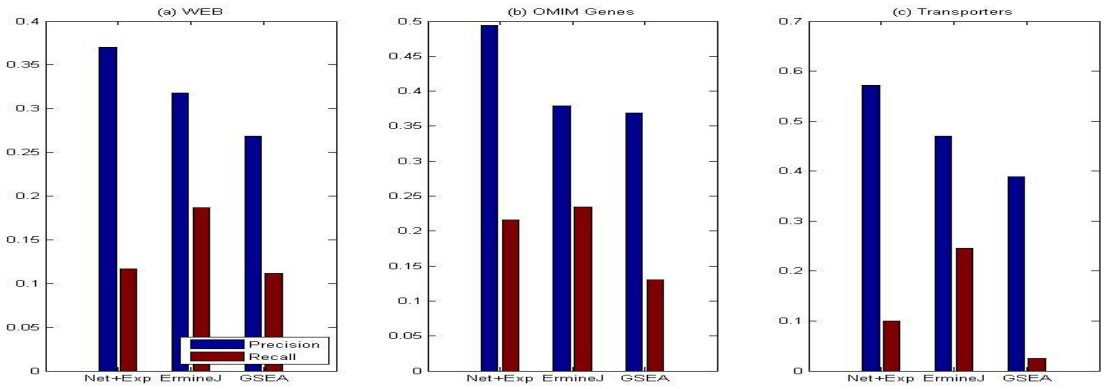
Supplementary Figure 10: Comparison of tissue-specificity data sources for genes and reactions.



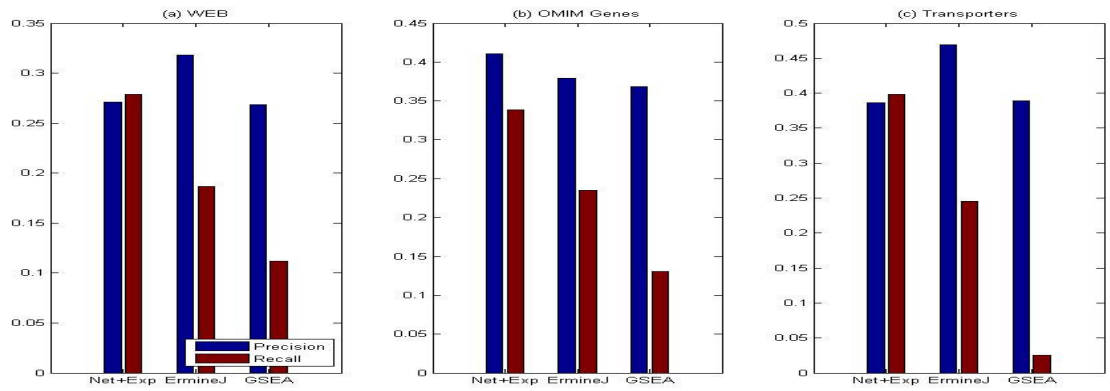
Supplementary Figure 11: Comparison of genes tissue-specificity predictions obtained by our method and with two functional enrichment-based methods, ErmineJ and GSEA. The prediction accuracy is computed over the set of genes that are predicted by our method to be active in at least a single tissue (see Supplementary Sections 3-4). The accuracy is calculated based on the tissue-specificity data described in Table 1 in the main text.



Supplementary Figure 12: Comparison of genes tissue-specificity predictions obtained by our method and with the functional enrichment-based methods, ErmineJ and GSEA. The prediction accuracy is computed over all metabolic genes in the model (excluding those on dead-end pathways and isozymes; see Supplementary Sections 3-4). The accuracy is calculated based on the tissue-specificity data described in Table 1 in the main text.



Supplementary Figure 13: Comparison of genes tissue-specificity predictions obtained by our method and with the functional enrichment-based methods, ErmineJ and GSEA. The prediction accuracy is computed over all metabolic genes in the model (excluding those on dead-end pathways; see Supplementary Sections 3-4). The accuracy is calculated based on the tissue-specificity data described in Table 1 in the main text.



Supplementary Figure 14: Comparison of genes tissue-specificity predictions obtained by our method and with the functional enrichment-based methods, ErmineJ and GSEA. The prediction accuracy is computed over all metabolic genes in the model (excluding those on dead-end pathways; see Supplementary Sections 3-4). The activity state of all genes that code for a certain isozymes are determined based on the activity state of the reaction catalyzed by the isozymes. The accuracy is calculated based on the tissue-specificity data described in Table 1 in the main text.

Noise in Boolean expression data	Error in predicted active reactions	Error in predicted inactive reactions
5%	0.9%	0.6%
10%	4%	7%
15%	8%	8%

Supplementary Table 1: Robustness of the predictions of reactions' activity to noise injected into the discretized expression data. Noise is simulated via random changes in the expression state of 5%, 10% or 15% of the genes. The analysis is performed using the liver tissue-specificity data.

Category	Precision	All genes		Potentially active		Predicted active in at least one tissue	
		Rec.	*p*-value	Rec.	*p*-value	Rec.	*p*-value
All genes	0.37	0.12	$1.3 \cdot 10^{-11}$	0.21	$1.6 \cdot 10^{-8}$	0.37	$2.6 \cdot 10^{-9}$
Disease genes	0.49	0.22	$1.1 \cdot 10^{-12}$	0.30	$2.1 \cdot 10^{-14}$	0.55	$<10^{-300}$
Transporter genes	0.57	0.10	$1.1 \cdot 10^{-8}$	0.26	$7.1 \cdot 10^{-9}$	0.41	$6.1 \cdot 10^{-7}$

Supplementary Table 2: Precision and Recall (Rec.) accuracy of gene tissue-specificity predictions computed over all genes in the model (excluding those on dead-end pathways), genes that may be potentially active (excluding dead-end pathways and isozymes), and genes that are predicted to be active in at least a single tissue (see Supplementary Section 3). The prediction accuracy is computed as described in the caption of Table 1 in the main text (Precision is equal in all sets by definition, while recall varies as shown).

Non-expressed reactions flux activity threshold	Deviations in predicted active reactions	Deviations in predicted inactive reactions
30%	0.04%	0%
60%	4.61%	4.77%
90%	4.2%	4.77%

Supplementary Table 3: Robustness of predicted reactions' activity to different thresholds on the flux activity of non-expressed reactions. Specifically, in the analysis presented in the main text, non-expressed reactions are considered to be inactive if they carry zero metabolic flux. The table shows that allowing non-expressed reactions to carry a small metabolic flux (i.e smaller than that used to determine the activity of expressed reactions) provides qualitatively similar results to those presented in the main text. Specifically, the method was applied using a flux activity threshold for non-expressed reactions that is 30%, 60% and 90% of that used for expressed reactions, and showed very small deviations from the predictions described in the main text. The analysis is performed using the liver tissue-specificity data.

References:

1. Duarte, N.C., Herrgard, M.J. & Palsson, B.O. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* **14**, 1298-1309 (2004).
2. Daran-Lapujade, P. et al. Role of transcriptional regulation in controlling fluxes in central carbon metabolism of Saccharomyces cerevisiae. A chemostat culture study. *J Biol Chem* **279**, 9125-9138 (2004).
3. Mahadevan, R. & Schilling, C.H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* **5**, 264-276 (2003).
4. Lee, H.K., Braynen, W., Keshav, K. & Pavlidis, P. ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics* **6**, 269 (2005).
5. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J.P. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23**, 3251-3253 (2007).
6. Shmueli, O. et al. GeneNote: whole genome expression profiles in normal human tissues. *C R Biol* **326**, 1067-1072 (2003).
7. Cline, M.S. et al. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2**, 2366-2382 (2007).