Reconciled Rat and Human Metabolic Networks for Comparative Toxicogenomics Analyses

A computational systems biology framework to facilitate preclinical drug development and biomarker discovery

A Dissertation

presented to the faculty of the School of Engineering and Applied Science

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

by

Edik Matthew Blais

May 2016

Department of Biomedical Engineering

University of Virginia

Reconciled Rat and Human Metabolic Networks for Comparative Toxicogenomics Analyses

---

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Edik Matthew Blais

May

2016

APPROVAL SHEET

The dissertation

is submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

AUTHOR

The dissertation has been read and approved by the examining committee:

Jason Papin, Ph.D.

Advisor

Jeffrey Saucerman, Ph.D.

Shayn Peirce-Cottler, Ph.D.

Young Hahn, Ph.D.

Christopher Holstege, M.D.

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, Dean, School of Engineering and Applied Science

May

2016

Contents

Introduction

Abstract

Rats serve an integral role in drug development and biomarker discovery, and understanding metabolic differences between rats and humans is critically important to minimize unexpected toxicities in clinical trials. Despite a high degree of physiologic and genomic similarities between rats and humans, several metabolic differences have been described that could affect whether a biomarker is elevated or whether a compound is toxic to the liver. A comprehensive knowledgebase of functional differences between rat and human metabolism would dramatically improve the translation of preclinical studies to human trials.

A genome-scale network reconstruction of metabolism serves as a repository for all known biochemical and transport reactions for an organism. In this dissertation, I have built the first genome-scale reconstruction of *Rattus norvegicus* metabolism, *iRno*, and a significantly improved reconstruction of *Homo sapiens* metabolism, *iHsa*. Comparative analyses with these models captured functional features known to distinguish rats from humans within purine, glycan, ascorbate, and bile acid metabolic pathways. Using reconciled biomass formulations, *iRno* and *iHsa* recapitulated realistic cellular growth rates under physiological constraints.

After extensive manual curation and network reconciliation, I demonstrated the use of *iRno* and *iHsa* in systems toxicology by generating biomarker predictions for rat and human hepatocytes treated with 76 pharmaceutical compounds and environmental toxicants from a comparative toxicogenomics database. I developed a novel gene expression integration algorithm to generate biomarker predictions that can be evaluated across metabolites, treatments, and organisms. Biomarker predictions were validated with literature-based evidence for antipyretic and antigout medicines. Comparative analyses provided mechanistic insights into the selection of metabolite biomarkers common to rats and humans. Using metabolomics and transcriptomics profiles from rat hepatocytes, I performed high-throughput validation of biomarker predictions. In the future, I anticipate that these models will serve as powerful computational platforms for contextualizing experimental data and making functional predictions consistent with rat and human biology for clinical and basic science applications.

Background and Significance

Pharmaceutical drug candidates often fail in preclinical or clinical trials due to unexpected drug-induced liver injury (DILI)[1]. The liver is a vital organ serving a variety of important functions in the body including carbohydrate, protein and fat metabolism, as well as diverse metabolic, vascular, immunological, secretory and excretory

functions. Several liver metabolic tasks are known to be disrupted in response to DILI, resulting in abnormal clinical chemistry values: gluconeogenesis (hypoglycemia), albumin synthesis (hypoalbuminemia), detoxification of ammonia (hyperammonemia), and bilirubin efflux (jaundice)[1-4]. In this dissertation research, we applied computational methods to investigate treatment-induced changes within rat and human metabolism.

Novel biomarkers of DILI are needed to facilitate the translation of preclinical studies in rats to clinical trials in humans. Candidate drugs are commonly screened in rats to identify potential biomarkers of therapeutic efficacy and toxicity. Traditional biomarkers used in preclinical drug development include proteins such as albumin, aspartate aminotransferase, and alkaline phosphatase and metabolites such as bilirubin, creatinine, and triglycerides[5]. Although these biomarkers can effectively diagnose DILI, traditional biomarkers are limited in scope to accurately predict future onset of DILI[6]. High-throughput metabolomics methods have the potential to significantly improve both diagnostic and predictive biomarkers of DILI by expanding the potential pool of biomarkers from a few dozen to thousands of metabolites[7-9]. In contrast to biomarkers like aspartate aminotransferase, which are released into the blood after hepatocellular membrane disintegration[6], we anticipate that metabolite biomarkers will require less catastrophic perturbations to be detected and will be more sensitive to the direct mechanisms of hepatocyte injury[6].

Preclinical efforts to identify biomarkers associated with toxicity typically use statistical methods that correlate changes in metabolite concentrations with toxicant exposure[7-13]. A major limitation of this conventional approach is that the metabolic response of rat hepatocytes will not necessarily reflect that of human hepatocytes. We anticipate that mechanistic modeling of metabolism with GENREs will substantially improve the ability to identify biomarkers that are predictive of toxicity. Human GENREs have provided mechanistic insights into metabolic biomarkers for inborn errors of metabolism and therapeutic strategies based on genomic alterations in cancers[14,15]. Constraint-based modeling algorithms have been successfully applied to human GENREs to study various aspects of human physiology and disease[16-21]. Other studies have utilized gene and protein expression measurements to create and analyze tissue-specific models[19,22]. A high-quality GENRE of rat metabolism would be a valuable resource for contextualizing high-throughput genomic datasets and bridging the knowledge gap that exists between humans and rats in clinical and basic science applications.

Using computational methods GENREs can be: 1) systematically perturbed using drugs and manipulation of growth conditions to predict essential or growth-reducing genes[23,24], 2) interrogated with high-throughput data to

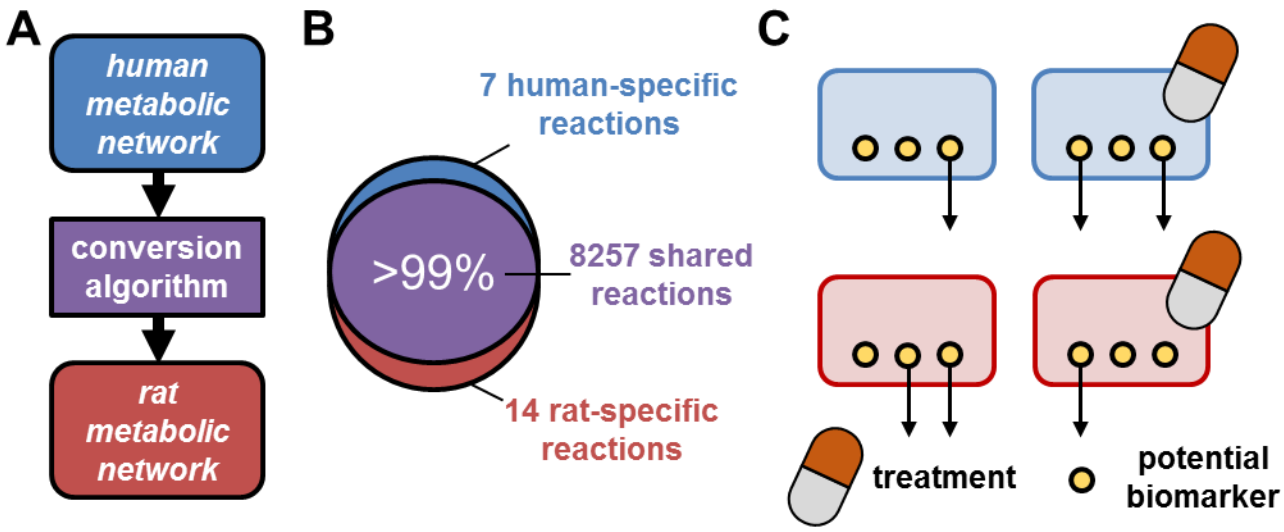
provide an understanding of metabolic off-target effects, 3) used to predict metabolite biomarkers indicative of the underlying metabolic state of a cell or tissue in response to an environmental toxin or the toxicological response to a drug compound[25]. Furthermore, with physiological constraints from experiments, GENREs can be used to make quantitative predictions of growth rates and byproduct secretion profiles[26,27].

Dissertation Aims

Specific Aim 1: Develop a method for constructing reconciled mammalian metabolic networks. We have constructed the first GENRE of *Rattus norvegicus* metabolism, *iRno*, based on a newly improved GENRE of *Homo sapiens* metabolism, *iHsa*. We developed an improved semi-automated method to generate a high-quality draft of *iRno* from *iHsa* using orthology annotations between human and rat genes (**Specific Aim A**). As a result, these draft GENREs were more suitable for comparative analyses and had fewer differences in the numbers of genes assigned to each reaction. We anticipate that this method will facilitate the construction of high-quality drafts for organisms.

Specific Aim 2: Capture known functional differences using rat and human metabolic networks. To better characterize the limitations of rats as a model of human biology, we assembled a compendium of known metabolic differences between rats and humans (**Specific Aim B**). We comprehensively captured known species-specific functions that could be distinguished within the computational frameworks of *iRno* and *iHsa*. As the first pair of reconciled mammalian GENREs, *iRno* and *iHsa* will be highly valuable resources for improving drug development from bench to bedside.

Specific Aim 3. Validate treatment-induced metabolic biomarkers predictions for hepatocytes. We developed a novel algorithm, TIMBR (Transcriptionally-Inferred Metabolic Biomarker Response), to predict potential biomarkers by integrating gene expression data into metabolic networks. We applied this algorithm to generate biomarker predictions in response to various treatments from a toxicogenomics database for rat and human hepatocytes which we validated with literature support (**Specific Aim C**). Additionally, we generated metabolomics and transcriptomics profiles of rat hepatocytes to perform high-throughput validation of TIMBR predictions. We anticipate that the computational framework developed here will provide mechanistic insights into hepatocyte metabolism and facilitate drug development and biomarker discovery.



Specific Aims – Dissertation Outcomes. (**A**) An automated method for converting a human GENRE (*iHsa*) into a draft rat GENRE (*iRno*). (**B**) A compendium of species-specific reactions that functionally distinguish *iRno* from *iHsa*. (**C**) A novel algorithm for predicting metabolic biomarkers in response to treatments using *iRno* and *iHsa*.

Dissertation Preview

In **Chapter 1** of this dissertation, we provided an overview of basic principles of genome-scale network reconstructions and demonstrated the use of metabolic modeling methods with a toy metabolic network of energy metabolism named Toycon1. In **Chapter 2**, we created a draft GENRE of rat metabolism (*iRno*) based on an existing human GENRE (*iHsa*) using a new metabolic network conversion method that maintains consistent network properties between the new and original models. In **Chapter 3**, we further resolved differences between *iRno* and *iHsa* through the process of manual curation and network reconciliation. In **Chapter 4**, we defined physiological constraints and comparative biomass formulations for rat and human hepatocytes and quantitatively captured realistic cellular growth rates and ATP yields with *iRno* and *iHsa*. In **Chapter 5**, we describe detailed methods for applying metabolic network modeling methods to guide genome annotation improvements and provide examples of annotation refinements for the rat genome based on *iRno*.

The reconstruction and reconciliation of rat and human metabolic networks described in **Chapters 1-5** provided a foundation for the comparative toxicogenomics analyses and biomarker predictions described in **Chapters 6-8**. In **Chapter 6**, we integrated gene expression data from a comparative toxicogenomics database into *iRno* and *iHsa* to better understand the functional response of hepatocytes to various compounds. In **Chapter 7**, we developed a new algorithm called TIMBR to generate biomarker predictions based on gene expression changes from comparative toxicogenomics data. In **Chapter 8**, we applied the TIMBR algorithm to validate biomarker predictions for rat hepatocytes using a high-throughput metabolomics dataset.

Chapter 1: Systems applications of metabolic networks

Synopsis

A genome-scale network reconstruction (GENRE) of metabolism consists of two major components: the stoichiometric matrix (S-matrix) and gene-protein-reaction (GPR) relationships. The S-matrix is comprised of stoichiometric relationships between metabolites and the reactions that occur in an organism (**Figure 1.1**). GPR relationships use Boolean logic to define rules between genes and reactions in the S-matrix (**Figure 1.2**). In **Chapters 2-5** of this dissertation, we describe the reconstruction of rat and human metabolic networks with species-specific GPR rules and a unified S-matrix that enables comparative analyses. Through the process of network reconciliation, we resolved differences between the numbers of genes in GPR rules for individual reactions and accounted for species-specific reactions within the S-matrix. In **Chapters 6-8**, we integrated comparative toxicogenomics data to analyze the physiological responses of rat and human hepatocytes to various pharmaceutical compounds and environmental toxicants (**Figure 1.3A**). To interrogate the use these models in biomarker discovery, we developed a novel algorithm to generate comparative biomarker predictions for rat and human hepatocytes (**Figure 1.3B**). We anticipate that these rat and human GENREs, collectively referred to as the Ratcon1 database, can improve the translation of preclinical studies into successful clinical trials. In this chapter, we provide an overview of fundamental methods, assumptions, and considerations that will be applied extensively throughout this dissertation. To promote the understanding of these methods in a biologically-relevant context, we created a toy metabolic network named Toycon1 that functionally captures central energy metabolism with a thousand-fold fewer reactions than Ratcon1.

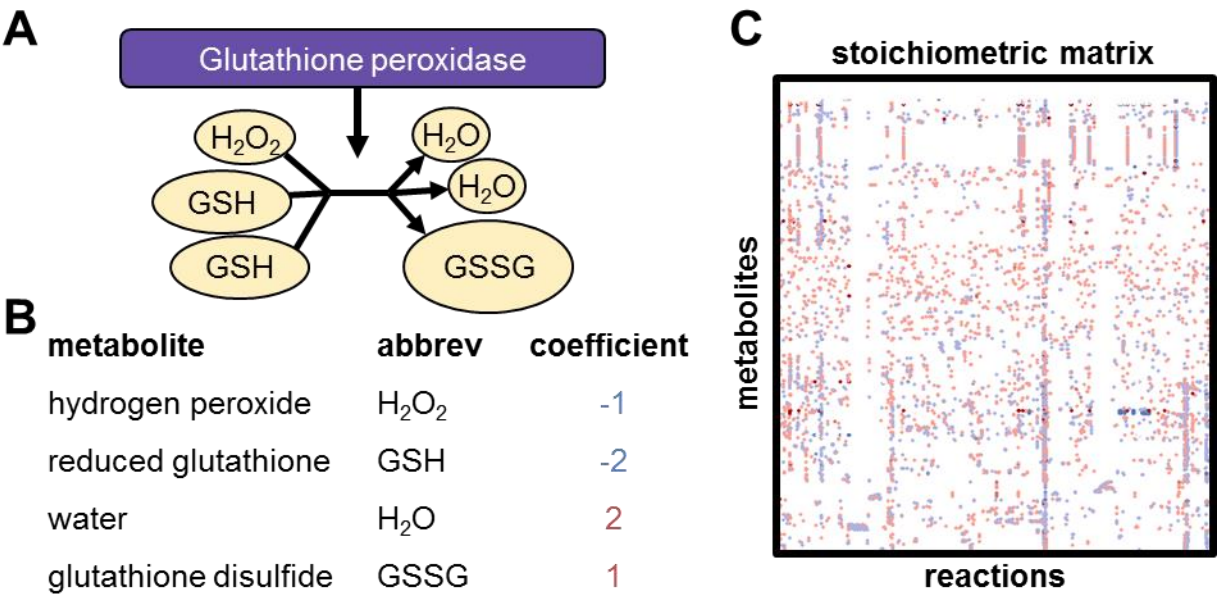


Figure 1.1 – The stoichiometric matrix describes relationships between reactions and metabolites. (**A**) Visual representation of the reaction, glutathione peroxidase, catalyzing the conversion of one unit of hydrogen peroxide and two units of reduced glutathione into two water molecules and one unit of glutathione disulfide (oxidized). (**B**) Stoichiometric coefficients describing the amounts of metabolites consumed (blue) and produced (red) by glutathione peroxidase. (**C**) The stoichiometric matrix (S-matrix) is a mathematical collection of reactions like the example in **A** formatted as a sparse matrix where each column represents a reaction and each row represents a metabolite. Each point in the sparse matrix accounts for the stoichiometric coefficient of each reaction-metabolite pair. The S-matrix in **C** represents thousands of reactions within the rat and human metabolic networks that will be discussed extensively throughout this dissertation.

Constraint-based modeling and flux balance analysis

Constraint-based methods have been developed to functionally interrogate metabolic networks. as mathematical models. A constraint-based method used extensively in the metabolic modeling community is flux balance analysis (FBA). By applying constraints to reaction fluxes known as reaction bounds, FBA determines the maximum possible flux through a specified reaction known as the objective function. Given a stoichiometric matrix (S), lower and upper bound (v_{lb} and v_{ub}) constraints for reaction fluxes (v), and an objective (v_{obj}), FBA uses linear programming software to solve the following optimization problem:

$$maximize\ v_{objective} \qquad \text{Equation 1.1}$$

$$v_{lb} \leq v \leq v_{ub} \qquad \text{Equation 1.2}$$

$$S \cdot v = \frac{dC}{dt} = 0 \qquad \text{Equation 1.3}$$

Equation 1.1 describes the objective function for the FBA problem, where $v_{objective}$ is the flux vector (v) multiplied by a vector of objective coefficients which are zero for all reactions except the objective reaction. As we will discuss in **Chapter 7**, advanced methods based on FBA principles can modify $v_{objective}$ to assign non-zero objective coefficients to multiple reactions at a time. **Equation 1.2** describes reaction constraints applied to the FBA problem. Arbitrarily large lower and upper bound values (e.g. -10^3 to 10^3 or -10^6 to 10^6) are typically assigned as constraints to all reactions except for exchange reactions. **Equation 1.3** describes the steady-state assumption made during FBA simulations which is critically important in formulating a linear programming problem that can be solved by optimization software. Kinetic parameters for enzyme reaction rates are not necessary to perform FBA because the rate of change in concentration for a metabolite with respect to time is zero. Thus, all FBA solutions provide feasible solutions to the problem described by **Equations 1.1**-**1.3** with requirement that metabolites cannot accumulate or be depleted. Furthermore, because of the law of mass action, all matter entering the metabolic network via metabolic exchange reactions must also exit the system in some form, assuming that no reactions create or destroy matter.

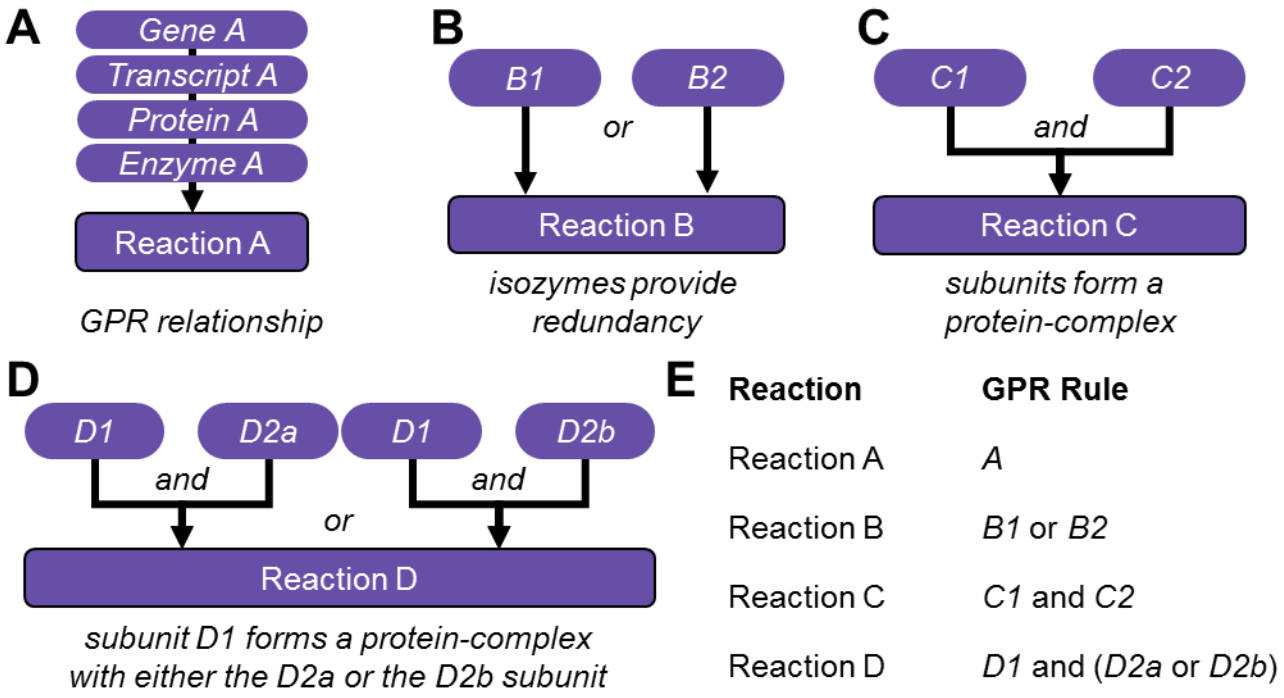


Figure 1.2 – Gene-protein-reaction (GPR) relationship rules describe the relationship between genotype and phenotype. (**A**) Example of a GPR rule representing an enzymatic reaction catalyzed by the protein-product of a single gene. Although GPR rules are represented with Entrez gene identifiers in this dissertation, Ensembl transcripts, UniProt proteins, and Enzyme Commission numbers would be suitable alternatives. (**B**) Example of a redundant GPR rule where either *B1* or *B2* can independently catalyze the same function. In this case, these isozymes are separated by an "or" statement in the GPR rule. (**C**) Example of a complex GPR rule where both

C1 and C2 are required for this catalytic reaction to occur. In this case, two non-redundant subunits that form a protein complex are separated by an “and” statement. (**D**) Example of a complex GPR with redundancies where D1 can form a protein complex with either D2a or D2b. In this case, the GPR rule can be separated by unique protein complexes or first by subunits then by redundancies as represented in **E**. (**E**) Table summarizing genotype-phenotype relationships from **A**-**D** as Boolean GPR rules.

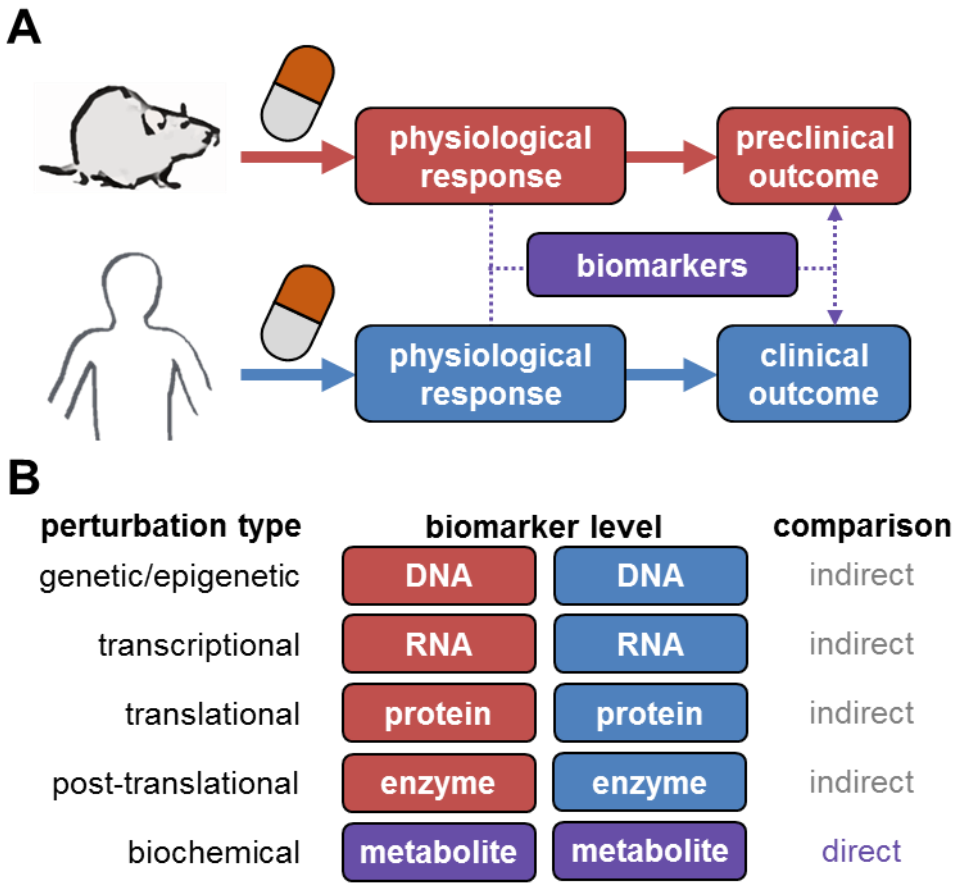


Figure 1.3 – A genome-scale rat metabolic network is needed to facilitate preclinical discovery of metabolic biomarkers. (**A**) Rats serve an integral role in preclinical drug development and biomarker discovery. Furthermore, understanding physiological responses to therapies in rats and humans can facilitate the rational selection of biomarkers that are indicators of a clinical outcome such as efficacy or toxicity. (**B**) Metabolic biomarkers generated by rats and human can be compared directly whereas genomic biomarkers rely on nucleotide and amino acid sequences that can vary across species and between individuals.

Toycon1: a simple metabolic network for educational purposes

We created a toy metabolic network reconstruction named Toycon1 that captures core aspects of central metabolism (**Figure 1.4A**). Toycon1 includes 4 metabolic reactions and 9 unique metabolites within cytosolic and mitochondrial compartments (**Figure 1.4B**). At the cellular boundary, five exchange reactions facilitate the consumption or secretion of glucose, lactate, O_2, H_2O, and CO_2. The first metabolic reaction, glycolysis, represents the net process of glucose fermentation to form lactate which also drives the phosphorylation of ADP to regenerate ATP. The second reaction, respiration, represents the net process of glucose oxidation via glycolysis and the tricarboxylic acid (TCA) cycle to form CO_2 and H_2O which also drives ATP regeneration and the shuttling of protons out of the mitochondria via the electron transport chain. The third reaction, ATP synthase, utilizes the proton gradient generated by the electron transport chain (ETC) to regenerate ATP. The fourth reaction, ATP demand, represents the process of spending energy stored as ATP via hydrolysis to ADP. The default objective of Toycon1 was set to maximize flux through the ATP demand reaction. It is important to note that the other three metabolic reactions facilitate the regeneration of ATP to fuel the ATP demand reaction. Throughout this chapter, we analyze Toycon1 to illustrate various methods that can be applied to metabolic networks. Unlike most toy networks based on hypothetical reactions, Toycon1 captures two commonly taught core biochemical pathways within central metabolism using stoichiometric relationships that are mass balanced, albeit in a simplified manner that bypasses many intermediate steps.

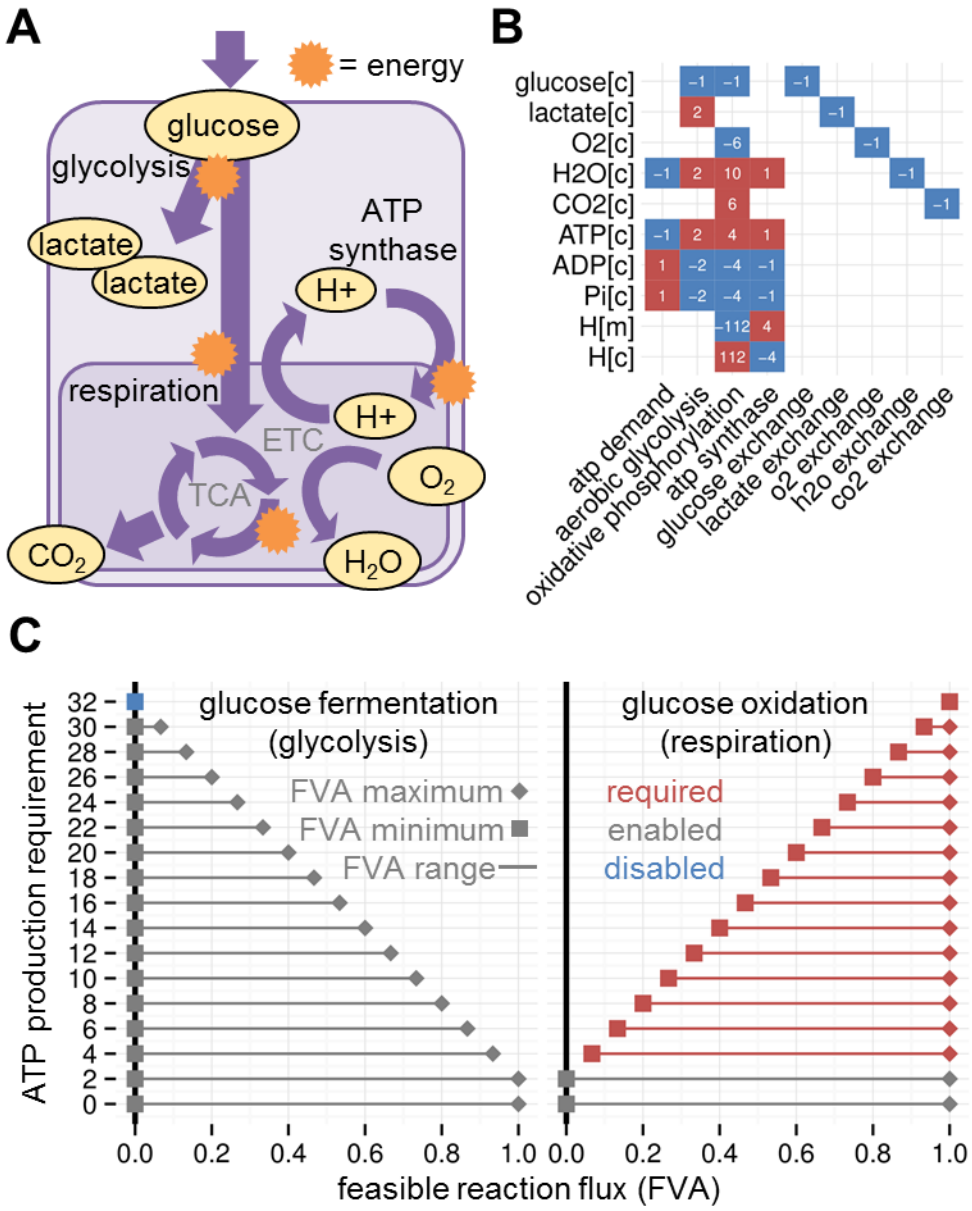


Figure 1.4 – Toycon1, a toy metabolic network that captures energy metabolism. (**A**) Toycon1 represents two major catabolic pathways for glucose that generate cellular energy via rephosphorylation of ADP to ATP: glucose fermentation to lactate via glycolysis and lactate dehydrogenase (LDH) and glucose oxidation to carbon dioxide via glycolysis, the tricarboxylic acid cycle (TCA cycle), and the electron transport chain (ETC). (**B**) Stoichiometric matrix for Toycon1, where each column represents a reaction and each row represents a metabolite. Within the matrix, negative values (blue) represent the relative amount of metabolite consumed and positive values (red) represent the relative amount of metabolite produced by each reaction. The first 4 columns represent biochemical reactions while the last 5 columns represent exchange reactions that allow metabolites to enter or leave the cell. (**C**) Flux variability analysis (FVA) was performed requirement increasing amounts of ATP demand flux along the y-axis. As ATP production yields requirements increased, maximum possible flux values for glucose fermentation decreased while minimum required flux values for glucose oxidation increased.

Recapitulating biological functions with metabolic tasks

Metabolic tasks can be formulated to capture a known biological processes such as glycolysis. In mammalian cells, glycolysis consumes glucose and produces pyruvate which can be secreted out of the cell as lactate. This process can be formulated as a metabolite task by allowing 1 unit of glucose to enter the metabolic network via the glucose exchange reaction and by setting the lactate exchange reaction as the objective function. Using FBA, we can predict that the maximum amount of lactate that can be produced by 1 unit of glucose. As a result, we found that 2 units of lactate can be secreted with 1 unit of glucose entering the system. To formulate a metabolic task based on this information, we can set exchange reaction parameters that allow uptake of 1 unit of glucose and require secretion of 2 units of lactate. This simple example highlights how glycolysis could be represented in the form of a metabolic task; however, we also know that cells often utilize glycolysis to produce energy. To more comprehensively capture glycolysis in the form of a task, we can also require minimum activity through the ATP demand reaction that consumes energy generated by glycolysis. Based on the directionality of each reaction in Toycon1, we can represent glycolysis as a task by setting the lower bound values of glucose exchange to -1, lactate exchange to +2, and ATP demand to +2. [19].

Metabolic tasks can also be used to specify functions that should not be possible to perform in a metabolic network. By modifying the glycolysis task described above to require 10 units of flux through the ATP demand reaction, Toycon1 would no longer be able to complete the task. If we relaxed the requirement for the secretion of lactate and allowed O_2 to enter the system, Toycon1 could redirect glucose to fuel the electron transport train that drives ATP synthase, producing more energy to satisfy the ATP demand requirement.

It is important to note that complete oxidation of glucose in the respiration pathway also requires secretion of CO_2 and H_2O. When using the RAVEN toolbox, this can be done explicitly by setting large positive upper bound values for CO_2 and H_2O exchange reactions or by specifying ALLMETS as an output parameter in the task. In **Table 1.1**, we provide examples of the types of parameters that can be used to formulate metabolic tasks related to glucose fermentation and glucose oxidation. Tasks that could not successfully be completed by Toycon1 were described as infeasible due to unrealistic expectations of energy production. Using Toycon1, we estimated that the maximum possible energy production yields given 1 unit of glucose were 2 in the absence of oxygen and 32 in the presence of oxygen. The yield for cellular respiration can be increased by adjusted by requiring fewer protons to drive the ATP synthase reaction based on the estimated efficiency of ATP synthase.

Table 1.1 – Toycon1 metabolic tasks that simulate cellular functions by specifying lower bound values (v_{lb}) and upper bound values (v_{ub}) for individual reactions. (*) Minimum required ATP yields for feasible metabolic tasks are sub-optimal (less than maximum). (**) Infeasible metabolic tasks require unrealistic ATP yields that should fail when simulated.

metabolic task	reaction	v_{lb}	v_{ub}
anaerobic glycolysis	glucose exchange	-1	∞
	lactate exchange	2	∞
	ATP demand*	1	∞
anaerobic glycolysis (infeasible)	glucose exchange	-1	∞
	lactate exchange	2	∞
	ATP demand**	10	∞
cellular respiration	glucose exchange	-1	∞
	O_2 exchange	-∞	∞
	CO_2 exchange	0	∞
	H_2O exchange	0	∞
	ATP demand*	10	∞
cellular respiration (infeasible)	glucose exchange	-1	∞
	O_2 exchange	-∞	∞
	CO_2 exchange	0	∞
	H_2O exchange	0	∞
	ATP demand**	100	∞

Genetic perturbations and gene essentiality

Constraint-based methods can be used to simulate the consequence of inhibiting an enzyme or deleting a gene within a metabolic network. Computational perturbations can be performed more rapidly than conventional experimental methods, providing the ability to screen identify all genes that are essential for a specific function. Using the ATP demand reaction as the objective, we simulated a genome-wide essentiality screen by deleting each reaction one-by-one. After each reaction was disabled, we performed FBA to determine the importance of that reaction on ATP production from one unit of glucose. Maximum possible flux values for the ATP demand

reaction after deleting individual reactions are shown in **Table 1.2**. Disabling glucose exchange completely blocked ATP production in Toycon1 which was expected considering glucose was the only fuel source entering the system. Inhibiting respiratory enzymes (TCA cycle and ETC) decreased the ATP production rate from 32 to 2. Although disabling the glycolytic fermentation pathway had no effect on ATP production, the importance of glycolysis in glucose catabolism via respiration was not captured because Toycon1 does not separate lactate dehydrogenase from the rest of glycolysis.

Table 1.2 – Maximum possible flux through the ATP demand reaction predicted by FBA after deleting individual reactions from Toycon1.

reaction deletion	ATP demand
glucose exchange	0
lactate exchange	32
O2 exchange	2
H2O exchange	2
CO2 exchange	2
glucose fermentation	32
glucose oxidation	2
ATP synthase	2
ATP demand	0

Metabolic networks enable systems-wide screening of genetic perturbations. Predicting the importance of individual reactions for a biological objective like ATP demand can be easily extended to the genome-scale with computational modeling. Furthermore, single, double, or triple knock-out simulations might reveal potential combinatorial effects that would otherwise be too resource-intensive to examine experimentally. None of the reactions in Toycon1 reduced ATP demand flux to 0 with the exception of glucose uptake and the ATP demand reaction itself (**Table 1.2**). To identify potential combinatorial strategies that abolish ATP production, we simulated a systems-wide double-knockout screen for 36 unique pairs of reactions. By comparing double versus single knockout predictions, we identified 10 double knockout strategies that reduced ATP production more than either single knockout simulation (**Table 1.3**). As expected, inhibiting critical components within both the glycolytic fermentation and the cellular respiration pathways more significantly reduced ATP production than the inhibition of either pathway alone. While the examples from Toycon1 were designed to be simple and intuitive to understand, predicting the consequences of single, double, and triple knockout strategies becomes exceedingly difficult for metabolic networks consisting of thousands of reactions.

Table 1.3 – Maximum possible flux through the ATP demand reaction predicted by FBA after performing a pairwise reaction deletion screen with Toycon1. Of the 36 possible reaction pairs (9 choose 2), 10 double reaction knockouts strategies that inhibited ATP production more than either single knockout strategy are shown.

deletion A	deletion B	A	B	A+B
fermentation	ATP synthase	32	2	0
fermentation	CO2 exchange	32	2	0
fermentation	H2O exchange	32	2	0
fermentation	O2 exchange	32	2	0
fermentation	respiration	32	2	0
lactate exchange	ATP synthase	32	2	0
lactate exchange	CO2 exchange	32	2	0
lactate exchange	H2O exchange	32	2	0
lactate exchange	O2 exchange	32	2	0
lactate exchange	respiration	32	2	0

Flux variability analysis

Flux variability analysis (FVA) is constraint-based technique that can be easily implemented in two steps. In the first step, FVA constraints a metabolic network to require a minimum amount of flux through an objective function, which is usually a percentage of the maximum value determined by running FBA. In the second step, FVA performs FBA to identify minimum and maximum flux values for other reactions in the network, providing feasible flux ranges that are consistent with the constraints imposed in the first step of FVA. Using the ATP demand reaction from Toycon1 as the primary objective (**Figure 1.4B**), we obtained a maximum value ATP turnover yield of 32 per unit of glucose entering the network. By requiring 16 units of flux (50% of the maximum value) through the ATP demand reaction, we determined feasible flux ranges for each of the 8 other reactions in the network. As a result, the fermentation reaction was capable of carrying between 0.000 and 0.533 units of flux while the respiration reaction was required to maintain flux values between 0.467 and 1.000. Interestingly, these values were not necessarily obvious despite the simplicity of Toycon1.

To further delineate the relationships between ATP production yields and glucose catabolic pathways, we performed FVA using incremental requirements of flux through the ATP demand reaction (**Figure 1.4**). As a result, we found that flux variability decreased for both glycolytic and oxidative catabolism of glucose when ATP yield requirements were larger than 2. As the ATP yield requirement increased to 32, the ability to convert glucose into lactate decreased to 0, completely disabling any activity through the glycolytic pathway. It is important to note that FVA can provide additional information compared to the methods described in the previous

section of this chapter where we mentioned that 32 flux through the ATP demand objective was feasible after knocking out the fermentation reaction (**Table 1.2**). In general, FVA can determine feasible flux ranges for any constrained network, providing the ability to explore other questions. For example, performing FVA after deleting the fermentation reaction would reveal which other reactions are affected by this knockout strategy. Overall, FVA is a powerful tool within the systems biologist's toolbox that can be used to interrogate the global impact of performing and/or perturbing metabolic functions.

Minimization of total flux principle

When maximizing an objective function, flux values returned for individual reactions from FBA are not necessarily useful because they represent a single possible solution of many feasible solutions that meet the maximum objective flux (**Figure 1.5A**). The constraint-based method called parsimonious flux balance analysis[28] (pFBA) can be used to identify a solution vector of reaction fluxes that meets the objective requirement and also meets a second requirement to further restrict the feasible solution space. Like FVA, pFBA first uses FBA to identify a maximum flux value through the objective and forces the objective to require at least a specified percentage (typically 80%) of the maximum. Second, pFBA minimizes the sum of fluxes across all reactions while also maintaining flux through the objective, providing a single value for the minimized total flux and a vector of fluxes that satisfy the solution (**Figure 1.5B**). Although multiple solutions can exist that satisfy the minimum total flux constraint, a pFBA solution vector represents the simplest solution that meets the demands of the original objective function whereas FBA solution vectors often include arbitrary values that may not necessarily be related to the objective function. Compared to FVA, pFBA provides a single vector of fluxes instead of minimum and maximum possible flux values that may not be informative if redundant routes are available.

FBA

$$\text{maximize } v_{objective}$$

s.t.

$$S \cdot v = 0$$

$$v_{lb} \leq v \leq v_{ub}$$

pFBA

$$\text{minimize } \sum v$$

s.t.

$$S \cdot v = 0$$

$$v_{lb} \leq v \leq v_{ub}$$

$$v_{objective\_pct} \leq v_{objective}$$

TIMBR

$$\text{minimize } \sum v \cdot w_{expression}$$

s.t.

$$S \cdot v = 0$$

$$v_{lb} \leq v \leq v_{ub}$$

$$v_{objective\_pct} \leq v_{objective}$$

Figure 1.5 – Comparison of optimization problems defined by FBA, pFBA, and TIMBR. (**A**) Flux balance analysis (FBA) determines the maximum feasible flux through an objective reaction ($v_{objective}$) while adhering to

upper (v_{ub}) and lower bound (v_{lb}) constraints for all reactions. In ToxExchange, default constraints include physiologically-relevant hepatocyte-specific flux ranges for metabolites that are available for consumption. (**B**) Parsimonious enzyme usage flux balance analysis (pFBA) first requires that the objective reaction is set to a percentage of the maximum possible value ($v_{objective\_pct}$) determined by FBA. Second, pFBA minimizes the sum of fluxes through all enzymatic reactions to obtain a value that can be interpreted as the total cost to perform a function (note that the model must be in an irreversible format so that no flux values are negative; this is equivalent to minimizing the sum of the absolute flux values). (**C**) TIMBR (Transcriptionally-Inferred Metabolic Biomarker Response) extends pFBA by assigning reactions weights based on gene expression changes ($w_{expression}$), which will be discussed in **Chapter 7**. Higher expression generally translates into a lower reaction weight, decreasing the contribution of flux through the reaction to the total sum of weighted fluxes. In some cases, pFBA can also incorporate reaction weights based on the gene length, minimizing the total amount of protein used, hence the name, parsimonious enzyme usage FBA. However, the novel feature introduced by TIMBR is the ability to summarize relative gene expression changes into reaction weights instead of absolute values like gene length or transcript counts per million.

Metabolic biomarker prediction strategies

Constraint-based modeling approaches have been developed that can predict known biomarkers for inborn errors of metabolism (IEMs) using previous human metabolic networks[14]. To simulate healthy conditions, FVA can be used to determine feasible flux ranges for exchange reactions while forcing non-zero flux through reactions (**Figure 1.6A**). To simulate metabolic deficiencies, FVA can be used to determine feasible flux ranges for exchange reactions while disabling genes associated with an IEM (**Figure 1.6B**). Directional changes can be determined by comparing how flux ranges changed between the two conditions. Although this strategy is useful for predicting biomarkers of IEMs, we are interested in predicting biomarkers in response to therapies with the rat and human metabolic networks discussed in this dissertation. With knowledge of specific genes targeted by a drug, FVA can be used to simulate biomarkers by treating drug targets like IEMs (**Figure 1.6C-D**). However, FVA-based biomarker predictions are not necessarily applicable to high-throughput toxicogenomics datasets when many genes are differentially expressed in response to a treatment, especially when mechanisms of action are not known.

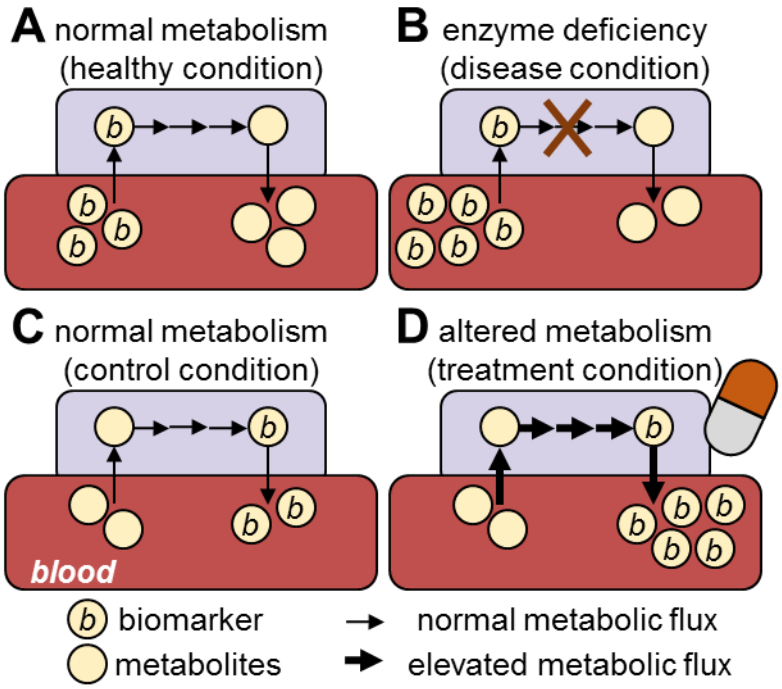


Figure 1.6 – Metabolic network modeling strategies for predicting biomarkers discussed in this dissertation. (**A**) In healthy patients, a potential biomarker metabolite is removed from the blood via a series of degradation reactions performed by the liver. (**B**) Accumulation of this metabolite in a patient's blood could serve as a biomarker of an inactivating mutation within the degradation pathway (or by a treatment that disrupts one or more steps in this pathway). (**C**) Under normal physiological conditions, a potential biomarker metabolite is produced in moderate quantities by the liver. (**D**) In response to a treatment, enzymes that synthesize this biomarker metabolite could be upregulated, increasing the production of this metabolite leading to accumulation detectable in the blood. Interestingly, biomarker levels were increased for both scenarios despite decreased enzymatic activity in **B** and increased enzymatic activity in **D**. It is important to note that levels of the non-biomarker metabolite in **D** will not necessarily change if its blood concentration is more tightly regulated than the biomarker metabolite.

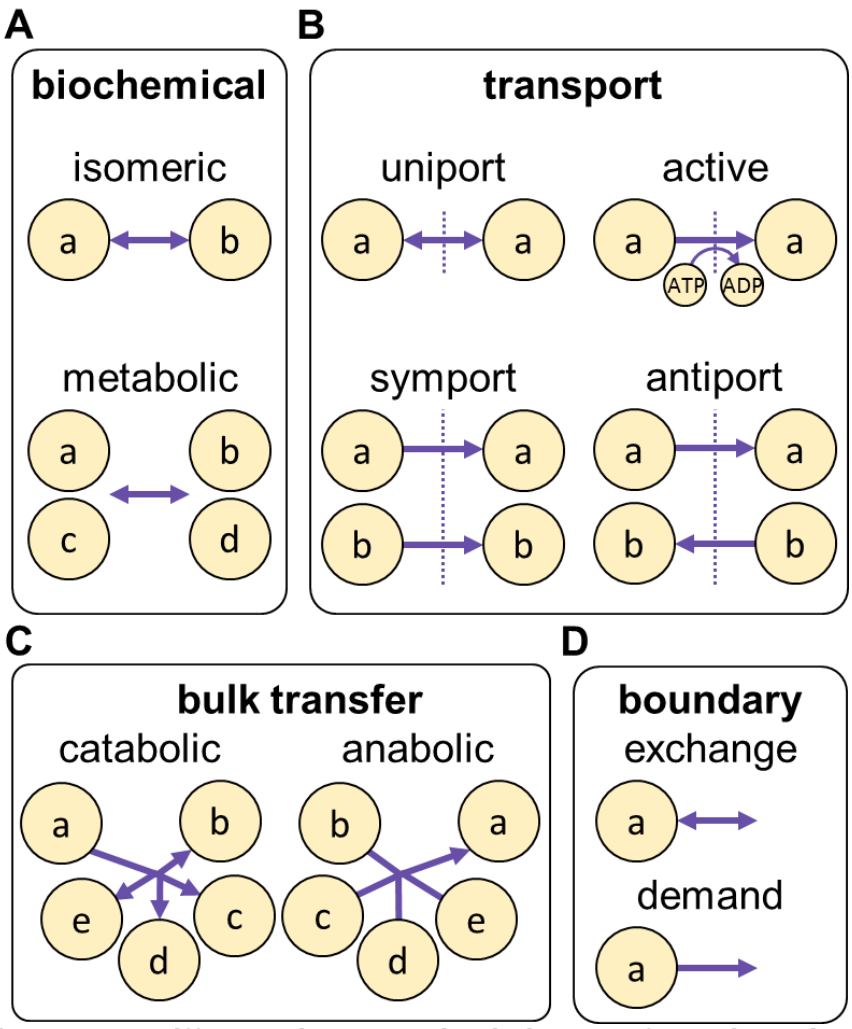


Figure 1.7 – Different classes and subclasses of reactions that can be included in a metabolic network. (**A**) The biochemical class of reactions facilitates that interconversion of metabolites that are not identical on both sides of the reactions. Metabolic reactions can be considered to include all biochemical reactions, although we added the isomeric subclass to describe reactions that involve the consumption of one metabolite to produce another metabolite. (**B**) Transport reactions include uniporters, symporters, antiporters, and active transporters. Although active transport reactions are technically biochemical reactions, we classify them as transporters due to their transport activity. (**C**) Bulk transfer reactions represent the aggregation of multiple metabolites into one metabolite or vice versa. Because these reactions typically do not occur in biology as a single step, bulk reactions are useful for summarizing relative ratios of metabolites such as the nucleotide composition of DNA as a component of biomass (see **Chapter 4** for biomass formulations that utilize bulk transfer reactions). (**D**) At the outermost boundary of the cell, exchange reactions represent the accrual or removal of metabolites from the system and are technically the reactions that should violate mass balance rules within the S-matrix. Boundary reactions also include demand reactions that function like exchange reactions but occur within cellular compartments. While exchange reactions can represent the uptake of nutrients and secretion of waste, demand reactions should only remove metabolites from the system in order to provide a sink for a metabolite that is consumed within the cell for a specific biological process such as cellular growth via the biomass sink reaction.

In **Chapter 7**, we describe a novel algorithm named Transcriptionally-Inferred Metabolic Biomarker Response (TIMBR) that expands on the minimization of total flux principle from pFBA to generate biomarker predictions. The primary objective of TIMBR is to calculate the minimum total flux needed to simulates the synthesis and secretion of a metabolic biomarker. By multiplying the reaction fluxes by reaction weights based on relative changes in gene expression (see **Figure 1.6C**), we can interrogate whether the minimum total weighted flux needed to produce a biomarker is increased or decreased in treatment versus control conditions. To integrate gene expression with TIMBR, reaction weights contribute a less to the total flux if genes are upregulated in the treatment condition. In **Chapter 7**, we discuss biomarker predictions generated by integrating comparative toxicogenomics data into rat and human metabolic networks with TIMBR. In **Chapter 8**, we integrate RNA-seq data in response to toxicants and attempt to validate biomarker predictions in a high-throughput manner.

Chapter 2: Reconstruction of a rat metabolic network based on a human metabolic network

Synopsis

In this dissertation, we present the first genome-scale network reconstruction (GENRE) of *Rattus Norvegicus* metabolism (*iRno*). In this chapter, we developed a computational method to generate a draft GENRE of rat metabolism based on a GENRE of *Homo sapiens* metabolism (*iHsa*). As a result, our approach provided initial drafts of *iRno* and *iHsa* that maintained consistent metabolic network properties that were introduced in the previous chapter. In the next chapter, we will describe the manual curation process used to reconcile differences between the initial rat and human metabolic networks created using our automated approach.

Prior to manual curation, we assembled a draft of *iRno* based on a draft of *iHsa* using orthology information (**Figure 2.1**). Inferring function through orthology can be difficult[29] when individual human genes are annotated to multiple rat orthologs (**Figure 2.2**). We developed a consensus approach to prioritize orthology annotations from five online resources: the Rat Genome Database (RGD)[30], Homologene[31], Ensembl[32], the Kyoto Encyclopedia of Genes and Genomes (KEGG)[33,34], and the Universal Protein Resource (UniProt)[35] (**Figure 2.3**). By prioritizing ortholog annotated in multiple databases, we assigned confidence to 4768 ortholog pairs between 2588 human genes and 2897 rat orthologs. Instead of using an arbitrary cutoff, we identified a minimal subset of orthologs that preserved the network properties of *iHsa* in the draft of *iRno*. Ultimately, we selected a subset of 2629 ortholog pairs between 2499 human genes and 2575 rat orthologs that were annotated in at least two databases (**Figure 2.3**). Without filtering orthology data, the draft rat network was difficult to compare with the original human network due to differences in the numbers of redundant enzymes associated with each reaction (**Figure 2.4**). The computational methods described in this chapter were critically important in preparing a draft of *iRno* that could be improved in parallel with *iHsa*, which will be discussed in **Chapter 3**. Our improved approach also provides a novel platform for the rapid generation of GENREs for other organisms that are used to study human diseases.

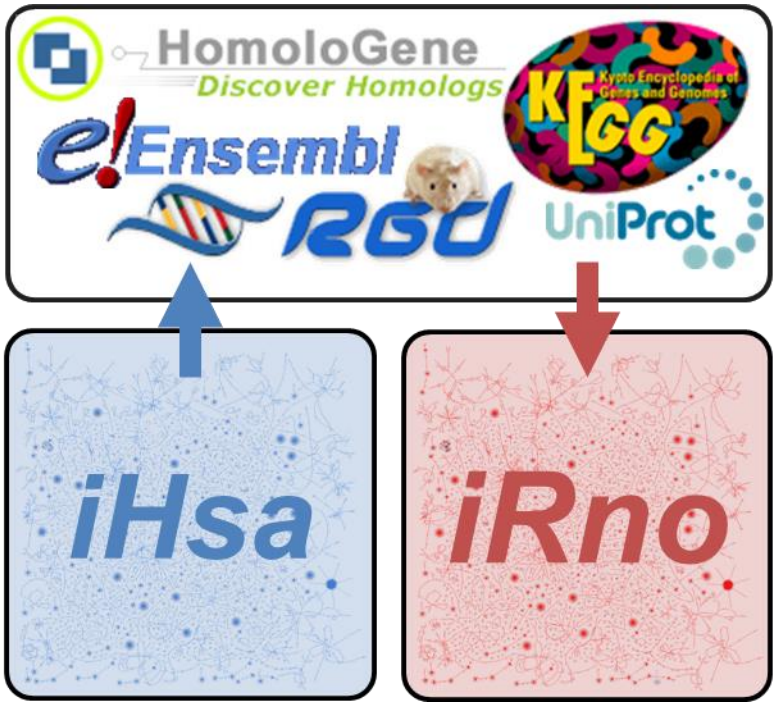


Figure 2.1 – The human metabolic network, *iHsa*, was adapted from existing human GENREs and then converted into a rat metabolic network, *iRno*, using orthology information. A consensus approach was developed using annotations from five orthology databases: Homologene, Ensembl, the Rat Genome Database (RGD), the Kyoto Encyclopedia of Genes and Genomes (KEGG), and the Universal Protein Resource (UniProt).

Table 2.1 – Comparison between rat and human genome sizes and characteristics. Statistics of GRCh38.p5 and Rnor_6.0 assemblies were obtained from Ensembl[32]. Gene counts and orthology statistics were obtained from orthology databases. Orthology statistics Abbreviations: Mb (Mega base pairs); GC content (guanosine cytosine percent content in genomic DNA).

	Human Genome	Rat Genome
chromosome pairs	23	21
genome size (Mb)	3547	3042
GC content	41.5%	42.3%
protein-coding genes	20,688	22,066
with orthology	94.9%	93.8%

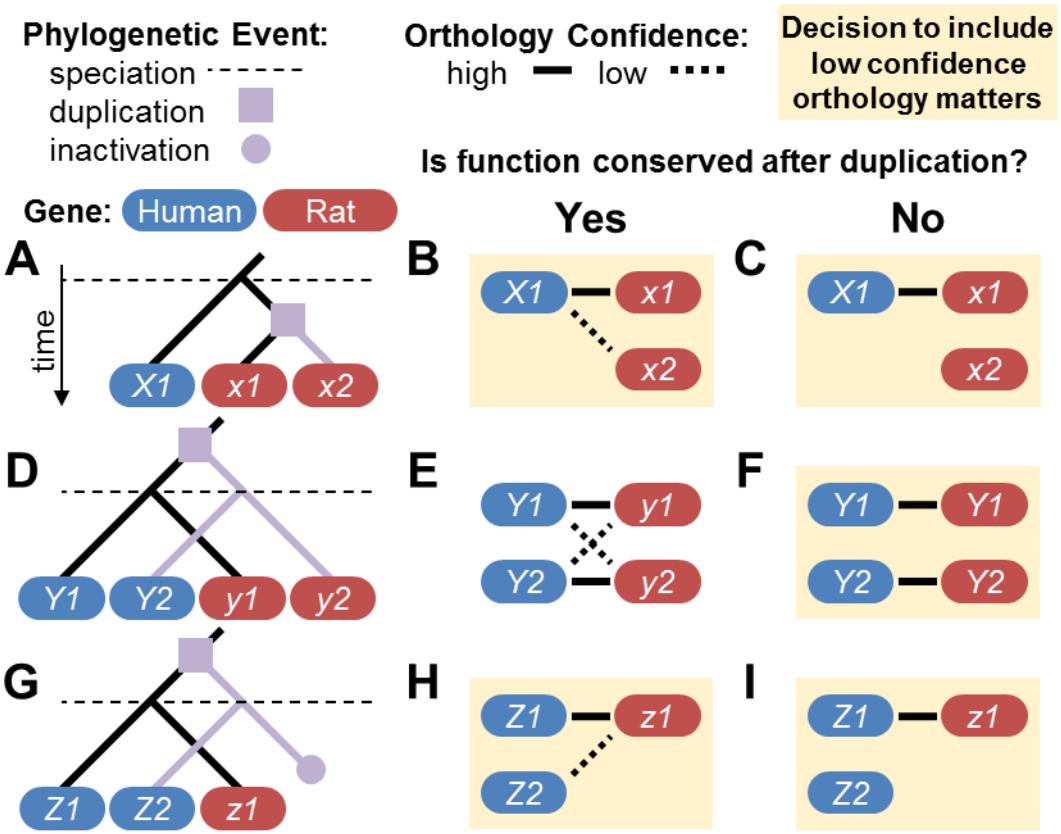


Figure 2.2 – Inferring function between orthologs is not trivial. (**A**) Example of the evolutionary history of a single gene in the most recent common ancestor of rats and humans that underwent a duplication event in rats but not humans after speciation. From this evolutionary relationship, *X1* may be annotated to two orthologs, *x1* and *x2*. Orthologous pairs of rat and human genes separated by shorter evolutionary distances were classified as high confidence and assigned the same number. (**B**) Assuming that function is conserved across *X1*, *x1*, and *x2* after speciation and duplication events, metabolic reactions associated with *X1* in a human GENRE should be associated with *x1* and *x2* as isozymes in a rat GENRE. This example highlights the importance of including multiple orthology annotations when converting GPR rules between species, even when *X1* and *x1* has stronger evidence for orthology than *X1* and *x2*. (**C**) Assuming *X1* and *x1* catalyze the same metabolic function but *x2* evolved an affinity for a different substrate after duplication, metabolic reactions associated with *X1* in a human GENRE should only be associated with *x1* and not *x2*. This example suggests that some orthology annotations may need to be discarded during the GPR conversion process (and potentially assigned to a new rat-specific reaction). (**D**) Evolutionary history of a single ancestral gene that was duplicated before speciation resulting in two human genes, *Y1* and *Y2*, and two rat genes, *y1* and *y2*. (**E**) Assuming that function is conserved across *Y1*, *y1*, *Y2*, and *y2* after duplication and speciation events, metabolic reactions associated with *Y1* and *Y2* as isozymes in a human GENRE should be also be associated with *y1* and *y2* as isozymes in a rat GENRE. (**F**) If the ancestral gene of *Y2*/*y2* evolved a novel function shortly before speciation and after duplication from the ancestral gene of *Y1*/*y1*, integrating low confidence orthology annotations between *Y1*/*y2* and *Y2*/*y1* into the GPR conversion process could generate GPR rules with twice as many rat genes as human genes. (**G**) Evolutionary history of a single ancestral gene that was duplicated before speciation resulting in two human genes, *Z1* and *Z2*, but only one rat gene, *z1*, after a loss of function mutation in the rat descendent of *Z2*'s

ancestral gene. (**H**) Assuming that function is conserved between *Z1*, *z1*, and *Z2*, metabolic reactions associated with *Z1* and *Z2* as isozymes in a human GENRE would only be catalyzed by *z1* in a rat GENRE. (**I**) If *Z1* and *Z2* were known to catalyze distinct reactions in a human GENRE, low confidence orthology annotations between *Z2*/*z1* might inappropriately suggest the addition of a human-specific reaction to a rat GENRE.

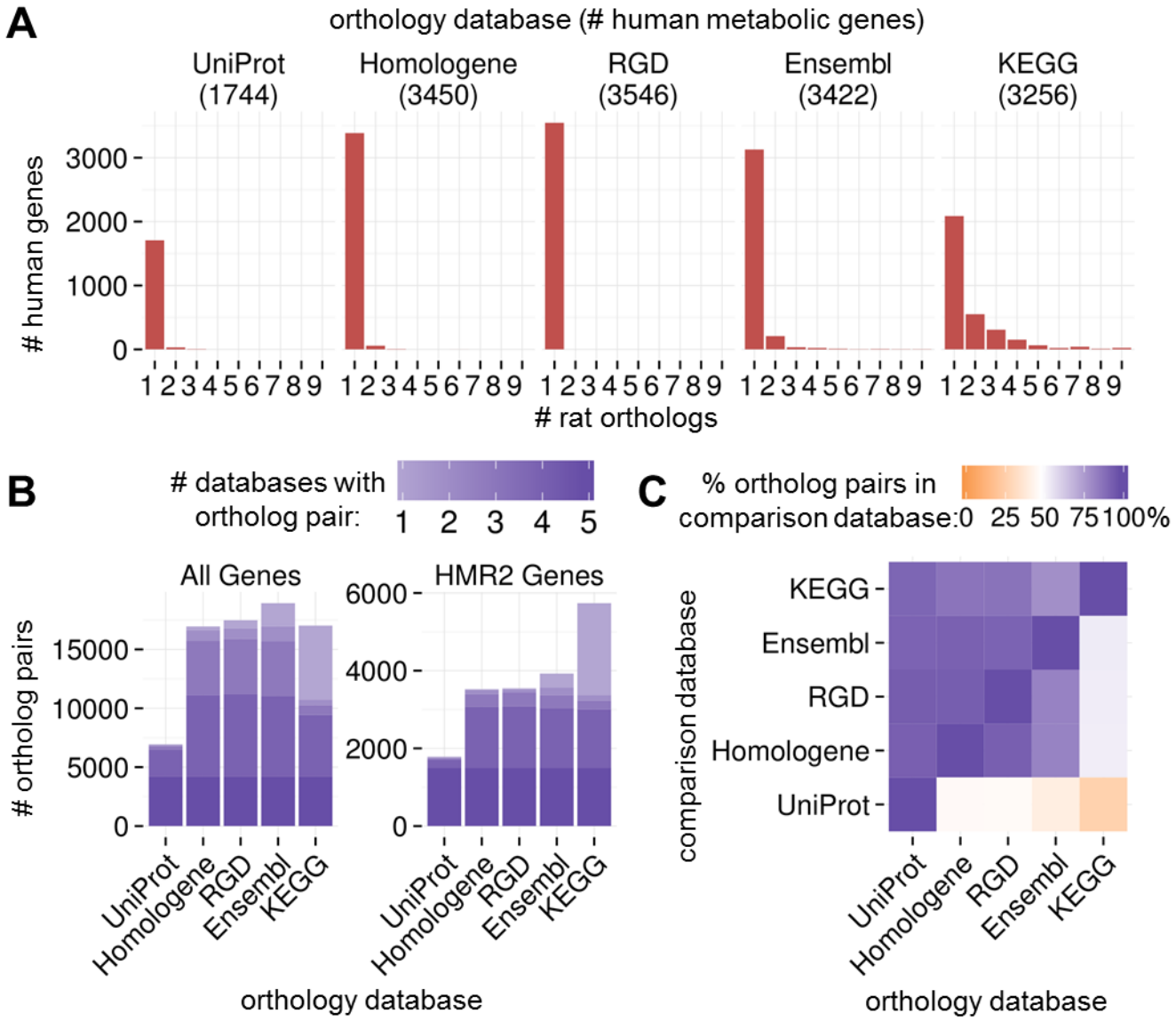


Figure 2.3 – Summary of orthology annotations between rat and human genes from five orthology databases. (**A**) Distributions of the numbers of rat orthologs annotated to individual human genes from each database. Numbers below each database name indicate the total numbers of human metabolic genes from HMR2 with at least one rat ortholog. Human genes with more than 9 orthologs are not shown. (**B**) Numbers of ortholog pairs from each orthology database that are also annotated in other orthology databases. Lighter and darker purple represent weaker and stronger consensus among databases, respectively. (**C**) Percent of ortholog pairs in each database (x-axis) that overlapped with in orthology annotations in other databases (y-axis).

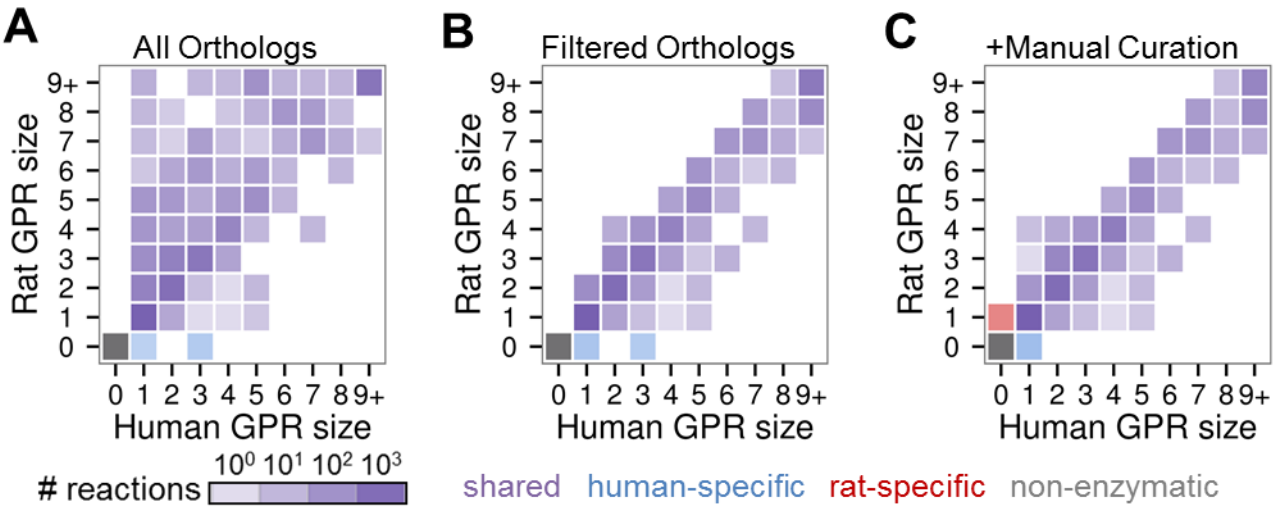


Figure 2.4 – GPR size comparison throughout the reconstruction process. (**A**) When using orthology annotations from all 5 databases, we observed that rat GPR sizes were frequently larger human GPR sizes, likely due false-positive annotations. (**B**) After filtering orthology information that preserved network properties between the two models, rat and human GPR sizes appeared much more balanced. (**C**) In **Chapter 3**, we further curated the draft networks generated in this chapter and added new rat-specific reactions (red) while still preserving balance across GPR rules.

Motivation to reconstruct the first genome-scale rat metabolic network

Rats serve an important role as model organisms in preclinical drug development and biomarker discovery[36-38]. Candidate drugs are routinely tested in rats to assess safety and efficacy prior to human clinical trials. Metabolomics profiling of rat serum and urine has been used to quantify potential metabolic biomarkers of drug activity or side effects seen in drug-induced liver injury (DILI) models[36-38]. Despite a high degree of genomic and physiologic similarities between rats and humans (**Table 2.1**)[39,40], functional differences within non-pharmacokinetic metabolism have been described that could influence whether a compound induces toxicity or elevates a biomarker[41-43]. Understanding species-specific differences between rats and humans will be important for the interpretation of preclinical animal studies in toxicology, drug development, and biomarker discovery[44,45].

A high-quality genome-scale network reconstruction (GENRE) of rat metabolism is needed to bridge the knowledge gap that exists between humans and rats in clinical and basic science applications. A GENRE acts as a repository for all known biochemical and transport reactions for an organism. Several GENREs with thousands of human genes have been published[16-19] while only core metabolic networks with dozens of genes are available for rat[46,47]. Human GENREs have been used to predict metabolic biomarkers for inborn errors of metabolism[14,17] (IEMs) and to analyze the metabolic effects of therapeutic strategies in the context of cancers,

toxicology, and diabetes[15,18-20]. Computational methods for integrating gene and protein expression measurements into GENREs have been developed to generate context-specific metabolic networks and enable comparative predictions across individual patients, treatment conditions, and tissue-types[19,22,48,49]. Furthermore, resolving metabolic differences between rat and human GENREs would enable cross-species comparisons as previously described for bacterial GENREs[50,51].

We have constructed the first GENRE of rat metabolism based on an existing GENRE of human metabolism. In this chapter, we developed a novel method to generate an automated draft of the rat metabolic network that maintains similar network properties as the initial human metabolic network. In the next chapter, we describe the iterative process of manually curating this new rat metabolic network in parallel with the initial human metabolic network. Unlike existing methods that have been applied to the reconstruction of mouse metabolic networks based on human models[52,53], the automated approach described in this chapter was designed to facilitate subsequent manual curation by minimizing unsupported differences and to improve the ability to make cross-species comparisons.

Survey of automated reconstruction approaches

Reconstructing a rat metabolic network from the bottom-up would be painstaking process at the genome-scale. One-by-one, reactions would need to be formulated as mathematical equations for the stoichiometric matrix and annotated with gene-protein-reaction (GPR) relationship rules consisting of rat enzymes using literature evidence. In **Chapter 1**, we provided an example of reconstructing a simplified toy network, named Toycon1, from the bottom-up. Although Toycon1 only included 4 metabolic and 5 transport reactions, formulating stoichiometric relationships between metabolites was time-consuming because metabolic processes are not necessarily described in literature as mass balanced reaction formulas. To overcome the technical hurdle of reformulating a reaction stoichiometry for each metabolic reaction in an organism, newer GENREs rely heavily on previously published GENREs and metabolic reaction databases. As we discussed in **Chapter 1**, reaction formulas can be adapted from one species to another because metabolic transformations typically involve metabolites found across many species. Even the first global reconstruction of human metabolism[16], *Homo sapiens* Recon 1, was built from the bottom-up after pulling the majority of its reaction formulas from a GENRE of *Saccharomyces cerevisiae* metabolism[54] and the KEGG database[34]. Overall, manually reconstructing

GENREs from the bottom-up by adding components one-by-one are intractable, especially for mammalian genomes that contain tens of thousands of protein-coding genes.

Computational methods have been developed to facilitate the initial reconstruction of a draft GENRE for an organism using a top-down approach. Fully-automated software such as Model SEED[55] and PathoLogic[56] in the Pathway Tools software suite assign genes from an annotated genome to reactions in a comprehensive reaction database such as MetaCyc[57,58]. With a top-down approach, a reaction can be removed if no genome sequences are similar to sequences from other organisms associated with that reaction. Gaps identified in the network can be filled by adding back reactions without GPR rules in order to maintain basic functionalities in the draft model. A limitation of using a top-down approach based on reaction databases is that differences in annotation quality between two similar organisms can result in substantially different draft GENREs, as recently described for two species of *Burkholderia*[51]. Automated methods are needed to facilitate the construction of a draft GENRE that are also robust to differences in quality across annotated genomes.

Outlook

In this chapter, we utilized an orthogonal approach to generate a draft GENRE of rat metabolism based on an existing GENRE of human metabolism. An orthogonal approach is similar to existing approaches used to establish mouse and human GENREs where reactions can be transferred from existing GENREs and/or reaction databases [59,60]; however, an important distinction is that our orthogonal approach is based on the assumption that network properties should be consistent between GENREs of similar species. By maintaining consistent network properties, we avoided annotating too many or too few genes to reactions shared by the draft GENRE of rat metabolism and the original GENRE of human metabolism. With the explicit intention to reconcile differences between rat and human GENREs, we avoided introducing unnecessary differences that would need to be manually resolved in the network reconciliation process, which will be discussed in **Chapter 3**. In this chapter, we provide detailed methods for generating draft metabolic networks based on existing models using an orthogonal approach that we used to generate initial draft GENREs of rat and human metabolism discussed throughout this dissertation. We also describe important considerations that motivated the development of the orthogonal approach by analyzing existing pairs of mouse and human metabolic networks.

Converting human GPR rules into rat GPR rules using orthology annotations

The first step in transforming a human GENRE into a rat GENRE involved assigning GPR rules consisting of rat genes to reactions associated with human GPR rules (**Figure 2.5**). Metabolic networks are typically comprised of two fundamental components: the stoichiometric relationships between metabolites and reactions, termed the stoichiometric matrix or S-matrix, and Boolean relationships between enzymes that catalyze a reaction, termed gene-protein-reaction GPR association rules (**Figure 2.5A**). Reactions and metabolites in the S-matrix were assumed to be organism-independent because the molecular building blocks of a cell such as amino acids, nucleic acids, and membrane lipids are generally consistent across multiple species. In contrast, enzymes catalyzing reactions within the GPR rules of a human GENRE were specifically encoded by the human genome, necessitating distinct formulations of GPR rules for the rat genome.

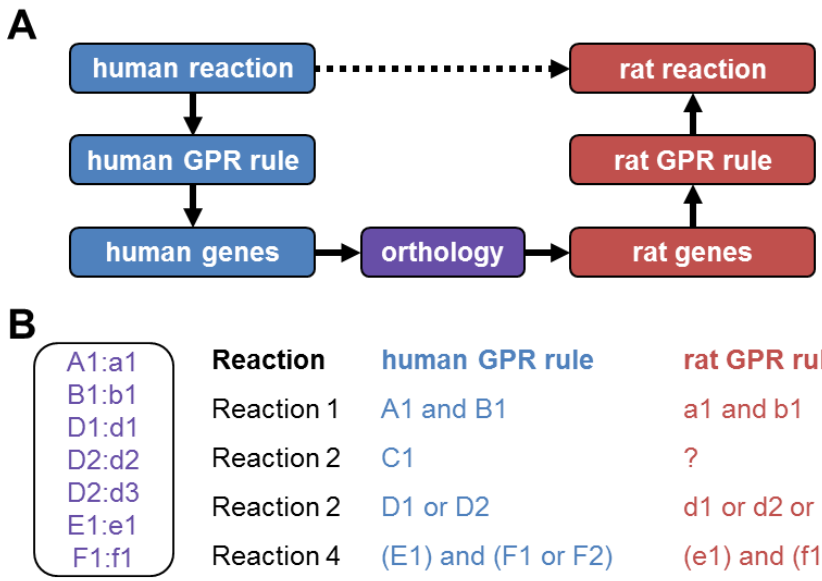


Figure 2.5 – Orthology information can be used to transform a human GENRE into a rat GENRE. (**A**) Reactions can be reassigned after transforming human GPR rules into rat GPR rules. Human genes can be replaced by known rat orthologs to reconstruct a rat GPR rule, assuming that function is conserved across orthology. (**B**) Orthology annotations describe individual pairs of human and rat genes that can be used to convert human GPR rules into rat GPR rules. GPR rules for 4 hypothetical reactions are shown as examples to highlight how rat GPR rules may include different numbers of genes compared to the original human GPR rule. In the first reaction, the rat genes, *a1* and *b1*, each map to one of two human genes, *A1* and *B1*, which encode distinct subunits in a multiprotein complex. In the second reaction, the human gene *C1* does not have any known rat orthologs, which might suggest that rat cells cannot carry out this function. In the third reaction, the rat GPR rule included more redundant genes than the original human GPR rule because the human gene, *D2*, has more than one known rat ortholog. In the fourth reaction, the human gene, F2, did not map to any known rat orthologs; however, the rat GPR rule can be considered functional because *F1* and *F2* redundantly encode the same component within the multiprotein complex.

To reconstruct a rat GPR rule based on an existing human GPR rule, human genes assigned to a reaction were replaced with orthologous rat genes. Inferring metabolic function through orthology is not trivial because orthologs descended from the most recent common ancestor of rats and humans have endured more than 50 million years of evolutionary pressures (**Figure 2.2**). Additionally, mutations involving the duplication and/or inactivation of gene after speciation can lead to one-to-many, many-to-many, or many-to-one orthology annotations between rat and human genes[61] (**Figure 2.5B**). If none of the human genes were annotated to any rat orthologs, the reaction was disabled in the rat GENRE and flagged as a potential human-specific reaction. Individual human genes mapped to multiple rat orthologs were replaced with multiple rat isozymes joined by "or" relationships. In **Figure 2.5**, we provide examples of how the number of genes in a GPR rule can change after replacing human genes with rat orthologs.

Survey of mammalian metabolic networks

Several published metabolic networks were considered for the basis of *iHsa* and *iRno*. These included the Human Metabolic Reaction 2.0 database (HMR2)[19], *Homo sapiens* Recon 1[16] and Recon 2[17], HumanCyc[57], and Hepatonet1[20]; ultimately, the largest network, HMR2, was chosen for its inclusivity. Additionally, HMR2 was capable of performing 256 well-curated metabolic tasks relevant to hepatocyte metabolism in both humans and rats. Despite the availability of multiple generations of human GENREs (**Figure 2.6A**), only metabolic networks representing core metabolic pathways have been described for rats[46,47]. Constructing *iRno* based on a mouse metabolic network[52,53,62] was also considered; however, each mouse model was derived from one of the human models described above (**Figure 2.6**).

A draft of *iHsa* was adapted from HMR2 by replacing human GPR rules consisting of Ensembl gene identifiers with protein-coding Entrez genes. Ensembl genes without equivalent protein-coding Entrez genes were discarded prior to GPR conversion. Additionally, several reactions with excessively large GPR rules such as generic protein kinase reactions were simplified or discarded to streamline the GPR conversion process. Prior to GPR conversion, annotations mapping a human gene to 10 or more rat orthologs were excluded to avoid inferring function from non-specific annotations. A major limitation of HMR2 was the lack of complex GPR rules where "and" relationships between genes can be used to describe subunits in a multiprotein complex. In the next chapter, we address this limitation by manually curating complex GPR rules for *iRno* and *iHsa*.

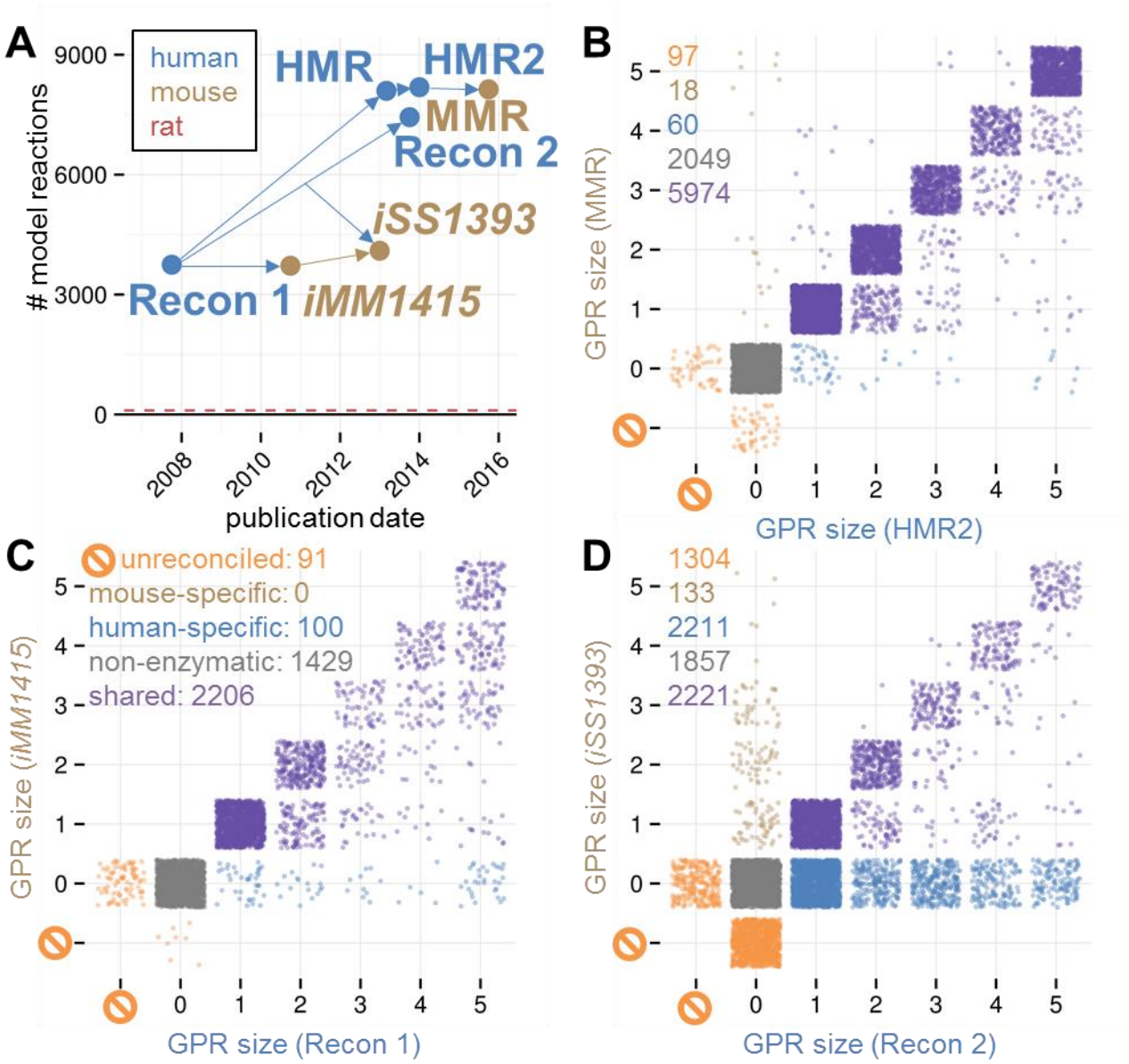


Figure 2.6 – History and comparability of mammalian metabolic networks. (**A**) Historical timeline shows the size and relationships between human metabolic networks (blue) and mouse metabolic networks (gold). HMR and Recon 2 were both derived from Recon 1 but published by separate groups. Each mouse model was based in part on a human model. Although Recon 2 was published after *iSS1393*, *iSS1393* uses similar reaction names as the published version of Recon 2 that were not used by Recon 1. *iSS1393* also included reactions that were unique to *iMM1415* but not originally reconciled with Recon 1. (**B-D**) Comparisons of GPR sizes (number of unique genes in a GPR rule) between pairs of human and mouse metabolic networks. Unreconciled reactions (orange) represent reactions that were non-enzymatic in one species and either enzymatic or absent in the other. Species-specific reactions were enzymatic in one species and absent in the other, which is the only appropriate representation for a reconciled difference.

Comparing mouse and human metabolic networks

We compared mouse and human GENREs to estimate the degree of consistency we should expect from rat and human GENREs after automated conversion (**Figure 2.6**). Surprisingly, we found significant numbers of unreconciled differences between the mouse models and relevant human models (**Figure 2.6A**). Unreconciled differences are defined as reactions that were non-enzymatic in one species and either enzymatic or absent in the other. The fewest discrepancies were found between HMR2 and the newest mouse metabolic reaction (MMR) GENRE (**Figure 2.6B**), where over 98% of enzymatic reactions were classified as shared. Recon 1[16] and *iMM1415*[53] did not include any properly assigned mouse-specific reactions (**Figure 2.6C**) because reactions newly added to *iMM1415* lacked valid mouse GPR rules[62]. For the second generation mouse GENRE, *iSS1319*, most of unreconciled differences between *iMM1415* and Recon 1 were resolved[62]; however, many of the changes incorporated into Recon 2 after Recon 1 were not applied to *iSS1319* (**Figure 2.6D**). Because *iSS1319* utilizes naming conventions from Recon 2 instead of Recon 1, comparative analyses between *iSS1319* and either Recon 1 or Recon 2 cannot be considered reliable.

Despite the fact that HMR2 and MMR shared a relatively higher proportion of reactions compared to previous pairs of human and mouse GENREs, we observed that mouse GPR rules were often smaller than human GPR rules for shared reactions (purple dots below the diagonal in **Figure 2.6B**) and that human-specific reactions were more common than mouse-specific reactions. These trends were consistent between other human and mouse GENREs, where only a handful mouse GPR rules from *iMM1415* and *iSS1319* were ever larger human GPR rules in Recon 1 (**Figure 2.6C**) or Recon 2 (**Figure 2.6D**). Based on these empirical observations, we would expect that a GENRE of rat metabolism would also have fewer redundant isozymes for shared metabolic functions and fewer unique functions overall compared to a GENRE of human metabolism.

To test the hypothesis that rats have fewer redundancies than humans for shared metabolic functions, we compared the numbers of rat and human genes annotated to reactions in the KEGG database[34]. Using a similar approach used to compare GPR sizes between GENREs, we found that KEGG reactions were annotated to relatively proportional numbers of rat and human genes (**Figure 2.7**). While some KEGG reactions were often associated with fewer rat genes than human genes, other KEGG reactions were associated with more rat genes than human genes, contradicting our preliminary observations. These results suggested that discrepancies between rodent and human GPR sizes were likely an artifact of the reconstruction process used to generate

mouse GENREs based on human GENREs. We suspected that rodent GPR rules were smaller in previous mouse GENREs due to limited availability of orthology annotations between mice and humans. To overcome this potential limitation when generating a rat GENRE based on a human GENRE, we developed a novel consensus approach to utilize orthology annotations from five orthology databases instead of choosing a single database.

Inferring metabolic function through orthology annotations is not trivial

Aggregating orthology annotations from multiple databases increases the risk of inappropriately replacing human genes with rat orthologs that do not perform the same function. A consensus approach was used to filter out low quality orthology annotations during the GPR conversion process. Each orthologous pair of rat and human genes was assigned a score of 1-5 corresponding to the number of databases in which that pair was annotated. Individual genes were also assigned confidence scores determined by protein-level evidences and annotation scores from Ensembl[32] and Uniprot[35]. For human genes mapped to multiple rat genes, orthologs were prioritized first by database scores then confidence scores to assign orthology ranks. Sensitivity analysis was performed to explore how filtering out ortholog pairs based on different cutoff values for database scores and orthology ranks affected the distributions of rat and human GPR sizes. Ultimately, a subset of 2629 ortholog pairs were selected that were annotated in at least 2 of 5 databases and limited each human gene to a maximum of 2 orthologs. This filtering step was important because methods that integrate gene expression data or simulate the impact of genomic alterations rely heavily on the number of redundant enzymes associated with a reaction.

Converting all orthology annotations present in any of the 5 orthology databases generated rat GPR rules with disproportionately more genes compared to the original human GPR rules (**Figure 2.4A**). This result directly contradicted our previous observations that rodent GPR rules should generally be smaller than human GPR rules (**Figure 2.6**) and was inconsistent with our observation that rat and human genes annotated to reactions in the KEGG database were relatively proportional (**Figure 2.7**). Based on these seemingly conflicting observations, we discovered that too many or too few orthology annotations could result in the over-assignment or under-assignment of rat genes to GPR rules. To avoid these potential pitfalls, we developed a consensus-based strategy that filters low quality orthology annotations from multiple databases and maintains moderately proportional rat and human GPR sizes after replacing human genes with rat orthologs (**Figure 2.4B**).

Furthermore, humans and rats have similarly sized genomes with approximately 20,000 genes, so genome-scale properties were assumed to be consistent unless literature evidence suggested otherwise.

Survey of five distinct orthology databases

We compared orthology databases to highlight the advantages and disadvantages of using a consensus approach versus an individual database (**Figure 2.3**). Surprisingly, the distribution of rat orthologs annotated to each human gene varied substantially between orthology databases (**Figure 2.3A**). Over a third of human genes were annotated to two or more rat orthologs in KEGG while RGD was restricted to one rat ortholog per human gene. Despite this limitation, RGD had the highest coverage of human genes with orthology annotations compared to KEGG which had the second least. UniProt covered the fewest human genes but most orthology annotations were consistently found in at least 4 of 5 orthology databases (**Figure 2.3B**). Based on this information, orthology annotations from UniProt alone might not be sufficient to carry out the GPR conversion process at the genome-scale; however, UniProt could be useful in a consensus-based GPR conversion method by reinforcing confidence in a core subset of well-annotated ortholog pairs. Most ortholog pairs were annotated in multiple databases although KEGG also included a large number of unique ortholog pairs (**Figure 2.3B**). Each database added between 12 and 514 unique human genes originally represented in HMR2 and between 16 and 2371 unique ortholog pairs not found in any of the other four databases. The percentage of non-overlapping ortholog pairs between any two databases was less than 50% with the exception of UniProt (**Figure 2.3C**). Despite a moderate degree of overlap, these results suggested that no consensus has been established among orthology annotation resources. As an alternative to choosing a single orthology database, a consensus approach would reduce the number of unconverted human genes and potentially capture more evolutionary differences between rats and humans (**Figure 2.2**).

An optimization algorithm to preserve complexity of rat and human GPR rules

To construct a rat metabolic network from a human metabolic network (**Figure 2.1**), human GPR rules were converted into rat GPR rules by replacing human genes with known rat orthologs (**Figure 2.5**). Pairs of orthologous genes between humans and rats were downloaded from five separate genomics databases: the Rat Genome Database (RGD)[30], Ensembl[32], the Kyoto Encyclopedia of Genes and Genomes (KEGG)[33,34], the Universal Protein Resource (Uniprot)[35], and Homologene[31]. Each individual database contained varying amounts of orthology information and unique orthology pairs, suggesting a lack of consensus (**Figure 2.3**). All five

databases were utilized in the GPR conversion process because inferring function through orthology can be especially difficult when considering genes with multiple orthologs (**Figure 2.2**).

A consensus approach was used to assign a confidence score to each pair of human and rat orthologs for initial construction of the rat GENRE, *iRno*. We developed a quantitative method to infer metabolic function by incorporating the collective efforts of multiple genome annotation communities: UniProt, Homologene, RGD, Ensembl, and KEGG. Using these five orthology databases, we assigned a confidence score to deprioritize the conversion of ortholog pairs that were annotated in fewer databases. This prioritization step was implemented to filter out low confidence orthology annotations due to the possibility that function may not be conserved as described in **Figure 2.2**. We anticipate that using orthology annotations derived from multiple computational methods will be more robust than quantitative methods such as BLAST because a single point mutation could be sufficient to alter the basic function of a metabolic enzyme while orthologs with low sequence similarity can catalyze similar reactions[61].

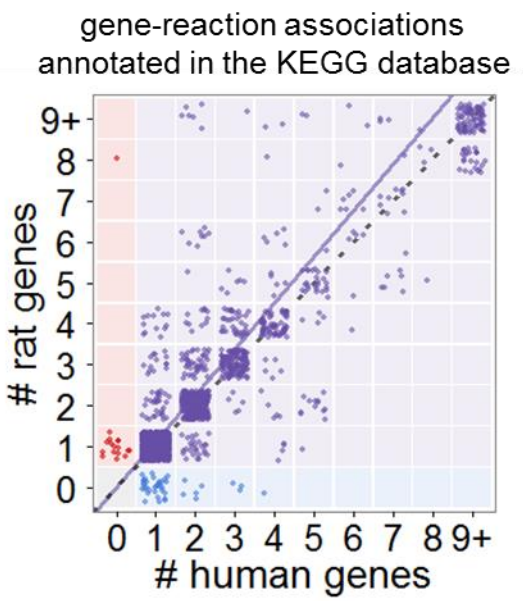
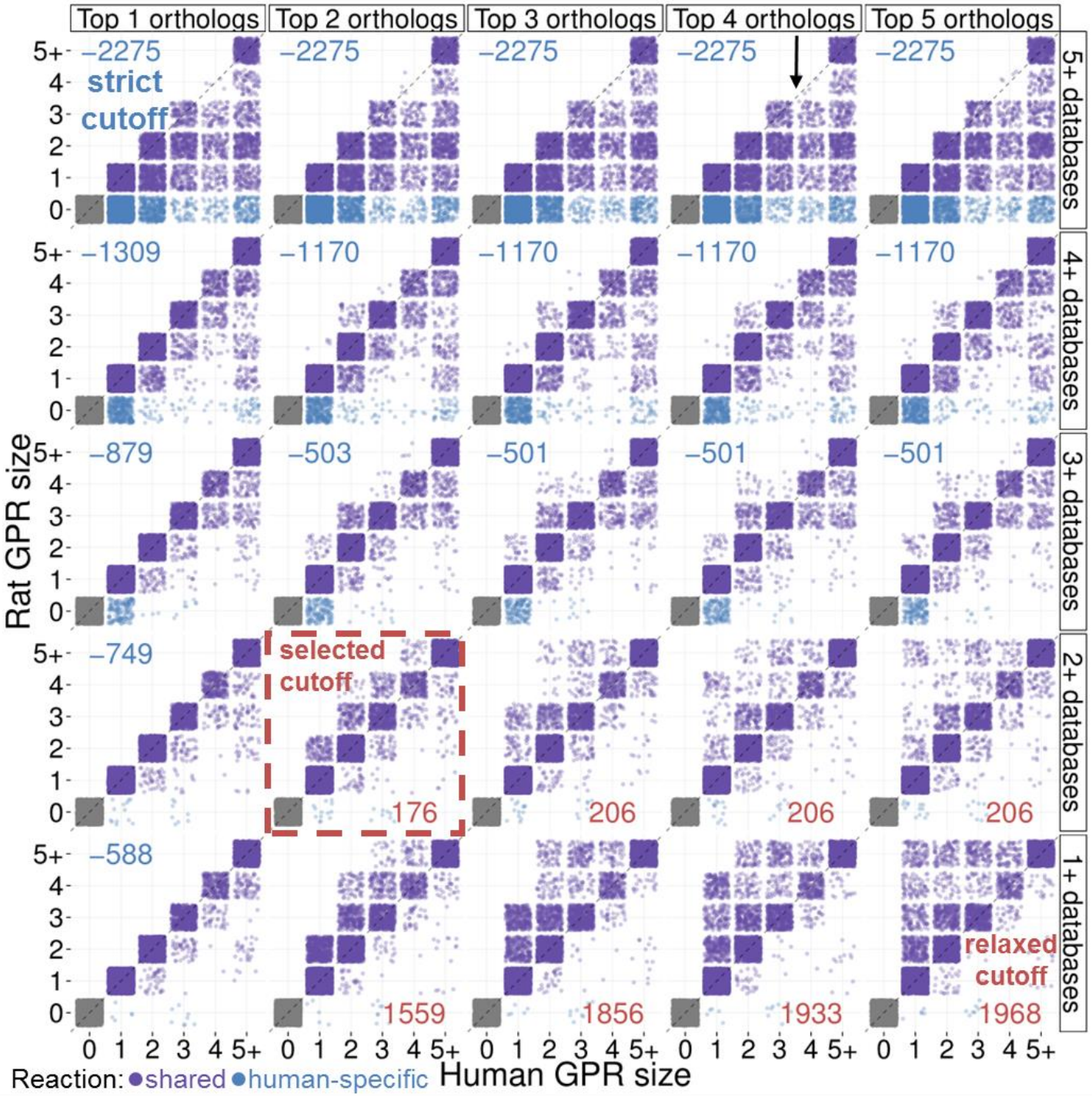


Figure 2.7 – Comparison between the numbers of rat and human genes associated with reactions in the KEGG database. Each dot represents a single KEGG reaction annotated to human and/or rat genes. The solid line shows the least squares regression fit with a slope of 1.1 which deviated slightly from the dashed unit line.

Aggregating orthology annotations from multiple databases increases the risk of inappropriately replacing human genes with rat orthologs that do not perform the same function. To identify a high-quality subset of orthology information that preserved functionality and GPR sizes between drafts of *iRno* and *iHsa*, we developed a consensus-based GPR conversion algorithm that required orthologs to be annotated in at least 1-5 databases and limited individual human genes to be replaced by a maximum of 1-5 rat orthologs (**Figure 2.8**). We found that converting all orthology annotations present in any of the 5 orthology databases generated rat GPR rules with disproportionately more genes compared to the original human GPR rules (bottom row of panels in **Figure 2.8**). We assumed that this difference was more likely explained by a large number of false positive orthology annotations than an actual genome-scale

difference in the redundancies between rat and human metabolic enzymes. Alternatively, requiring orthologs to be annotated in all 5 databases generated much smaller rat GPR rules for shared reactions and introduced a large number of human-specific reactions that would substantially reduce the functionality of *iRno* (top row of panels in **Figure 2.8**). We found a balanced relationship between the numbers of reactions with larger rat GPR rules than human and numbers of reactions with larger human GPR rules than rat by removing orthology annotations found in only one database and restricting the replacement of each human gene to two rat orthologs (selected cutoff panel in **Figure 2.8**). Additionally, the selected cutoff preserved the same functionalities as the relaxed cutoff when evaluating metabolic tasks from HMR2[19]. This filtering step was important because methods that integrate gene expression data or simulate the impact of genomic alterations rely heavily on the number of redundant enzymes associated with a reaction[63,64].

Figure 2.8 – Converting GPR rules using a consensus approach. Using a consensus approach, a subset of high-quality orthology annotation from 5 databases was obtained to generate an automated draft of a rat metabolic network based on a human metabolic network. Sensitivity analysis of two parameters used to filter orthology annotations on the relative sizes of rat and human GPR rules: minimum database count (by row) and maximum orthology rank (by column). Smaller orthology rank thresholds limited fewer rat orthologs to be replaced by individual human genes. Larger database count thresholds removed ortholog pairs that were not annotated frequently across multiple databases. Dots represent the relative rat and human GPR sizes for individual reactions and the number highlighted in each panel represents the number of shared reactions with larger rat than human GPR sizes minus the number of shared reactions with larger human than rat GPR sizes. Using orthology annotations from any database (bottom row) generated rat GPR rules that were frequently larger than the original human GPR rules, unless each human gene was limited to one ortholog (bottom left panel). Requiring orthology annotations to be described by at least 3 different databases (middle row) increased the numbers of reactions automatically annotated as human-specific (blue dots) and provided human GPR rules that were frequently larger than rat GPR rules. Ultimately, the pair of selected cutoff parameters (boxed panel) used in the GPR conversion process provided balanced numbers of disproportionately sized GPR rules between draft rat and human networks.


Top 1 orthologs
Top 2 orthologs
Top 3 orthologs
Top 4 orthologs
Top 5 orthologs
5+ databases
4+ databases
3+ databases
2+ databases
1+ databases
–2275
strict
cutoff
–1309
–1170
–879
–503
–501
–749
selected
cutoff
176
206
–588
1559
1856
1933
1968
relaxed
cutoff
Rat GPR size
Human GPR size
0 1 2 3 4 5+
Reaction: shared human-specific

Chapter 3: Reconciliation of rat and human metabolic networks

Synopsis

In **Chapter 2** chapter, we described a semi-automated approach to generate a draft genome-scale reconstruction (GENRE) of rat metabolism based on an existing GENRE of human metabolism. In this chapter, we manually curated the rat metabolic network (*iRno*) in parallel with the original human metabolic network (*iHsa*) in a process called network reconciliation. As a result, these models successfully captured known metabolic features that distinguish humans and rats. By reconciling species-specific differences, we substantially improved purine degradation and bile acid synthesis pathways in both models. In **Chapter 4**, we develop reconciled biomass formulations and demonstrate that these models can quantitatively predict cellular growth rates and ATP yields unlike previous human GENREs. In **Chapter 5**, we discuss how manually metabolic networks can be used to refine genome annotations using examples identified during the manual curation process described in this chapter. By applying network reconciliation principles throughout the reconstruction process, these important efforts also facilitated the comparative toxicogenomics that will be discussed in **Chapters 6-7**.

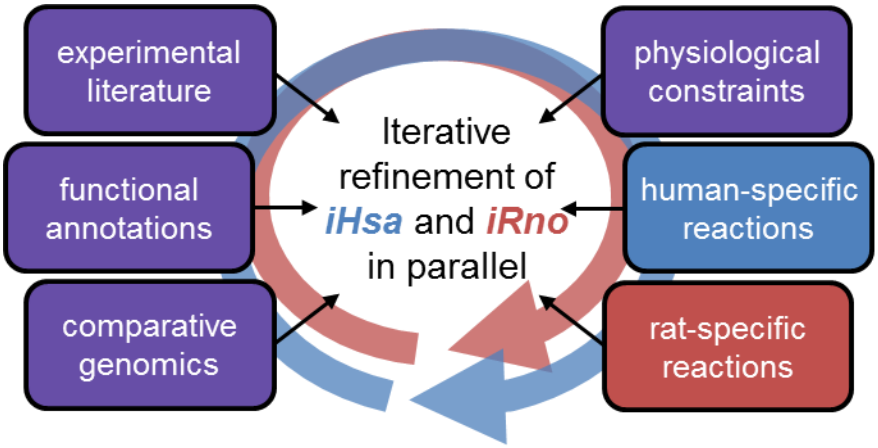


Figure 3.1 – *iRno* and *iHsa* were manually curated in parallel to reconcile species-specific differences. The efforts described in this chapter contributed to the development of the first curated genome-scale rat metabolic network and a substantially improved human metabolic network based on draft reconstructions described in **Chapter 2**.

In the previous chapter, we transformed *iHsa* into *iRno* by replacing human gene-protein-reaction (GPR) relationship rules with orthologous rat GPR rules. Here, we extensively curated *iRno* and *iHsa* by incorporating 169 new reactions, manually updating 1103 GPR relationship rules, and annotating 5000 new references to experimental literature and external databases[33,34,57,65] (**Figure 3.1**). Compared to previous human and mouse GENREs[16,17,19,52,53], *iRno* and *iHsa* captured the highest numbers of total reactions, enzymatic reactions, reactions associated with complex GPR rules, and annotations to external databases (**Table 3.1**). Furthermore, reactions were reconciled for potential differences between rat and human networks which has not previously

been described for existing mammalian networks (**Table 3.1**). As a result, *iRno* and *iHsa* represent two of the most comprehensive metabolic reconstructions and the first pair of mammalian metabolic networks reconciled for comparative analyses to date. These models will serve as powerful computational platforms for contextualizing experimental data and making functional predictions consistent with rat and human biology for clinical and basic science applications.

Genome-scale differences after network reconciliation

Network reconciliation was emphasized throughout the entire reconstruction process for *iHsa* and *iRno* to facilitate cross-species predictions. Oberhardt *et al.*[50] compared GENREs of two closely related *Pseudomonas* species developed independently, and found that cross-species predictions were unrealistic without extensive network reconciliation. The percentage of reactions shared between the two bacterial models increased from 33% to 86%, achieved mostly by resolving differences in nomenclature used to describe reactions and metabolites themselves. Between *iRno* and *iHsa*, a much higher percentage (>99%) of reactions was shared, likely as a result of bypassing the need to reconcile terminology-based differences between species. In contrast, the percent of reactions that overlapped between previous mouse and human GENREs was closer to 98% due in part to reactions that were not reconciled (**Table 3.1**).

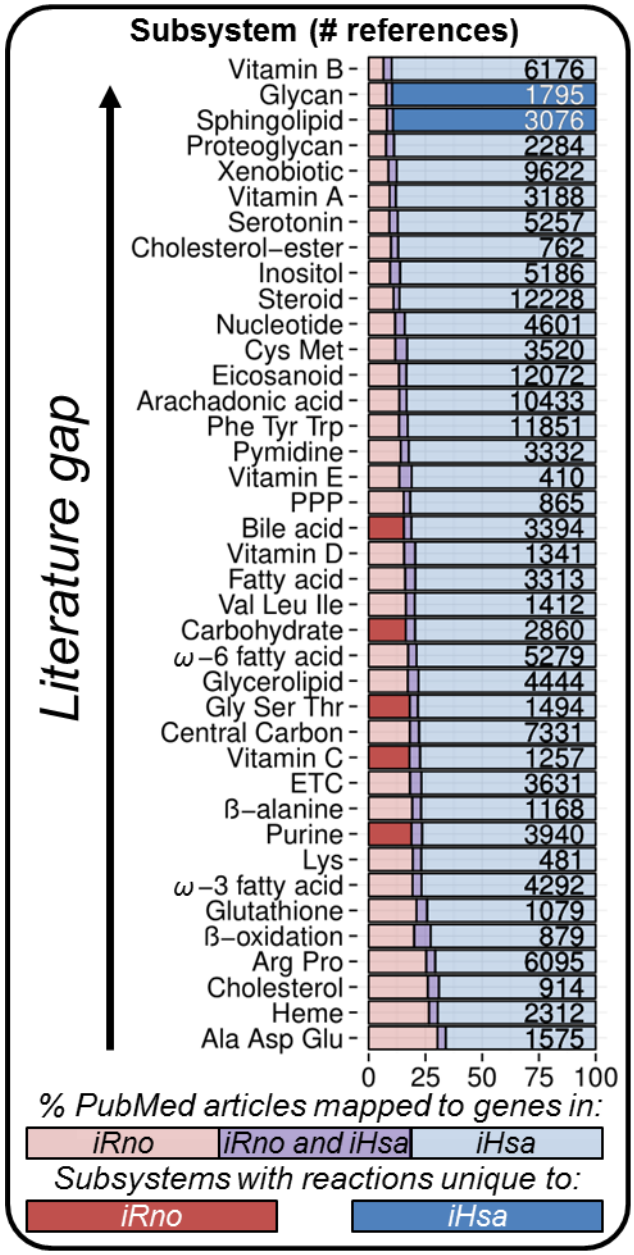


Figure 3.2 – Subsystem-level comparison of the knowledge gap that exists between rats and humans. Stacked bars represent the percentages of PubMed articles mapped to rat and/or human genes for all metabolic genes represented in a subsystem. PubMed articles referenced human genes more frequently than rat genes within all subsystems, but the knowledge gap was larger for pathways that included one or more human-specific reactions. Abbreviations for subsystems refer to: electron transport chain (ETC); pentose phosphate pathway (PPP).

Table 3.1 – Comparison of reconciled rat and human GENREs with previous mammalian GENREs. *iRno* and *iHsa* expand upon HMR2 with curated GPR rules that account for enzyme complexes, updated annotations to external databases, and no unreconciled** differences between species-specific models. (*) Genes and reactions associated with highly redundant GPR rules (10+ isozymes) such as generic signaling processes were excluded in these model comparisons. (**) Unlike species-specific reactions which are enzymatic in one species and absent in the other, unreconciled reactions can either be enzymatic in one species and non-enzymatic in the other, or non-enzymatic in one species or absent in another. (***) Species-specific tasks are explicitly designed to succeed in one species and not the other.

Database	**Ratcon**		**HMA**		**BiGG**		**VHM**
Organism	Rat	Human	Mouse	Human	Mouse	Human	Human
Model	*iRno*	*iHsa*	MMR	HMR2	*iMM1415*	Recon 1	Recon 2
Genes*	**1994**	**1991**	**1902**	**1986**	**1332**	**1435**	**1617**
Reactions	**8271**	**8264**	**8140**	**8181**	**3726**	**3742**	**7440**
enzymatic*	5745	5738	5597	5604	2204	2297	4446
isozymic*	2863	2691	3013	3135	776	832	1647
enzyme complex*	620	620	0	0	237	250	461
annotated in KEGG	2412	2406	-	1527	-	-	-
species-specific	14	7	18	60	0	100	-
unreconciled**	0	0	62	85	83	17	-
Metabolites	**5620**	**5620**	**5516**	**5546**	**2775**	**2766**	**5063**
unique metabolites	3200	3200	3170	3155	1503	1509	2626
compartments	8	8	8	8	8	8	8
biomass metabolites	184	169	117	117	41	41	41
annotated in KEGG	3353	3353	-	689	-	-	2601
Metabolic tasks	**327**	**327**	**56**	**256**	**254**	**260**	**354**
species-specific***	12	2	-	-	-	-	-

After extensive manual curation and network reconciliation, there was a high degree of confidence in the conserved metabolic functionality of *iRno* and *iHsa*. At the genome-scale, 99.6% of all gene-associated biochemical and transport reactions were annotated with both rat and human genes. Despite extensive efforts to identify metabolic activities unique to rat or human genomes, most metabolic subsystems included zero species-specific reactions (**Figure 3.2**). To approximate whether sufficient literature information was available to identify known species-specific differences, we compared how frequently PubMed articles referenced rat and human genes within individual subsystems. We found that rat genes were referenced less frequently compared to human genes, although the literature gap between rat and human genes varied substantially by subsystem

(**Figure 3.2**). Subsystems with human-specific reactions included fewer references to rat genes relative to human genes compared to other subsystems. Interestingly, the number of reactions classified as human-specific based on orthology annotations decreased from 19 to 7 after performing network reconciliation, suggesting a higher degree of consistency between rat and human metabolic capabilities than currently annotated. Alternatively, rat-specific reactions were not identified for any of the poorly studied subsystems in rats (**Figure 3.2**), suggesting that additional studies may reveal undiscovered differences between rat and human metabolism.

Metabolic enzymes unique to either rats or humans more frequently contributed to increased redundancies rather than new functionalities when comparing the relative sizes of rat and human GPR rules across shared reactions (**Figure 3.3**). GPR sizes were consistent (along the diagonal of **Figure 3.3A**) for nearly 80% of reactions associated with both rat and human genes (an example is illustrated in **Figure 3.3B**); however, known differences in the numbers of redundant rat and human genes have been described[66] for reactions like the example shown in **Figure 3.3C**. Capturing variability between rat and human GPR rules is important because the numbers of redundant isozymes or subunits in a protein complex affect the relative robustness of reactions to genetic perturbations. Despite individual variations in GPR sizes and a handful of species-specific reactions, rat and human GPR rules remained relatively balanced at the genome-scale (**Figure 3.3A**) and were not suggestive of any global differences in robustness within metabolism.

Functional differences captured by rat and human metabolic networks

We assembled a comprehensive collection of 327 metabolic tasks that captured known functions within rat and/or human metabolism. Each task represented a known biological process such as producing glucose from lactate during gluconeogenesis or breaking down glutamine into CO_2 and urea. We described the process of formulating metabolic tasks in **Chapter 1** and also provide detailed methods for identifying species-specific differences at the end of this chapter. As a result, we recapitulated 14 new species-specific tasks, 42 new shared tasks, and 271 shared tasks previously described in the validation of human metabolic network reconstructions[17,19,20]. Species-specific tasks were well represented across multiple subsystems including ascorbate, purine, glycan, and bile acid metabolism, and each task was characterized by one or two unique enzymatic reactions (**Figure 3.4**). Below, we showcase the importance of capturing these differences with *iRno* and *iHsa* in the contexts of human biology and disease.

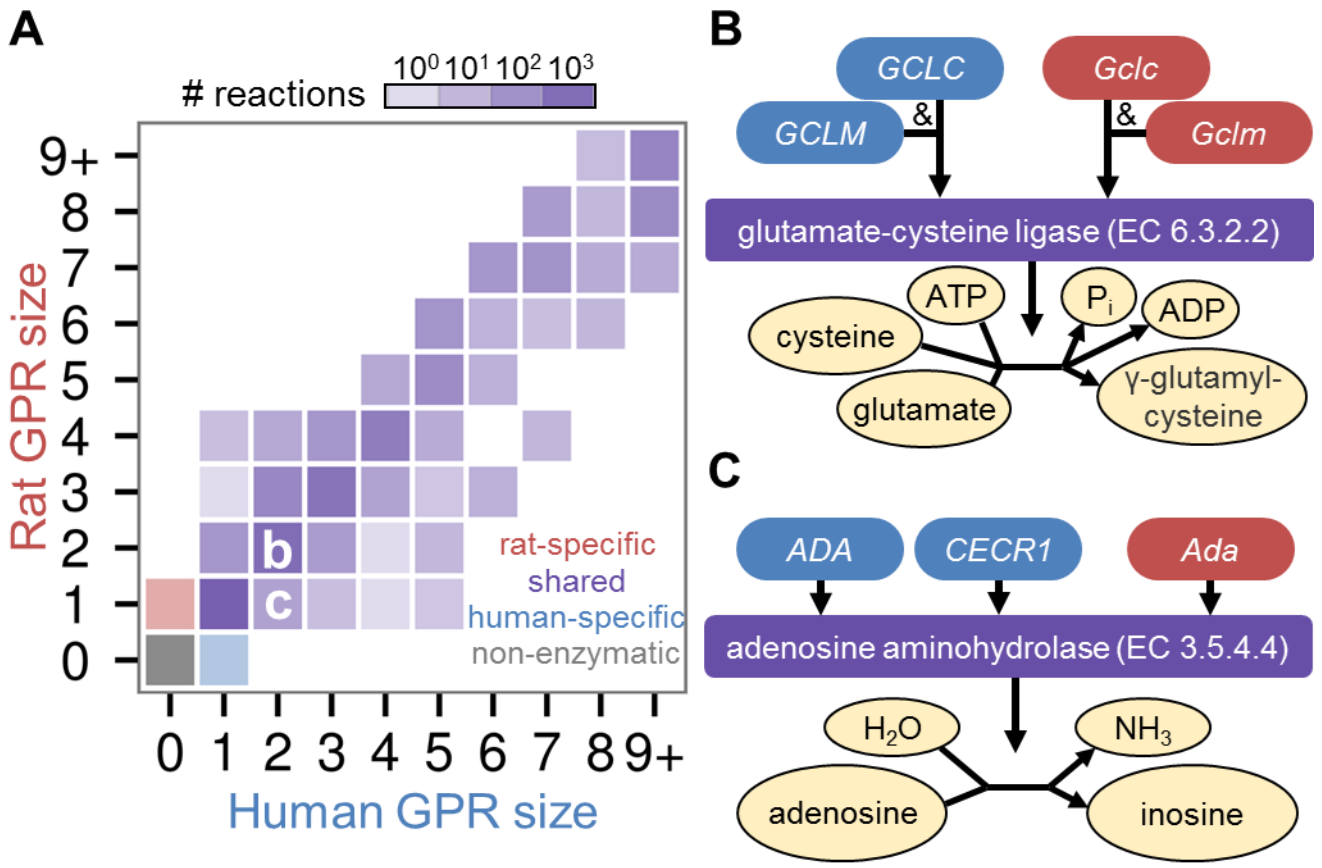


Figure 3.3 – Reconciled gene-protein-reaction (GPR) relationships between iRno and iHsa allow for varying degrees of redundancy. (**A**) Comparison of the number of genes catalyzing each reaction in *iRno* and *iHsa*. Gene-associated reactions capable of catalysis in both *iRno* and *iHsa* were classified as "shared" reactions (purple). Reactions associated with GPR rules in only one organism were classified as species-specific (red and blue for rat-specific and human-specific, respectively). Reactions present in both models that had no known GPR rule assignments were classified as non-enzymatic (gray). Each tile's color intensity represents the (log-scaled) frequency of reactions in that bin. The annotated letters **b** and **c** refer to the individual tiles from which the reactions in panels **B** and **C** are binned. (**B**) Example of a shared reaction with balanced GPR rules in *iRno* and *iHsa*. This reaction, glutamate-cysteine ligase (EC 6.3.2.2), requires both a catalytic subunit (*Gclc* / *GCLC*) and a regulatory subunit (*Gclm* / *GCLM*) to join glutamine with cysteine and form γ-glutamyl-cysteine, a precursor of glutathione. This reaction was manually curated because the original HMR2 did not contain information related to protein-complexes in GPR rules. (**C**) Example of a shared reaction, adenosine aminohydrolase (EC 3.5.4.4), that is involved in purine degradation and can be catalyzed by two redundant human isozymes or one rat enzyme.

Unlike humans, rats are capable of producing vitamin C (ascorbate) (**Figure 3.4A**) and are thus resistant to scurvy[6767676261]. *iRno* and *iHsa* captured this species-specific phenomenon with a new task that simulated *de novo* vitamin C synthesis in a glucose minimal media environment. The rat-specific enzyme, *Gulo*, is known to

be responsible for this functional difference which has limited the use of rat as a model organism for scurvy since the early 20th century[68]. Simulated deletion of *Gulo* in *iRno* was sufficient to block this function in rats, consistent with a *Gulo*-deficient strain of rat developed to study scurvy (**Figure 3.5A-B**). Furthermore, adding a functional copy of the enzyme L-gulonolactone oxidase enabled *iHsa* to complete this task *in silico*, similar to a previous study that restored vitamin C synthesis in a human cell line using the murine ortholog of *Gulo*[69] (**Figure 3.5C-D**).

In humans, the purine degradation pathway yields urate as the end byproduct, which can accumulate and cause gout[70]. Rats can further degrade urate into allantoin (**Figure 3.4B**) and are resistant to gout formation[71]. *iRno* and *iHsa* captured this functional difference with new metabolic tasks that simulated the production of urate from purines, which is common to both species, and allantoin from purines which is absent in humans. The first two steps involved in converting urate into allantoin, urate oxidase (EC 1.7.3.3) and 5-hydroxyisourate hydrolase (EC 3.5.2.17), are catalyzed by the rat-specific enzymes, *Uox* and *Urah*, respectively. The human orthologs for these two rat-specific genes, *UOXP* and *URAHP*, are nonfunctional pseudogenes; however, the third and last enzymatic step, OHCU (2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline) decarboxylase (EC 4.1.1.97) is a shared reaction catalyzed by *Urad* in rats and *URAD* in humans. By resolving species-specific differences in the purine degradation pathway, we also removed reactions that allowed Recon 2[17] and HMR2[19] to degrade urate into urea, a function known to be absent in mammals but present in other vertebrates including fish[71]. As a result, the ability to study gout or make biomarker predictions of urate production with *iHsa* was improved over previously published human reconstructions[17,19].

Most mammals, including rats, can synthesize N-glycolylneuraminic acid (Neu5Gc), a sialic acid found in glycolipids and glycoproteins[72,73], via the rat enzyme *Cmah*, cytidine monophosphate-N-acetylneuraminic acid hydroxylase (EC 1.14.18.2) (**Figure 3.4C**). Humans cannot produce Neu5Gc from N-acetylneuraminic acid (Neu5Ac), a prevalent sialic acid in humans, because *CMAHP* is a nonfunctional pseudogene in humans[73]. Despite this difference, human sialyltransferases can incorporate nonhuman sialic acids into glycans obtained through the consumption of red meat[72], which we also captured as a shared task in rats and humans.

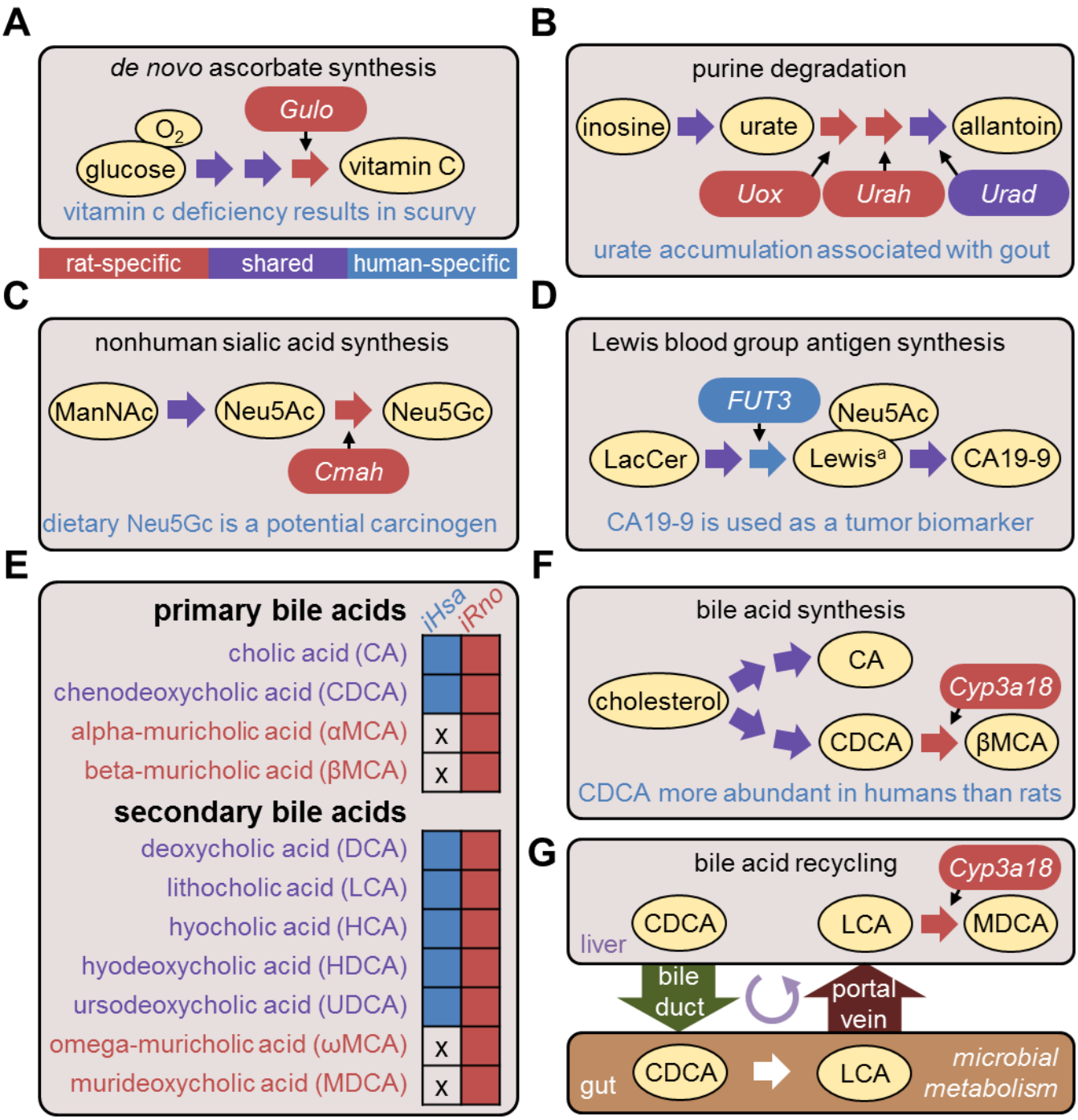


Figure 3.4 – Functional differences known to distinguish rat and human metabolism. (**A**) Rats are capable of synthesizing vitamin C from limited substrates, providing an inherent resistance to scurvy. The rat-specific enzyme, *Gulo*, catalyzes the last enzymatic step of the vitamin C synthesis pathway: L-gulonolactone oxidase (EC 1.1.3.8). The human ortholog of *Gulo* is a nonfunctional pseudogene (*GULOP*). (**B**) Gout formation is associated with accumulation and crystallization of urate, the end-product of purine catabolism in humans. Rats are resistant to gout because urate can be further degraded into allantoin, which is more soluble than urate. (**C**) Most mammals can synthesize the monosaccharide, Neu5Gc, which is known as a nonhuman sialic acid that is incorporated into glycoproteins. (**D**) The human-specific enzyme *FUT3* synthesizes the Lewis[a] antigen (Le[a]) which is involved in Lewis blood group determination and is the precursor for the pancreatic cancer biomarker, CA19-9. (**E**) Metabolic tasks simulating the production of primary and secondary bile acids were consistent with

bile acid species reported in a previous study[78] that compared rat and human liver samples (X's). (**F**) Summary of rat-specific and shared primary bile acid synthesis routes from cholesterol. *Cyp3a18* was hypothesized as the critical enzyme enabling rats to produce rat-specific primary bile acids, which had not been previously described. (**G**) In the process of "bile acid recycling", primary bile acids secreted by the liver into the gut are transformed by bacteria and subsequently reabsorbed by the liver. Synthesis of secondary bile acids were accounted for in *iRno* and *iHsa* by including extracellular reactions associated with gut bacteria. Interplay between rat liver and gut metabolism was necessary for *iRno* to simulate the synthesis and secretion of the rat-specific bile acid, murideoxycholic acid (MDCA). Abbreviations: N-acetylmannosamine (ManNAc); N-acetylneuraminic acid (Neu5Ac); N-glycolylneuraminic acid (Neu5Gc); lactosylceramide (LacCer); carbohydrate antigen 19-9 (CA19-9).

The human-specific enzyme, *FUT3*, encodes a fucosyltransferase involved in the Lewis blood group system. An individual with a functional copy of *FUT3* can produce the Lewis a antigen[74] (Le^a) and sialyl-Le^a, a clinical biomarker for pancreatic cancer commonly referred to as carbohydrate antigen 19-9 (CA19-9) (**Figure 3.4D**). Despite the inability of rats to synthesize Le^a, we expect that CA19-9 could be produced from exogenous Le^a by the orthologous sialyltransferases, *St3gal3* and *ST3GAL3*[75,76]. Surprisingly, *FUT3* was the only functional difference attributed to human-specific enzymes after performing network reconciliation between *iRno* and *iHsa*.

Improvements within bile acid metabolism

The bile acid metabolic pathway was expanded in *iRno* and *iHsa* to include bile acids that may serve as biomarkers in rats and humans[77] (**Figure 3.4E**). Human hepatocytes can synthesize chenodeoxycholic acid and cholic acid from cholesterol to facilitate dietary lipid absorption (**Figure 3.4F**). In addition to these two primary bile acids, rat hepatocytes also produce large quantities of α-muricholic and β-muricholic acid that are absent in humans[77,78] (**Figure 3.4F**). After extensive manual curation of this pathway (described below), metabolic tasks simulating the synthesis of four primary and seven secondary bile acids were consistent with previous experiments directly comparing bile acids detected in the livers and sera of rats and humans[77,78] (**Figure 3.4G**).

The mammalian intestinal microbiome plays an important role in converting primary bile acids synthesized by hepatocytes into secondary bile acids which can be reabsorbed and further metabolized in the liver[79]. This "bile acid recycling" expands the global pool of metabolites encountered by humans and rats beyond what their individual genomes allow. In order to account for bile acid recycling in *iRno* and *iHsa*, we introduced a new extracellular subsystem of "gut" metabolic reactions that converted primary bile acids into secondary bile acids

(**Figure 3.4G**). This simplified system representing the intestinal microbiome was necessary to capture the synthesis of murideoxycholic acid, a rat-specific bile acid derived from the secondary bile acid, lithocholic acid (**Figure 3.4G**).

While curating the bile acid synthesis pathway, we discovered that the critical enzymatic step involved in the production of rodent-specific bile acids was not annotated to any rat or mouse genes (**Figure 3.4G**). In **Chapter 5**, we discuss how we identified *Cyp3a18* as a candidate gene for performing this function and highlight how reconciling metabolic network reconstructions can guide the improvement of genome annotations[80].

Identifying species-specific reactions

During the manual curation process, new reactions were added to *iRno* and/or *iHsa* that were not previously in HMR2. Identification of species-specific reactions was prioritized because rat-specific reactions were unlikely to be included in a human GENRE. *iRno* and *iHsa* were expanded and updated in parallel when possible to maintain consistency in the reconstruction process. For each new reaction added, rat and human GPR rules were constructed manually using evidence from experimental literature and functional annotation databases. Evidence supporting the presence of a reaction in one organism and not the other was necessary for classifying a reaction as species-specific. Otherwise, reactions directly associated with rat and human enzymes or indirectly through orthology annotations were assumed to be shared. In total, 69 biochemical, 32 transport, 40 exchange reactions were added that were not present in HMR2. All transport and exchange reactions were shared by *iRno* and *iHsa* and nine biochemical reactions were unique to *iRno*. Additionally, most reactions originally annotated as human-specific after the automated reconstruction process described in **Chapter 2** were considered shared reactions after identifying appropriate rat enzymes. Ultimately, 14 rat-specific and 7 human-specific metabolic reactions were included in *iRno* and *iHsa* (**Table 3.1**) in addition to 16 artificial reactions involved in the formation of species-specific components used in the biomass formulation (see **Chapter 4**). Below, we provide detailed methods and additional considerations that were critical in identifying and annotating species-specific reactions.

Updating annotations to external databases

iRno and *iHsa* were expanded to include species-specific reactions from the KEGG database and literature sources. Lists of reactions and modules linked to genes annotated in humans (hsa) and rats (rno) were obtained using KEGG's Representational state transfer (REST)-style interface (http://www.kegg.jp/kegg/rest/keggapi.html). Reactions linked to humans and not rats or to rats and not humans

were manually investigated for their feasibility as actual species-specific reactions. To identify potential differences between the metabolic capabilities of rats and humans from literature, various searches were performed using PubMed (http://www.ncbi.nlm.nih.gov/pubmed) with combinations of the keywords: rat, human, comparative genomics, cross-species, species-specific, metabolism, metabolic deficiency. No comprehensive comparative analyses were identified other than original publication of the rat genome[40].

We initially explored the Kyoto Encyclopedia for Genes and Genomes[33,34] (KEGG) database as a starting point for identifying species-specific differences between rat and human metabolism. Prior to adding potential new reactions to *iRno* and *iHsa*, existing reactions and metabolites were updated with annotations to external databases. Throughout the entire reconstruction process, 742 reactions and 354 metabolites were assigned new or updated KEGG annotations to replace empty, incorrect, generic, or obsolete KEGG identifiers. Overall, the numbers of unique KEGG REACTION and KEGG COMPOUND identifiers represented across *iRno* and *iHsa* increased from 1376 and 1650 to 1702 and 1721, respectively, compared to the initial human GENRE, HMR2[19]. These updated annotations were necessary to avoid creating duplicate entries of unique reactions or metabolites and to facilitate assigning GPR rules to 122 reactions not previously associated with any genes.

Interestingly, some rat-specific functions were already included as non-gene associated reactions in HMR2 and deleted from *iHsa* as part of the reconciliation process. For example, the metabolic reaction catalyzing the conversion of threonine to glycine and acetaldehyde was originally present in HMR2 as a spontaneous (non-enzymatic) reaction with no external annotations[19]. This reaction was identified in the KEGG database as the rat-specific reaction, threonine aldolase (R00751). As a result, this reaction was assigned a new GPR rule in *iRno* and disabled in *iHsa*. Of 18 rat-specific KEGG reaction annotations: 5 were already represented and removed from *iHsa*; 11 were added as new rat-specific reactions *iRno*; and 2 redundant with other rat-specific reactions were ignored. Of 75 human-specific KEGG annotations: 4 were already represented in *iHsa*, and had been disabled in *iRno* as a result of the GPR conversion process described in **Chapter 2**; 14 were re-classified as shared reactions after identifying suitable rat orthologs; and 57 involved in peripheral pathways such as xenobiotic metabolism were ignored.

Curating GPR rules to include complex relationships

In addition to resolving differences at the reaction-level, *iRno* and *iHsa* were further reconciled for comparative analyses by manually updating GPR rules. A major disadvantage of using HMR2 for the basis of *iRno* and *iHsa* was the absence of complex GPR relationship rules with multiple subunits in a protein complex (**Table 3.1**). Manually curated GPR rules containing protein-complexes were based primarily on GPR relationships from *Homo sapiens* Recon 2[17], experimental literature, and genome annotation databases. As a result, GPR rules with multiple subunits were constructed for 620 reactions in *iRno* and *iHsa* (see **Figure 3.3B** for example). After network reconciliation and extensive manual curation, the numbers of rat and human genes mapped across shared reactions remained balanced (**Figure 3.3A**; **Figure 3.4C**).

GPR rules comprised of more than one gene were initially limited to isozymic "or" relationships because none of the 1390 unique GPR rules in HMR2 described relationships between subunits in a protein complex[19]. To overcome this limitation, GPR rules were manually constructed to include "and" logical operators for both rat and human models when possible. Evidence supporting the requirement of multiple enzymatic subunits to perform a metabolic function were obtained from functional annotation databases and experimental literature. We also compared complex human GPR rules from the second largest human GENRE, *Homo sapiens* Recon 2 (versions 2.0.3 and 2.0.4), with an early draft of *iHsa* in order to convert isozymic relationships into complex relationships.

Formulating metabolic tasks

Metabolic tasks representing known biological functions of rats and humans were simulated in *iRno* and *iHsa*. To formulate new metabolic tasks, we implemented methods discussed in **Chapter 1** that were previously applied to HMR2[19]. Overall, curated rat and human models successfully performed 327 tasks including 256 from HMR2[19] and 15 adapted from Recon 2[17]. An additional set of 57 new tasks were defined including 14 species-specific tasks (**Table 3.2**) such *de novo* synthesis of vitamin C (**Figure 3.4A**). Identifying species-specific tasks and tasks related to hepatocyte metabolism was prioritized to capture biological functions that might be important for the use of these models in studying toxicology which will be discussed in later chapters. In the next chapter, we will also discuss several tasks related to biomass synthesis and energy maintenance that were formulated to quantitatively capture cellular growth rates under physiological constraints. Below, we discuss strategies that were implemented as part of the iterative network reconciliation process (**Figure 3.1**).

Table 3.2 – Summary of metabolic tasks added as a result of the network reconciliation process. The numbers of tasks intended to fail within each category are shown for tasks that cannot be completed by either *iRno* (red), *iHsa* (blue), or both models (purple).

Metabolic Task Category	Tasks Added	Failure Intentional
Biomass synthesis and energy maintenance	16	1, 5
Bile acid metabolism	29	7
Purine, glycan, and ascorbate metabolism	12	2, 4, 2
Various tasks adapted from Recon 2	14	-

Because none of the original 271 metabolic tasks were considered unique to humans, metabolic task simulations were expected to be consistent between *iRno* and *iHsa* unless directly supported by literature evidence. To avoid changes that negatively impact the performance of *iRno* and *iHsa*, we simulated all available metabolic tasks during each iterative round of model updates. Removal of reactions from *iHsa* as part of the reconciliation process did not affect the completion of any metabolic tasks that had already been defined. Furthermore, the addition of new reactions to *iRno* and *iHsa* did not enable completion of 19 tasks explicitly intended to fail such as the *de novo* synthesis of essential amino acids. In cases where model changes introduce unexpected task results, it is imperative to justify any inconsistencies with literature support, undo changes that were responsible for model discrepancies, or provide an alternative solution with additional model changes.

Prior to manual curation, the automated draft of *iRno* originally failed to complete 3 human tasks related to bile acid synthesis that have been described as functional in rats[78]. Assigning the rat gene *Akr1c14* to 3α-hydroxysteroid dehydrogenase (EC 1.1.1.50) was sufficient to resolve all 3 inconsistent metabolic task predictions between *iRno* and *iHsa*[66]. In contrast, the mouse GENRE, *iMM1415*, required the addition of 95 reactions to complete 260 metabolic tasks after automated conversion from *Homo sapiens* Recon 1[16,53] (**Table 3.1**). Because orthology between *Akr1c14* and *AKR1C4* had not previously been annotated in any of the 5 orthology databases discussed in **Chapter 2**, we manually investigated all reactions that were annotated as human-specific after automated reconstruction of *iRno* to resolve differences attributed to missing orthology annotations. As a result, nearly all species-specific enzymes were also supported by functionally important species-specific tasks. Additionally, some of these changes led to genome annotation refinements which will be discussed in **Chapter 5**.

Formulating species-specific tasks

After resolving inconsistencies for previously defined functions, we searched for species-specific differences that could be implemented as metabolic tasks. We queried presence or absence of all KEGG MODULEs for rats and humans, revealing 2 human-specific modules and 1 rat-specific module. Each module described the ability of an organism to synthesize a product *de novo* from a starting substrate. In the rat-specific module, M00129, vitamin C can be synthesized from glucose; however, we found that *iRno* also required oxygen uptake to complete this task, highlighting an advantage of GENREs as a centralized repository to annotate known metabolic functions. In KEGG, the human-specific modules for chenodeoxycholic acid synthesis from cholesterol and degradation of heparan sulfate into disaccharides were each characterized by a single missing enzyme in rats. For chenodeoxycholic acid synthesis, the blocked reaction, 3-alpha-hydroxysteroid dehydrogenase (EC 1.1.1.50) was not annotated for any rat enzymes in KEGG. As mentioned above, the rat gene, *Akr1c14*, is known demonstrate to this activity[66], suggesting that this module be reclassified as complete in both rats and humans. For heparan sulfate degradation, manually assigning the rat enzyme, *Hgsnat*, to its known function (EC 2.3.1.78) was able to recapitulate this previously annotated human-specific function in rats[81]. In **Chapter 5**, we will discuss how the manual curation process for a metabolic network can also be leveraged to guide genome annotations refinements using these examples.

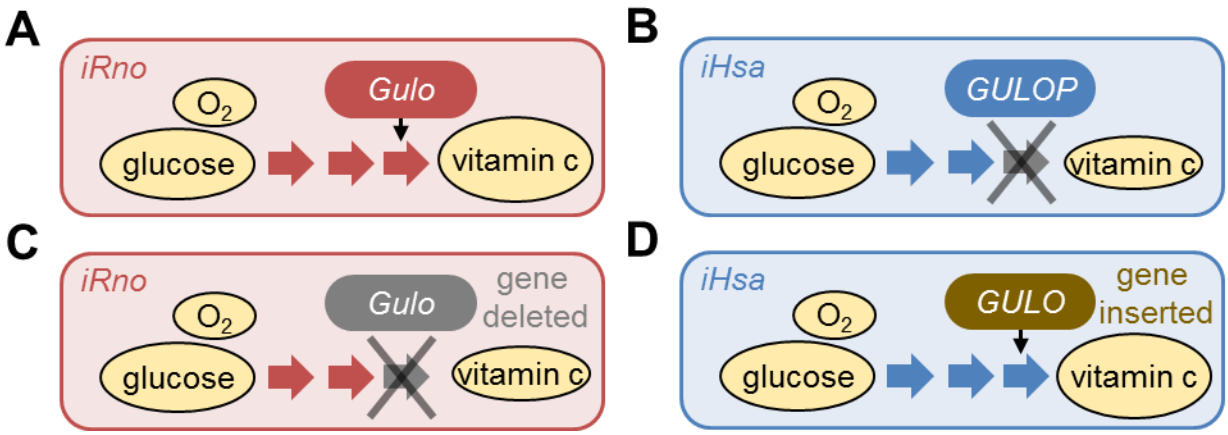


Figure 3.5 – Simulating genetic engineering strategies with *iRno* and *iHsa*. (**A**) Rats are capable of synthesizing vitamin C from limited substrates which was captured by *iRno*. The last enzymatic step of this process is known to be catalyzed by L-gulonolactone oxidase (*Gulo*). (**B**) Humans cannot synthesize vitamin C from limited substrates which was captured by *iHsa*. The human ortholog of the rat gene, *Gulo*, is a non-functional pseudogene. (**C**) By simulating the deletion of *Gulo* with *iRno*, rats were no longer predicted to be capable of synthesizing vitamin C given glucose and oxygen, consistent with the result of the simulation shown in **B**. (**D**) By simulating the knock-in of a functional equivalent to *Gulo* in humans, *iHsa* was capable of performing *de novo* vitamin C synthesis, consistent with the result of the simulation shown in **A**.

Genetic engineering strategies with metabolic networks

Through the process of network reconciliation, we have established rat and human metabolic networks that can be used to develop new genetic engineering strategies that bridge the gap between rat and human biology. L-gulonolactone oxidase (*Gulo*) has been described as the critical enzyme for vitamin C synthesis that differentiates rats from humans[68] (**Figure 3.5A**). Using flux variability analysis (FVA)[82], 10 reactions were required by *iRno* to synthesize vitamin C under glucose minimal media conditions. The only enzymatic reaction required for vitamin C synthesis in *iRno* that was also absent in *iHsa* was L-gulonolactone oxidase (EC 1.1.3.8) (**Figure 3.5B**). In agreement with the KEGG MODULE, "ascorbate biosynthesis, animals" (M00129), L-gulonolactone oxidase (K00103) was annotated as the only enzymatic step missing in humans (**Figure 3.5C**). Additionally, deleting *Gulo* blocked the ability of *iRno* to produce vitamin C, consistent with a *Gulo*-deficient strain of rat developed to study scurvy[83]. Artificially adding L-gulonolactone oxidase to *iHsa* enabled the human model to successfully complete the vitamin C synthesis task (**Figure 3.5D**), as previously described in a study that restored vitamin C synthesis in a human cell line using the murine ortholog of *Gulo*[69].

Acknowledgements

This chapter was adapted from a manuscript is currently under revision at *Nature Communications* titled "Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions" and written by Edik M. Blais, Kristopher D. Rawls, Zhuo I. Li, Glynis L. Kolling, Ping Ye, Anders Wallqvist, and Jason A. Papin. I thank all co-authors who contributed to work. I also thank Issac Li and Michael Carter for improving annotations in the rat and human networks and Kevin D'Auria, Jennifer Bartell, and Arvind Chavali for providing feedback on an earlier version of the manuscript.

Chapter 4: Quantitative growth rate predictions

Synopsis

Cellular proliferation is a universal feature shared across all domains of life. The process of transforming raw materials into cellular biomass is a fundamental step required for the growth of an organism. Genome-scale network reconstructions (GENREs) can capture the metabolic machinery of an organism that salvage, synthesize, and organize the macromolecular building blocks within a cell. In this chapter, we describe comprehensive biomass formulations for rat and human hepatocytes based on experimental literature. To facilitate comparative analyses between the GENREs of rat (*iRno*) and human (*iHsa*) metabolism in accordance with the network reconciliation process described in **Chapter 3**, we established a unified biomass subsystem of reactions that also takes into account species-specific differences in biomass formulations. By applying physiologically-relevant constraints to metabolic exchange reactions, quantitative growth predictions were remarkably consistent with growth rates previously reported for hepatocytes. In this chapter, we discuss important quality control measures implemented throughout the manual curation process that contributed to the ability of *iRno* and *iHsa* to accurately capture growth.

Defining metabolic objectives for mammalian cell types

Computational methods such as flux balance analysis (FBA) rely on specifying a biological objective to interrogate a metabolic network. Biomass synthesis can be considered a fundamental objective in biology. Without the drive to proliferate, a species will eventually die out; however, growth is not always the objective of a cell. As multicellular organisms, humans and rats have developed highly specialized tissues that perform distinct metabolic functions. For example, cardiomyocytes constantly regenerate ATP from ADP to fuel the molecular motors that drive cardiac muscle contractions. Using a human cardiac-specific metabolic network, a recent study defined ATP turnover as an objective in order to systematically explore how different substrates affect the efficiency of cardiac function[84]. For a kidney-specific metabolic network, objective functions associated with kidney filtration were defined such as the removal excess waste products like urea and the reabsorption of amino acids[85]. These examples underscore the importance of understanding physiology when using GENREs to study metabolism, particularly in complex organisms like rats and humans.

The mammalian liver is the serves a variety of functions that are critical for maintaining metabolic homeostasis and staying healthy. These include gluconeogenesis, glycogen storage, detoxification of ammonia to urea, breakdown of lipids, and bile acid synthesis. The human liver-specific metabolic network, HepatoNet1[20], captured

these and other known liver functions in a collection of metabolic tasks. In **Chapter 2**, we verified that the initial drafts of *iHsa* and *iRno* were capable of performing tasks described in HepatoNet1 and the newer human GENRE, the Human Metabolic Reaction database[19] 2.0 (HMR2). In **Chapter 3**, we defined additional species-specific metabolic tasks that are also performed in the liver such as *de novo* vitamin C synthesis and rodent-specific bile acid production. Despite covering a wide range of hepatic functions, analyzing each metabolic task independently does not capture the complexity associated with the tradeoffs between different metabolic objectives.

In this chapter, we defined quantitative parameters for simulating the metabolic activity of hepatocytes at the systems-scale. We compiled experimental values reported in literature to formulate a comprehensive representation of hepatocyte biomass and implemented a new subsystem of biomass reactions to facilitate comparative analyses of growth across *iRno* and *iHsa*. By applying physiological constraints to metabolite exchange reactions, hepatocyte functions such as urea, albumin, and glucose production could be maintained at realistic rates while taking into account the limitations of nutrient consumption rates. By specifying biomass synthesis as the objective function, growth rate prediction under physiological constraints were consistent with known growth rates reported in literature for hepatoma cell lines.

Formulating biomass from heterogeneous datasets

In a metabolic network, the biomass formula represents the recipe for a cell simplified into a list of ingredients. Although the quantities of individual ingredients may vary extensively by cell type and by species, most lists will include DNA, RNA, proteins, lipids, and glycans. These macromolecular building blocks are represented in GENREs as individual amino acids, nucleic acids, fatty acids, and carbohydrates. Unfortunately, experimental studies quantifying all of these ingredients are not readily available, so we implemented a piecemeal approach to construct biomass formulas for rat and human hepatocytes. Throughout this chapter, we discuss important considerations for defining biomass formulas using incomplete knowledge and sparsely populated datasets.

When gathering data from experimental literature, it is paramount to consider the units in which biomass will be formulated. The biomass formula represents each component as a molar quantity per cell quantity (e.g. µmol per million cells; mmol / g dry weight; mmol / mm^3 tissue). In order to reconcile different types of measurements across datasets, metabolite quantities represented by mass can be divided by the molar mass ratio to obtain a molar quantity and cellular quantities can be interchangeable with appropriate conversion rates (cells per gram

dry weight; cells per mg protein; gram dry weight per mm^3 tissue). In **Table 4.1**, we show experimentally-derived conversion rates from a study that compared quantification methods to determine cellularity for rat and human hepatocytes.

Table 4.1 – Experimentally reported[86] hepatocyte cellularity (cell count density) using different methods. These values provide the ability to combine values from sparse datasets.

Hepatocytes	cells / µg liver	cells / µg protein
Human	139	1015
Rat	117	1522

In order to quantitatively estimate growth rates, the biomass should be formulated with units that are consistent with any experimental flux measurements (minus the time component) that will be applied as constraints to exchange reactions. By defining biomass as fmol per cell (1 femtomole = 10^{-15} moles) and exchange fluxes as fmol per cell per hour, growth rate predictions using the biomass reaction can be interpreted as growth per hour. To approximate the doubling time of cells growing exponentially at steady-state, divide the natural logarithm of 2 by the predicted growth rate.

In many cases, data are only available as relative quantities for metabolites within a specific group such as amino acids. To integrate relative quantities into a biomass formula, additional information such as the amount of protein in a cell can be used to infer absolute quantities per cell. As an example, the relative abundance of glutamate was reported as 10.34% in human liver microsomes and 10.19% in rat liver microsomes from a comparative study[87]. From a separate study[88], protein contributed to approximately 61.4% of the cellular mass by dry weight for a human hepatocyte cell line whose mass was estimated to be 419.1 pg dry weight per cell. Combining these values with a molar mass of 147.13 g / mol for glutamate, we can estimate that the cellular composition of glutamate in human hepatocytes is 174 fmol per cell by the following calculations:

$$\frac{10.34\ pg\ glutamate}{100\ pg\ protein} * \frac{59.1\ pg\ protein}{100\ pg\ dry\ wt\ cells} * \frac{419.1\ pg\ dry\ wt\ cells}{1\ cell} * \frac{1\ pmol\ glutamate}{147.13\ pg} = \frac{0.174\ pmol}{cell} \quad \text{Equation 4.1}$$

Consistent with the network reconciliation process described in **Chapter 3**, we assumed that introducing species-specific differences should be avoided unless supported by evidence. Without comparative quantities from both rat and human hepatocytes, we generally assumed that the mass contribution for each biomass component was roughly equal. For the cellular composition of glutamate in rat hepatocytes, we can replace 10.34 with 10.19 in the equation above and estimate a value of 172 fmol per cell. However, we identified a technical limitation of this

strategy when recalculating the total mass per cell across all amino acids which was primarily due to rounding errors. To maintain consistency in the total mass contributed by a single biomass subcomponent like amino acids, we created separate rat and human protein synthesis reactions within a new biomass subsystem that generate an average protein unit. With this formulation, each species has a distinct protein synthesis reaction with distinct substrate ratios but the metabolite produced is shared by both species and contributes an equal amount of protein mass to the biomass formula.

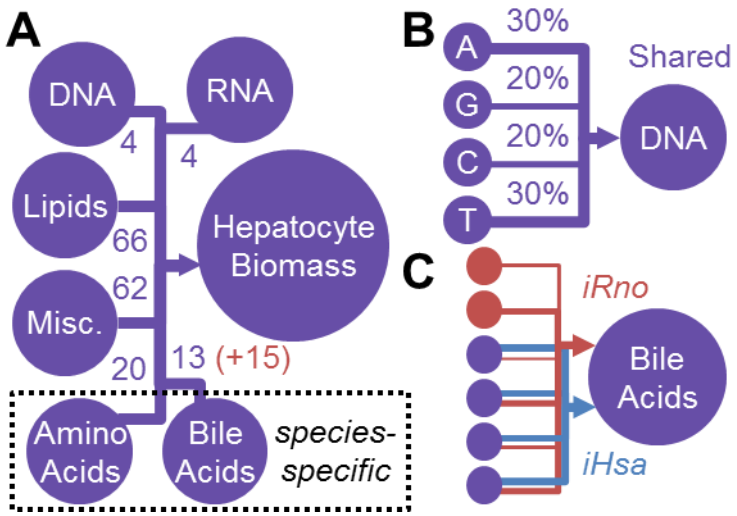


Figure 4.1 – A unified biomass reaction was created for rat and human hepatocytes to enable comparative predictions using *iRno* and *iHsa*. (**A**) Each metabolite consumed in the hepatocyte biomass reaction represents an "average" biomass subcomponent that can be synthesized with a separate reaction. DNA, RNA, lipids, and miscellaneous metabolites (abbreviated Misc.) like glycogen and vitamin C were assigned species-independent synthesis reactions. Bile acids and amino acids, which can vary significantly between species, were assigned species-specific synthesis reactions in *iRno* and *iHsa*. Numbers indicate how many unique metabolites are shared (purple) or rat-specific (red) within each biomass subcomponent. (**B**) The biomass subcomponent for an average DNA molecule is produced by consuming experimentally-derived ratios of individual deoxynucleotides. In this reaction, more adenine (A) and thymine (T) are incorporated into DNA than cytosine (C) and guanine (G) as indicated by percent labels and by line thickness. (**C**) Synthesis of an "average" bile acid was defined separately for *iRno* and *iHsa* in order to account for species-specific metabolites (muricholic acids) and relative abundances.

A unified biomass subsystem that accounts for species-specific differences

A novel system of reactions representing biomass synthesis was developed to enable cross-species predictions of growth between *iRno* and *iHsa* with a single biomass reaction (**Figure 4.1**). Using quantitative values from hepatocyte-based experimental literature, *iRno* and *iHsa* included reactions that consumed known quantities of

individual metabolites to produce an "average" biomass precursor metabolite that was species independent (**Figure 4.1A**). For biomass precursors with relatively similar compositions such as the average nucleotide incorporated into DNA (**Figure 4.1B**), macromolecular synthesis reactions were shared by *iRno* and *iHsa*. Species-specific macromolecular synthesis reactions were added to represent distinct compositions of bile acids (**Figure 4.1C**) and amino acids obtained from studies comparing metabolomics profiles of rat and human hepatocytes[77,78,87]. This generalized framework for biomass formulations was implemented in *iRno* and *iHsa* to simulate hepatocyte growth and can be extended to formulate species-specific compositions for groups of metabolites in any tissues with quantitative or comparative metabolomics data. Additionally, hepatocyte growth and production of each of the 8 macromolecular precursors under strict physiological conditions were simulated as separate metabolic tasks (see **Chapter 3** for detailed description of metabolic tasks).

Biomass formulations for rat and human hepatocytes

New biomass metabolites were defined for each macromolecular subcomponent present in a hepatocyte (**Figure 4.1A**). We estimated the relative contribution of each biomass metabolite to the overall biomass of a cell (percent of dry weight): DNA (2.3%), RNA (3.7%), lipids (17.2%), protein-incorporated amino acids (59.1%), free amino acids (3.8%), bile acids (.1%), glycogen (3.2%), and miscellaneous metabolites (10.5%). Miscellaneous metabolites included vitamins, and cofactors, and other metabolites present at high intracellular concentrations such as vitamin C, citrate, and glutathione. The relative abundances of individual metabolites within each subgroup were determined from several previously published studies[87-102]. Data directly comparing metabolite profiles between human and rat hepatocytes were available for amino acids and bile acids. To account for these differences within a generalized framework, species-specific reactions were added to *iRno* and *iHsa* for the synthesis of these two biomass metabolites (**Figure 4.1C**). For biomass components with similar compositions between rats and humans, shared reactions were used to produce estimated hepatocyte-specific compositions (**Figure 4.1B**). This new cross-species framework can be extended to formulate new biomass compositions for cross-species analyses within and between various cell or tissue types using the same centralized biomass precursor metabolites. For example, biomass metabolites for DNA, RNA, and protein could be reused when formulating a biomass reaction for adipocytes that incorporates a modified lipid composition higher percentage of lipids using and formulations would require modified lipid compositions can utilize the same biomass metabolites for DNA and RNA while modifying lipid contribution biomass reaction for adipocytes, biomass

metabolites such as DNA and RNA can easily be reused be easily adapted by increasing the contribution of lipids, metabolites

Physiological constraints for hepatocyte growth

Physiological ranges for exchange reactions were determined using a consensus approach (**Figure 4.2**). Experimentally measured metabolite consumption and secretion rates were obtained for rat liver cells and rat hepatocytes from 6 existing studies[19,46,47,78,103-105]. Exchange reaction equations were formulated such that negative and positive fluxes represented consumption and secretion, respectively. Flux measurements were standardized to units of fmol $cell^{-1}$ $hour^{-1}$ using previously described conversion rates[86]. In order to normalize quantitative measurements from different experimental systems, absolute flux measurements were median-scaled using metabolites measured in all 6 experiments to the average median value of the 3 *in vitro* experiments. To assign experimental observations as physiological constraints, lower and upper bounds for exchange reactions were determined based on minimum and maximum normalized values across all experimental observations.

Physiological constraints were applied to *iRno* and *iHsa* either as strict constraints for quantitative simulations of hepatocyte biomass (**Figure 4.2**) or relaxed constraints for treatment-induced simulations in **Chapters 6-8**. Under relaxed and strict physiological constraints, lower bound values less than zero were used to allow nutrient uptake of measured metabolites. Under strict physiological constraints, lower bound values greater than zero were also applied requiring secretion of urea, glucose, glutamate, aspartate, 3-hydroxybutyrate, and albumin. Upper bound values were also applied to require consumption or limit secretion of metabolites under strict physiological constraints (**Figure 4.2A**). Additionally, estimated uptake rates for 12 inorganic ions (**Figure 4.2B**) and 13 essential nutrients (**Figure 4.2C**) were assigned under both relaxed and strict physiological constraints. For metabolites with flux measurements available in rat and human hepatocytes, differences between species were considered negligible relative to feasible flux ranges with the exception of bile acids.

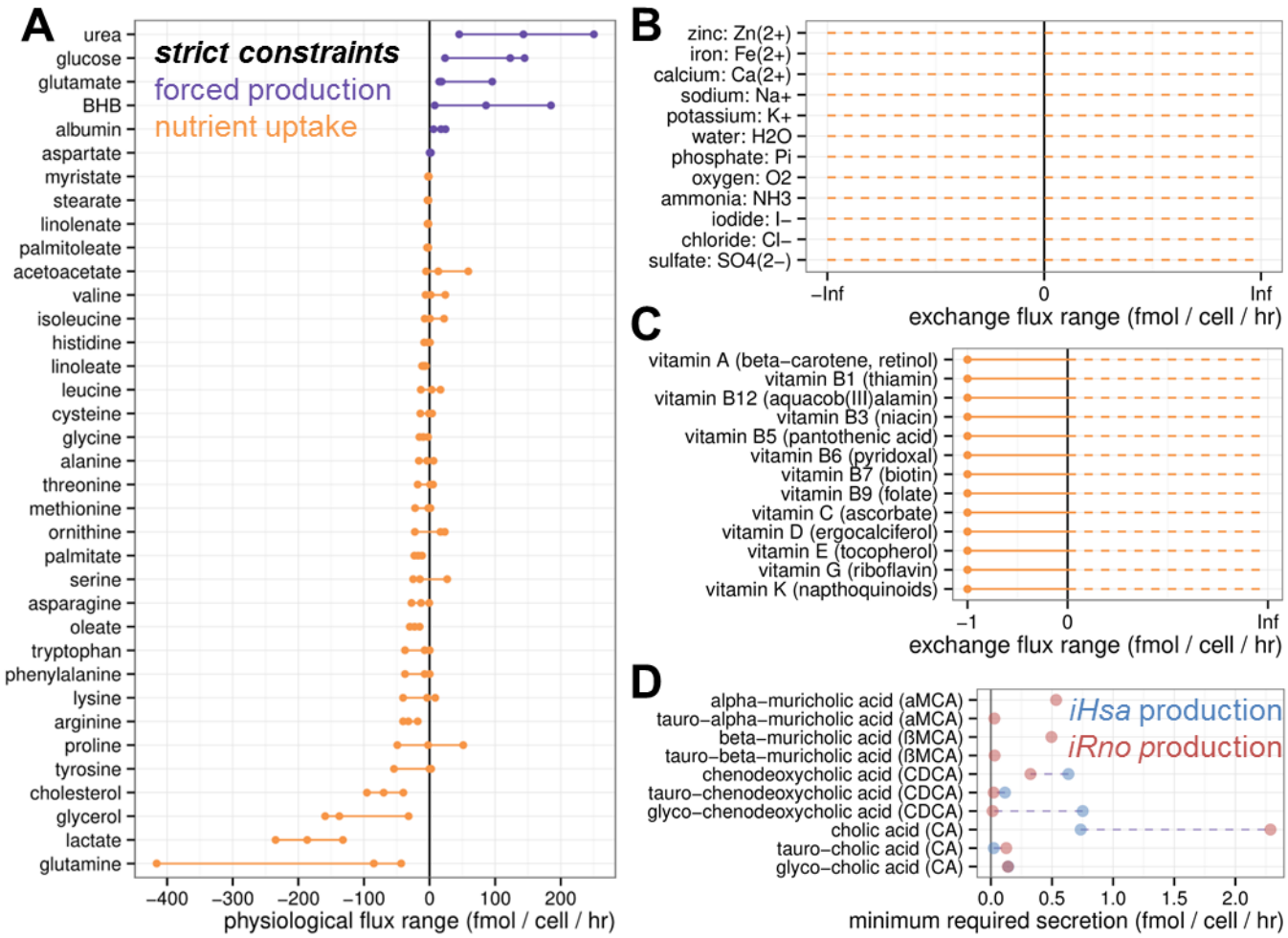


Figure 4.2 – Strict physiological constraints applied to *iRno* and *iHsa*. (**A**) Experimentally reported flux measurements from rat hepatocytes were obtained from six separate studies to constrain *iRno* and *iHsa* with exchange boundaries that represent physiological conditions. Minimum, median, and maximum reported exchange fluxes in fmol $cell^{-1}$ $hour^{-1}$ are shown for each metabolite. Reaction lower bounds for exchange metabolites were set to the minimum reported value (leftmost point) as strict physiological constraints. For simulations of hepatocyte growth using the biomass objective, the maximum value for each metabolite was also applied as the upper bound to exchange reactions. Exchange fluxes for albumin were scaled to represent the secretion of an average amino acid from albumin because the albumin metabolite represents a full-length protein with 608 amino acids (in rats). (**B**) Experimentally reported flux measurements from **A** were also applied as relaxed physiological constraints for toxicogenomics biomarker predictions. Reaction lower bounds with positive values (forced production) to were set to zero and upper bounds with negative values (forced consumption) were set to positive infinity (10^6). (**C**) Species-specific constraints required distinct quantities of bile acids to be produced under strict physiological conditions in hepatocytes. Each point represents the lower bound applied to either *iRno* (red) or *iHsa* (blue) based on serum concentrations in rats and humans, respectively. Synthesis and secretion of α- and β-muricholic acids in both taurine-conjugated and unconjugated forms were only requirements for *iRno*. (**D**) Inorganic metabolites were allowed unconstrained consumption rates of -10^6 fmol $cell^{-1}$ $hour^{-1}$. (**E**) Cofactors and vitamins considered essential in humans were set to an uptake value of 1 fmol $cell^{-1}$ $hour^{-1}$ in rat and human networks.

Under strict physiological conditions, species-specific constraints were formulated for the export of bile salts by hepatocytes (**Figure 4.2D**). In addition to the unique ability of rats to synthesize muricholic acids, the relative abundances of bile acids differed significantly in a recent study that compared serum bile acid profiles of rats and humans[78]. Rat-specific and human-specific reactions were defined to produce an average bile salt measured for each organism, similar to species-specific reactions formulated for biomass synthesis. Under strict physiological conditions, a minimum flux of 0.4 fmol cell^{-1} hour^{-1} was required through a unified exchange reaction representing average bile salt production.

Quantitatively validated growth rate predictions

To interrogate the use of these reconstructions for making cell-specific predictions, we defined quantitative biomass compositions for rat and human hepatocytes. (**Figure 4.1**). Using flux balance analysis[106] (FBA) with biomass production as the objective, *iRno* and *iHsa* predicted maximum growth rates of 0.048 hour $^{-1}$ and 0.040 hour $^{-1}$, respectively, under strict physiological constraints (**Figure 4.2**). These predicted doubling times of approximately 16 hours were remarkably consistent with reported doubling times of 16.9 hours in rat[107] and 17.8 hours in human[108] hepatocyte cell cultures. Because biomass compositions and boundary conditions were independently formulated from different resources, these quantitative biomass predictions served as validation for these models and their comprehensive representations of hepatocellular growth. Physiological constraints also enable off-the-shelf use of *iRno* and *iHsa* for integration of comparative genomics data and systems-level analyses of hepatocyte metabolism. In **Chapter 6**, we discuss toxicology applications of *iRno* and *iHsa* that implement the physiological constraints and biomass formulations described in this chapter.

In **Chapter 3**, we discussed how *iRno* was uniquely capable of de novo vitamin C synthesis. We formulated a metabolic task to simulate growth in the absence of an external source of vitamin C and confirmed that vitamin C consumption was required for biomass synthesis in *iHsa* but not in *iRno*. We further investigated the functional impact of vitamin C deficiency on cellular growth by constraining the uptake of the vitamin C exchange reaction to values between 1 fmol cell^{-1} hour^{-1} (physiological consumption rate[96]) or 0 (vitamin C deficiency). The maximum theoretical flux through the biomass reaction of each model, containing equimolar amounts of vitamin C per cell (0.06 fmol cell^{-1}), was measured *in silico* using flux balance analysis (FBA)[106]. When the uptake rate

of vitamin C was decreased below 25% of normal physiological rates, the maximum possible growth rate was reduced exclusively in *iHsa* and not in *iRno*.

Despite this distinction, limiting the uptake of vitamin C within an order of magnitude of the physiological uptake rate had no effect on growth, suggesting that vitamin C is not likely a growth rate-limiting factor under normal conditions in either organism. Because vitamin C was available in excess, we used flux variability analysis (FVA) to predict which metabolic exchange reactions were constrained the most while requiring 100% flux through the biomass reaction. As a result, we found that maximum biomass synthesis rates in both species were limited by essential amino acid availability rather than other vitamins or nutrients. Because species-specific amino acid ratios were incorporated into the biomass formulation, this result also explained minor differences in maximum growth rates between the species. This relatively simple exercise highlights the potential usefulness of a quantitatively accurate biomass formulation to generate hypotheses about rate-limiting substrates between species.

Thermodynamically infeasible futile cycles contribute to unreasonable ATP yields

As we mentioned in **Chapter 2**, *iRno* and *iHsa* were constructed as an expansion of the Human Metabolic Reaction 2.0 database[19] (HMR2). An important advantage of *iHsa* (and *iRno*) is that one unit of glucose regenerates 25.6 units of ATP with an unlimited supply of oxygen and 2 units of ATP in the absence of oxygen. In HMR2[19] and other human GENREs[16,17], one unit glucose yields a nearly infinite amount of ATP regeneration capacity due to thermodynamically infeasible loops. Below, we describe some of the improvements incorporated into *iRno* and *iHsa* that were necessary to overcome unrealistic ATP yields produced by previous human models.

In **Chapter 1**, we demonstrated that Toycon1, a toy metabolic network consisting of 9 reactions, was capable of accurately capturing theoretical ATP yields. An important reaction in energy metabolism is ATP synthase, which is represented in Toycon1 with the reaction formula: adp + pi + 4 h[c] ⇔ atp + h2o + 4 h[m], which describes the regeneration of ATP from ADP driven by flux moving down the proton gradient between mitochondrial and cytosolic compartments. To provide an example of how simple mistakes can lead to catastrophically inaccurate predictions in a minimalistic representation of energy maintenance, we used Toycon1 to predict maximum ATP yields after adding the following two reactions: a reversible transport reaction representing the passive diffusion of lactate across the mitochondrial membrane (lactate[c]⇔lactate[m]) and a reversible transport reaction representing the symport of lactate and a proton across the mitochondrial membrane (lactate[c] + h[c] ➔

lactate[m] + h[m]). When adding these two reactions which are also present in the first human GENRE (*Homo sapiens* Recon 1[16]), the sky's the limit for ATP regeneration (i.e. maximum ATP production will infinitely scale with any arbitrarily large boundary conditions set by the user to represent infinity). The problem is caused by a violation of thermodynamics: enzymes in the electron transport chain (ETC) primarily shuttle protons across the mitochondrial membrane to maintain a pH gradient that drives ATP synthase. This process requires energy from electron carriers like NADH because it is thermodynamically unfavorable. Thus, allowing lactate to freely enter into the inner mitochondrial space and exit with an additional proton would not be thermodynamically feasible under normal physiological conditions. To resolve this issue that bypasses the need for oxidative phosphorylation in driving ATP synthase, lactate should only be allowed to exit mitochondria by diffusion (or with active transport) and should also require a proton for entry because facilitated cotransporters typically move metabolites down an electrochemical gradient, not up. Under extreme conditions, these transport reactions might be able to reverse direction, but realistic physiological approximations should always be assumed when generating a genome-scale network representing for off-the-shelf use by other researchers. Surprisingly, an earlier version of the BiGG database (www.bigg.ucsd.edu) included a comment for the reversible lactate transporter (D_LACtm) in Recon 1 that mentioned it might contribute to free ATP regeneration. Although the comment is no longer available, the reaction has yet to be changed to irreversible.

Updated human GENREs have since resolved this exact problem that enabled Recon 1 to generate ATP without a carbon-based energy or "fuel" source like glucose; however, we found that Recon 2 and HMR2 included other thermodynamically infeasible reactions to augment ATP yields by shuttling protons across the mitochondrial membrane. Unlike the example described above, other problematic reactions participated in futile cycles involving dozens of reactions across multiple compartments. To ` problematic reactions, we used flux variability analysis (FVA) and parsimonious flux balance analysis (pFBA). Despite knowing which reactions were required for unrealistic ATP production, identifying a solution was not trivial because simulation results included reactions that participated in ATP regeneration but were not necessarily problematic. Furthermore, multiple problems contributed independently to unrealistic ATP yields and not every problem was related mitochondrial oxidative phosphorylation. It was especially difficult to identify which transport reactions were problematic because proton movement could be concealed when an acidic metabolite was transported between compartments and then converted into its conjugate base, releasing the proton.

To avoid introducing thermodynamically infeasible during the network reconciliation and manual curation process described in **Chapter 3**, we defined new tasks to ensure ATP production capabilities were consistent with known yields after updating *iRno* and *iHsa*. To avoid direct proton movement problems as we described above for lactate transporters, we defined a task to regenerate ATP from ADP without a carbon-based fuel source a task that was intended to fail. We also verified that excess oxygen and inorganic ions would not increase ATP yields above 50 given a unit of glucose, which we observed was possible with multiple human GENREs.

Bulk reactions can contribute to mass balance violations

We encountered an unexpected issue involving *de novo* synthesis of essential fatty acid by bulk reactions from HMR2, which we identified when designing and simulating metabolic tasks that synthesize biomass components. In addition to the quantitative task formulated for biomass as a whole, we simulated the synthesis of individual biomass components (**Figure 4.1**) given minimal nutrients and less than minimal nutrients. For example, we verified that protein synthesis was not possible without an external supply of essential amino acids and glucose. For membrane lipid synthesis, we utilized the existing lipid subsystem of reactions to generate averaged lipid metabolites that accounted for the relative proportions of saturated and unsaturated fatty acid chains. As a result, we found that one of the tasks that was supposed to fail lipid synthesis actually succeeded without the input of essential fatty acids. Using FVA, we found that a bulk lipid degradation reaction produced slightly different ratios of fatty acids than the bulk lipid synthesis reaction consumed, thus enabling a route for creating mass in the form of essential fatty acids at the cost of consuming nonessential fatty acids. Due to a minor accounting mistake in the relative ratios produced and degraded by these reactions from HMR2, *iRno* and *iHsa* were initially capable of synthesizing membrane lipids without external sources of essential fatty acids, which like amino acids, should have been required. We resolved the discrepancy between metabolic tasks and model predictions by removing the lipid degradation reaction because we were primarily interested in biomass synthesis.

Mass balance is also important for metabolic networks because imbalanced reactions can participate in futile cycles that infinitely increase the availability of metabolite that acts as a fuel source for ATP regeneration. Although mass balance violations are commonly found in metabolic networks and usually addressed as updates are published, mass balance violations associated with bulk metabolites like those used in our new system of biomass reactions are a new possible limitation that should considered when formulating metabolic tasks. We did not expect that bulk synthesis and degradation reactions could accidentally enable *de novo* fatty acid

synthesis because from HMR2 had been one of the first GENREs to utilize multi-step bulk synthesis reactions. It is important to note that we only encountered this issue because we had assembled additional tasks related to the biomass formulation described in this chapter. Furthermore, without separating the biomass reaction into separate subcomponents (**Figure 4.1**), we would not have formulated these tasks in the first place, highlighting both a distinct advantage and a potential pitfall associated with highly complex biomass formulations.

Outlook

As a result of the efforts described in this chapter, we captured two quantitatively important functions of cellular metabolism: growth rates and ATP production yields. It is important to consider that healthy mammalian cells generally do not attempt to maximize these functions as biological objectives; however, cancer cells that proliferate uncontrollably likely operate near an optimal growth rate and metastatic cancer cells require additional energy for migration[109]. We implemented this comparative biomass framework to study toxicology in **Chapter 6** by generating treatment-specific networks of rat and human hepatocyte metabolism based on a gene expression algorithm that requires a biological objective. Instead of requiring optimal growth rates, we constrained rat and human subnetworks to perform basic hepatocyte functions that required energy maintenance and sub-optimal growth rates as described in literature. Overall, these improvements demonstrate the quantitative capabilities of *iRno* and *iHsa* and increase our confidence in their abilities to investigate other biological objectives under physiological constraints such as biomarker production which we describe in **Chapter 7**.

Acknowledgements

This chapter was adapted from a manuscript is currently under revision at *Nature Communications* titled "Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions" and written by Edik M. Blais, Kristopher D. Rawls, Zhuo I. Li, Glynis L. Kolling, Ping Ye, Anders Wallqvist, and Jason A. Papin. I thank all co-authors who contributed to work. I also thank Issac Li and Michael Carter for improving annotations in the rat and human networks and Kevin D'Auria, Jennifer Bartell, and Arvind Chavali for providing feedback on an earlier version of the manuscript.

Chapter 5: Network-guided improvements to genome annotations

Synopsis

In this chapter, we describe computational methods for guiding the improvement of genome annotations using GENREs. Once constructed, a GENRE can be used to integrate data from heterogeneous sources and generate hypotheses for subsequent experimental validation. Constraint-based modeling techniques such as Flux Balance Analysis (FBA) can facilitate the identification of knowledge gaps within a metabolic network when computational predictions do not match experimental observations. Using topologically-based and biologically-inspired metabolic network refinement, we can better characterize enzymatic functions present in an organism and link these functional annotations to genome sequences. At the end of this chapter, we provide examples of how we applied these techniques during the reconstruction, refinement, and reconciliation of rat and human GENREs described in previous chapters.

Motivation to continually improve genome annotations

A genome-scale network reconstruction (GENREs) serves as a platform for organizing a network of biochemical and transport functions in the context of an annotated genome. Complete assemblies of genome sequences are currently available for over 4000 organisms including at least 300 eukaryotic and 30 mammalian species[110]; however, curated GENREs have only been established for a fraction of these genomes. The ability to generate high quality GENREs for bacterial and eukaryotic organisms depends on the availability of high quality genome annotations. From over 2000 genomes that have been sequenced, around 40% of the protein products are not annotated to a specific function[111]. Over 5000 enzymatic functions have been described across all species, approximately one third have no known corresponding genes or proteins[112,113]. Bridging these gaps of knowledge between gene and function is important to fully utilize data available in the post-genomic era.

Protein structure and function is often well conserved between organisms, so we can infer function between homologous genes or proteins across organisms[114]. This is useful for identifying potential functions of uncharacterized ORFs (see **Figure 5.1**), such as those in a newly sequenced genome. However, this approach can be limited because there may be multiple homologous sequences when comparing a query sequence against millions of sequences from other organisms, and in reality only one of the many homologous sequences may actually demonstrate the appropriate enzymatic activity. The error rate for function assigned by sequence similarity may be as high as 49% [115]; thus, a more guided approach is necessary for assigning function to ORFs.

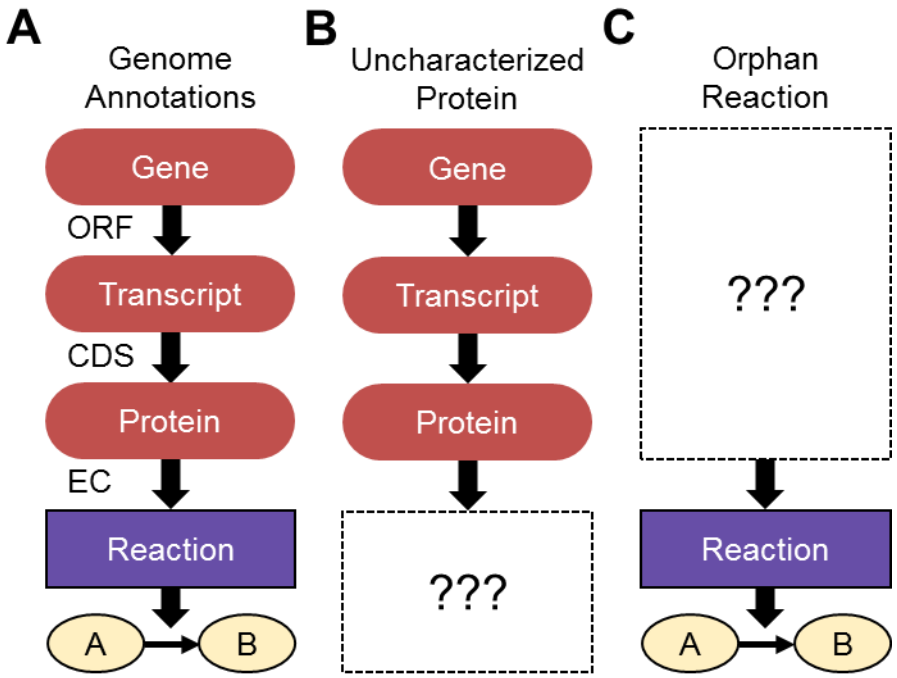


Figure 5.1 – Knowledge gaps between structural and functional annotations. (**A**) Multiple types of information can be annotated to the genome of an organism. An Open Reading Frame (ORF) describes a sequence within the genome that is transcribed from a gene into a transcript. A Coding DNA Sequence (CDS) describes the coding region of an ORF that is translated into a protein after the transcript is processed. An Enzyme Commission (EC) number describes the enzymatic activity of a protein. EC numbers is an example of a functional annotation and ORFs and CDSs are examples of structural annotations. (**B**) Structurally annotated proteins that have no known functional annotations are considered uncharacterized proteins. (**C**) Enzymatic functions that are known to occur in an organism but are not annotated to any genes are considered orphan reactions.

Genome-scale metabolic reconstructions serve as a platform for integrating data from heterogeneous sources and generating hypotheses for further experimental validation. Metabolic networks are constructed from existing genome annotations and manually expanded from literature-based sources and biochemical information contained in publicly available databases[116]. Resulting models contain a comprehensive set of known biochemical reactions and their associated ORFs. Implementing a systems-level approach allows for the identification of potential gaps in knowledge based on discrepancies between model predictions and experimental data (e.g., gene essentiality screens) as well as topological features of the network (e.g., pathways resulting in dead-ends). With the assistance of semi-automated algorithms and manual inspection, we can fill in these knowledge gaps by modifying the network to include additional biochemical reactions that were previously missing, or by removing functions that were improperly added by previous annotators. By finding ORFs encoding

for enzymes orthologous to those that catalyze the same functions in other organisms, we can improve both the structural and functional annotation of the genome for an organism of interest, while also creating a higher quality metabolic network.

Genome annotations can be categorized into two major types: structural annotations and functional annotations (**Figure 5.1**). Structural annotations describe the locations and sequences of genomic regions that encode an ORF for a transcribed gene or a CDS for a translated transcript. Functional annotations catalogue biological knowledge for a protein product such as catalytic activity, enzyme regulation, cofactor binding sites, and numerous other attributes represented by Gene Ontology (GO) terms[117]. Several computational techniques are available to structurally annotate ORFs directly for a newly sequenced organism[118,119], but functional annotation can be more challenging to assign as the sequence of an ORF alone does not necessarily describe its biological function[120]. Even once a genome has been completely sequenced, determining the structure of ORFs can be difficult with complex initiation, termination and splicing rules, and imperfect gene-calling algorithms[121]. Additionally, only around 1% of protein sequences have experimentally-derived annotations[122]; thus, computational techniques are necessary for feasibly assigning functional annotations.

Methods in this chapter are geared towards improving the annotation of organisms for which a genome-scale reconstruction is available. Automated methods are available to generate a draft GENRE based on an annotated genome[123,124] as discussed in **Chapter 2**. In this chapter, we describe methods to:

1. Predict missing and mis-annotated biochemical reactions for a given organism using metabolic network reconstructions. These include biologically-inspired refinements, which bridge the gap between model predictions and experimental data, as well as topologically-based algorithms that find and fill blocked pathways in a given network. These methods help improve the functional annotation of the genome for the organism of interest as well as improve the predictive ability of the metabolic model.

2. Assign candidate ORFs to novel functions as well as to existing functions that lack ORFs (orphan reactions, see **Figure 5.1**). These relationships provide the link between functional and structural annotation, which are both important to a higher quality annotation and metabolic model.

3. Use a systems approach to decide on which network modifications to include (and further validate) when posed with multiple gap filling solutions.

4. Perform experiments to verify existence of candidate ORFs. This will help strengthen our confidence in both the structural and functional annotation of the genome in the organism of interest.

Survey of computational resources for network-driven annotation improvements

The Enzyme Commission (EC) classification system is used to define enzymatic activities that can occur within different organisms, and an EC number characterizes in part the functional annotation for an enzyme (and correspondingly for the catalyzed reaction(s)). EC's are classified according to the following hierarchical scheme: EC-1 (oxidoreductases), EC-2 (transferases), EC-3 (hydrolases), EC-4 (lyases), EC-5 (isomerases) and EC-6 (ligases). There can be several sub-classes under these six categories. For example, the enzyme hexokinase, which is associated with an EC number of 2.7.1.1, belongs to the class on 'transferases' (enzymes that aid in the transfer of a functional moiety from one metabolite to another) and the subclass on 'transferring phosphorous-containing groups'. Other enzymes in the same subclass include glucokinase (2.7.1.2) and galactokinase (2.7.1.6). The ENZYME database (http://enzyme.expasy.org/) contains detailed information on EC numbers.

BLAST (Basic Local Alignment Search Tool) calculates sequence similarity scores between sequences of amino acids or nucleic acids[125]. This bioinformatics tool enables quantitative, high-throughput comparisons between an individual sequence of interest (in FASTA format) and a reference database of sequences. BLAST can facilitate the identification of paralogs within a species as well as orthologs between species. Although variations of the BLAST algorithm have been developed for specific biological contexts, BLAST results typically contain a quantitative similarity score between sequence pairs known as an E-value[114,126] (Expectation-value). A common application of a BLAST result is to assign EC numbers between genes that share high sequence similarity. BLAST implementations are available through web-interfaces at NCBI (http://blast.ncbi.nlm.nih.gov/), ExPASy (http://web.expasy.org/blast/), and UniProt (http://www.uniprot.org/).

In this chapter, we describe two strategies that utilize BLAST to assign functional annotations to uncharacterized ORFs.

- **Forward BLAST**. The first strategy, called forward BLAST, compares the sequence of an uncharacterized ORF against a database of ORFs with known functional annotations from other organisms. The forward BLAST strategy can be used to infer functional annotations from genes or proteins that are orthologous to the input sequence. In cases where multiple distinct functions are suggested by forward BLAST, protein annotations from closely related species should be prioritized.

- **Reverse BLAST**. The second strategy, called reverse BLAST, identifies potential new metabolic reactions that can be added to fill in knowledge gaps within a metabolic network. Reverse BLAST compares the sequence of a functionally annotated ORF against genome sequences from an organism with a metabolic network. Because reverse BLAST starts with a known function instead of a structurally annotated sequence from an organism, this approach can be used to identify candidate ORFs that have not previously been annotated. Manually identify proteins in other organisms that share the same enzymatic function (EC number) from MetaCyc or similar database. Choose proteins from phylogenetically similar organisms first. Perform BLAST for each candidate protein against the whole genome of the modeled organism to identify similar sequences. Additional details on different BLAST techniques can be found in a previous study[114]. Results are limited to local orphans as they rely on enzymatic functions linked with ORFs in other organisms.

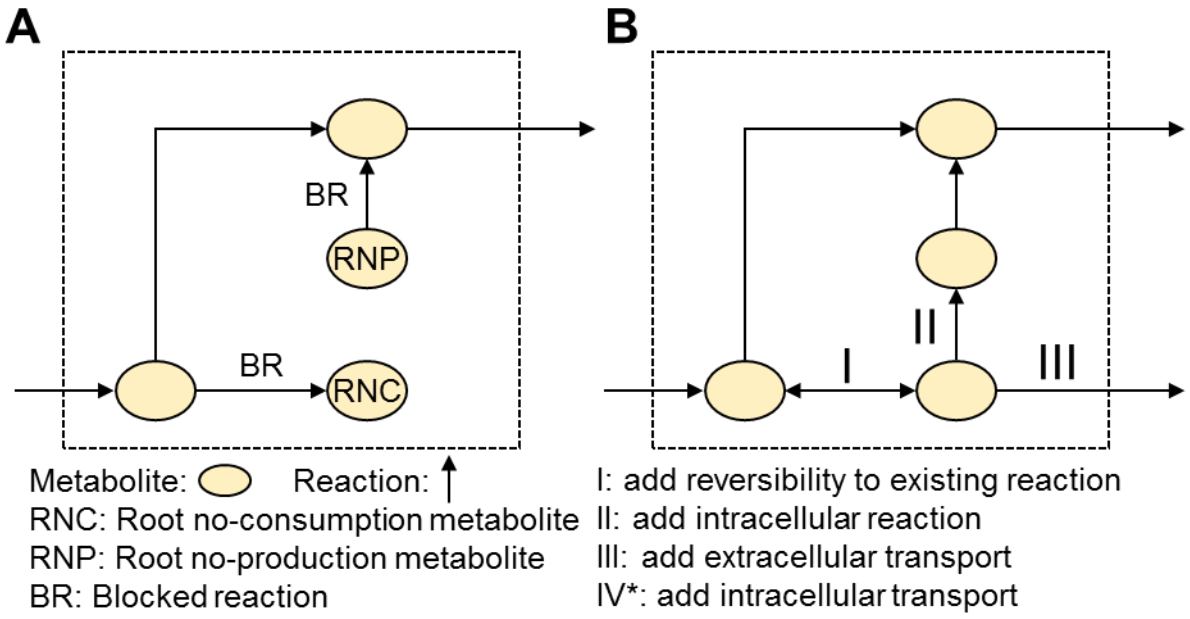


Figure 5.2 – Gap filling solutions that can restore flux through dead-end metabolite or a blocked reaction. (**A**) Toy network diagram depicting two blocked reactions (BR) caused by root no-consumption (RNC) and root no-production (RNP) metabolites. (**B**) Different categories of reactions can be added to the network in order to restore flux through the RNC metabolite: (Category I) Added reversibility to an existing reaction, (Category II) Added intracellular reaction, (Category III) Added extracellular transport reaction, (Category IV, not shown) Added intracellular transport reaction. Note that in this case the Category II solution restores flux through both dead-end metabolites.

Reaction databases

In the process of curating a GENRE, four categories of network modifications can be applied to the stoichiometric matrix (**Figure 5.2B**): modifying reversibility to an existing reaction in the network (**Category I**); adding or removing enzymatic or spontaneous reactions within a single compartment (**Category II**); adding or removing extracellular transport reactions and/or exchange reactions that enable a metabolite to be consumed or secreted (**Category III**); and adding or removing an intracellular transport reaction between intracellular compartments (**Category IV**). Detailed explanations for classifying reactions into Categories I-III were described previously[127]. Category IV reactions were introduced in this study to distinguish transport reactions from exchange reactions because mammalian GENREs frequently include multiple intracellular compartments.

Metabolic network often incorporate established biochemical relationships between reactions and metabolites from various publicly available databases and experimental literature sources. Below, we provide a list of some important publicly available biochemical databases that contain information on genome, enzymes, reactions and/or pathways. The list below is not intended to be comprehensive; rather, it provides a flavor for the kinds of publicly available resources that can be used in the genome-scale metabolic reconstruction and modeling process.

- **KEGG** (Kyoto Encyclopedia of Genes and Genomes) database contains comprehensive data on known enzymatic reactions that occur across various organisms[33,128,129].
    - Availability: http://www.genome.jp/kegg
- **ExPASy** (Expert Protein Analysis System) contains comprehensive information on EC numbers and protein structure[130,131]. For more information, visit http://expasy.org/.
- **SEED** allows for quickly generating automated draft metabolic networks for prokaryotic organisms of interest[132]. For more information, visit http://www.theseed.org.
- **MetaCyc** contains comprehensive information on pathways and enzymes across many organisms[133,134]. For more information, visit http://metacyc.org/.
- **GeneDB** is a pathogen genome database maintained by the Wellcome Trust Sanger Institute[135]. For more information, visit http://www.genedb.org/Homepage.

- **MetRxn** allows queries of a comprehensive metabolite/reaction database and comparisons of metabolites/reactions between KEGG, MetaCyc, several metabolic reconstructions, and more[136]. For more information, visit http://metrxn.che.psu.edu/.
- **UniProt** is a comprehensive knowledgebase of annotated protein sequences across many organisms[137]. For more information, visit http://www.uniprot.org/.
- **Metabase:** A wiki database of biological databases[138]. For more information, visit http://metadatabase.org.
- **Published GENREs**: Metabolic networks are typically available as SBML files or in a Spreadsheet format[139]. Published GENRE can usually be found in the supplementary material of the original article or on a website hosted by the corresponding author (e.g. http://bme.virginia.edu/csbl/Downloads1.html). Additional repositories for GENREs include the BiGG database[140,141] (http://bigg.ucsd.edu/), MetaCyc[58] (http://metacyc.org/), MEMOSys[142], MetaNetX[143] (http://metanetx.org/), and SEED[132] (http://modelseed.org/).

Constraint-based methods

The **Constraints Based Reconstruction and Analysis (COBRA) Toolbox**[144] (https://github.com/opencobra/cobratoolbox) is a software package available for the MATLAB programming environment that includes several functions that can be useful for identifying and filling knowledge gaps in a GENRE. An alternative to the MATLAB version of the COBRA toolbox is **COBRApy**[145] (https://github.com/opencobra/cobrapy), a software package that implements similar functions in the Python programming environment. COBRA features such as writing Systems Biology Markup Language (SBML) formatted metabolic models[139] require installation of the **SBML Toolbox**[146] (http://sbml.org/Software/SBMLToolbox).

- **Flux Balance Analysis (FBA)** identifies a flux distribution through the reaction network that produces an optimal flux through the objective function. Availability: COBRA Toolbox 2.0 under optimizeCbModel()
- **Flux Variability Analysis (FVA).** The COBRA function, *fluxVariability*, implements FVA and computes the ranges of possible fluxes for all reactions in a network while still maintaining a primary objective flux value such as optimal biomass production[82,147]. Using FVA, reactions produce minimum and maximum flux values of zero are considered blocked.
- **GapFind**. The COBRA function, *gapFind*, identifies dead-end metabolites in a network including root no-production and root no-consumption metabolites[148] (**Figure 5.2**).

- **DetectDeadEnds**. The COBRA function, *detectDeadEnds*, identifies dead-end metabolites that participate in only one reaction within the S-matrix. Supplying *detectDeadEnds* with a COBRA model will return a list of all metabolites that participate in only one reaction (optionally excluding extracellular metabolites)[144]. This method does not return all possible dead-end gaps, but it detects metabolites that participate in only one reaction: including reversible reactions (Category I solution of a dead-end), which *GapFind* will not detect because these are not technically dead-ends. It may be preferred to identify Category II solutions to incorporate the metabolite into metabolic pathways.
- **SMILEY**. To restore activity to blocked reactions, the COBRA function, *growthExpMatch*, implements the SMILEY algorithm and predicts a minimum set of enzymatic or transport reactions for addition that will restore flux through a blocked reaction[149]. For each blocked reaction, set the objective function to the blocked reaction, and then execute *growthExpMatch* with a relevant reaction database such as KEGG. For unblocking reversible reactions, solutions that restore flux through a blocked reaction in one direction will not necessarily restore flux in the opposite direction. To generate solutions that unblock the reaction in both directions, separately set the objective function to minimize and maximize flux through the blocked reaction. For unblocking multiple reactions with one suggested set of reactions, set the objective to maximize flux through a set of reactions that are blocked. Although SMILEY cannot suggest reactions for removal, SMILEY can suggest Category I, II, III, and if not de-compartmentalized, Category IV reactions. SMILEY has been extensively used for gap filling purposes as previously reported[127,150].

The Pathway Tools software environment (http://biocyc.org/download.shtml) integrates genome, pathway, and regulatory data for analysis and visualization[151,152]. These functions often utilize data from MetaCyc and UniProt as a reaction database and can be used as an alternative to the COBRA toolbox:

- **MetaFlux**. The MetaFlux algorithm utilizes multiple gap-filling methods to aid in developing metabolic models and defining a feasible biomass reaction[153]. This algorithm suggests a maximum subset of biomass metabolites that can be produced given a minimum set of network modifications in the form of added/removed reactions. This algorithm accommodates an initial biomass reaction that may include some metabolites that are unable to be produced. Network changes can be manually inspected for feasibility using visualization software that is integrated with the Pathway Tools platform.

- **PHFiller**. The PHFiller (Pathway Hole Filler) algorithm finds genes for orphan reactions using BLAST and protein databases[154]. Execute PHFiller to identify proteins that catalyze the orphan reaction of interest. PHFiller returns lists of candidate ORFs from the organism's sequenced genome that may catalyze each orphan reaction. This algorithm semi-automates performing Reverse BLAST using existing protein databases, and results are limited to local orphans as they rely on enzymatic functions linked with ORFs in other organisms.
- **PHFiller-GC**. The PHFiller-GC (Pathway Hole Filler – Genome Context) algorithm extends the PHFiller algorithm to use a context-specific prediction of genes for orphan reactions based on shared pathways, shared operons between proteins, shared proteins in a complex, and regulatory interactions[155]. PHFiller-GC improves upon PHFiller by considering not only BLAST sequence similarities, but also similarity based on other associations such as shared protein complexes, shared operons, regulatory elements, and transcription factors. This allows the algorithm to explore more complex components of genome annotations to identify candidate sequences for global orphan reactions.

Additional gap filling methods are available as stand-alone software tools:

- **GapFill**. The GapFill algorithm suggests adding reactions from a reaction database to restore flux through dead-end reactions[148]. Supply GapFill with a reaction database, a list of all reactions from the network reconstruction, and root no-production/consumption metabolites. Only category I, II, and III suggestions will be returned. GapFill requires an irreversible GENRE which includes reactions in the forward direction. Reversible reactions are represented in an irreversible network by adding a new forward reaction that consumes/produces the same metabolites in the opposite direction and by constraining lower reaction bounds to at least 0.
- **GrowMatch**. Using a reaction database, the GrowMatch algorithm suggests adding or removing reactions to reconcile differences between model predictions and gene essentiality screens and nutrient utilization assays[156]. GrowMatch also avoids the creation of new inconsistencies in other conditions. A modified version of GrowMatch is also implemented in Model SEED[132].
- **OMNI**. The OMNI (optimal metabolic network identification) method compares metabolic flux analysis (MFA) data and in silico predictions of flux distributions and suggests reactions to add/remove to better correlate predictions and experimental data[157]. Supply OMNI with a library of reactions from a reaction database such

as KEGG, a list of existing reactions that are allowed to be removed, or fluxomics data. OMNI suggestions include both added and removed reactions that can improve the consistency between fluxomics data and metabolic predictions.

- **BNICE**. The BNICE algorithm suggests reactions that can consume or produce metabolites based on reaction rules from the EC classification system[158]. Providing BNICE with pairs of dead-end metabolites will suggest feasible biochemical reactions that link the two. Suggestions may include novel reactions that may not be characterized in other organisms. Preferably, only add novel reactions when a high quality network has a few gaps that cannot be filled by other methods.

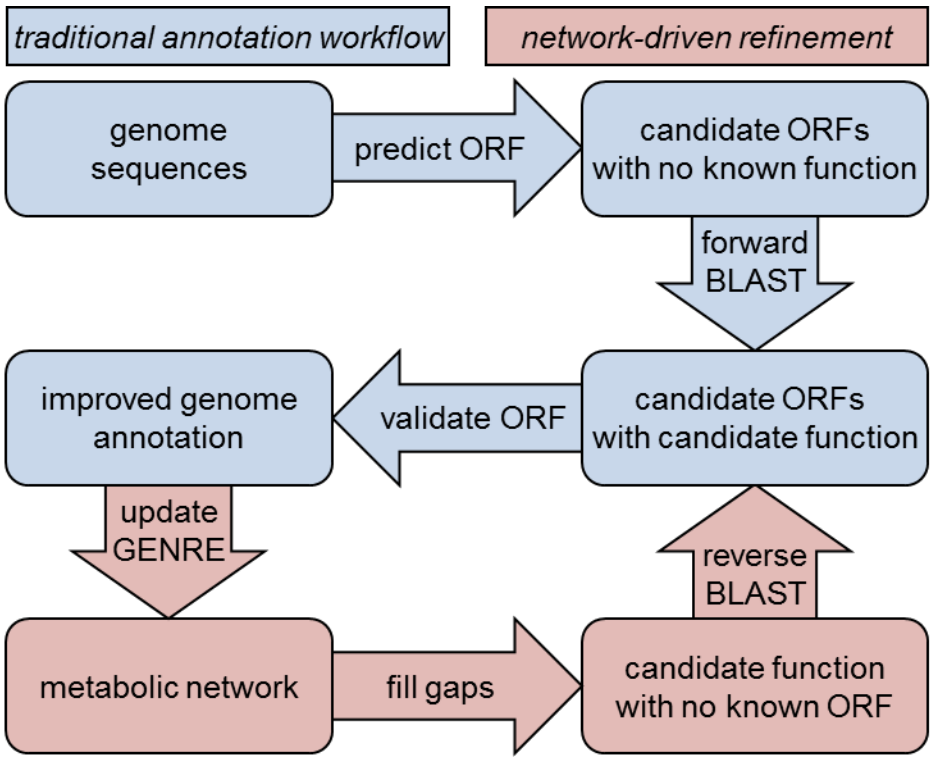


Figure 5.3 – A conceptual framework for iteratively updating metabolic networks and refining genome annotations. A traditional genome annotation workflow (blue) uses bioinformatics and experimental methods to assign new structural and functional annotations to previously unannotated sequences. Candidate functions can be inferred using forward BLAST by calculating sequence similarity between the uncharacterized ORF and functionally annotated ORFs across from other organisms. Metabolic network modeling provides an alternative computational method for assigning functional annotations to uncharacterized ORFs (red). Gap filling methods can be used to identify missing reactions in a metabolic network using existing reaction databases. Known sequences from any species that are functionally annotated to a candidate reaction can be compared via reverse BLAST to structurally annotated ORFs from the genome mapped to the metabolic network. Although forward and reverse BLAST apply similar bioinformatics algorithms, forward BLAST compares a single sequence of interest to potentially millions of annotated genes. In contrast, reverse BLAST compares functionally annotated sequences from one or more organisms to several potential sequences of interest in the original organism. By identifying knowledge gaps in metabolism networks, assigning candidate genes via reverse BLAST, and updating models and genomes with new genome annotations, this process can be performed iteratively as long as gaps exist in the updated network.

Iterative steps for refining metabolic networks and genome annotations

Here, we describe iterative steps for improving genome annotations using metabolic network modeling. Unlike traditional approaches that identify genome annotations based on a sequence of unknown function, we implemented a systems approach to identify sequences of unknown function that can be functionally annotated to novel reactions suggested by metabolic network modeling predictions[159]. This network-driven discovery process involves four major steps which are iteratively compatible with each other (**Figure 5.3**):

1. Suggest modifications to the stoichiometric matrix by adding or removing reactions based on:
    a. Manual inspection of literature evidence
        i. Early-stage examination of central metabolism
        ii. Mid-stage multi-pathway refinement
        iii. Late-stage network validations
    b. Semi-automated analysis of network topology
        i. Gap filling based on dead-end metabolites
        ii. Gap filling based on blocked reactions
    c. Semi-automated analysis of experimental data
    d. Adding or removing reactions based on prediction/experimental discrepancies
2. Identify candidate ORFs/CDSs that catalyze candidate reactions and existing orphan reactions
3. Manually choose suggested network modifications at the systems-level
4. Experimentally verify the presence and structure of candidate ORFs

Overall, the methods outlined in this chapter can be used to identify knowledge gaps in metabolic networks. Throughout and after the process of reconstructing a metabolic network, these methods serve as guidelines for updating metabolic networks, identifying new knowledge gaps, fillings those knowledge gaps, and refining genome annotations. Because genome annotation refinements also lead to improvements of the metabolic network, applying semi-automated algorithms combined with manual inspection and experimental validation

define will ultimately lead to high-quality metabolic models for the systems biology community as well as contribute to improved genome annotations for an even larger research community.

Biologically-inspired metabolic network refinement

At all stages in the reconstruction process, it is important to evaluate the functionality of the network model so that any subsequent modifications consistently lead to a higher quality model. Any deficiencies in model functionality should be manually examined so as to identify enzymes/reactions to fill in these gaps in knowledge.

1. *Early-stage examination of central metabolism.* While drafting a reconstruction, examine central metabolic pathways (e.g. glycolysis, TCA cycle and pentose phosphate pathway) with literature support for completeness. Visualize these pathways in biochemical reaction databases (see Subsection 2.2) such as KEGG. For example, to evaluate functionality of glycolysis in a newly reconstructed metabolic network, use FBA to optimize for pyruvate production via the pyruvate kinase reaction under glucose-only nutrient conditions. If zero flux is obtained for the objective, manually check for gaps or deficiencies in the pathway.

2. *Mid-stage multi-pathway refinement.* At intermediate stages of model building, expand this process from individual pathways to include multiple pathways (such as amino acid and nucleotide metabolism). For example, use FBA to simulate the production of individual amino acids and nucleotides given a particular carbon source (e.g. glucose). It is important to utilize experimental literature evidence in this process (for example, not all organisms can synthesize all 20 amino acids *de novo* and may need to scavenge necessary amino acids from the environment).

3. Late-stage network validations. In the later stages of the reconstruction process, ensure basic functionality of the model in the context of a biologically-inspired objective function such as biomass or ATP production. A semi-automated algorithm for establishing a feasible biomass is MetaFlux.

Identification of knowledge gaps using network topology

At any stage in the reconstruction process, there may be blocked reactions, which are reactions that cannot carry flux in a metabolic model, usually as a result of containing, being upstream (root no-consumption), or being downstream (root no-production) of a dead-end metabolite (see **Figure 5.2**). Unblocking these reactions can be performed at the metabolite level, aimed at restoring fluxes that utilize the dead-end metabolite, or at the reaction level, aimed at restoring flux through the blocked reaction. These two general methods may provide different gap filling solutions, though they share the same root causes.

For network-topology based methods in the COBRA toolbox, limit nutrient consumption to a small value (-1 assuming that negative exchange fluxes represent nutrient uptake) and constrain default bounds of all other reactions to a large number such as 100000. Setting intracellular reactions to extremely large values, and opening all exchange reactions may improve the functionality of network topology-based methods[144]. More specifically, set the [lower bounds, upper bounds] of: forward-only reactions [0, 10000]; reverse-only reactions to [-10000, 0]; reversible reactions to [-10000, 10000]; and exchange reactions to [-1, 10000]. If an algorithm incorrectly identifies gaps, then increasing the magnitude of non-uptake bounds by 10, 100, or 1000 can fix some issues.

At the metabolite level: restoring flux through dead-end metabolites

1. (optional) De-compartmentalize the model into only intracellular and extracellular compartments (i.e. replace non-redundant reactions within sub-cellular compartments such as mitochondria, endoplasmic reticulum, or nucleus with cytosolic reactions). Dead-end metabolites may block additional reactions based on compartmentalization, and in one instance this was addressed by de-compartmentalizing the human metabolic model[127]. Substantially fewer reactions were blocked suggesting that this approach can be used when too many reactions are blocked. De-compartmentalization removes the possibility of addition of Category IV reactions.
2. Identify dead-end metabolites using GapFind or DetectDeadEnds:
3. Suggest reactions that include dead-end metabolites as products for root no-production cases or as substrates for root no-consumption cases. Semi-automated algorithms for identifying candidate reactions directly based on input metabolites include GapFill and BNICE.

At the reaction level: restoring flux to blocked reactions

1. Identify blocked reactions using FVA. Consider all reactions that have lower and upper flux ranges of approximately [0, 0] as blocked reactions. Specifying different media conditions for exchange reactions can be useful for identifying blocked reactions in specific contexts. Loops can be allowed or avoided when running FVA by setting the parameter, *allowLoops*, to 1 or 0, respectively; allowing loops will hide reactions that cannot carry flux unless a potentially thermodynamically infeasible loop is carrying flux, which may or not be relevant to the biology of the model. FVA calculates the full possible ranges of flux values for all reactions

while maintaining a set percentage (default: 100%) of maximal flux through an objective. If this percentage is set to zero, the objective function will not be important. For identifying blocked reactions in the context of simulating biomass, set the parameter, *optPercentage*, to a lower value such as 10%. In this case, maintaining 100% flux through the biomass reaction can often limit flux through alternative pathways. If all flux from a rate-limiting carbon source is allocated to biomass, then alternative reactions in non-optimal pathways that utilize carbon will result in flux ranges of [0, 0] flux, yielding unnecessarily blocked reactions.

2. Suggest reactions that can be added to restore flux through each blocked reaction using SMILEY. In this case, solutions directly alleviate blocked reactions and may indirectly resolve dead-end metabolites.

Gene essentiality screens, nutrient utilization assays, and fluxomics data suggest experimentally-inspired model refinements

Implementing FBA on metabolic network reconstructions provides the ability to predict growth yields for organisms under different substrate conditions and genetic perturbations[160]. Implementing FBA on the *iAF1260* metabolic reconstruction of *E. coli* yielded an accuracy of 91% for gene essentiality predictions as compared to experimental observations [161].

Adding (or removing) reactions to reconcile predictions with experimental data:

1. Define a biologically relevant objective function for the metabolic model [124,162]. Consider a biomass function of nucleic acids, amino acids, lipids, and energy maintenance when comparing predictions to growth assays. Some experimental screens measure secretion of metabolites or other phenotypic properties, so adjust the objective of the metabolic model accordingly.

2. Define nutrient conditions relevant to a biological setting. Consider carbon, nitrogen, phosphorus, and sulfur sources as well as presence of oxygen (See Ref. [150] for an example of establishing a minimal media relevant to biological conditions).

3. Manipulate the model to emulate any experimental perturbations. For example, remove relevant reactions from a COBRA model using *removeRxns* or *deleteModelGenes* for essentiality screens.

4. (optional) Ensure feasibility of the biomass function for at least one condition manually by FBA. Using algorithms in this section with an objective that requires a large number of additional reactions is usually neither computationally feasible nor biologically relevant.

5. Identify knowledge gaps by comparing experimental data to model data. Perform FBA for each condition and compare output to observed result (See **Table 5.1**). Possible discrepancies occur when:

    a. Model predicts growth (or other relevant output) when no growth is experimentally observed; as a result, remove reactions from network

    b. Model predicts no growth (or other relevant output) when growth is experimentally observed; as a result, add reactions to network

6. Suggest reactions to add (or remove) using one or more of the following semi-automated algorithms: SMILEY, GrowMatch, or OMNI.

Proceed to identifying candidate ORFs for these reactions and narrow down choices of network modifications to experimentally validate.

Table 5.1 – Problem-driven methods to identify and reconcile knowledge gaps in metabolic models

Type of knowledge gap	Methods to identify knowledge gaps	Methods to resolve knowledge gaps	Type of network modification
dead-end metabolites cannot be consumed or produced	GapFind DetectDeadEnds	GapFill BNICE	add new reactions; category II preferred
blocked reactions cannot carry flux due to dead-end metabolites	FVA Flux Sampling	SMILEY	add new reactions; category II preferred
model predictions are inconsistent with experimental knowledge (e.g. fluxomics data, gene essentiality or nutrient utilization assays)	FBA predicts no growth when growth is observed	SMILEY GrowMatch OMNI	add new reactions; category II preferred
	FBA predicts growth when growth is not observed	GrowMatch OMNI	remove existing reactions; category I or III preferred
orphan reactions have no known GPR relationships and are not spontaneous	findOrphanRxns	Reverse BLAST PHFiller PHFiller-GC	assign ORF to orphan reaction

Annotating candidate ORFs for orphan reactions

Orphan Reactions are enzymatic reactions that have no assigned ORFs or proteins that catalyze this reaction (**Figure 5.1**). Disconnect between structural annotation and functional annotation occurs either in the scope of an individual organism (local orphan) or across all organisms (global orphan). For local orphan reactions, utilizing BLAST and BLAST-related algorithms are useful for identifying potential ORFs, but global orphan reactions have no known enzymes which we can compare sequence similarity. In a metabolic model, an orphan reaction has no relationship defined in the GPR, but the stoichiometry is defined in the S-matrix. The following steps describe the process of annotating candidate ORFs to newly added orphan reactions as well as orphan reactions already included in a GENRE.

1. After adding candidate reactions to a COBRA model, use *findOrphanRxns* to locate all orphan reactions.

2. Identify whether each reaction is a local orphan reaction or a global orphan reaction. Query a reaction database such as UniProt, KEGG, or MetaCyc to see a list of annotated gene or protein sequences across all organisms. Reactions annotated to genes or sequences in other species are considered local orphan reactions.

3. Use the following algorithms to identify candidate ORFs that may encode for enzymes that catalyze an orphan reaction of interest. Reverse BLAST, PHFiller, or PHFiller-GC.

Choosing reactions to experimentally validate

Before proceeding to experimental methods for verifying candidate ORFs and validating their function, efforts should be made to manually curate the metabolic network. The algorithms described in this chapter are defined as semi-automated because manual inspection is essential to ensuring that only biologically-relevant and plausible reactions are added or removed from the metabolic network. Using all gap filling methods described in this chapter would yield an exorbitant number of suggested network modifications, so this section outlines general considerations for selecting reactions to subsequently validate.

Use a systems approach to add or remove reactions: consider simpler solutions that resolve the most knowledge gaps with the fewest network changes before addressing each problem on its own. Instead of adding all reactions from these methods to the model at once, select only a few solutions and iterate through the model and rerun

gap-filling techniques. This iterative, systems approach to gap filling will ensure higher quality models and better genome annotations.

Considerations for adding reactions based on suggestions from semi-automated algorithms:

- When multiple solutions are available for the various gap filling conditions, choose small subsets of reactions that resolve model predictions with the most experimental conditions.
- Perform several iterations of each algorithm to produce a comprehensive set of possible reactions. When solutions are limited, try other databases where available.
- Choose reactions that have candidate ORFs when possible over those that will remain local orphan reactions.
- Examine all solutions for biological feasibility. Take into consideration literature suggestions.
- Prioritize adding Category II reactions when possible (See Table 1).
- Category I and III reactions are simplified solutions (e.g. exporting a root no-consumption metabolite out of the cell to unblock a reaction). Add category I reactions only when literature supports the reversibility of an enzymatic reaction, or a separate distinct enzyme can catalyze the opposite reaction. Add category III reactions only when literature supports the concept that a metabolite is excreted and/or a transporter or relevant channel has been characterized.
- Calculate feasible ranges of ΔG (free energy) for each reaction in the organism of interest and exclude thermodynamically infeasible reactions.
- Be cautious of creating thermodynamically infeasible loops when adding reactions. For example, the H+ gradient across the mitochondrial membrane fuels ATP regeneration from ADP through ATP synthase. The electron transport chain maintains this gradient by utilizing energy generated from other metabolic processes in order to pump H+ out of the mitochondrion, but additional reactions in a metabolic model may have contribute to this process. In a published reconstruction of human metabolism[16], a reversible mitochondrial transporter allows symport of lactate and H+ ions, while a separate reaction allows lactate to freely diffuse between mitochondria and the cytosol. These two reactions effectively serve as a thermodynamically

infeasible alternative to the electron transport chain, fueling infinite ATP production through ATP synthase. While this loop may have been created to unblock a reaction involving lactate, it inadvertently reduces the quality of the model with respect to processes that involve ATP regeneration.

Considerations for removing reactions based on suggestions from semi-automated algorithms:

- Some metabolic models have confidence scores associated with annotated GPR relationships, such as those in the BiGG database[141]. Choose to remove reactions with low confidence (annotated based on inferred function made only by sequence homology) over those with high confidence (supported with biochemical and literature evidence).

- Be cautious in removing reactions permanently. Enzymatic function may be possible at the genome-scale under the right conditions, but enzyme expression or function may be dependent on specific environmental, signaling, regulatory, or time-dependent factors that prevent the organism from adapting to a particular growth condition.

- Not all reactions need to be unblocked. Evolutionary trends may have caused a loss of a key metabolic enzyme in a pathway, leaving the other members structurally and functionally intact at the molecular level, but rendered loss in functionally in the larger scope of the metabolic network. These are called biological gaps instead of knowledge gaps [159]. Thus, be cautious and consider adding reactions only when justified in the scope of biology (See Subsection 3.6).

- Prioritize removing Category I and III reactions over Category II reactions unless literature evidence strongly supports lack of a transport or reversible reaction.

Structural annotation validation for candidate ORFs

The process of adding new reactions and adopting orphan reactions (and consequently incorporating new ORFs into the model) yields an opportunity to structurally annotate candidate ORFs through experimental validation. For each candidate transcript identified in this workflow, we can experimentally verify the presence and sequence of the ORF with the following methods:

1. Verify the presence of candidate ORFs using RT-PCR (reverse transcription polymerase chain reaction). Forward and reverse primers for 5' and 3' ends of transcript should be based on the known sequence of a candidate ORF.

2. If only a partial sequence of an ORF is available or if the previous step failed due to errors in termini, perform RACE (rapid amplification of cDNA ends) to define the transcript boundaries. Primers should be based on inner-regions of candidate ORF. Upon identification of transcript boundaries, repeat step 1 with new primers

3. Clone ORF vector into *E. coli*. The amplicons generated from RT-PCR can be cloned using a Gateway donor vector and transfected into *E. coli*.

4. Sequence ORF amplicons using either high-throughput Sanger sequencing for 5' and 3' end verification or using a Roche 454FLX Titanium sequencing system. Successful cloning and matching of the sequence from a given ORF to its predicted gene would experimentally validate the presence of the hypothesized transcript.

Additional details related to these methods[80] and applications of this experimental approach to verify candidate transcripts of the alga *Chlamydomonas reinhardtii*[123,163] have been described in recent literature.

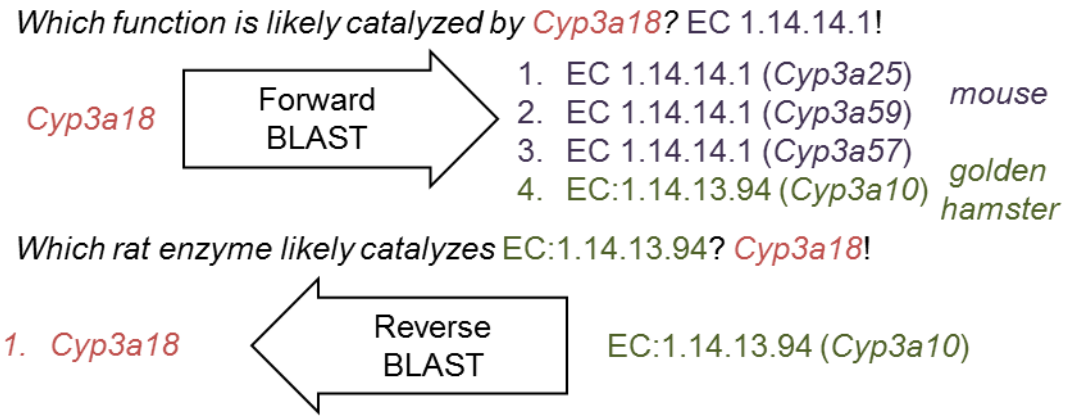


Figure 5.4 – Orphan reaction identified in rats led to updated annotation for *Cyp3a18*. When curating *iRno*, we found an experimental study demonstrating the ability of rats to perform 6-beta hydroxylation of lithocholic acid (EC 1.14.13.94) and 6-beta hydroxylation chenodeoxycholic acid (no EC number found), which was not assigned to any rat sequences. Based on knowledge that the golden hamster gene *Cyp3a10* (green) catalyzes reactions, we performed a reverse BLAST search of the *Cyp3a10* sequence against the rat genome, providing *Cyp3a18* as the top candidate for this function. In contrast, the top 3 candidates from other organisms of a forward BLAST included 3 mouse genes (purple) before *Cyp3a10* that would likely lead to the conclusion that *Cyp3a18* was an unspecified monooxygenase (EC 1.14.14.1).

Updated functional annotations for *Cyp3a18* and *Akr1c14*

In **Chapter 2**, we converted an initial human GENRE into a draft of *iRno* by inferring functional annotations across orthologous genes. Ideally, this approach can be applied to other organisms based on existing GENREs for evolutionarily related species; however, extensive manual curation was necessary to fill in several knowledge gaps despite the availability of high quality annotations for rat and human genomes. Organisms with lower quality genome annotations will likely require significant manual curation as experienced by authors of the original mouse GENRE[53]. Below, we highlight two examples of important knowledge gaps that were identified during the reconstruction, curation, and reconciliation processes described in previous chapters.

While curating the bile acid synthesis pathway, we discovered that the critical enzymatic step involved in the production of rodent-specific bile acids was not annotated to any rat or mouse genes (**Figure 5.4**). A previous study hypothesized that an unknown cytochrome P450 family 3 member could produce β-muricholic acid and murideoxycholic acid via 6-beta hydroxylation of chenodeoxycholic acid and lithocholic acid, respectively[164]; however, without a specific genome annotation assignment this reaction has remained an orphan for nearly two decades. Using Basic Local Alignment Search Tool (BLAST) (http://www.uniprot.org/blast/), we compared the Golden Hamster (*Mesocricetus auratus*) gene, *Cyp3a10* (UniProt ID: Q64148), which was reported to perform 6-beta hydroxylation of lithocholic acid[165] (EC 1.14.13.94), to rat genes. We identified *Cyp3a18* as the best candidate with the highest sequence identity to *Cyp3a10* and with protein-level evidence of expression in rat hepatocytes[166]. Additionally, *Cyp3a18* had not yet been isolated[167] until after original attempts to annotate an enzyme to this orphan reaction[164]. Furthermore, *Cyp3a18* was the only potential match with no known human orthologs, consistent with the absence of this function in humans[78]. In contrast, a traditional BLAST comparison of *Cyp3a18* against other mammalian genomes resulted in three mouse genes with higher sequence identity but different functional annotations compared to *Cyp3a10* (**Figure 5.4**), highlighting how reconciling metabolic network reconstructions can guide the improvement of genome annotations[80]. Although we have not experimentally validated this suggested reannotation, we have provided strong computational evidence and have consulted with the Rat Genome Database[30] to annotate 6-beta hydroxylation of chenodeoxycholic and lithocholic acids as enzymatic functions for *Cyp3a18* in rats (**Figure 5.5**). These annotations are also likely appropriate for *Cyp3a25* in mice based on high sequence homology with *Cyp3a18* (**Figure 5.4**).

Prior to manual curation of *iRno* and *iHsa*, the automated draft of *iRno* originally failed to complete 3 human metabolic tasks related to bile acid synthesis that have been described as functional in rats[78]. We identified that reactions catalyzed by 3α-hydroxysteroid dehydrogenase (EC 1.1.1.50) were blocked in *iRno* and not in *iHsa* because none of the 3 human genes assigned to this function were annotated to any rat orthologs. We searched literature for evidence of the presence or absence of this function in rats and found a comparative study[66] between rat and human enzymes experimentally verifying that not only did the rat enzyme, *Akr1c14*, have this function, but that it had a higher specificity than the 3 human enzymes examined. Assigning the rat gene *Akr1c14* to 3α-hydroxysteroid dehydrogenase (EC 1.1.1.50) reactions was sufficient to resolve all 3 inconsistent metabolic task predictions between *iRno* and *iHsa*[66].

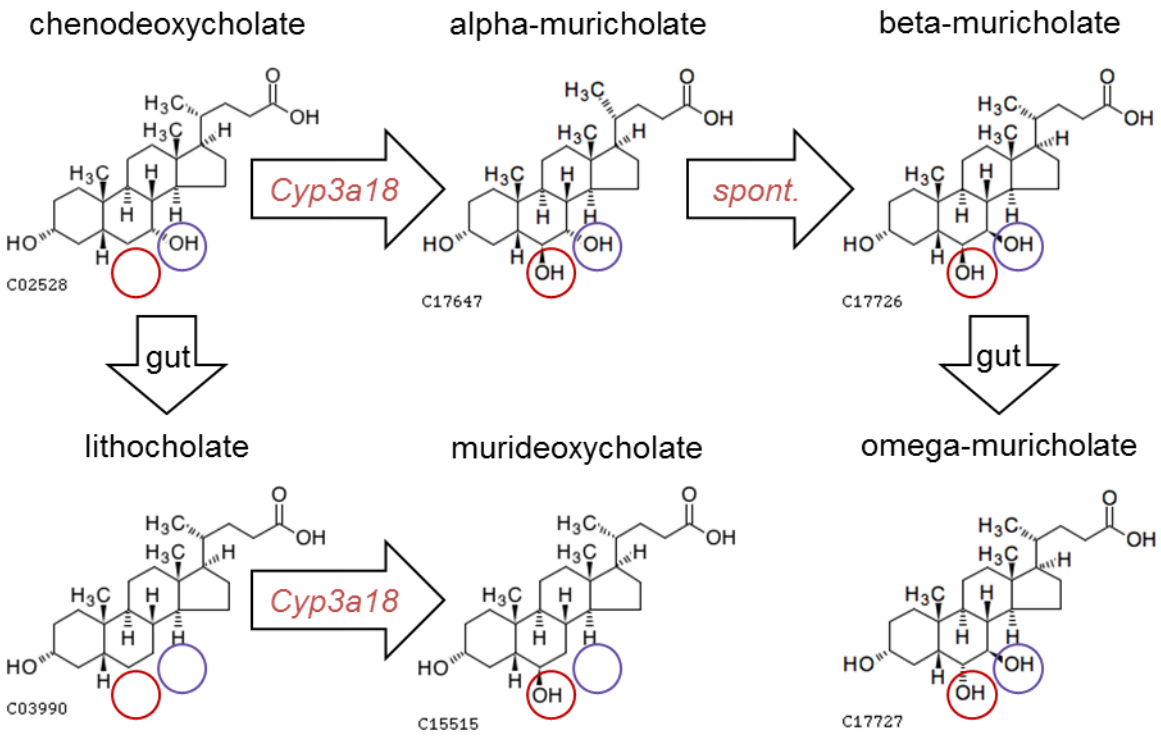


Figure 5.5 – Chemical structures of rodent-specific bile acids and precursor metabolites. *Cyp3a18* likely catalyzed 6-beta hydroxylation of both lithocholate and chenodeoxycholate based on similar structural properties of each substrate. We assumed that conversion of alpha-muricholic acid into beta-muricholic acid was likely spontaneous (abbreviated as *spont.*) or performed by the *Cyp3a18* enzyme because beta-muricholic acids are more highly abundant than alpha-muricholic acids[102]. This multi-step process requires a 7a-dehydroxylase (EC: 1.1.1.159) in *iRno* but is annotated as an orphan because we were uncertain about whether this reaction should or should not also be functional in *iHsa*. Red circles indicate rodent-specific hydroxylation sites. Purple circles indicate sites of 7-alpha hydroxylation. Chemical structures were obtained from KEGG and include KEGG compound identifiers[34].

Missing functional annotation for *Akr1c14*

Interestingly, these examples highlight two distinct types of knowledge gaps that can lead to genome annotation refinements. For *Cyp3a18*, the knowledge gap likely existed because researchers did not have the entirety of the rat genome[40] available despite their accurate speculations[164] that the rat enzyme was likely a family 3 member of the Cytochrome P450 class of enzymes. For *Akr1c14*, the knowledge gap was not due to a lack of experimental knowledge but due to missing annotations across all 5 orthology databases that we used to generate the initial draft rat network in **Chapter 2**. We expected to find more knowledge gaps involving orphan reactions like *Cyp3a18* that would contribute to more differences between rats and human; however, we encountered several knowledge gaps due to missing annotations like *Akr1c14.* Because orthology between *Akr1c14* and *AKR1C4* had not previously been annotated in any of the 5 orthology databases, we manually investigated all reactions that were annotated as human-specific after automated reconstruction of *iRno* to resolve differences attributed to missing orthology annotations. As a result of the reconciliation process described in **Chapter 3**, curated rat and human networks were even more similar than we anticipated because nearly all of the human-specific reactions in the draft rat network were resolved with experimental evidence despite a lack of orthology annotations.

Outlook

Metabolic modeling can be used to improve genome annotation by filling in network gaps and linking biological functions to gene and protein sequences. Throughout this process we want to focus on using modeling as a semi-automated tool for generating hypotheses and using manual inspection, biological reasoning, and experimental validation to revise both functional and structural annotations. Iterating through these computational and experimental steps can facilitate the curation of a genome-scale network reconstruction that is also built on the foundation of high quality genome annotations. The potential applications of the methods described in this chapter will improve as genome annotations for other organisms become more accurate and complete, providing the opportunity to guide and accelerate annotation efforts across bacterial and mammalian genomes.

Acknowledgements

Methods for refining genome annotations with metabolic networks described in this chapter were adapted from the book chapter: Linking genome-scale metabolic modeling and genome annotation. *Methods in Molecular Biology*. Jan 17, 2013[80]; written by Edik M Blais, Arvind K. Chavali, and Jason A. Papin. Discussions related to re-annotating the rat genome was partially adapted from a manuscript currently under revision at *Nature*

Communications titled “Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions” written by Edik M. Blais, Kristopher D. Rawls, Zhuo I. Li, Glynis L. Kolling, Ping Ye, Anders Wallqvist, and Jason A. Papin. I thank all co-authors who contributed to these works, notably Arvind Chavali who contributed significantly to development and description of the experimental protocols outlined in this chapter.

Chapter 6: Comparative toxicogenomics analyses

Synopsis

The laboratory rat has been used as a surrogate to study human biology for more than a century. Rats are often used as a surrogate model for understanding human hepatotoxicity; consequently, it is critically important to understand species-specific responses to toxic compounds in order to efficiently translate preclinical studies. In previous chapters, we described the creation of the first genome-scale network reconstruction of rat metabolism (*iRno*) and a substantially updated human metabolic network reconstruction (*iHsa*). By resolving known differences between rat and human metabolism in **Chapter 3**, these reconciled metabolic networks can generate predictions that are suitable for comparative analyses. To investigate systems-level differences between rats and humans, we simulated the toxicological effects of 70 compounds on rat and human hepatocytes using high-throughput gene expression data from a large-scale toxicogenomics study. Comparative analyses revealed that treatment-induced changes were inconsistent within bile acid and glutathione metabolism and generally consistent within amino acid and central carbon metabolism. Additionally, negatively correlated predictions within fatty acid metabolism were directly associated with the known mechanisms of toxicity for three hepatotoxic compounds. We also demonstrated how metabolic networks can serve as powerful computational platforms for contextualizing experimental data by highlighting important differences between model predictions and gene expression changes alone. Overall, we anticipate that the predictive framework described in this chapter will provide mechanistic insights into the consistency of rat and human biology for clinical and basic science applications.

Understanding pharmacodynamics with transcriptomics profiling

High-throughput measurements of gene expression have been extensively used to study the physiological response of healthy cells to perturbations. Within the field of toxicology, these toxicogenomics experiments provide a snapshot into molecular changes that correlate with the development of toxicity for a pharmaceutical drug. By understanding the pharmacodynamics of a drug, researchers can identify potential biomarkers of toxicity or develop counteractive measures that could reduce toxicity. In this chapter, we developed a novel comparative toxicogenomics strategy to integrate high-throughput measurements of gene expression from rat and human hepatocytes into metabolic networks to better understand potential species-specific differences within the context of pharmacodynamics.

Survey of toxicogenomics resources

Two large-scale toxicogenomics databases containing thousands of high-throughput gene expression profiles have been made publically available over the past decade: the Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system (Open TG-GATEs)[44,168] which was sponsored by the Japanese Toxicogenomics Project (http://toxico.nibiohn.go.jp) and DrugMatrix (https://ntp.niehs.nih.gov/drugmatrix) which was acquired by National Toxicology Program. Both databases include samples of primary rat hepatocytes, rat liver tissue, and rat kidney tissue exposed to pharmaceutical compounds and environmental toxicants at various doses for several hours up to multiple days. DrugMatrix includes gene expression data for a total of 638 compounds and also measures heart and thigh tissue expression for a subset of these compounds. The Open TG-GATEs includes gene expression data for a total of 130 compounds and also measures gene expression in primary human hepatocytes. To showcase the toxicogenomics applications of metabolic network modeling, we selected Open TG-GATEs because it provided an excellent resource for comparative toxicogenomics analyses. In future studies, we anticipate that more comprehensive systems toxicogenomics analyses could be performed by integrating rat expression data from the DrugMatrix database into *iRno*.

Survey of gene expression integration methods for metabolic networks

Metabolic networks serve as a platform for integrating high-throughput omics data. Several algorithms have been developed to constrain metabolic networks based on gene expression[63]; however, a systematic comparison in the predictive ability across several gene expression integration methods came to the conclusion that the appropriateness of each method depended on the biological question being asked[64]. In this chapter, our goal was to investigate whether physiological changes in response to pharmaceutical compounds and environmental toxicants were consistent between rat and human hepatocytes. When surveying various methods that were available for integrating gene expression data into rat and human metabolic networks, we found that most methods integrated absolute gene expression while we were interested in integrating relative gene expression. Implementing a method that integrates relative changes in expression was more suitable for comparative toxicogenomics analyses because we would assume that the baseline metabolic function of rat and human hepatocytes were roughly similar; however, differences in the metabolic response to a specific perturbation could be more pronounced between rat and human hepatocytes.

For the analyses described in this chapter, we utilized a method called Metabolic Adjustment by Differential Expression (MADE) that is available within the Toolbox for Integrating Genome-scale metabolism, Expression,

and Regulation (TIGER) (www.github.com/pauljensen/tiger)[48,169]. This method was chosen because it was designed to integrate differential expression rather than absolute expression. Similar to other methods, MADE attempts to disable metabolic genes with relatively low expression values. Unlike most available gene expression integration methods, expression is considered low when p-values are significantly different between two conditions. For a typical toxicogenomics study, genes that were downregulated or upregulated would turned off in the treatment condition and control condition, respectively.

An important feature of MADE is the ability to specify a minimum metabolic requirement through an objective function. In this chapter, we utilize the biomass function defined in **Chapter 4** to ensure that treatment and control models remain functional to a reasonable degree after integrating gene expression changes; however, in some rare cases, we found that the models returned by MADE were not actually function. This technical issue remains unsolved but may be caused by inappropriate rounding errors introduced during the process of solving the mixed-integer linear programming problem defined by MADE. To overcome this technical hurdle, we developed a straightforward workflow to identifying potential genes that were accidentally turned off by the MADE algorithm. First, the entire list of deleted genes suggested by MADE should be deleted using the original COBRA model to verify that the metabolic objective does not meet the pre-determined requirement using flux balance analysis (FBA). Second, iterate through the deleted list and select one gene that will not be deleted and perform FBA after deleting all of the other genes in the list. Ideally, at least one gene will restore flux to the metabolic objective when it was not deleted providing a subset of the deleted gene list that satisfies the intended requirements of the MADE algorithm. In cases where multiple genes restore the objective flux, either the gene with the least significant p-value for differential expression should be chosen based on the formulation of the MADE algorithm. Alternatively, if more than one model generated by MADE was negatively impacted by this technical issue, the fewest genes that fix the most models should be applied as a parsimonious solution. Using the latter approach, we found that un-deleting 3 genes was sufficient to restore biomass synthesis in 9 models that could not produce biomass after running MADE.

We reduced the number of non-functional models initially generated by MADE to 9 out of 280 down from a much larger number by specifying optional parameters when using MADE. ‘round_states’ was set to false because very small non-integer values were rounded down to zero resulting in deleted genes that should not be deleted. ‘set_IntFeasTol’ was set to 1e-12 to ensure that all small flux values that may be important are not interpreted

as zero. In general, a higher incidence of non-functional models generated by MADE are possible when using higher values for the 'p_thresh' parameter which encourages MADE to turn off more genes. Additionally, a higher value for the 'obj_frac' parameter was also associated with increased incidence non-functional models when running MADE because more gene-associated reactions were needed to maintain stricter biomass requirements. Despite these caveats, MADE remains one of the best methods available for integrating differential gene expression and these recommendations may be helpful when using similar algorithms that implement mixed-integer linear programming problems.

Preprocessing gene expression microarrays from rat and human hepatocytes

To explore the effects of exposure to pharmaceutical compounds and environmental toxicants on normal metabolic functions, high-throughput gene expression measurements in rat and human hepatocytes were obtained from the Open TG-GATEs database and analyzed within the computational frameworks of *iRno* and *iHsa*. Gene expression profiles from this large-scale toxicogenomics database were used to analyze reaction-level changes in response to individual pharmaceutical compounds or environmental toxicants. Raw microarray data of rat and human hepatocytes treated with 119 different compounds were available were downloaded from ArrayExpress (E-MTAB-797 for rat hepatocytes; E-MTAB-798 for human hepatocytes)[170] and pre-processed using the *oligo* package[171] in the R/Bioconductor programming environment[172,173].

Calculating gene expression changes between treatment and control conditions

Expression changes in response to each dose at the 24-hour time point were evaluated independently for each compound and organism using the *limma* package[174]. Genes with a false discovery rate (FDR)-corrected q-value less than 0.1 were considered significantly differentially expressed. For compounds tested at multiple doses, the dose resulting in the largest number of differentially expressed genes was chosen for toxicogenomics analyses independently for each compound and each organism. Expression changes evaluated in response to 70 individual compounds mapped to 1945 and 2225 metabolic genes in *iRno* and *iHsa*, respectively. Of the 119 compounds with data available for both rat and human hepatocytes, 70 were selected for model integration that significantly altered at least 15 human and 15 rat metabolic genes.

Integrating gene expression changes using MADE

Context-specific models of rat and human hepatocytes were generated independently for each compound in each organism using the MADE algorithm. For each set of expression fold changes and FDR-adjusted q-values, MADE produced a control-specific subnetwork and a treatment-specific subnetwork by assigning binary gene-

states consistent with downregulation (disabled in treatment; enabled in control) and upregulation (enabled in treatment; disabled in control) observations across 1945 rat genes or 2225 human genes (**Figure 6.1A-B**). Each condition-specific model generated by MADE was also required to maintain a minimum growth rate of one doubling per week using the hepatocyte biomass reaction as the objective under physiological constraints (see **Chapter 4** for detailed descriptions of biomass formulations and physiological conditions for hepatocytes).

By integrating gene expression changes into *iRno* and *iHsa* using MADE, we generated binary gene-states for 140 treatment and 140 control models (**Figure 6.1B**). Each condition-specific model generated was a subset of *iRno* or *iHsa* and was capable of maintaining a biomass synthesis rate of one doubling per week under physiological conditions. The proportion of gene-associated reactions altered in each network correlated with the proportion of differentially expressed metabolic genes (Pearson's r = 0.86 for *iRno* and 0.91 *iHsa*; p-value = 2.2e-16 for *iRno* and *iHsa*) (**Figure 6.2**). As a result of gene expression integration, gene-level observations were transformed into reaction-level predictions that were directly comparable between *iRno* and *iHsa*, unlike gene expression changes which are difficult to analyze across species. Additionally, predicted reaction-level changes between treatment and control conditions could be interpreted analogously to upregulated and downregulated expression changes.

Analyzing condition-specific subnetworks generated by MADE

Metabolic changes in response to each compound were simulated at the reaction-level using flux variability analysis (FVA). Binary gene-states determined by MADE were applied to each condition-specific model of *iRno* and *iHsa* using the *sybil* package[175] in the R/Bioconductor programming environment[172,173]. Using FVA, minimum and maximum feasible flux values were calculated for each reaction in each condition-specific model while maintaining a growth rate consistent with one doubling per week under physiological conditions. For each reaction in each condition-specific model, one of three reaction-states was assigned: disabled, enabled, or required. Reactions with a minimum flux greater than zero or a maximum flux less than zero were classified as required. Enabled reactions were capable of carrying non-zero fluxes in one or both directions but were not essential for growth. Disabled reaction-states were assigned to reactions incapable of carrying non-zero flux. Flux values were rounded to 6 decimal places prior to classifying reaction-states to minimize rounding errors produced during FVA.

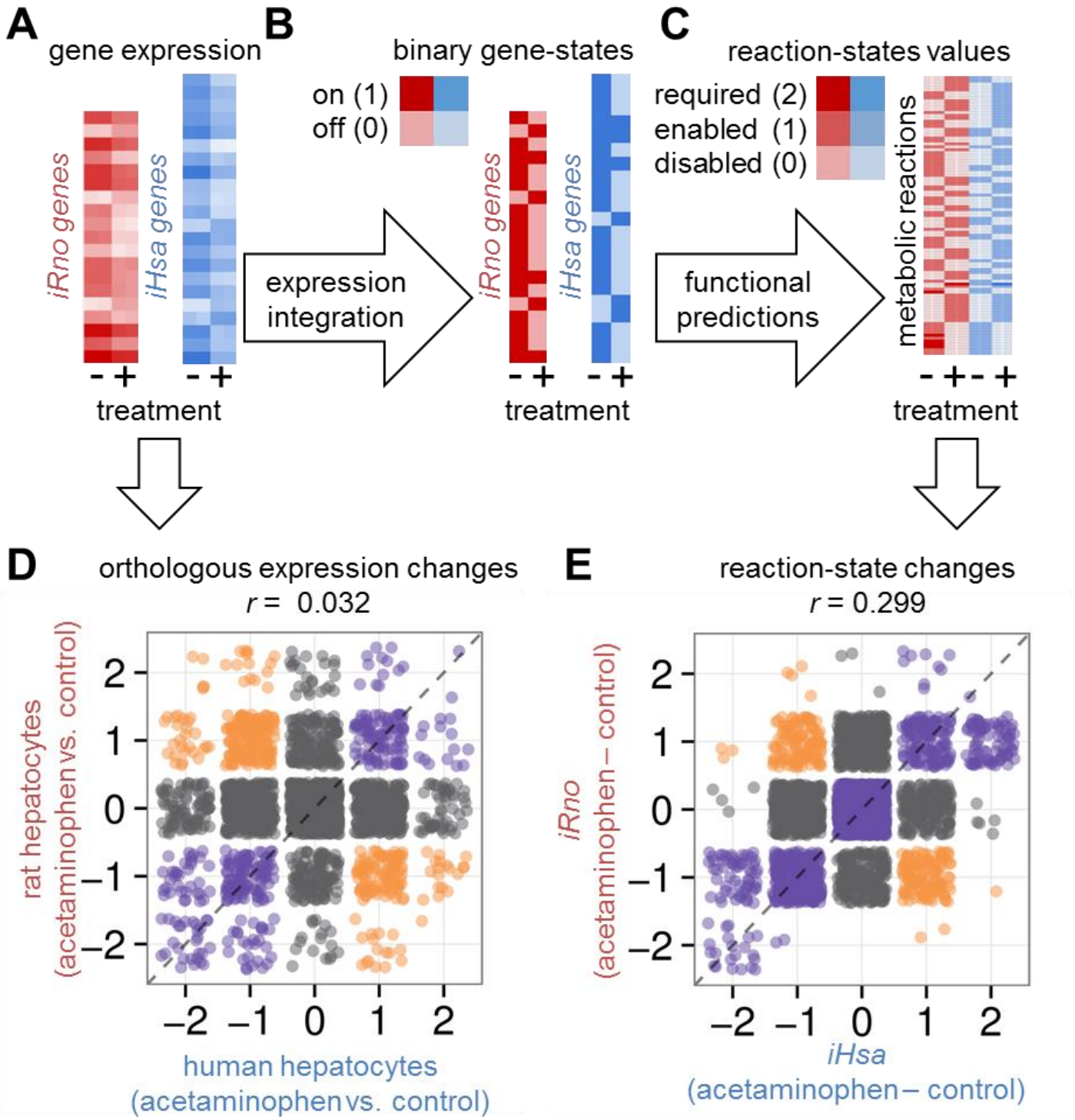

A
gene expression
iRno genes
iHsa genes
- + - +
treatment
B
binary gene-states
on (1)
off (0)
expression integration
iRno genes
iHsa genes
- + - +
treatment
C
reaction-states values
required (2)
enabled (1)
disabled (0)
functional predictions
metabolic reactions
- + - +
treatment
D
orthologous expression changes
r = 0.032
rat hepatocytes (acetaminophen vs. control)
human hepatocytes (acetaminophen vs. control)
E
reaction-state changes
r = 0.299
iRno (acetaminophen – control)
iHsa (acetaminophen – control)

Figure 6.1 - Reaction-state values were calculated across all reactions in *iRno* and *iHsa* based on gene expression data from the Japanese Toxicogenomics Database's Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (Open TG-GATEs). (**A**) Gene expression changes in response to 70 compounds were evaluated in hepatocytes across 1945 rat and 2234 human metabolic genes. Heatmap intensities represent rat and human expression values for a subset of rat and human genes that were significantly differentially expressed between acetaminophen-treated and control-treated hepatocyte samples. (**B**) Gene expression $\log_2$ fold changes and FDR-adjusted q-values for 70 compounds were integrated into *iRno* and *iHsa* using the MADE algorithm. For each set of differentially expressed genes, treatment-specific and control-specific subnetworks were generated for *iRno* or *iHsa* that captured significant changes in expression by deleting genes in models with decreased expression relative to the other condition. It is important to note that not all differentially expressed genes were translated into differences in binary gene-states between treatment and control models because each subnetwork generated by MADE must also be capable of producing hepatocyte growth. (**C**) Flux variability analysis (FVA) was used to assess to the functional impact of gene expression integration for all 280 models generated in **B**. Reaction-states were calculated and classified as required (non-zero flux necessary for hepatocyte growth under physiological constraints), enabled (can carry non-zero or zero flux), or disabled (cannot carry non-zero flux) for 2802 biochemical reactions across all conditions. For each compound in each organism, reaction-states were transformed into reaction-state changes by subtracting the reaction-state value of the control model from the reaction-state value of treatment model. (**D**) Gene expression changes in response to acetaminophen, an over-the-counter drug with well-documented hepatotoxicity in humans and rats, were not correlated when comparing orthologous pairs of rat and human genes. Discretized values representing the direction (-1, 0, 1) of the $\log_2$ fold change between acetaminophen- and control-treated hepatocytes was multiplied by 0 for insignificant changes (FDR > 0.1) and multiplied by 2 for genes with absolute $\log_2$ fold change greater than 1. Orthologous gene expression changes were discretized to maximize comparability with reaction-state changes where values range from -2 to 2. (**E**) Reaction-state changes predicted in response to acetaminophen exposure were correlated across shared biochemical reactions between *iRno* and *iHsa*.

The functional responses of rat and human hepatocytes to treatments were evaluated *in silico* by comparing predicted reaction-states between treatment and control models. Each condition-specific reaction-state was first assigned an integer value representative of its functional importance: disabled (0), enabled (1), or required (2) (**Figure 6.1C**). Reaction-state changes were calculated by subtracting the value assigned to the reaction-state (0 = disabled, 1 = enabled, 2 = required) of the control model from the value of the treatment model, providing a quantitative estimate of relative importance between -2 (decreased importance after treatment) and 2 (increased importance after treatment). With this method, positive and negative reaction-state changes could be interpreted similarly as upregulation and downregulation events at the gene-level (**Figure 6.1D-E**). As a result, reaction-

state changes were observed across a 2802 distinct biochemical reactions at least once. These reaction-state changes were comparable across species with the exception of 167 and 105 biochemical reactions exclusively altered in *iRno* or *iHsa*, respectively.

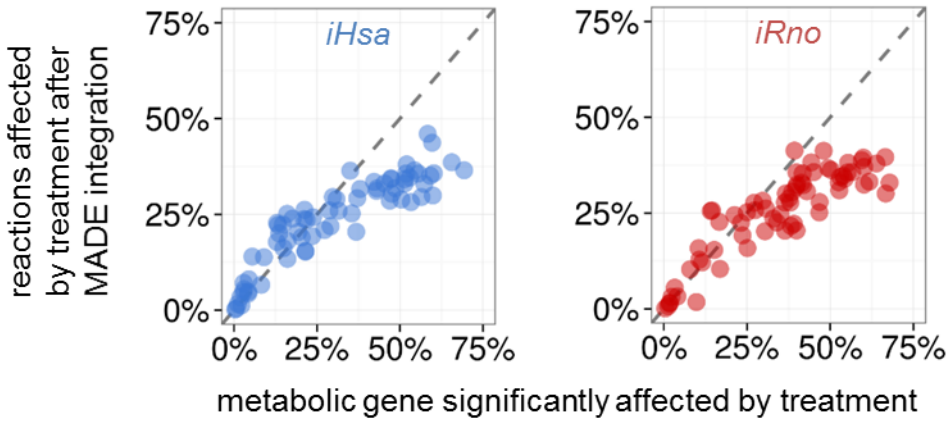


Figure 6.2 – The relationship between gene expression changes and reaction-state changes across 70 compounds. The numbers of significantly differentially expressed genes integrated into *iHsa* (blue, left) and *iRno* (red, right) were correlated with the numbers of reaction-state changes between treatment and control models in response to each compound. As more genes are considered differentially expression, MADE cannot further constrain the metabolic networks because normal physiological functions must also remain active.

Comparing predicted reaction-state changes across species

To determine whether rat and human hepatocytes were more or less similar in their metabolic response to individual compounds, reaction-state changes were compared across all reactions shared between *iRno* and *iHsa*. To simplify reaction-level comparisons across species, reaction-state changes were interpreted as increasing (1), unchanged (0), or decreasing (-1). Reaction-level similarity was assessed by calculating the correlation coefficient between directional changes predicted in rat versus the directional change predicted in human across all compounds. Similarly, compound-level consistencies were determined by calculating the correlation coefficient between rat and human directional changes across all reactions. A similar cross-species comparison was performed for each compound at the pathway-level for directional differences using KEGG pathway annotations.

Rat and human predictions were generally correlated at the reaction-level

Reaction-state changes in *iRno* and *iHsa* that co-occurred in response to the same toxicant were more often altered in the same direction (19591) than in opposite directions (12140). Because predictions made by *iRno* and *iHsa* were independent and could change in either direction, consistent and inconsistent reaction-state

changes were each expected to have a 50% chance of co-occurring. This result demonstrated the capability of *iRno* and *iHsa* to independently generate comparable and generally consistent predictions between rats and humans.

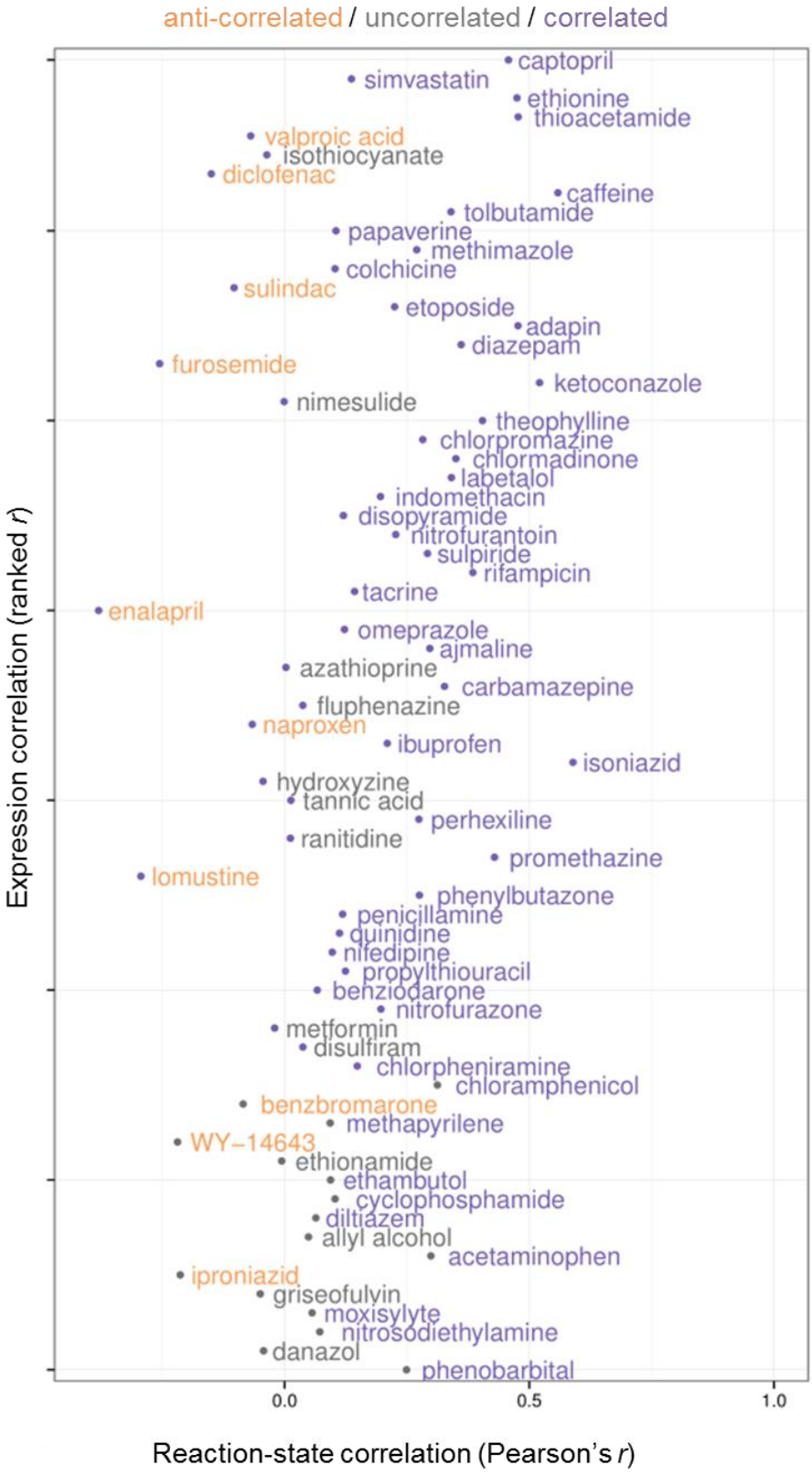


Figure 6.3 – Comparative analysis of reaction-state changes between *iRno* and *iHsa* revealed distinct correlations between species not seen at the level of gene expression. Individual toxicants were classified as correlated (purple, FDR < 0.1), uncorrelated (gray), or anti-correlated (orange) based on the Pearson's correlation between rat and human orthologous gene expression changes (dot color). Reaction-state changes were classified similarly in response individual toxicants (text color). Compounds were ordered by relative rank (y-axis) and the value (x-axis) of the Pearson's correlation coefficient for expression-level changes and reaction-state changes, respectively. These correlation analyses demonstrated that metabolic gene expression changes integrated into *iHsa* and *iRno* can produce distinct results compared to gene expression changes alone. Acetaminophen is used as an example in **Figure 6.1**. Expression correlation coefficients displayed were positioned by rank to avoid overlapping compound identifiers.

At the global-level rat and human hepatocytes responded similarly across a wide range of perturbations, but consistency between *iRno* and *iHsa* predictions varied between individual compounds. Pearson's correlations were determined for each compound by comparing reaction-state changes between *iRno* and *iHsa* across 2802 shared biochemical reactions (**Figure 6.1**). In general, positive correlations were associated with higher numbers of consistent co-occurring reaction-state changes and consistently unchanging reaction-state changes. Co-occurring reaction-state changes in opposite changes contributed to negative correlations, and non-overlapping reaction-state changes weakened correlations. Of the 70 compounds evaluated, 47 were significantly correlated (FDR < 0.1) compared to 13 and 10 that were uncorrelated or anti-correlated, respectively. Correlation coefficients between reaction-state changes in *iRno* and *iHsa* varied between -0.37 and 0.58 for the metabolic response to each compound. These predicted changes in response to individual compounds were distinguishable from orthologous gene expression changes alone (**Figure 6.3**). Two prominent examples uncorrelated at the expression-level (**Figure 6.1D**) but positively correlated at the reaction-level (**Figure 6.1E**) were acetaminophen and phenobarbital (**Figure 6.3**). Both compounds have been studied extensively in rats and humans to understand drug-induced liver injury [176]. These results demonstrated how mapping gene expression changes onto functional models can provide a framework for analyzing toxicogenomics data.

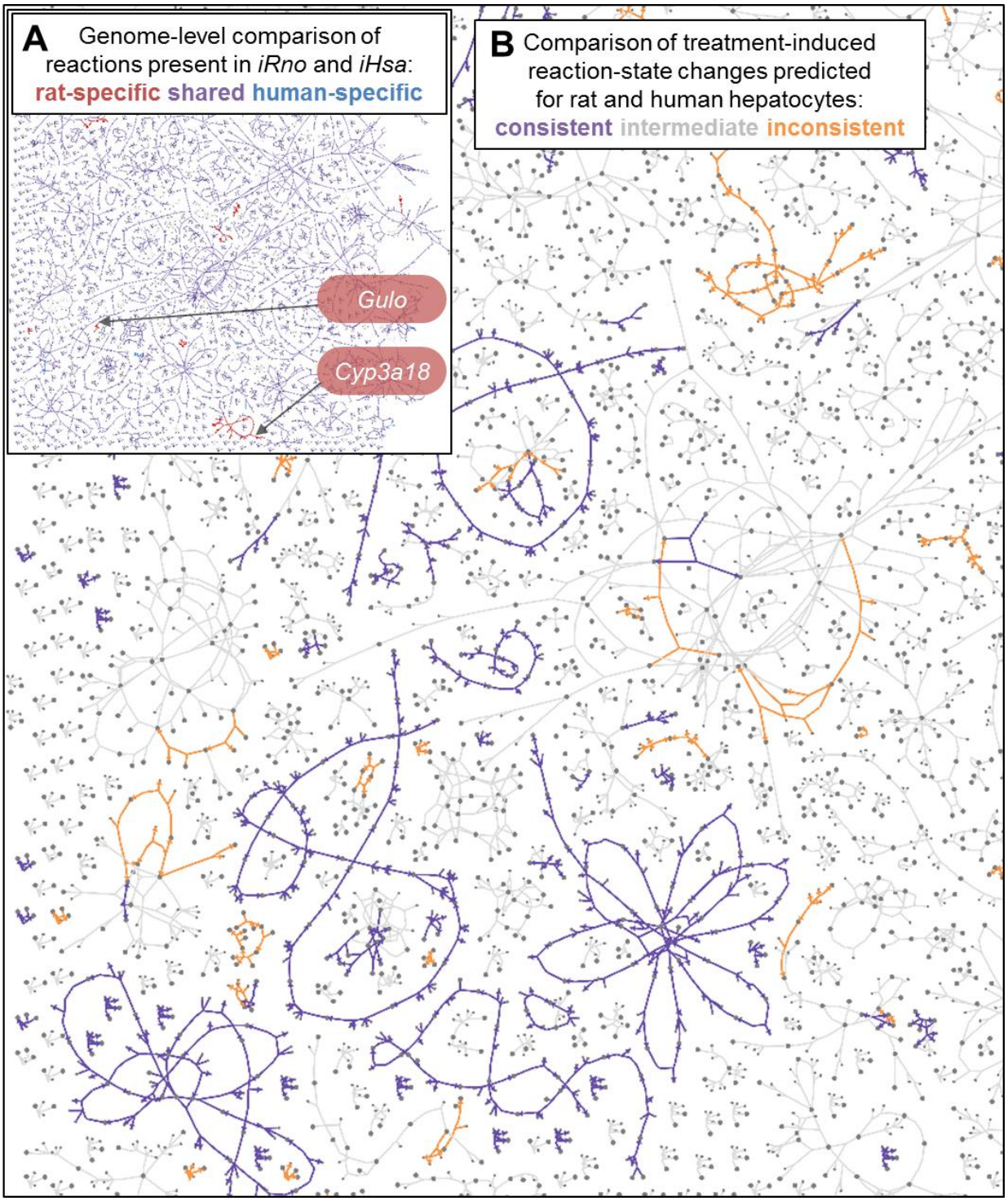


Figure 6.4 – Network visualization of cytosolic reactions from *iRno* and *iHsa* at the genome- and expression-levels. (**A**) Relatively few reactions were unique to *iRno* and *iHsa* outside of bile acid metabolism and vitamin C biosynthesis. Arrows point to rat-specific reactions catalyzed by *Gulo* and *Cyp3a18*. (**B**) High-throughput gene expression changes were integrated into *iRno* and *iHsa* to analyze the reaction-level responses of rat and human hepatocytes to a panel of 70 toxicants. Reaction-state changes were evaluated for 70 pairs of treatment and control models across all reactions in each organism. For each reaction, the Pearson's correlation coefficient between reaction-state changes of *iRno* and *iHsa* was evaluated across all 70 toxicants. Reaction-level correlation coefficients were z-score normalized and classified as either inconsistent (z-score < -1), intermediate (-1 < z-score < 1), or consistent (z-score > 1).

Comparative toxicogenomics analyses of species-specific predictions

Pearson's correlations were determined for individual reactions by comparing reaction-state changes between *iRno* and *iHsa* across all 70 compounds. Relative correlations for reactions were categorized as consistent, intermediate, or inconsistent and visualized at the network-level (**Figure 6.4**). A metabolic network map of reactions and metabolites from *iRno* and *iHsa* generated using MetDraw (www.metdraw.com) [177] in the Python programming environment (http://www.python.org). An SBML file containing the superset of reactions capable of carrying flux in either *iRno* or *iHsa* under physiological constraints was input into MetDraw for visualization. Using the MetColor feature available within MetDraw, reactions that were capable of carrying flux in only one organism were highlighted as species-specific reactions. Although species-specific reactions at the genome-scale were relatively sparse across cytosolic reactions in *iRno* and *iHsa* (**Figure 6.4A**), toxicogenomics predictions revealed interconnected regions of consistent and inconsistent treatment-induced changes between *iRno* and *iHsa* (**Figure 6.4B**).

Differences in treatment-induced metabolic changes between *iRno* and *iHsa* were prominent within 49 individual KEGG pathways (**Figure 6.5**). Pearson's correlation coefficients were determined for reaction-state changes within each KEGG pathway in response to 70 individual toxicants. Positive and negative correlations were significant (FDR < 0.1) for 400 and 147 pathway-toxicant pairs, respectively. Changes in central carbon and amino acid metabolic pathways were more frequently correlated than anti-correlated, suggesting that rat hepatocytes are better suited for studying these pathways within the context of human toxicity (**Figure 6.5**). In contrast, toxicant-induced metabolic changes were often anti-correlated for reaction-state changes involved in bile acid metabolism and glutathione metabolism (**Figure 6.5**). Interestingly, pyrimidine metabolism was highly correlated while purine metabolism was less correlated, consistent with known differences in purine degradation that are described in **Chapter 3**. Overall, these results highlight pathways in which rat could serve as a relatively good or poor model of human metabolism.

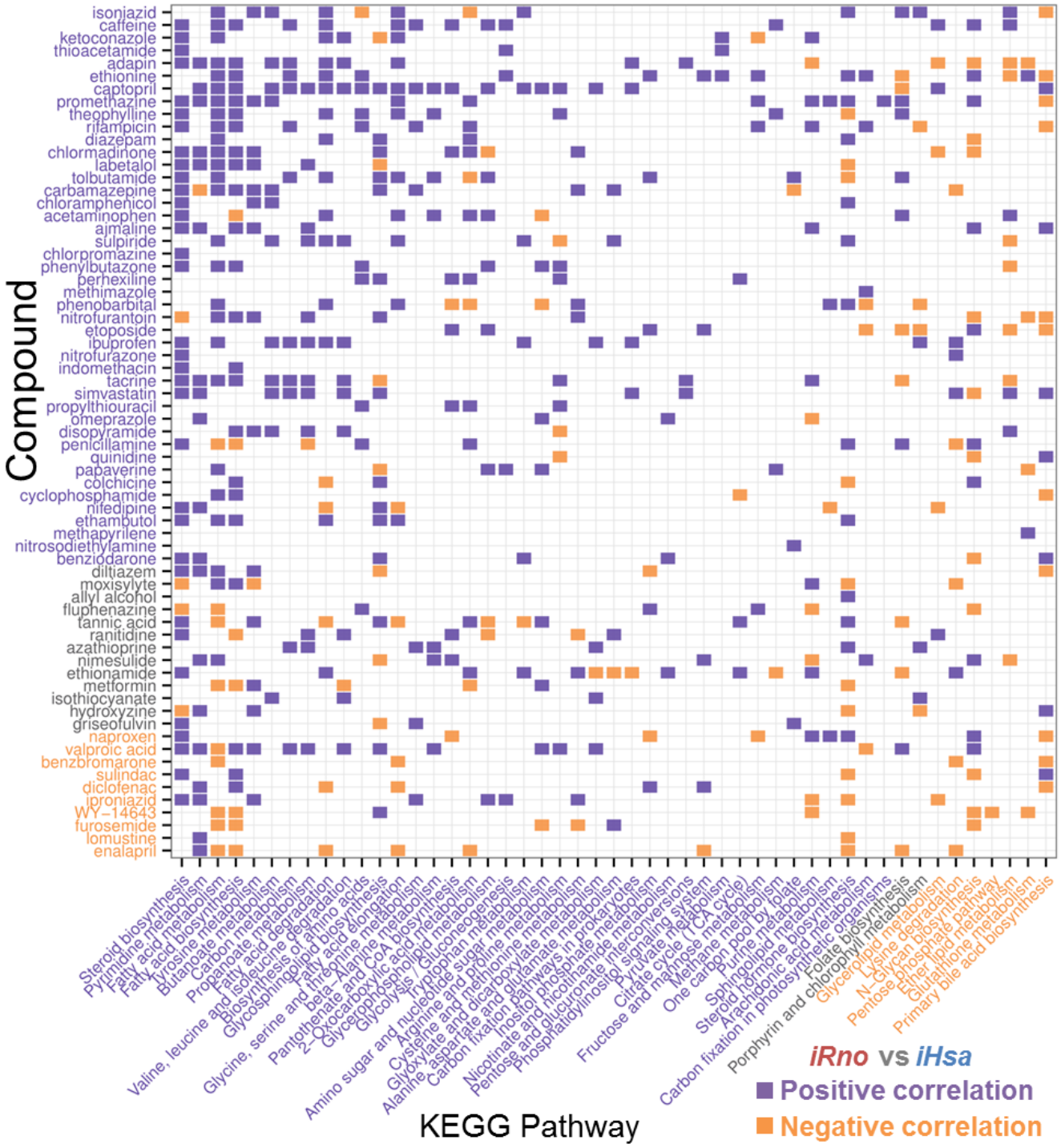


Figure 6.5 – Comparative pathway-level analysis of metabolic changes predicted by *iRno* and *iHsa* in response to individual toxicants. Pearson's correlation coefficients are displayed as individual tiles for each KEGG pathway in response to each toxic compound determined by comparing reaction-state changes induced by *iRno* and *iHsa* for all reactions within that pathway (number of model reactions shown in parentheses). Compounds were ordered by the correlation coefficient determined across 2802 metabolic reaction-state changes. Pathways were ordered by the number of significant positive correlations (FDR < 0.1) minus the number of negative correlations observed across the 70 compounds.

Pathway-level correlations of rat and human predictions for individual compounds

Our analyses revealed that valproic acid and diclofenac, drugs whose gene expression changes were highly correlated between rat and human orthologs, induced reaction-state changes that were significantly anti-correlated (**Figure 6.3**). For valproic acid, we predicted that reactions within fatty acid metabolic pathway were negatively correlated which was surprising given the known role of valproic acid as a substrate for fatty acid metabolism[178] (**Figure 6.5**). We explored known mechanisms of valproic acid-induced hepatotoxicity and found that interference with fatty acid beta oxidation can cause mitochondrial stress that is directly linked to the mechanism of valproic acid-mediated toxicity[178]. For diclofenac, our predictions suggested that reaction-state changes related to fatty acid degradation were anti-correlated, and again we discovered that the molecular mechanism of diclofenac is directly associated with lipid peroxidation[179]. These findings provides examples of how model predictions can serve as a guide for investigating potential pathways that contribute to unexpected toxicities in humans. Furthermore, the potential for unexpected toxicities from these compounds can easily be overlooked in the context of orthologous gene expression changes alone. Another compound that was classified as having negatively correlated reaction-state changes within the fatty acid metabolic pathway as well as a known mechanism of hepatotoxicity related to mitochondrial stress was benzbromarone[180].

Outlook

The analyses described in this chapter represent the first large-scale integration of gene expression changes metabolic networks for two species. The ability to interrogate genome-scale metabolic networks has been largely limited to the availability of reconciled models. Prior to *iRno* and *iHsa*, the only two pairs of bacterial species of metabolic networks have been constructed. Additionally, the ability to integrate differential expression lags behind methods that integrate absolute expression, which is surprising considering that absolute expression does not necessarily correlate with enzyme activity due to differences in kinetics, localization, regulatory motifs, and numerous other factors. Inferring metabolic changes based on differential expression makes no assumptions about the basal level of activity of an enzyme, only that changes in homeostasis may be occurring. As we emphasize in the next chapter, improved algorithms are needed to integrate relative gene expression for other metabolic network analyses such as biomarker predictions.

Acknowledgements

This chapter was adapted from a manuscript that was initially reviewed at *Molecular Systems Biology*, modified substantially after removing the analyses described in this chapter, and currently under revision at *Nature Communications* titled “Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions” and written by Edik M. Blais, Kristopher D. Rawls, Zhuo I. Li, Glynis L. Kolling, Ping Ye, Anders Wallqvist, and Jason A. Papin. I thank all co-authors who contributed to work, specifically Kristopher Rawls for providing renderings of metabolic networks with MetDraw. I also thank Issac Li and Michael Carter for improving annotations in the rat and human networks and Kevin D’Auria, Jennifer Bartell, and Arvind Chavali for providing feedback on the version of the manuscript submitted to *Molecular Systems Biology*. I also thank the two anonymous peer-reviewers who supplied many useful suggestions that were taken into consideration as we improved the manuscript and developed the methods described in the next chapter.

Chapter 7: Comparative biomarker predictions

In previous chapters, we described the creation, curation, and reconciliation of genome-scale network reconstructions (GENREs) of rat (*iRno*) and human (*iHsa*) metabolism. To demonstrate the use of these metabolic networks in systems toxicology, we integrated high-throughput toxicogenomics data to predict biomarker changes in response to 76 environmental and pharmaceutical compounds for rat and human hepatocytes. Using literature-based evidence, we validated biomarker predictions related to known mechanisms of action for antipyretic and anti-gout medications. Comparisons of rat and human biomarker predictions revealed mechanistic insights into a human-specific side effect caused by theophylline distinct from that of the structurally-related compound, caffeine. Overall, the comparative network analyses between rat and human metabolism presented here provide a novel framework for improving the translation of future preclinical studies in rats to humans.

Benchmarking biomarkers predictions for inborn errors of metabolism

Before performing comparative toxicogenomics analyses, we benchmarked the ability of *iHsa* to generate consistent biomarker predictions validated in previous human GENREs. Metabolic biomarkers are routinely screened to pinpoint genetic deficiencies in metabolic enzymes and to diagnose IEMs[14]. Known associations between genes and metabolites for various IEMs were evaluated as previously described in the validation of *Homo sapiens* Recon 2[14,17]. Biomarkers changes for each IEM were estimated by comparing feasible flux ranges via flux variability analysis (FVA) for metabolite exchange reactions between healthy and unhealthy conditions. Unhealthy and healthy conditions were simulated by disabling and forcing flux through reactions associated with an IEM, respectively. For the healthy condition, individual reactions associated with IEM genes were constrained to 90% of the maximum possible flux value determined by FBA under open exchange conditions as described previously[17]. Open exchange conditions were formulated to allow unconstrained uptake (-1000 units) of 12 inorganic ions and limited uptake (-1 units) of exchangeable metabolites. Biomarker prediction performance was measured by the sensitivity to detect known biomarkers of IEMs.

Table 7.1 – Sensitivity of *iHsa* in predicting known biomarkers of IEMs compared to previous human reconstructions.

IEM Biomarker	Count	*iHsa*	HMR 2.0	Recon 2.04	Recon 2.00
Elevated	83	80	78	77	66
Reduced	16	3	2	4	10
Total	99	83.8%	80.8%	82.8%	76.8%

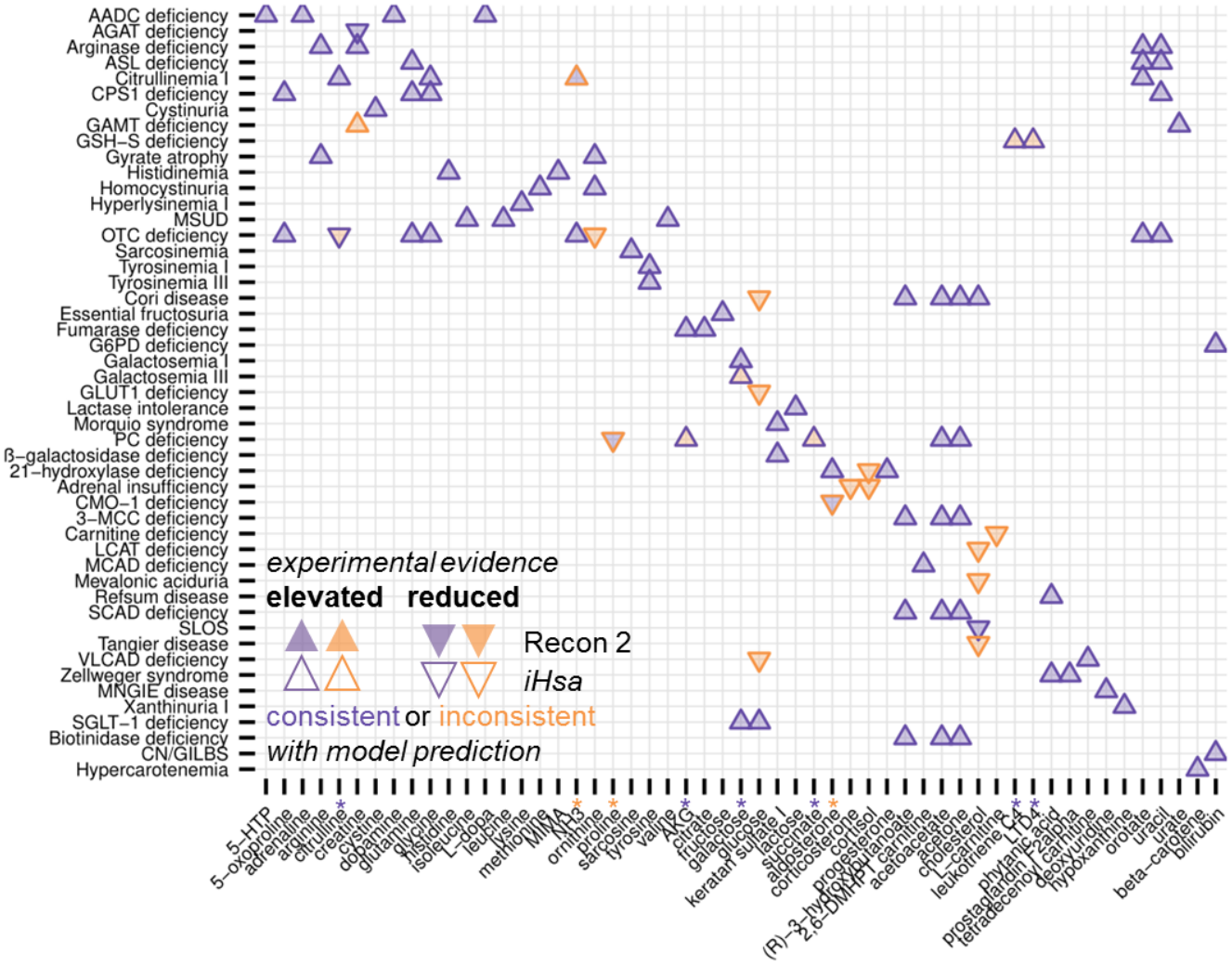


Figure 7.1 – Biomarker predictions for inborn errors of metabolism (IEMs). Comparisons between biomarker predictions generated by *iHsa* and *Homo sapiens* Recon 2[17] against known metabolite biomarkers for IEMs[14,17]. Triangles pointing up and down represent biomarkers known to be elevated or reduced in patients with IEMs. Purple and orange colors represent predictions that were either consistent or inconsistent, respectively, with *iHsa* (triangle outline) and/or Recon 2.04 (triangle filling). Biomarker predictions with increased (purple asterisks) or decreased (orange asterisks) performance in *iHsa* compared to Recon 2 (version 2.04) are highlighted for individual metabolites. Metabolites abbreviations: methyl-imidazole acetic acid (MIMA); 5−hydroxy−L−tryptophan (5-HTP); 2,6−dimethylheptanoyl−carnitine (2,6-DMHPT-crn).

We evaluated the ability of *iHsa* to predict known metabolic biomarkers for 49 IEMs (**Figure 7.1**). The performance of *iHsa* to predict 99 biomarker/IEM pairs was compared to Recon 2[17] (version 2.04) and HMR2[19]. Metabolites were predicted as elevated, reduced, or unchanged using *iHsa*, HMR2, and Recon 2 as previously described in the validation of Recon 2[17]. *iHsa* correctly predicted 83% of 99 IEM biomarkers compared to 81%

for HMR2 and 82% for the most recent iteration of Recon 2 (**Table 7.1**). For IEM predictions, we applied open constraints to exchange reactions that were more consistent across *iHsa*, HMR2, and Recon 2 than default constraints (see below). Compared to predictions described in the original Recon 2 publication[17] (**Table 7.1**), predictions for all three human GENREs were slightly more sensitive for elevated biomarkers but less sensitivity for reduced biomarkers. We also explored possible species-specific differences and found that most GPR rules associated with IEM mutations were closely mirrored by equivalent rat GPR rules.

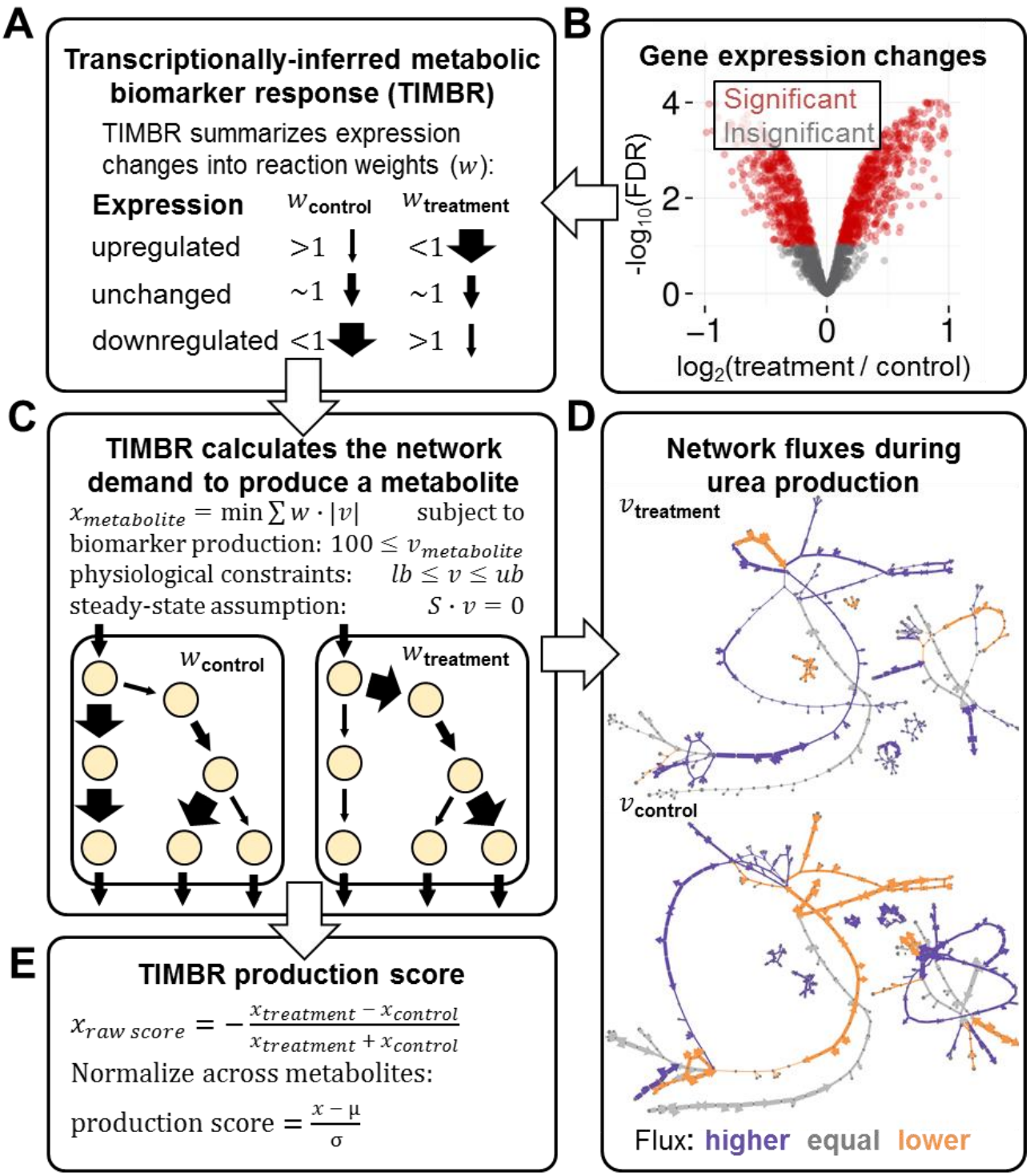


Figure 7.2 – TIMBR is a novel method for predicting treatment-induced biomarkers by integrating gene expression changes into metabolic networks. (**A**) TIMBR calculates reaction weights using $\log_2$ fold changes of significantly (FDR < 0.1) differentially expressed genes. For each reaction, $\log_2$ fold changes are averaged across isozymes after assigning a value of 0 to any insignificant changes. For reactions associated with protein complexes, the subunit with the largest value after averaging is selected. Summarized values are then transformed into larger (or smaller) reaction weights for representing relative expression between treatment and control conditions. (**B**) Caffeine-induced gene expression changes are displayed as a volcano plot for rat hepatocytes. (**C**) Optimization problem formulated by TIMBR to estimate the global network demand needed to produce a metabolite. The objective function minimizes the sum of all reaction fluxes (v) multiplied by TIMBR reaction weights (w). Treatment and control conditions were simulated separately for each potential biomarker under similar physiological constraints that assumed steady-state reaction fluxes. The minimum required production rate for each metabolite was set to either a rate of 100 fmol $cell^{-1}$ $hour^{-1}$ or 90% of the maximum possible flux value, whichever was smaller. (**D**) Optimal caffeine-weighted ($w_{treatment}$) and control-weighted ($w_{control}$) flux distributions ($v_{treatment}$ and $v_{control}$) for biomarker production of urea determined by integrating gene expression changes from **B** into *iRno*. Non-zero fluxes that were higher (purple), equal (gray), or lower (orange) relative to the other condition were displayed using MetDraw[177]. Arrow thickness represents the inverse reaction weight as described in **A**. In this example, the global network demand (sum of weighted fluxes) was smaller in the treatment condition than the control condition, indicating that caffeine induced expression changes that were more consistent with the production of urea compared to controls. (**E**) Raw production scores in response to individual treatment strategies were calculated for each metabolite separately by comparing global network demands determined in **C** for the treatment and control conditions. TIMBR production scores represent these raw production scores normalized across all relevant metabolites with biomarker predictions.

A novel gene expression integration method for generating biomarker predictions

Rats are often used as a surrogate model for understanding human hepatotoxicity; consequently, it is critically important to understand species-specific responses to experimental compounds in order to efficiently translate preclinical studies. To explore the effects of exposure to pharmaceutical compounds and environmental toxicants on normal metabolic functions, high-throughput gene expression profiles of rat and human hepatocytes were obtained from the Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system[44,168] (Open TG-GATEs) and analyzed within the computational frameworks of *iRno* and *iHsa*. We preprocessed raw microarray data from the Open TG-GATEs independently for 119 individual compounds and calculated gene expression changes between control samples and samples treated with a low, medium, or high dose for eight hours. Of 119 compounds with expression data available, 76 were considered suitable for comparative toxicogenomics

analyses after excluding treatments that did not significantly affect (FDR < 0.1) at least 1% of the 1927 or 2176 metabolic genes that mapped to *iRno* or *iHsa*, respectively.

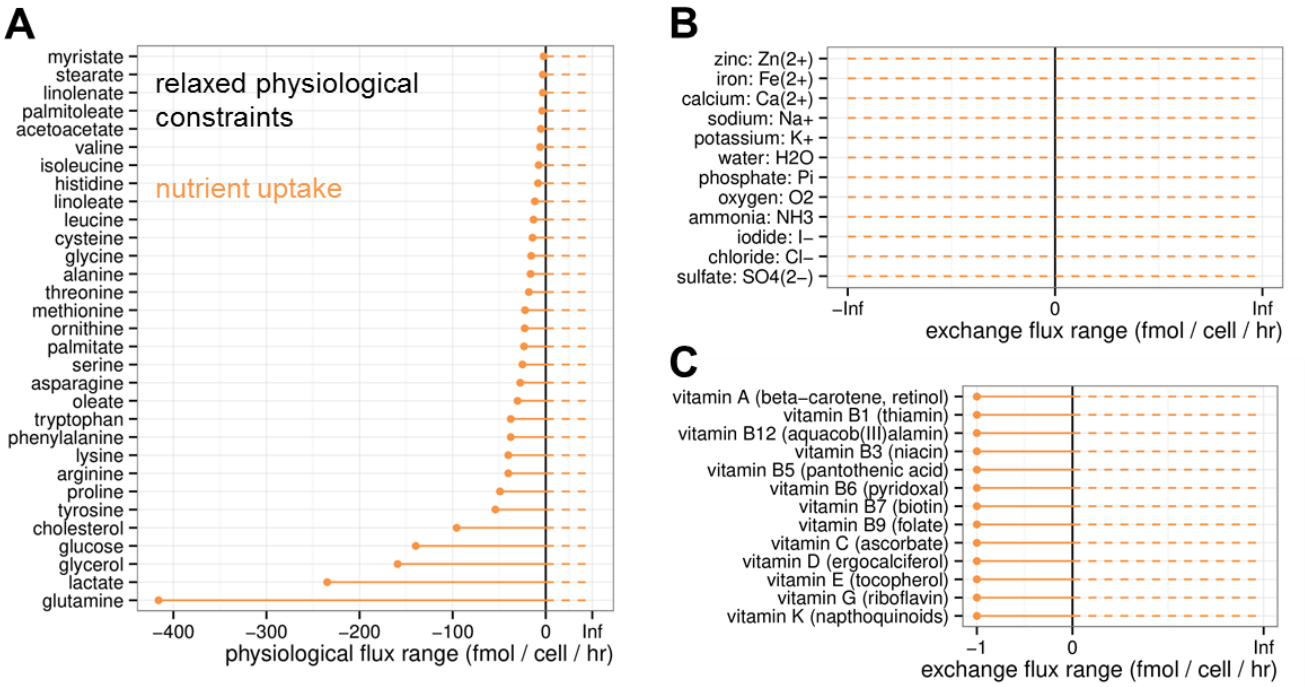


Figure 7.3 – Relaxed physiological constraints applied to *iRno* and *iHsa* for TIMBR predictions. (**A**) Experimentally reported flux measurements from rat hepatocytes were obtained from six separate studies described in **Chapter 4** were applied as relaxed physiological constraints. Reaction lower bounds were set to physiological values, allowing nutrient uptake, and reaction upper bounds with were set to positive infinity (10^6). (**B**) Inorganic ions were allowed unconstrained consumption rates of -10^6 fmol cell^{-1} hour^{-1}. (**C**) Cofactors and vitamins considered essential in humans were set to an uptake value of 1 fmol cell^{-1} hour^{-1} in rat and human networks.

To demonstrate the utility of *iRno* and *iHsa* in biomarker discovery for human toxicology, we generated biomarker predictions for rat and human hepatocytes exposed to these 76 environmental toxicants and pharmaceuticals. Species-specific gene expression changes in response to 76 compounds were integrated into *iRno* and *iHsa* using Transcriptionally-Inferred Metabolic Biomarker Response (TIMBR) (**Figure 7.2**), a new algorithm that estimates the feasibility of producing a metabolite given changes in gene expression. First, TIMBR summarizes gene expression $\log_2$ fold changes into reaction weights that represent the relative cost or demand of carrying flux through each reaction for treatment and control conditions (**Figure 7.2A-B**). Second, TIMBR calculates the global network demand required for biomarker production by minimizing the weighted sum of fluxes across all reactions for each condition (**Figure 7.2C-D**). This general approach, known as parsimonious enzyme usage

flux balance analysis (pFBA)[181], has been adapted for integrating absolute gene expression measurements (present or absent)[49,182] but not for relative gene expression changes (upregulated or downregulated) as done here with TIMBR. By integrating relative changes in gene expression, TIMBR predictions represented the relative propensity to produce metabolites in response to an individual compound. As a result, relative production scores were determined independently for each treatment by normalizing TIMBR predictions across all exchangeable metabolites (**Figure 7.2E**). By applying similar physiological constraints to *iRno* and *iHsa* (**Figure 7.3**) and requiring similar production rates for each metabolite (**Figure 7.2B**), TIMBR provided a novel framework for making biomarker predictions across metabolites, treatments, and organisms. In contrast, a similar approach that integrated absolute expression was described as capable of making comparisons across experimental conditions but not between individual metabolites[182]. Detailed methods and additional considerations of the TIMBR algorithm are available at the end of this chapter.

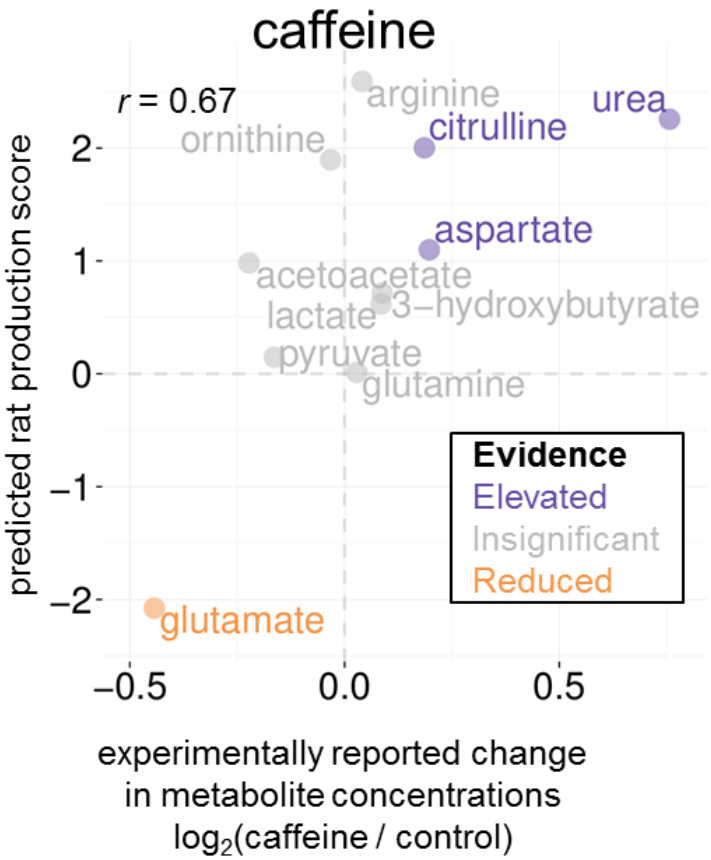


Figure 7.4 – Validation of caffeine-induced biomarker predictions for rat hepatocytes. Comparison of rat production scores calculated by TIMBR in response to caffeine with previously reported changes in metabolite concentrations after caffeine exposure[183]. All metabolites that were experimentally upregulated (3 purple) and downregulated (1 orange) with a reported p-value of less than 0.05 were also predicted in the top or bottom quartiles of production scores.

Validation of caffeine-induced biomarker predictions for rat hepatocytes

We validated TIMBR as a quantitative method for predicting relative metabolic changes in response to caffeine (**Figure 7.4**). Caffeine-induced gene expression changes from rat hepatocytes (**Figure 7.2B**) were integrated into *iRno* using TIMBR to generate production scores (**Figure 7.2E**). An increased production score for a metabolite like urea indicated that genes involved in urea synthesis and secretion were more consistently upregulated than downregulated by caffeine. Reaction weights and fluxes that contributed to urea production in caffeine treatment and control conditions are visualized in **Figure 7.2D**. We compared increased production scores for metabolites like urea (**Figure 7.4**) with previously reported metabolic changes in response to caffeine such as increased urea production in isolated rat hepatocytes or elevated serum levels of urea in rats[183]. We evaluated the ability of TIMBR to quantitatively predict relative caffeine-induced concentration changes by comparing production scores against reported $\log_2$ ratios between treatment and control sample concentrations for urea and ten additional metabolites measured in liver samples[183]. We found that rat production scores based on *in vitro* gene expression data significantly correlated (Pearson's $r = 0.667$; p-value = 0.0249) with caffeine-induced liver metabolic changes reported *in vivo* (**Figure 7.4**). Additionally, all metabolites that were experimentally elevated (urea, citrulline, and aspartate) or reduced (glutamate) by caffeine treatment were consistently predicted in the top or bottom 25% of production scores, respectively. For metabolites that were not significantly affected after *in vivo* treatment with caffeine, most TIMBR predictions were within the middle 50% of production scores, with the exception of ornithine and arginine. Overall, the TIMBR algorithm successfully predicted *in vivo* metabolite concentration changes in response to caffeine with a Matthew's correlation coefficient of 0.69, indicating both high sensitivity (100%) and specificity (71%). Because TIMBR predictions are based on transcriptional changes and do not rely on any knowledge of a compound's mechanism of action, we anticipate this computational approach will be broadly applicable to any compound that induces a detectable physiological response.

Comparative toxicogenomics biomarker predictions

Species-specific differences in the metabolic response to a drug candidate could hamper the successful translation of preclinical biomarkers of efficacy or toxicity from rats to humans. We compared TIMBR predictions generated by integrating gene expression changes into *iRno* and *iHsa* and found a weak but significant positive correlation (Pearson's $r = 0.1958$; p-value $< 10^{-11}$) between rat and human production scores across 286 metabolites and 76 compounds. We analyzed rat and human production scores predicted in response to

individual compounds and categorized 40 as positively correlated, 23 as uncorrelated, and 13 as negatively correlated using an FDR significance threshold of 0.1.

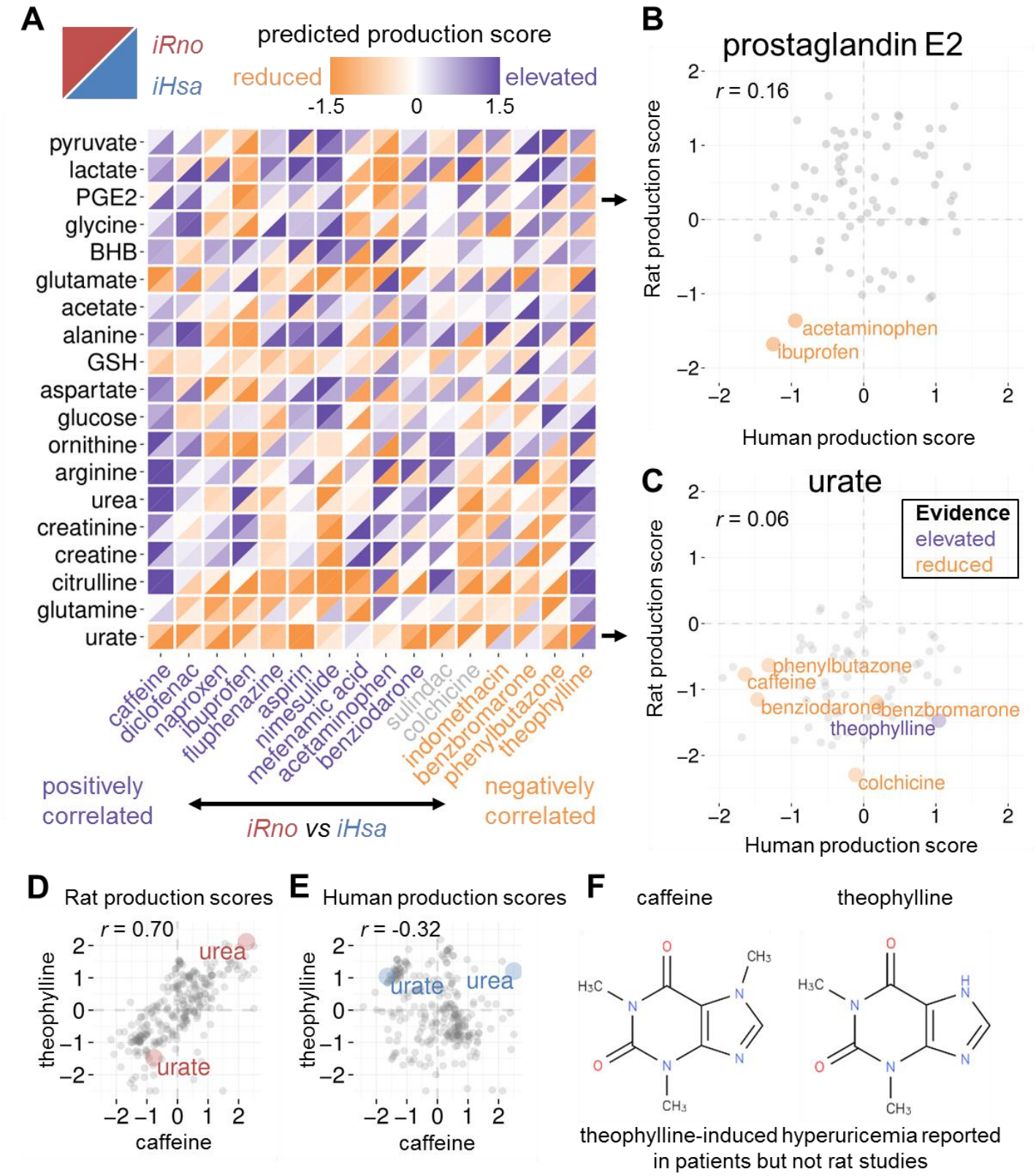


Figure 7.5 – Comparative biomarker predictions in response to anti-gout and antipyretic compounds. (**A**) Heatmap of 16 metabolite biomarkers predicted to increase (purple) or decrease (orange) in response to 16 individual compounds. Metabolite production scores for rat (upper left triangle) and human (upper left triangle) hepatocytes were generated by integrating treatment-induced gene expression changes into *iRno* and *iHsa* using TIMBR. Rat and human production scores across all 286 metabolites were classified as positively correlated (FDR < 0.1), uncorrelated, or negatively correlated (FDR < 0.1) for each individual compound. Compounds were ordered by correlation coefficients and metabolites were ordered by average production scores across all 76 compounds. Abbreviations: β-hydroxybutryate (BHB); prostaglandin E2 (PGE2). (**B**) Scatterplot comparing rat and human production scores for prostaglandin E2 across 76 compounds. Two antipyretic compounds with known cyclooxygenase inhibitor activities, acetaminophen and ibuprofen, were predicted to consistently decrease prostaglandin E2 production in both rat and human hepatocytes. (**C**) Scatterplot comparing rat and human production scores for urate across 76 compounds. Rat production scores for urate were consistently decreased by anti-gout medications that are known to reduce urate accumulation (colchicine, phenylbutazone, benziodarone, benzbromarone). Human production scores were also decreased for anti-gout compounds with the exception of benzbromarone. (**D**) Rat production scores in response to two xanthine derivatives, caffeine and theophylline, were strongly correlated. Urate and urea were consistently predicted to decrease and increase in response to both compounds for rat hepatocytes. (**E**) Human production scores in response to theophylline and caffeine were less correlated than rat production scores. Urate was predicted to increase in response to theophylline and decrease in response to caffeine for human hepatocytes, although human predictions for urea were consistent with increased urea production scores predicted for rat hepatocytes treated with caffeine and theophylline. Theophylline is known to cause increased serum levels of urate in patients but has not been reported in rats. (**F**) Chemical structures for theophylline and caffeine are identical with the exception of a single methyl group.

We validated TIMBR predictions against known metabolic changes related to the therapeutic efficacy for antipyretic and anti-gout medicines. Ibuprofen and acetaminophen are over-the-counter antipyretics that are known to inhibit cyclooxygenase enzymes (COX-1 and COX-2)[184]. We compared rat and human biomarker predictions for individual metabolites across all 76 compounds and found that rat and human production of prostaglandin E2, a metabolite synthesized downstream of COX1/2, was predicted to decrease in response to acetaminophen and ibuprofen (**Figure 7.5B**). For anti-gout compounds, we analyzed the predicted effects of benzbromarone[185], benziodarone[185], colchicine[186], and phenylbutazone[187] on urate production. Despite differences in chemistry, we found that both rat and human production scores for urate were decreased for three out of four anti-gout medications, consistent with their abilities to decrease urate accumulation (**Figure 7.5C**). Furthermore, urate production was predicted to increase in response to several compounds for human

hepatocytes but rarely for rat hepatocytes (**Figure 7.5C**), consistent with known species-specific differences in purine degradation described in **Chapter 3**.

Species-specific predictions for xanthine derivatives

TIMBR biomarker predictions were generally consistent between *iRno* and *iHsa* for caffeine but not for theophylline despite the fact that both compounds are structurally-similar derivatives of xanthine (**Figure 7.5A**). We investigated species-specific differences and found that theophylline-induced urate production was predicted to increase for human hepatocytes and decrease for rat hepatocytes (**Figure 7.5C**). Consistent with human production scores, theophylline has previously been reported to increase serum urate levels in patients (hyperuricemia)[188]. In contrast, caffeine was predicted to decrease urate production in both rat and human hepatocytes (**Figure 7.5C**) despite known differences in the purine degradation pathway. Interestingly, caffeinated beverages have been associated with a decreased incidence of hyperuricemia in patients[189] (**Figure 7.5C**). Using only toxicogenomics data as an input into *iRno* and *iHsa*, TIMBR provided comparative predictions that led to mechanistic insights into how two nearly indistinguishable compounds (**Figure 7.5F**) induced similar responses in rat (**Figure 7.5D**) but drastically different responses in human (**Figure 7.5E**). These results demonstrate the utility of computational tools in making functional predictions that could mitigate serious unexpected toxicities. Additionally, the toxicogenomics analysis pipeline using TIMBR provides a framework for streamlining preclinical drug development by highlighting potentially discrepant biomarkers.

Outlook

Overall, our results emphasize the importance in accounting for differences between rat and human metabolism. Although only a handful of reactions were unique to *iRno* or *iHsa*, we were able to capture functional differences that need to be considered when using rats as a surrogate for human biology. The consensus-based reconstruction and reconciliation approach described in this study can be extended to facilitate comparative analyses across other model organisms. Additionally, we anticipate that TIMBR will improve the ability to generate mechanistic biomarker predictions using gene expression changes from any organism with an available GENRE. As the first pair of reconciled mammalian GENREs, *iRno* and *iHsa* serve not only the comparative toxicogenomics community but also researchers interested in studying inborn errors of metabolism using rats or cancer in humans, among other possible pathologies.

Transcriptionally-inferred metabolic biomarker response (TIMBR) algorithm

Gene expression profiles of rat and human hepatocytes treated with 119 different compounds were obtained from the Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system (Open TG-GATEs) (http://toxico.nibiohn.go.jp)[44,168]. Raw microarray data were downloaded from ArrayExpress (E-MTAB-797 for rat hepatocytes; E-MTAB-798 for human hepatocytes) and pre-processed using the *oligo* package in the R/Bioconductor programming environment. Expression changes after 8 hours of toxicant exposure were independently determined for each compound and organism using the *limma* package. Genes with a false discovery rate (FDR)-corrected q-value less than 0.1 were considered significantly differentially expressed. Of the 119 compounds with data available for both rat and human hepatocytes, 76 were selected for model integration that significantly altered at least 1% of the 1925 rat genes or the 2177 human genes common to both microarrays and models.

Biomarker changes in response to 76 environmental compounds and pharmaceutical toxicants were predicted for rat and human hepatocytes with TIMBR, a novel constraint-based analysis algorithm (**Figure 7.2**). TIMBR integrates gene expression data from toxicant-treated and control samples and calculates a production score for each exchangeable metabolite under relaxed physiological constraints (**Figure 7.3**). TIMBR production scores represent the consistency between the reactions needed to synthesize and secrete a potential biomarker and the relative expression of genes associated with those reactions. For each metabolite, a production cost was calculated by minimizing the total weighted flux across all reactions while maintaining positive flux through its extracellular exchange reaction:

$$minimize\ [w \cdot |v|\ subject\ to\ S \cdot v = 0;\ v_{lb} \leq v \leq v_{ub}] \qquad (7.1)$$

In equation (7.1), v is a vector of reaction fluxes, w is a vector of reaction weights based on gene expression measurements, S is the stoichiometric matrix, v_{lb} and v_{ub} are vectors of lower and upper bound constraints for reaction fluxes. To simulate physiologically-relevant conditions, nutrient uptake was limited to physiological values by setting lower bound constraints (v_{lb}) of metabolite exchange reactions to quantitative values derived from experimental literature (**Figure 7.3A**). To simulate production of a potential biomarker, non-zero positive flux was forced by setting the lower bound (v_{lb}) through a metabolite's exchange reaction to 90% of the maximum possible secretion rate determined by FVA under relaxed physiological constraints or a value of 100 fmol cell^{-1} hour^{-1}, whichever was smaller. In order to solve this optimization problem with a linear programming solver,

metabolic networks were first converted into irreversible metabolic networks where each reversible reaction was represented by separate forward and reverse reactions in the stoichiometric matrix and v_{lb} is non-negative.

Reaction weights (w_i) based on relative gene expression were calculated independently for each treatment in each organism. Given a set of gene expression changes, TIMBR generates two sets of reaction weights (**Figure 7.2A**) for the optimization problem in equation (1) that represent the costs associated with carrying flux through reactions in toxicant-treated and control conditions (**Figure 7.2B**). In contrast to approaches that integrate gene expression measurements as weights for flux minimization[49,182], TIMBR uses differential expression instead of absolute expression values.

To transform relative gene expression changes into reaction weights for treatment and control conditions, TIMBR implements a novel approach for summarizing multiple expression changes through GPR relationships. In general, GPR rules use Boolean operators to describe multiple genes that encode redundant isozymes with an "OR" relationship and genes that encode subunits in an enzyme complex with an "AND" relationship. For a group of isozymes, the $\log_2$ fold change was averaged such that the effect of one upregulated isozyme could either be offset by downregulation in another isozyme or diluted by the presence of multiple unaffected isozymes. For subunits in an enzyme complex, the $\log_2$ fold change with the largest absolute value was used. Expression changes were summarized for $\log_2$ fold changes because the distributions of $\log_2$ fold changes were more evenly distributed (**Figure 7.2B**). Summarized reaction values based on $\log_2$ fold changes were inverse log transformed and multiplied to the default vector of reaction weights to represent the control condition (**Figure 7.2A**). For the treatment condition, default reaction weights were divided by reaction fold change values such that upregulated reactions contributed less to the sum of weighted fluxes than downregulated reactions and vice versa for controls (**Figure 7.2A**). Default weights of 1 for biochemical reactions and 2 for transport reactions were doubled for reactions with no gene associations or expression data available. Treatment and control condition weights were then applied separately to either *iRno* or *iHsa* in order to calculate the global network demand (sum of weighted fluxes) for the production of each potential biomarker (**Figure 7.2C**).

Production scores representing relative biomarker changes were determined by comparing biomarker production costs based on treatment and control reaction weights (**Figure 7.2E**). Raw production scores were calculated based on the relative global network demand defined in equation (1) between treatment and control conditions for each biomarker:

$$raw\ production\ score = \frac{treatment - control}{treatment + control} \quad (7.2)$$

Raw production scores from equation (7.2) across all potential metabolic biomarkers were normalized independently for each compound in each organism using a z-score transformation (**Figure 7.2E**). With this method, positive or negative production scores could be interpreted as the increased or decreased propensity for a metabolite to be synthesized and secreted in response to a treatment relative to other metabolites. To determine whether rat and human hepatocytes were more or less similar in their metabolic response to individual compounds, production scores were analyzed across all 286 potential biomarkers shared between *iRno* and *iHsa*. Biomarker-level similarity was assessed by calculating the correlation coefficient between rat and human production scores across all compounds. Similarly, consistencies were determined for individual compounds by calculating the correlation coefficient between rat and human production scores across all metabolites.

Unlike previous methods that integrate absolute gene expression, the TIMBR method requires a specific GPR rule format in order to summarize gene expression changes to the reaction-level. GPR rules involving redundant subunits in a protein complex were structured according to the following format: (*A1* or *A2*) and (*B1* or *B2*), where redundant enzymes are grouped together for each subunit. As a result, a TIMBR weight represents an average change in gene expression for the subunit that experienced the largest perturbation. Because TIMBR summarizes directional changes instead of absolute values, the following alternative Boolean representation could yield different results: (*A1* and *B1*) or (*A1* and *B2*) or (*A2* and *B1*) or (*A2* and *B2*), where non-redundant subunits are grouped together for each possible protein complex. With this alternative representation, a reaction weight would represent an average of the largest gene expression changes observed for each possible protein complex. Although both approaches are conceptually similar, TIMBR implements the former approach that summarizes gene expression changes independently for each subunit.

Metabolic network maps of reactions and metabolites from *iRno* and *iHsa* were generated using MetDraw[177] (www.metdraw.com) in the Python programming environment (http://www.python.org). An SBML file containing the superset of reactions capable of carrying flux in either *iRno* or *iHsa* under relaxed physiological conditions was input into MetDraw for visualization of global networks. SBML files containing the subset of rat reactions with non-zero flux values from TIMBR simulations of urea production in response to caffeine.

Acknowledgements

This chapter was adapted from a manuscript currently under revision at *Nature Communications* titled "Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions" written by Edik M. Blais, Kristopher D. Rawls, Zhuo I. Li, Glynis L. Kolling, Ping Ye, Anders Wallqvist, and Jason A. Papin. I thank all co-authors who contributed to this work, specifically Kristopher Rawls for providing renderings of metabolic networks using MetDraw. I also thank Michael Carter for network annotations and Kevin D'Auria, Jennifer Bartell, and Arvind Chavali for providing feedback on an earlier version of the manuscript.

Chapter 8: High-throughput metabolomics validation of transcriptionally-inferred biomarker predictions

Synopsis

In **Chapter 7**, we performed comparative toxicogenomics analyses of rat and human metabolic networks to predict potential biomarkers for various pharmaceutical compounds and environmental toxicants. To generate biomarker predictions based on changes in gene expression, we developed a novel algorithm called Transcriptionally-Inferred Metabolic Biomarker Response (TIMBR). In this chapter, we further evaluate the performance of TIMBR by comparing high-throughput metabolomics data with biomarker predictions based on gene expression changes from RNA-seq experiments. As a result, TIMBR production scores and metabolomics changes detected in spent cell culture media were significantly correlated for carbon tetrachloride (CCl_4) and slightly correlated for 2,3,7,8-tetrachlorodibenzodioxin (TCDD). Experimental validations of model predictions discussed in this chapter demonstrate the utility of the TIMBR algorithm and supports the use of metabolic network models in improving preclinical biomarker discovery in the context of toxicology.

Preclinical biomarker discovery of drug-induced hepatotoxicity

Pharmaceutical drug candidates often fail in preclinical or clinical trials due to unexpected drug-induced liver injury (DILI)[1]. The liver is a vital organ serving a variety of important functions in the body including carbohydrate, protein and fat metabolism, as well as diverse metabolic, vascular, immunological, secretory and excretory functions. Several liver metabolic tasks are known to be disrupted in response to DILI, resulting in abnormal clinical chemistry values: gluconeogenesis (hypoglycemia), albumin synthesis (hypoalbuminemia), detoxification of ammonia (hyperammonemia), and bilirubin efflux (jaundice)[1-4]. In this proposed research, we aim to apply computational methods to investigate toxicant-induced changes within rat and human metabolism.

Novel biomarkers of DILI are needed to facilitate the translation of preclinical studies in rats to clinical trials in humans. Candidate drugs are commonly screened in rats to identify potential biomarkers of therapeutic efficacy and toxicity. Traditional biomarkers used in preclinical drug development include proteins such as albumin, aspartate aminotransferase, and alkaline phosphatase and metabolites such as bilirubin, creatinine, and triglycerides[5]. Although these biomarkers can effectively diagnose DILI, traditional biomarkers are limited in scope to accurately predict future onset of DILI[6]. High-throughput metabolomics methods have the potential to significantly improve both diagnostic and predictive biomarkers of DILI by expanding the potential pool of

biomarkers from a few dozen to thousands of metabolites[7-9]. In contrast to biomarkers like aspartate aminotransferase, which are released into the blood after hepatocellular membrane disintegration[6], we anticipate that metabolite biomarkers will require less catastrophic perturbations to be detected and will be more sensitive to the direct mechanisms of hepatocyte injury[6].

Preclinical efforts to identify biomarkers associated with toxicity typically use statistical methods that correlate changes in metabolite concentrations with toxicant exposure[7-13]. A major limitation of this conventional approach is that the metabolic response of rat hepatocytes will not necessarily reflect that of human hepatocytes. We hypothesize that mechanistic modeling of metabolism with GENREs will substantially improve the ability to identify biomarkers that are predictive of toxicity. Human GENREs have provided mechanistic insights into metabolic biomarkers for inborn errors of metabolism and therapeutic strategies based on genomic alterations in cancers[14,15]. Constraint-based modeling algorithms have been successfully applied to human GENREs to study various aspects of human physiology and disease[16-21]. Here, we apply the TIMBR algorithm described in **Chapter 7** to further validate the use of the rat metabolic network, *iRno*, for future use in clinical and basic science applications.

To validate the ability of *iRno* to predict metabolic biomarkers using the TIMBR algorithm, we generated high-throughput metabolomics and transcriptomics profiles of rat hepatocytes treated with 5 chemical compounds (**Figure 8.1**). In the previous chapter, we described the validation of a handful of biomarker predictions based on gene expression microarrays using literature-based evidence. Here, we performed high-throughput validation of biomarker predictions based on RNA-seq data with metabolomics data. Four compounds were investigated experimentally (**Figure 8.1**), although after quality control analyses only TCDD and CCl_4 were suitable for validation of model predictions. An overview of the experimental protocol for generating these data at the 24 hour time point are provided in **Figure 8.2**. A similar approach was used for generating metabolomics and transcriptomics profiles of primary rat hepatocytes after 6 hours of treatment. Detailed descriptions of experimental protocols will be provided upon publication of a manuscript related to this project.

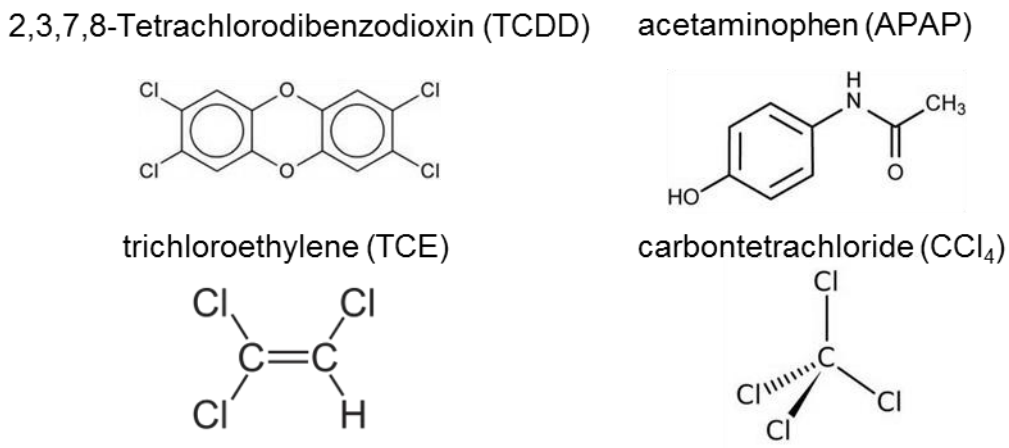


Figure 8.1 – Transcriptomics and metabolomics changes were evaluated in response to four classic hepatotoxicants. After quality control analysis, experimental data for TCDD and CCl_4 were considered suitable for further metabolic network analysis.

Data preprocessing and normalization

Raw RNA-seq data were preprocessed and normalized using kallisto (www.github.com/pachterlab/kallisto) and sleuth (www.github.com/pachterlab/sleuth) software (**Figure 8.3**). These methods quantify transcript abundance much faster than traditional methods enabling processing of treatment and control samples on a single computer rather than a cluster of computers. First, a rat transcriptome was obtained from RefSeq (www.ncbi.nlm.nih.gov/refseq) and filtered to include only known annotations by removing model predicted annotations (prefixes starting with X). After building a rat transcriptome index of sequences to which RNA-seq reads can be aligned, individual FASTQ formatted data containing short oligonucleotide reads were pseudoaligned by kallisto. Aligned transcript counts were imported into the sleuth package in the R/Bioconductor[172] programming environment and filtered to remove low abundance transcripts. After filtering, 10539 unique transcripts mapping to 10121 rat genes were detected in one or more conditions. Differential expression was calculated between treatment and control samples independently for each time point using the Wald test as implemented in sleuth. Resulting beta coefficients (analogous to a natural log fold change) and false-discovery rate-adjusted p-values were then analyzed using the TIMBR algorithm as defined in the previous chapter. TIMBR production scores were generated for each metabolite for which an exchange reaction could carry positive flux. In some cases, intracellular demand reactions were created for metabolites that could not be transported to the extracellular compartment, mostly to enable TIMBR predictions for additional metabolites detected in the metabolomics datasets.

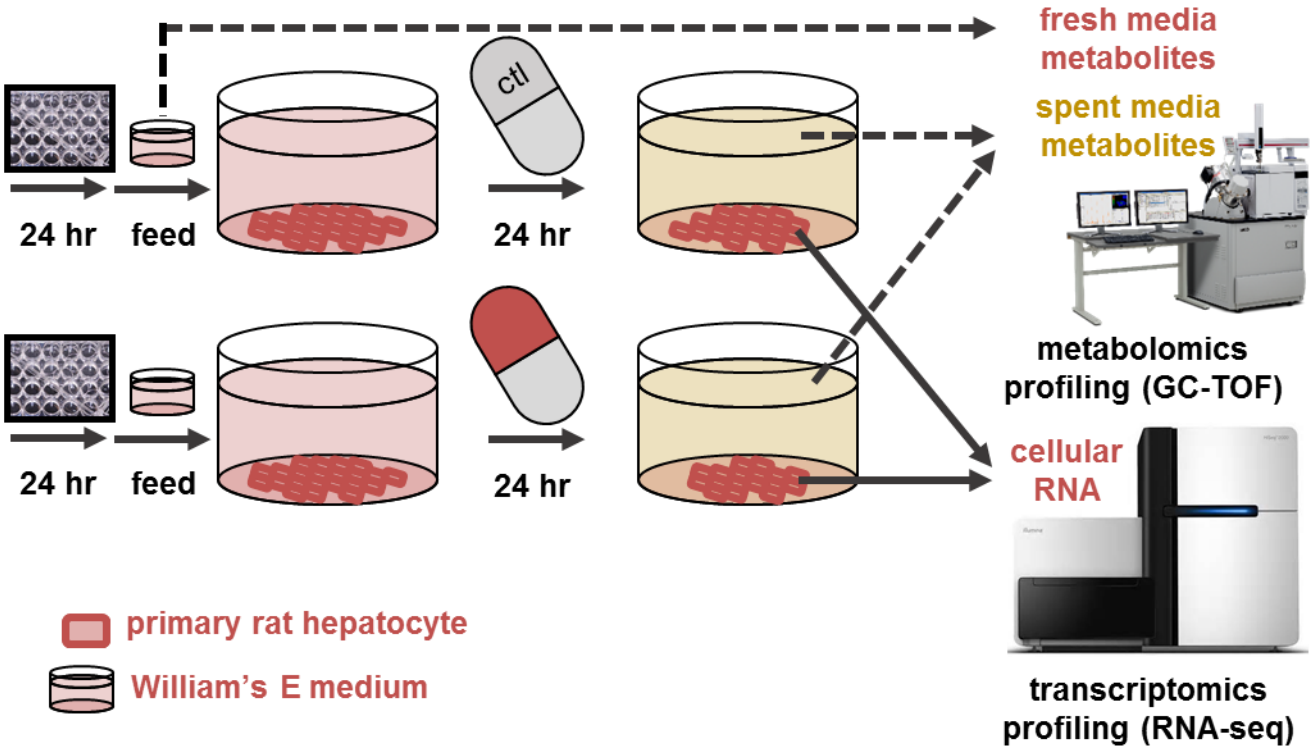


Figure 8.2 – Experimental protocol for generating transcriptomics and metabolomics profiles of primary rat hepatocytes. Spent media samples were obtained from cultured hepatocytes after 24 hours of exposure to hepatotoxicants or controls. Fresh media samples were obtained immediately after treatment began. Cellular RNA samples were obtained from cells cultured after 24 hours of treatment. Cryopreserved cells were plated 24 hours prior to treatment.

Untargeted metabolomics of primary metabolites was performed by West Coast Metabolomics using gas chromatography time of flight (GC-TOF) mass spectrometry. Metabolite abundance of metabolites was estimated as the peak intensity area under the curve as provided by West Coast metabolomics. Metabolite abundance was normalized between samples by dividing the log intensity value by the sum of the log intensity values across all metabolites. Normalized values were then compared between treatment and control conditions using the Student's t-test independently for each time point. Metabolomics data were preprocessed in the R/Bioconductor programming environment to calculate changes in relative metabolite abundances between experimental conditions that could potentially be translated in biomarkers of treatment exposure. Although the possibility that the concentrations chosen for these treatment may not be representative of an actual toxicological response, these data were appropriate for validating TIMBR predictions because metabolomics data were paired with transcriptomics profiles.

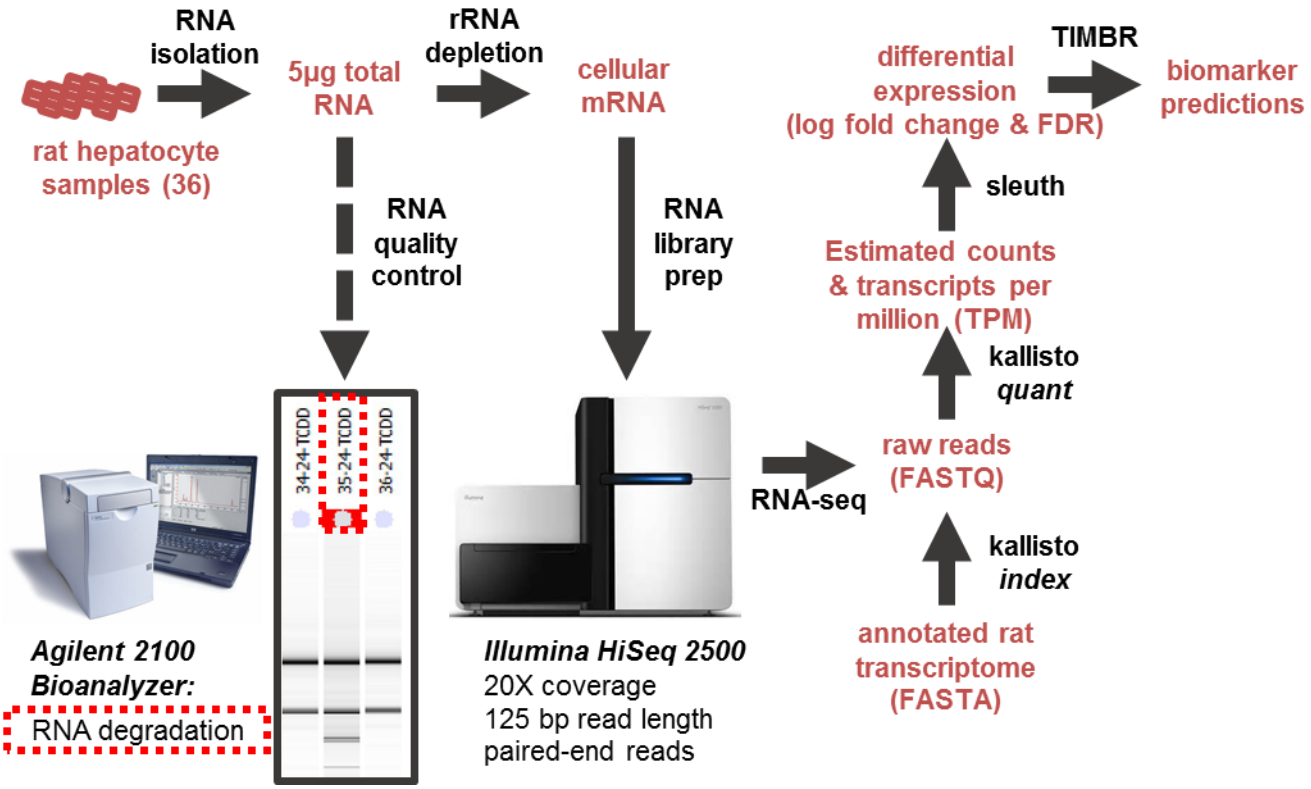


Figure 8.3 – RNA-sequencing pipeline from RNA preparation to biomarker predictions with the TIMBR algorithm. RNA sequencing was performed by the DNA Sciences Core which is supported by the University Of Virginia School Of Medicine.

Correlations between metabolomics changes and biomarker predictions

After quality control analysis of transcriptomics and metabolomics data, we found that only two drugs, CCl_4 and TCDD were suitable for validation of TIMBR predictions. The numbers of differentially expressed metabolic genes in response to each treatment are shown in **Table 8.1**. Acetaminophen-induced changes were indicative of RNAdegradation or poor sample quality rather than global transcriptional changes that could be interpreted as meaningful.

Table 8.1 – Numbers of significantly differentially expressed genes (FDR < 0.1) in response to individual compounds.

treatment	duration	upregulated	downregulated	% significant
APAP	6 hours	519	252	56.9
APAP	24 hours	265	747	71.6
CCL4	6 hours	71	163	17.3
CCL4	24 hours	22	110	9.3
TCDD	6 hours	233	324	41.1
TCDD	24 hours	85	92	12.5
TCE	6 hours	0	0	0.0
TCE	24 hours	89	179	16.6

We validated TIMBR as a quantitative method for predicting relative metabolic changes in response to CCl_4 (**Figure 8.4**). Metabolomics changes in spent media for 45 metabolites were comparable with TIMBR predictions after data preprocessing and quality control. For CCl_4, TIMBR productions scores were highly correlated with metabolomics changes, capturing a potential decrease in glutamine (not statistically significant experimentally) and a significant increase in glycine abundance at 24 hours. Despite a lack of power to highly significant metabolic changes, model predictions for CCl_4 were consistent with experimental observations. Correlation coefficients between TIMBR production scores and log fold changes in metabolite abundances were significant at 6 hours (Pearson's r = 0.375; p-value = 0.011) and 24 hours (Pearson's r = 0.396; p-value = 0.007). It is important to note that data for each time point was preprocessed and analyzed independently, supporting the reproducibility of the TIMBR algorithm in generating predictions that are consistent with biology. Furthermore, we found that model predictions for the 6 hour time point were even more consistent with experimental observations at the 24 hour time point (Pearson's r = 0.475; p-value = 0.00097) likely because metabolic changes in spent media concentrations are not instantaneous. It is important to note that data for each time point was preprocessed and analyzed independently, supporting the reproducibility of the TIMBR algorithm in generating predictions that are consistent with biology.

Biomarker predictions were generally less consistent for TCDD than for CCl_4 (**Figure 8.5**). TIMBR production scores were not significantly correlated with experimental observations; however, the trend was positive at both time points. Correlation coefficients between TIMBR production scores and log fold changes in metabolite abundances were not significant at 6 hours (Pearson's r = 0.144; p-value = 0.346; not shown) and 24 hours (Pearson's r = 0.012; p-value = 0.940; not shown). Biomarker predictions based on transcriptional changes at 6 hours were more predictive of metabolic changes at 24 hours (**Figure 8.5**) (Pearson's r = 0.192; p-value = 0.207) than at 6 hours, consistent with our observations for CCl_4 (**Figure 8.4**). Fortunately, these results are still valuable for iterative process of model prediction, experimental validation, and model refinement that we outlined in **Chapter 5**. We observed from these results that lactate was being secreted by hepatocytes in the control condition (relative to blank media); however, our default physiological constraints defined in **Chapter 4** were setup to allow lactate uptake. By updating exchange reaction constraints to more closely match the absolute changes in metabolite relative to blank media, we anticipate that model predictions can be dramatically improved.

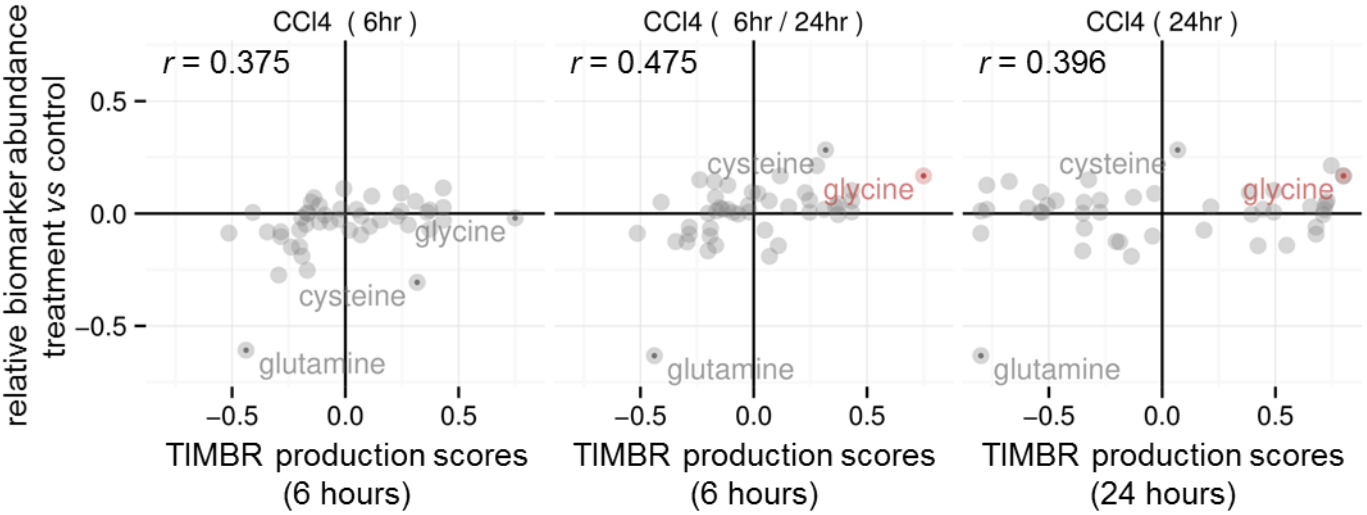


Figure 8.4 – Biomarker comparisons between predicted TIMBR production scores and metabolomics changes in response to CCl_4. Glycine abundance at 24 hours was the only significantly altered metabolite from CCl_4 experiments, although several insignificant changes had moderately large fold changes relative to controls. TIMBR predictions at the 6 hour time point and biomarker changes at the 24 hour time point were compared to highlight how cysteine is moderately increased at 24 hours (insignificant) after having been predicted to be elevated at 6 hours. Although not statistically significant, glutamine was the most downregulated biomarker metabolite. Red and gray circles respectively indicate metabolites with experimentally significant or insignificant changes in abundance between treatment and control conditions. These values are not reflective of whether the metabolite is consumed or produced (spent media versus fresh media).

Mechanistic insights with TIMBR predictions

To demonstrate the use of metabolic network modeling in providing mechanistic insights into biomarker predictions we investigated the upregulation of glycine in response to CCl_4 (**Figure 8.4**). Glycine was chosen because it was the only metabolite that was moderately altered relative to control treatment (FDR < 0.3) at 24 hours (none were significant at the earlier time point with this threshold). Glycine production scores were also the most upregulated metabolite at both time points for CCl_4. Because transcriptional changes do not necessarily result in immediate changes in metabolite concentrations, upregulated glycine TIMBR production scores after 6 hours of treatment were not inconsistent with the lack of change in metabolite levels at 6 hours. Interestingly, this result may actually suggest that TIMBR production scores may be predictive of metabolic changes before they actually occurred at 24 hours.

To provide mechanistic context for why glycine was upregulated in response to CCl_4 (**Figure 8.4**), we compared reaction fluxes between treatment and control TIMBR solutions generated by assigning reaction weights based on relative changes in gene expression. As a result, we found that the same 19 reactions carried equal amounts

of flux in both TIMBR solutions when the minimum glycine production rate was set to 1 fmol cell^{-1} hour^{-1}. Because both simulations utilized identical pathways, we investigated which reactions were assigned different TIMBR weights and found that a single reaction was different between the treatment and control conditions: a glycine-serine antiporter between the cytosol and the extracellular space. As a sanity check, we looked at transcriptional changes for genes within this transport reaction's GPR rule and found that the same gene was consistently upregulated (FDR < 0.01) at both time points (**Table 8.2**). Because this antiporter facilitated the export of glycine and import of serine, we hypothesized that changes in serine would proceed in the opposite direction of glycine. Looking back at experimental results for CCl_4, we confirmed that serine was slightly decreased in the metabolomics data (statistically insignificant log fold change of -0.19), confirming our suspicions that these two metabolites might mechanistically be linked by this specific reaction. In this example, we were able to establish a mechanistic connection between glycine and a specific transporter that implicated serine to better understand how the physiological response of hepatocytes exposed to CCl_4 translate into a potential biomarker change. This type of analysis highlights the specific advantage of applying the TIMBR algorithm to predict metabolic changes within a metabolic network that cannot be achieved based on purely correlative studies.

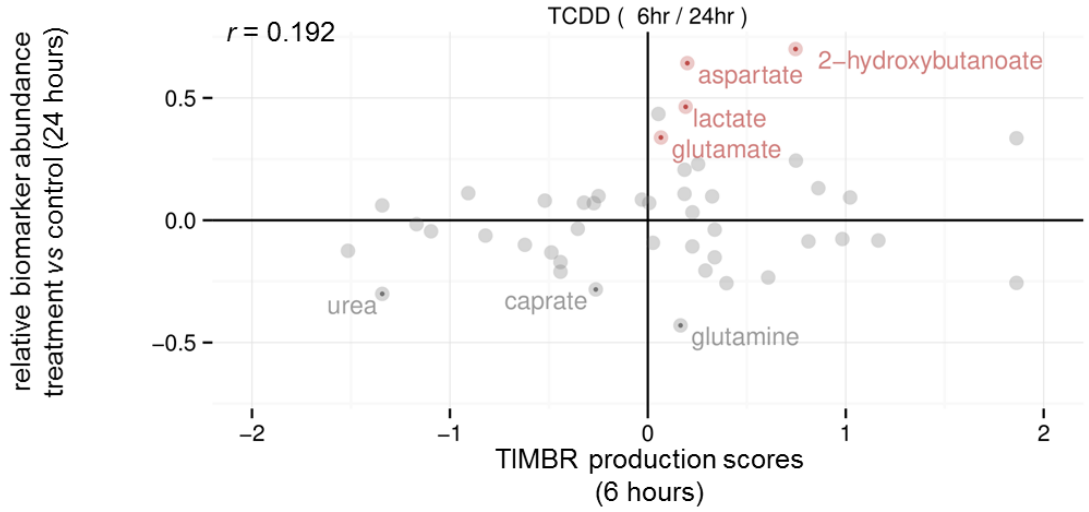


Figure 8.5 – Biomarker comparisons between predicted TIMBR production scores and metabolomics changes in response to TCDD. Aspartate, lactate, glutamate, and 2-hydroxyburanoate abundances were significantly increased (FDR < 0.3) at the 24 hour time point for TCDD experiments. Metabolites with the largest decreases in abundance are also highlighted although they are not statistically significant. Dots that appear close to the horizontal line can be interpreted as false positive predictions while dots close to the vertical line can be interpreted as false negative predictions, although we did not assign a quantitative cutoff value for a positive prediction. Red and gray circles respectively indicate metabolites with experimentally significant or insignificant changes in abundance between treatment and control conditions. These values are not reflective of whether the metabolite is consumed or produced (spent media versus fresh media).

Pinpointing exactly how glycine was predicted to be upregulated was remarkably simple considering the complexity of the metabolic network. A key contributing factor was that we only forced a small amount of flux through the biomarker exchange reaction. By increasing the minimum required flux through this reaction to 100, we found that optimal solution fluxes or TIMBR weights were different between across 106 reactions. These results highlight how it might be important to avoid forcing too much activity through the exchange reaction when running the TIMBR algorithm, considering that only a small number of genes were differentially expressed for CCl_4. For the result presented in this chapter, we forced a minimum required flux of 1 through each exchange reaction which achieved better performance than a value of 100 (data not shown). In the previous chapter, we applied the TIMBR algorithm with a minimum required flux of 100 which warrants further investigation into sensitivity analyses for various thresholds to increase the performance of the analyses described in this dissertation.

Table 8.2 – Relative gene expression changes for the rat gene, *Slc3a2*, in response to CCl_4 treatment after 6 and 24 hours.

treatment	duration	Entrez gene	gene symbol	logFC	FDR	p-value
CCl_4	6 hours	50567	Slc3a2	0.33	1.10E-08	1.04E-10
CCl_4	24 hours	50567	Slc3a2	0.25	7.11E-06	7.41E-08

Outlook

In this capstone analysis of the dissertation, we have provided evidence that computational models of metabolism can predict metabolic changes using gene expression data. Future studies are warranted in applying the algorithms and models described in throughout this dissertation to studying toxicology and metabolism in general. We plan on implementing the TIMBR algorithm to *in vivo* transcriptomics and fluxomics data generated by collaborators of this project that is funded by the United States Department of Defense. Ultimately, we anticipate that the rat and human metabolic networks described in this dissertation will be a valuable resource for contextualizing high-throughput genomic datasets and bridging the knowledge gap that exists between humans and rats.

Acknowledgements

This chapter was partially adapted from a manuscript currently under revision at Nature Communications titled “Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions” written by Edik M. Blais, Kristopher D. Rawls, Zhuo I. Li, Glynis L. Kolling, Ping Ye, Anders Wallqvist, and Jason A. Papin. This chapter was also adapted from an unsubmitted manuscript with that includes similar authorship with the addition of Bonnie Dougherty. I thank all co-authors who contributed to this work, specifically Glynis Kolling and Kristopher Rawls for performing the cell culture experiments and generating transcriptomics and metabolomics data that I used in computational analyses.

Chapter 9: Dissertation Discussions

Synopsis

In this dissertation, I have provided the first comprehensive summary of functional differences between rat and human metabolism. Metabolism is often described as conserved across species, but the degree of conservation for two mammalian species has never been systematically investigated. I was surprised to discover that nearly all of their biochemical capabilities were shared at the genome-scale. This finding underscores the important role of rats as a surrogate for understanding human biology and disease. The computational framework described throughout this dissertation can be used to highlight potential limitations of rats as a model organism and to further improve the effectiveness of rats in the preclinical stages of drug development and biomarker discovery. By establishing the foundation for *in silico* alternative, I hope that this research will facilitate the use of live animals more efficiently.

Historical perspective

The laboratory rat has been used as a surrogate to study human biology for more than a century[190]. Nearly a century ago, the first functional difference between rat and human metabolism was discovered when scurvy could not be reproduced in rats. Helen Parsons investigated[191] whether rats were capable of performing *de novo* vitamin C synthesis or if the nutritional requirements for rats were different than other organisms that were susceptible to scurvy like Guinea pigs. In a fascinating experiment, Guinea pigs on a scurvy-inducing diet were cured when fed liver tissue from rats on a similar scurvy diet. Remarkably, Parsons described that her finding was consistent with an account from the arctic explorer, Elisha Kane, who attributed his freedom from scurvy at sea during the winter by eating rats that infested the ship. I found this study intriguing because it embodied several concepts that are discussed throughout this dissertation: understanding the metabolic requirements for the healthy state is important; the ability to control nutritional inputs can be incredibly useful; and species-specific differences can limit the applicability of a model organism to study human disease.

The first genome-scale comparison between humans and rats was achieved when the rat genome was sequenced in 2004, shortly after the completion of the human genome project. However, the first systems-level analysis of rat biology was published in 1915[190] and updated in 1922[192], in which Donaldson provided a comprehensive reference datasets complete with tables and data related to the anatomical, physiological, and pathological properties of rats. Following a similar paradigm, this dissertation marks the first genome-scale collection of the metabolic capabilities of rats. Each of these studies provides comparative analyses discussing

the similarities and differences between rats and humans. Additionally, my goal, like theirs, was to comprehensively catalogue an important feature of rat biology and provide a rich resource for other researchers to build on. Interestingly, I encountered similar challenges with assembling heterogeneous pieces of information into a cohesive collection, as Donaldson's said[192]: "It is hardly necessary to add that in most directions our information is fragmentary." This dissertation, while limited to the scope of metabolism, provides an additional systems-level perspective to rat biology.

Species-specific differences

Despite a surprisingly small number of species-specific differences at the genome level, differences at the individual gene-level can alter network functionality and have the capacity to limit the use of rats to study human pathology and disease. Unlike rats, humans exclusively rely on dietary sources of vitamin C, which may obfuscate the clinical translation of rat studies that have described vitamin C as a potential biomarker[41,43]. The abundance and absence of β-muricholic acids in rats and humans, respectively, can have substantial implications within the context of toxicology because bile acids are frequently used as blood-based biomarkers of liver damage[42]. In **Chapter 7**, I described how human-specific deficiencies in the purine degradation could lead to unexpected side effects like gout in response to theophylline when extrapolating experimental results from rats.

From an evolutionary perspective, the consideration that humans may actually have fewer metabolic capabilities than rats was certainly intriguing. Although I exhaustively searched experimental literature to identify features that distinguish humans from rodents in **Chapter 3**, *FUT3* was the only example of a human-specific enzyme was distinguished humans from rats. In the future, I anticipate that more will be discovered that can be integrated into future iterations of the Ratcon database. One explanation for why more unique enzymes have been characterized for rats than humans is that metabolic enzymes first characterized in rats are likely followed up in humans whereas functions discovered other organisms may not be evaluated in rats. Furthermore, enzymes that are unique to humans may not have been functionally characterized yet because functional studies *in vivo* are rarely performed in humans first. Additionally, differences between rat and human metabolism are likely more pronounced for individual tissues where enzymes can be uniquely expressed in one organism and not the other.

Network reconciliation

Through the process of network reconciliation[50], I developed highly curated rat and human metabolic networks that recapitulate known specific-specific metabolic functions, quantitatively capture cellular growth rates, and generate comparative biomarker predictions. Reconciling differences between *iRno* and *iHsa* also improved the quality of each model individually. For example, I identified several rat-specific reactions that needed to be removed from *iHsa*. I also discovered that most reactions originally annotated as human-specific were capable of being catalyzed by rat enzymes. These examples highlight how reconciling differences between two species models can lead to valuable improvements that may be overlooked when focusing on a single species.

Although *iRno* is not the first animal adaptation of a human GENRE, this parallelized refinement approach has only previously been applied to bacterial GENREs[50,51]. Furthermore, by incorporating biochemical functions known to distinguish rats from humans, *iRno* and *iHsa* can facilitate the translation of therapeutic strategies and biomarkers that are supported by comparative model predictions. This feature does not apply to previous mouse GENREs which were based on, but not reconciled with, human GENREs[52,53]. Because these mouse GENREs were missing at least 60 reactions from each of their human counterparts, I anticipated that *iRno* and *iHsa* would contain a much larger number of species-specific reactions. Surprisingly, network reconciliation revealed that rat and human metabolism was much more conserved than my original expectations. This discovery is important for the reconstruction of other mammalian metabolic networks because previous mouse networks are likely missing functions due to insufficient orthology annotations, further supporting the orthogonal approach that I developed in **Chapter 2**.

Together, *iRno* and *iHsa* serve as a computational resource for understanding of rat metabolism within the context of human biology. *iRno* and *iHsa* have the capacity to improve the effectiveness of rat as a model organism in drug development and biomarker discovery. Potential applications include identifying combinatorial therapeutic strategies against cancers that minimize toxicity in normal cells[15], optimizing cell culture media formulations or experimental diets for specific diseases, and exploring potential genetic engineering strategies for new rat strains that better mimic human biology. Resources can be prioritized for simulated experiments that minimize species-specific differences *in silico* over those not supported by the mechanistic framework of these reconciled rat and human metabolic networks.

Rats serve an integral role in drug development, and understanding differences between rats and humans is critically important for the translation of preclinical studies into successful clinical trials[36-38]. Despite a high degree of physiologic and genomic similarities between rats and humans[39], several metabolic differences have been described that could affect whether a biomarker is elevated or whether a compound is toxic to the liver[41,44]. The rat and human metabolic networks described in this dissertation represent a comprehensive knowledgebase of functional differences between rat and human metabolism. In the future, we anticipate the toxicogenomics studies described in **Chapters 6-8** has the potential to dramatically improve the translation of preclinical studies to human trials.

Chapter 10: Future Directions

High-throughput screening and refining genome annotation

Disrupting enzymes is a fundamental strategy used in biology to understand relationships between genotypes and phenotypes. Experimental techniques that disrupt gene function via knockdown or knockout have enabled the genome-scale identification of essential genes in cancer cells[193,194]. With recent advances in genome editing tools such as the CRISPR/Cas9 system, extensive libraries with tens of thousands of sequences targeted to nearly 20,000 genes are now available that will dramatically increase the productivity of researchers in annotating the human genome[195]. **Chapter 5** in this dissertation provides a conceptual framework for leveraging new experimental techniques like CRISPR to refine genome annotations iteratively with a metabolic network.

Although genome-wide knockout screens that require running tens of thousands of samples are financially feasible with current technology, a genome-scale double-knockout screen for would require enough resources to run hundreds of millions of samples (20,000 choose 2 = 199,990,000 samples per replicate). Computational modeling provides an excellent opportunity to perform screening of combinatorial knock-out strategies because *in silico* experiments can be performed much faster and at a lower cost. Although computational modeling cannot entirely replace experimental methods, model predictions can be informative in narrowing down millions or billions of potential experiments to a number that fits within a researcher's budget. Additionally, high-throughput experimental screening methods are typically limited to a handful of readouts like cell number. Metabolic networks can be used to investigate the effect of each knockout strategy on multiple outputs. For example, **Chapter 7** described the use of human metabolic networks to generate biomarker predictions for several metabolic disorders known as inborn errors of metabolism (IEMs) by simulating the knockout of genes known to be associated with each disease[14].

Strain-specific models

By genetically knocking out specific genes in rats, researchers can develop strains that may better represent human diseases. As an example, genetically modified strains of rodents missing metabolic enzymes such as glutathione transferases have been developed as model systems for pharmacology and toxicology[196]. The functional representations of rat and human metabolism developed in this dissertation provide an excellent tool that can be used to predict the systems-wide effects of disrupting metabolic enzymes. The metabolic tasks defined in **Chapter 3** offer quality control measures to reduce the chance of introducing unintended side effects. Additionally, the comparative nature of these rat and human metabolic networks also ensure that human

physiology is kept in perspective. Estimating potential benefits and identifying potential negative outcomes will be highly valuable to researchers that aim to develop more clinically relevant models with limited resources.

Before applying Ratcon1 framework to design genetic engineering strategies for new rat strains, an immediately applicable next step would be to characterize existing rat strains. A recent publication characterized the genomic landscape of 40 rat strains that with over 12 million genomic variants. These strains are used for a variety of human diseases and metabolic networks specifically for individual strains would be highly valuable for researchers to get the most out of the Ratcon1 database. Additionally, in the future I anticipate that integrating the Ratcon1 database with the Rat Genome Database[197] would further promote the impact of the work of this dissertation and be of significant use to a large community of researchers that use rats to study human biology.

Biomarker discovery and personalized medicine

Eventually, personalized metabolic networks based on patient-specific data will be capable of analyzing all potential treatment options for a clinical condition in order to provide mechanistic insights into the combination of drugs that will provide the patient with the best balance between therapeutic efficacy and adverse toxicity. Treatment regimens can include a personalized panel of diagnostic and prognostic biomarkers that facilitate decision-making during the course of treatment. For example, biomarkers of efficacy can be monitored to determine when a drug stops working and an alternative treatment strategy should be implemented. With effective biomarkers of drug-induced liver, kidney, and heart toxicity, clinicians can significantly reduce the risk of serious injury due to unexpected toxicities. Furthermore, personal genomics profiles could inform the selection criteria for biomarkers by avoiding those that are not consistent with a patient's metabolic capabilities. As an example, the pancreatic cancer biomarker, CA19-9, is a human-specific metabolite not found in rats (see **Chapter 3**); however, genetic variants in the enzyme responsible for CA19-9 synthesis, *FUT3*, limit the applicability of this biomarker to patients with the specific Lewis blood phenotypes[198].

Whole-body metabolic networks

Most the applications discussed in this dissertation were limited to cell culture experiments. Translating *in vitro* studies to *in vivo* can be challenging, especially when considering that measuring fluxes *in vivo* is incredibly difficult. GENREs can facilitate the rapid interrogation and interpretation of high-throughput genomics and metabolomics data to fill in knowledge gaps that would otherwise be difficult to measure. Furthermore, metabolic networks can be used to extrapolate changes measured in one organ system to another via whole-body

metabolic network modeling (**Figure 10.1**). Future applications of the metabolic networks in this dissertation include systems toxicology analyses of the interplay between liver, kidney, and heart to establish biomarkers that are prognostic of life-threatening injuries that overcome the body's compensatory abilities. By understand how multiple organs respond to a given drug, more specific biomarkers can aid in accurate diagnoses of unknown toxic exposures, which are of significant interest to military and industrial communities.

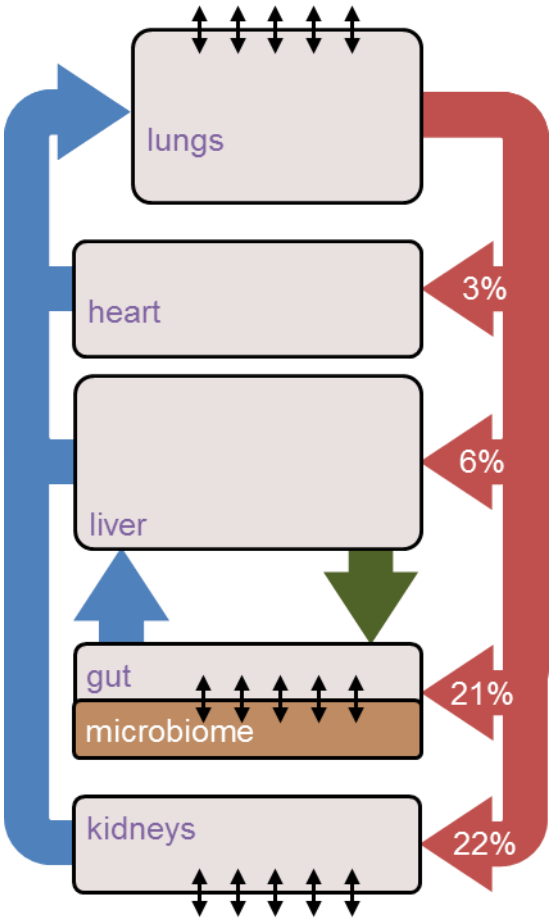


Figure 10.1 – Schematic for a whole-body representation using genome-scale metabolic networks. Organ systems can be represented as compartments in a metabolic network and can be connected by arteries (red) and veins (blue). Manual curation of the bile synthesis pathway described in **Chapter 3** specified gut-specific reactions that should occur in the mammalian microbiome compartment (brown) after transport through the bile duct (green array). Host exchange reactions (arrows) representing kidney filtration, dietary consumption, gastrointestinal waste removal, and breathing can enable systems-scale analyses that account for all metabolites that enter or leave the body. Approximating percentages of blood flow that enter each compartment could potentially be used to inform quantitative biomarker predictions.

References

1 Tujios, S. & Fontana, R. J. Mechanisms of drug-induced liver injury: from bedside to bench. *Nat Rev Gastroenterol Hepatol* **8**, 202-211, doi:10.1038/nrgastro.2011.22 (2011).

2 Lheureux, P. E. & Hantson, P. Carnitine in the treatment of valproic acid-induced toxicity. *Clin Toxicol (Phila)* **47**, 101-111, doi:10.1080/15563650902752376 (2009).

3 Watkins, P. B. Biomarkers for the diagnosis and management of drug-induced liver injury. *Seminars in liver disease* **29**, 393-399, doi:10.1055/s-0029-1240008 (2009).

4 McKenzie, R. *et al.* Hepatic failure and lactic acidosis due to fialuridine (FIAU), an investigational nucleoside analogue for chronic hepatitis B. *The New England journal of medicine* **333**, 1099-1105, doi:10.1056/nejm199510263331702 (1995).

5 Vangala, S. & Tonelli, A. Biomarkers, metabonomics, and drug development: can inborn errors of metabolism help in understanding drug toxicity? *The AAPS journal* **9**, E284-297, doi:10.1208/aapsj0903031 (2007).

6 Tonomura, Y. *et al.* Diagnostic and predictive performance and standardized threshold of traditional biomarkers for drug-induced liver injury in rats. *J Appl Toxicol* **35**, 165-172, doi:10.1002/jat.3053 (2015).

7 Sun, X. *et al.* Metabonomics evaluation of urine from rats administered with phorate under long-term and low-level exposure by ultra-performance liquid chromatography-mass spectrometry. *J Appl Toxicol* **34**, 176-183, doi:10.1002/jat.2848 (2014).

8 Martin, J. C. *et al.* Can we trust untargeted metabolomics? Results of the metabo-ring initiative, a large-scale, multi-instrument inter-laboratory study. *Metabolomics : Official journal of the Metabolomic Society* **11**, 807-821, doi:10.1007/s11306-014-0740-0 (2015).

9 Cobb, J. *et al.* A novel fasting blood test for insulin resistance and prediabetes. *Journal of diabetes science and technology* **7**, 100-110 (2013).

10 Zhang, Z. H. *et al.* General toxicity of Pinellia ternata (Thunb.) Berit. in rat: a metabonomic method for profiling of serum metabolic changes. *Journal of ethnopharmacology* **149**, 303-310, doi:10.1016/j.jep.2013.06.039 (2013).

11 Lu, F. *et al.* Urinary metabonomics study on toxicity biomarker discovery in rats treated with Xanthii Fructus. *Journal of ethnopharmacology* **149**, 311-320, doi:10.1016/j.jep.2013.06.040 (2013).

12 Zheng, T. *et al.* The metabolic impact of methamphetamine on the systemic metabolism of rats and potential markers of methamphetamine abuse. *Molecular bioSystems* **10**, 1968-1977, doi:10.1039/c4mb00158c (2014).

13 Zheng, T. *et al.* Metabolic phenotype of rats exposed to heroin and potential markers of heroin abuse. *Drug and alcohol dependence* **127**, 177-186, doi:10.1016/j.drugalcdep.2012.06.031 (2013).

14 Shlomi, T., Cabili, M. N. & Ruppin, E. Predicting metabolic biomarkers of human inborn errors of metabolism. *Mol Syst Biol* **5**, 263, doi:10.1038/msb.2009.22 (2009).

15 Folger, O. *et al.* Predicting selective drug targets in cancer through metabolic networks. *Mol Syst Biol* **7**, 501, doi:10.1038/msb.2011.35 (2011).

16 Duarte, N. C. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A* **104**, 1777-1782, doi:10.1073/pnas.0610772104 (2007).

17 Thiele, I. *et al.* A community-driven global reconstruction of human metabolism. *Nat Biotechnol* **31**, 419-425, doi:10.1038/nbt.2488 (2013).

18 Mardinoglu, A. *et al.* Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Mol Syst Biol* **9**, 649, doi:10.1038/msb.2013.5 (2013).

19 Mardinoglu, A. *et al.* Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat Commun* **5**, 3083, doi:10.1038/ncomms4083 (2014).

20 Gille, C. *et al.* HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol Syst Biol* **6**, 411, doi:10.1038/msb.2010.62 (2010).

21 Ma, H. *et al.* The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* **3**, 135, doi:10.1038/msb4100177 (2007).

22 Agren, R. *et al.* Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput Biol* **8**, e1002518, doi:10.1371/journal.pcbi.1002518 (2012).

23 Joyce, A. R. & Palsson, B. O. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* **7**, 198-210, doi:10.1038/nrm1857 (2006).

24 Zomorrodi, A. R. & Maranas, C. D. Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data. *BMC Syst Biol* **4**, 178, doi:10.1186/1752-0509-4-178 (2010).

25 Jerby, L. & Ruppin, E. Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling. *Clin Cancer Res* **18**, 5572-5584, doi:10.1158/1078-0432.CCR-12-1856 (2012).
26 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, doi:10.1038/nature02456 (2004).
27 Edwards, J. S., Ibarra, R. U. & Palsson, B. O. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol* **19**, 125-130, doi:10.1038/84379 (2001).
28 Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nat Biotechnol* **28**, 245-248, doi:10.1038/nbt.1614 (2010).
29 Searls, D. B. Pharmacophylogenomics: genes, evolution and drug targets. *Nat Rev Drug Discov* **2**, 613-623, doi:10.1038/nrd1152 (2003).
30 Shimoyama, M. *et al.* The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res* **43**, D743-750, doi:10.1093/nar/gku1026 (2015).
31 Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **43**, D6-17, doi:10.1093/nar/gku1130 (2015).
32 Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res* **43**, D662-669, doi:10.1093/nar/gku1010 (2015).
33 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
34 Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42**, D199-205, doi:10.1093/nar/gkt1076 (2014).
35 UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212, doi:10.1093/nar/gku989 (2015).
36 Carvajal-Zarrabal, O. *et al.* Avocado oil supplementation modifies cardiovascular risk profile markers in a rat model of sucrose-induced metabolic changes. *Dis Markers* **2014**, 386425, doi:10.1155/2014/386425 (2014).
37 Sun, J. *et al.* Identification of a metabolic biomarker panel in rats for prediction of acute and idiosyncratic hepatotoxicity. *Comput Struct Biotechnol J* **10**, 78-89, doi:10.1016/j.csbj.2014.08.001 (2014).
38 Lin, Z. *et al.* Joint MS-based platforms for comprehensive comparison of rat plasma and serum metabolic profiling. *Biomed Chromatogr* **28**, 1235-1245, doi:10.1002/bmc.3152 (2014).
39 Mattes, W. *et al.* Detection of hepatotoxicity potential with metabolite profiling (metabolomics) of rat plasma. *Toxicol Lett* **230**, 467-478, doi:10.1016/j.toxlet.2014.07.021 (2014).
40 Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521, doi:10.1038/nature02426 (2004).
41 Bando, K. *et al.* GC-MS-based metabolomics reveals mechanism of action for hydrazine induced hepatotoxicity in rats. *J Appl Toxicol* **31**, 524-535, doi:10.1002/jat.1591 (2011).
42 Luo, L., Schomaker, S., Houle, C., Aubrecht, J. & Colangelo, J. L. Evaluation of serum bile acid profiles as biomarkers of liver injury in rodents. *Toxicol Sci* **137**, 12-25, doi:10.1093/toxsci/kft221 (2014).
43 Shannahan, J. H. *et al.* Pulmonary oxidative stress, inflammation, and dysregulated iron homeostasis in rat models of cardiovascular disease. *J Toxicol Environ Health A* **73**, 641-656, doi:10.1080/15287390903578208 (2010).
44 Uehara, T. *et al.* The Japanese toxicogenomics project: application of toxicogenomics. *Mol Nutr Food Res* **54**, 218-227, doi:10.1002/mnfr.200900169 (2010).
45 Clayton, T. A. *et al.* Pharmaco-metabonomic phenotyping and personalized drug treatment. *Nature* **440**, 1073-1077, doi:10.1038/nature04648 (2006).
46 Chan, C., Berthiaume, F., Lee, K. & Yarmush, M. L. Metabolic flux analysis of cultured hepatocytes exposed to plasma. *Biotechnol Bioeng* **81**, 33-49, doi:10.1002/bit.10453 (2003).
47 Lee, K., Berthiaume, F., Stephanopoulos, G. N. & Yarmush, M. L. Profiling of dynamic changes in hypermetabolic livers. *Biotechnol Bioeng* **83**, 400-415, doi:10.1002/bit.10682 (2003).
48 Jensen, P. A. & Papin, J. A. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics* **27**, 541-547, doi:10.1093/bioinformatics/btq702 (2011).
49 Song, H. S., Reifman, J. & Wallqvist, A. Prediction of metabolic flux distribution from gene expression data based on the flux minimization principle. *PLoS One* **9**, e112524, doi:10.1371/journal.pone.0112524 (2014).
50 Oberhardt, M. A., Puchalka, J., Martins dos Santos, V. A. & Papin, J. A. Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. *PLoS Comput Biol* **7**, e1001116, doi:10.1371/journal.pcbi.1001116 (2011).

51 Bartell, J. A., Yen, P., Varga, J. J., Goldberg, J. B. & Papin, J. A. Comparative metabolic systems analysis of pathogenic Burkholderia. *J Bacteriol* **196**, 210-226, doi:10.1128/jb.00997-13 (2014).
52 Mardinoglu, A. *et al.* The gut microbiota modulates host amino acid and glutathione metabolism in mice. *Mol Syst Biol* **11**, 834, doi:10.15252/msb.20156487 (2015).
53 Sigurdsson, M. I., Jamshidi, N., Steingrimsson, E., Thiele, I. & Palsson, B. O. A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC Syst Biol* **4**, 140, doi:10.1186/1752-0509-4-140 (2010).
54 Duarte, N. C., Herrgard, M. J. & Palsson, B. O. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome research* **14**, 1298-1309, doi:10.1101/gr.2250904 (2004).
55 Devoid, S. *et al.* Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. *Methods Mol Biol* **985**, 17-45, doi:10.1007/978-1-62703-299-5_2 (2013).
56 Romero, P. *et al.* Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* **6**, R2, doi:10.1186/gb-2004-6-1-r2 (2005).
57 Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **42**, D459-471, doi:10.1093/nar/gkt1103 (2014).
58 Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* **44**, D471-480, doi:10.1093/nar/gkv1164 (2016).
59 Radrich, K. *et al.* Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Syst Biol* **4**, 114, doi:10.1186/1752-0509-4-114 (2010).
60 Agren, R. *et al.* The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. *PLoS Comput Biol* **9**, e1002980, doi:10.1371/journal.pcbi.1002980 (2013).
61 Gabaldon, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nature reviews. Genetics* **14**, 360-366, doi:10.1038/nrg3456 (2013).
62 Heinken, A., Sahoo, S., Fleming, R. M. & Thiele, I. Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut microbes* **4**, 28-40, doi:10.4161/gmic.22370 (2013).
63 Blazier, A. S. & Papin, J. A. Integration of expression data in genome-scale metabolic network reconstructions. *Front Physiol* **3**, 299, doi:10.3389/fphys.2012.00299 (2012).
64 Machado, D. & Herrgard, M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput Biol* **10**, e1003580, doi:10.1371/journal.pcbi.1003580 (2014).
65 Scheer, M. *et al.* BRENDA, the enzyme information system in 2011. *Nucleic Acids Res* **39**, D670-676, doi:10.1093/nar/gkq1089 (2011).
66 Penning, T. M., Jin, Y., Heredia, V. V. & Lewis, M. Structure-function relationships in 3alpha-hydroxysteroid dehydrogenases: a comparison of the rat and human isoforms. *J Steroid Biochem Mol Biol* **85**, 247-255 (2003).
67 Parsons, H. T. THE ANTISCORBUTIC CONTENT OF CERTAIN BODY TISSUES OF THE RAT: THE PERSISTENCE OF THE ANTISCORBUTIC SUBSTANCE IN THE LIVER OF THE RAT AFTER LONG INTERVALS ON A SCORBUTIC DIET. *Journal of Biological Chemistry* **44**, 587-602 (1920).
68 Kawai, T., Nishikimi, M., Ozawa, T. & Yagi, K. A missense mutation of L-gulono-gamma-lactone oxidase causes the inability of scurvy-prone osteogenic disorder rats to synthesize L-ascorbic acid. *The Journal of biological chemistry* **267**, 21973-21976 (1992).
69 Ha, M. N. *et al.* Functional rescue of vitamin C synthesis deficiency in human cells using adenoviral-based expression of murine l-gulono-gamma-lactone oxidase. *Genomics* **83**, 482-492, doi:10.1016/j.ygeno.2003.08.018 (2004).
70 Wu, X. W., Muzny, D. M., Lee, C. C. & Caskey, C. T. Two independent mutational events in the loss of urate oxidase during hominoid evolution. *Journal of molecular evolution* **34**, 78-84 (1992).
71 Ramazzina, I., Folli, C., Secchi, A., Berni, R. & Percudani, R. Completing the uric acid degradation pathway through phylogenetic comparison of whole genomes. *Nature chemical biology* **2**, 144-148, doi:10.1038/nchembio768 (2006).
72 Samraj, A. N. *et al.* A red meat-derived glycan promotes inflammation and cancer progression. *Proc Natl Acad Sci U S A* **112**, 542-547, doi:10.1073/pnas.1417508112 (2015).
73 Chou, H. H. *et al.* A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. *Proc Natl Acad Sci U S A* **95**, 11751-11756 (1998).
74 Costache, M. *et al.* Evolution of fucosyltransferase genes in vertebrates. *The Journal of biological chemistry* **272**, 29721-29728 (1997).

75 Carvalho, A. S. *et al.* Differential expression of alpha-2,3-sialyltransferases and alpha-1,3/4-fucosyltransferases regulates the levels of sialyl Lewis a and sialyl Lewis x in gastrointestinal carcinoma cells. *The international journal of biochemistry & cell biology* **42**, 80-89, doi:10.1016/j.biocel.2009.09.010 (2010).
76 Kitagawa, H. & Paulson, J. C. Cloning and expression of human Gal beta 1,3(4)GlcNAc alpha 2,3-sialyltransferase. *Biochemical and biophysical research communications* **194**, 375-382, doi:10.1006/bbrc.1993.1830 (1993).
77 Marion, T. L., Perry, C. H., St Claire, R. L., 3rd & Brouwer, K. L. Endogenous bile acid disposition in rat and human sandwich-cultured hepatocytes. *Toxicol Appl Pharmacol* **261**, 1-9, doi:10.1016/j.taap.2012.02.002 (2012).
78 Garcia-Canaveras, J. C., Donato, M. T., Castell, J. V. & Lahoz, A. Targeted profiling of circulating and hepatic bile acids in human, mouse, and rat using a UPLC-MRM-MS-validated method. *J Lipid Res* **53**, 2231-2241, doi:10.1194/jlr.D028803 (2012).
79 Swann, J. R. *et al.* Systemic gut microbial modulation of bile acid metabolism in host tissue compartments. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4523-4530, doi:10.1073/pnas.1006734107 (2011).
80 Blais, E. M., Chavali, A. K. & Papin, J. A. Linking genome-scale metabolic modeling and genome annotation. *Methods Mol Biol* **985**, 61-83, doi:10.1007/978-1-62703-299-5_4 (2013).
81 Bame, K. J. & Rome, L. H. Acetyl-coenzyme A:alpha-glucosaminide N-acetyltransferase. Evidence for an active site histidine residue. *The Journal of biological chemistry* **261**, 10127-10132 (1986).
82 Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* **5**, 264-276 (2003).
83 Mizushima, Y., Harauchi, T., Yoshizaki, T. & Makino, S. A rat mutant unable to synthesize vitamin C. *Experientia* **40**, 359-361 (1984).
84 Karlstadt, A. *et al.* CardioNet: a human metabolic network suited for the study of cardiomyocyte metabolism. *BMC Syst Biol* **6**, 114, doi:10.1186/1752-0509-6-114 (2012).
85 Chang, R. L., Xie, L., Xie, L., Bourne, P. E. & Palsson, B. O. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput Biol* **6**, e1000938, doi:10.1371/journal.pcbi.1000938 (2010).
86 Sohlenius-Sternbeck, A. K. Determination of the hepatocellularity number for human, dog, rabbit, rat and mouse livers from protein concentration measurements. *Toxicol In Vitro* **20**, 1582-1586, doi:10.1016/j.tiv.2006.06.003 (2006).
87 Benga, G. & Ferdinand, W. Amino acid composition of rat and human liver microsomes in normal and pathological conditions. *Biosci Rep* **15**, 111-116 (1995).
88 Niklas, J., Noor, F. & Heinzle, E. Effects of drugs in subtoxic concentrations on the metabolic fluxes in human hepatoma cell line Hep G2. *Toxicol Appl Pharmacol* **240**, 327-336, doi:10.1016/j.taap.2009.07.005 (2009).
89 Olsson, J. M., Eriksson, L. C. & Dallner, G. Lipid compositions of intracellular membranes isolated from rat liver nodules in Wistar rats. *Cancer Res* **51**, 3774-3780 (1991).
90 Yamada, M. *et al.* Biochemical characteristics of isolated rat liver stellate cells. *Hepatology* **7**, 1224-1229 (1987).
91 Bartles, J. R., Feracci, H. M., Stieger, B. & Hubbard, A. L. Biogenesis of the rat hepatocyte plasma membrane in vivo: comparison of the pathways taken by apical and basolateral proteins using subcellular fractionation. *J Cell Biol* **105**, 1241-1251 (1987).
92 Yang, L. Y., Kuksis, A., Myher, J. J. & Steiner, G. Origin of triacylglycerol moiety of plasma very low density lipoproteins in the rat: structural studies. *J Lipid Res* **36**, 125-136 (1995).
93 Gibbons, G. F., Khurana, R., Odwell, A. & Seelaender, M. C. Lipid balance in HepG2 cells: active synthesis and impaired mobilization. *J Lipid Res* **35**, 1801-1808 (1994).
94 Barle, H. *et al.* The concentrations of free amino acids in human liver tissue obtained during laparoscopic surgery. *Clin Physiol* **16**, 217-227 (1996).
95 Triguero, A. *et al.* Liver intracellular L-cysteine concentration is maintained after inhibition of the trans-sulfuration pathway by propargylglycine in rats. *Br J Nutr* **78**, 823-831 (1997).
96 Lykkesfeldt, J., Hagen, T. M., Vinarsky, V. & Ames, B. N. Age-associated decline in ascorbic acid concentration, recycling, and biosynthesis in rat hepatocytes--reversal with (R)-alpha-lipoic acid supplementation. *FASEB J* **12**, 1183-1189 (1998).

97 Turunen, M., Olsson, J. & Dallner, G. Metabolism and function of coenzyme Q. *Biochim Biophys Acta* **1660**, 171-199 (2004).
98 Zhao, M. *et al.* Vitamin B-6 restriction impairs fatty acid synthesis in cultured human hepatoma (HepG2) cells. *Am J Physiol Endocrinol Metab* **304**, E342-351, doi:10.1152/ajpendo.00359.2012 (2013).
99 Momchilova, A. *et al.* Resveratrol alters the lipid composition, metabolism and peroxide level in senescent rat hepatocytes. *Chem Biol Interact* **207**, 74-80, doi:10.1016/j.cbi.2013.10.016 (2014).
100 Qin, X. Y. *et al.* The effect of acyclic retinoid on the metabolomic profiles of hepatocytes and hepatocellular carcinoma cells. *PLoS One* **8**, e82860, doi:10.1371/journal.pone.0082860 (2013).
101 da Silva, V. R. *et al.* Targeted metabolomics and mathematical modeling demonstrate that vitamin B-6 restriction alters one-carbon metabolism in cultured HepG2 cells. *Am J Physiol Endocrinol Metab* **307**, E93-101, doi:10.1152/ajpendo.00697.2013 (2014).
102 Setchell, K. D. *et al.* Bile acid concentrations in human and rat liver tissue and in hepatocyte nuclei. *Gastroenterology* **112**, 226-235 (1997).
103 Yang, H., Roth, C. M. & Ierapetritou, M. G. Analysis of amino acid supplementation effects on hepatocyte cultures using flux balance analysis. *OMICS* **15**, 449-460, doi:10.1089/omi.2010.0070 (2011).
104 Banta, S., Yokoyama, T., Berthiaume, F. & Yarmush, M. L. Effects of dehydroepiandrosterone administration on rat hepatic metabolism following thermal injury. *J Surg Res* **127**, 93-105, doi:10.1016/j.jss.2005.01.001 (2005).
105 Banta, S. *et al.* Contribution of gene expression to metabolic fluxes in hypermetabolic livers induced through burn injury and cecal ligation and puncture in rats. *Biotechnol Bioeng* **97**, 118-137, doi:10.1002/bit.21200 (2007).
106 Varma, A. & Palsson, B. O. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Applied and environmental microbiology* **60**, 3724-3731 (1994).
107 Kim, B. H. *et al.* Dedifferentiation of conditionally immortalized hepatocytes with long-term in vitro passage. *Experimental & molecular medicine* **32**, 29-37, doi:10.1038/emm.2000.6 (2000).
108 Chen, Y. *et al.* Transplantation of immortalized human fetal hepatocytes prevents acute liver failure in 90% hepatectomized mice. *Transplantation proceedings* **42**, 1907-1914, doi:10.1016/j.transproceed.2010.01.061 (2010).
109 Yizhak, K., Chaneton, B., Gottlieb, E. & Ruppin, E. Modeling cancer metabolism on a genome scale. *Mol Syst Biol* **11**, 817, doi:10.15252/msb.20145307 (2015).
110 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462, doi:10.1093/nar/gkv1070 (2016).
111 Blaby-Haas, C. E. & de Crecy-Lagard, V. Mining high-throughput experimental data to link gene and function. *Trends in biotechnology* **29**, 174-182, doi:10.1016/j.tibtech.2011.01.001 (2011).
112 Hanson, A. D., Pribat, A., Waller, J. C. & de Crecy-Lagard, V. 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list--and how to find it. *The Biochemical journal* **425**, 1-11, doi:10.1042/BJ20091328 (2010).
113 Pouliot, Y. & Karp, P. D. A survey of orphan enzyme activities. *BMC bioinformatics* **8**, 244, doi:10.1186/1471-2105-8-244 (2007).
114 Emes, R. D. Inferring function from homology. *Methods in molecular biology* **453**, 149-168, doi:10.1007/978-1-60327-429-6_6 (2008).
115 Jones, C. E., Brown, A. L. & Baumann, U. Estimating the annotation error rate of curated GO database sequence annotations. *BMC bioinformatics* **8**, 170, doi:10.1186/1471-2105-8-170 (2007).
116 Thiele, I. & Palsson, B. O. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5**, 93-121, doi:10.1038/nprot.2009.203 (2010).
117 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29, doi:10.1038/75556 (2000).
118 Rombel, I. T., Sykes, K. F., Rayner, S. & Johnston, S. A. ORF-FINDER: a vector for high-throughput gene identification. *Gene* **282**, 33-41 (2002).
119 Lamesch, P. *et al.* hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics* **89**, 307-315, doi:10.1016/j.ygeno.2006.11.012 (2007).
120 Frishman, D. Protein annotation at genomic scale: the current status. *Chemical reviews* **107**, 3448-3466, doi:10.1021/cr068303k (2007).
121 Jones, S. J. Prediction of genomic functional elements. *Annual review of genomics and human genetics* **7**, 315-338, doi:10.1146/annurev.genom.7.080505.115745 (2006).

122 Erdin, S., Lisewski, A. M. & Lichtarge, O. Protein function prediction: towards integration of similarity metrics. *Current opinion in structural biology* **21**, 180-188, doi:10.1016/j.sbi.2011.02.001 (2011).
123 Manichaikul, A. *et al.* Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nature methods* **6**, 589-592, doi:10.1038/nmeth.1348 (2009).
124 Chavali, A. K., Whittemore, J. D., Eddy, J. A., Williams, K. T. & Papin, J. A. Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major. *Molecular systems biology* **4**, 177, doi:10.1038/msb.2008.15 (2008).
125 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).
126 Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic acids research* **36**, W5-9, doi:10.1093/nar/gkn201 (2008).
127 Rolfsson, O., Palsson, B. O. & Thiele, I. The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *BMC systems biology* **5**, 155, doi:10.1186/1752-0509-5-155 (2011).
128 Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic acids research* **32**, D277-280, doi:10.1093/nar/gkh063 (2004).
129 Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic acids research* **36**, D480-484, doi:10.1093/nar/gkm882 (2008).
130 Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research* **31**, 3784-3788 (2003).
131 Schneider, M., Tognolli, M. & Bairoch, A. The Swiss-Prot protein knowledgebase and ExPASy: providing the plant community with high quality proteomic data and tools. *Plant physiology and biochemistry : PPB / Societe francaise de physiologie vegetale* **42**, 1013-1021, doi:10.1016/j.plaphy.2004.10.009 (2004).
132 Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology* **28**, 977-982, doi:10.1038/nbt.1672 (2010).
133 Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research* **40**, D742-753, doi:10.1093/nar/gkr1014 (2012).
134 Karp, P. D. & Caspi, R. A survey of metabolic databases emphasizing the MetaCyc family. *Archives of toxicology* **85**, 1015-1033, doi:10.1007/s00204-011-0705-2 (2011).
135 Hertz-Fowler, C. *et al.* GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic acids research* **32**, D339-343, doi:10.1093/nar/gkh007 (2004).
136 Kumar, A., Suthers, P. F. & Maranas, C. D. MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC bioinformatics* **13**, 6, doi:10.1186/1471-2105-13-6 (2012).
137 Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic acids research* **32**, D115-119, doi:10.1093/nar/gkh131 (2004).
138 Bolser, D. M. *et al.* MetaBase--the wiki-database of biological databases. *Nucleic acids research* **40**, D1250-1254, doi:10.1093/nar/gkr1099 (2012).
139 Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524-531 (2003).
140 King, Z. A. *et al.* BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res* **44**, D515-522, doi:10.1093/nar/gkv1049 (2016).
141 Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. O. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics* **11**, 213, doi:10.1186/1471-2105-11-213 (2010).
142 Pabinger, S., Rader, R., Agren, R., Nielsen, J. & Trajanoski, Z. MEMOSys: Bioinformatics platform for genome-scale metabolic models. *BMC systems biology* **5**, 20, doi:10.1186/1752-0509-5-20 (2011).
143 Ganter, M., Bernard, T., Moretti, S., Stelling, J. & Pagni, M. MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics* **29**, 815-816, doi:10.1093/bioinformatics/btt036 (2013).
144 Schellenberger, J. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature protocols* **6**, 1290-1307, doi:10.1038/nprot.2011.308 (2011).
145 Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol* **7**, 74, doi:10.1186/1752-0509-7-74 (2013).
146 Keating, S. M., Bornstein, B. J., Finney, A. & Hucka, M. SBMLToolbox: an SBML toolbox for MATLAB users. *Bioinformatics* **22**, 1275-1277, doi:10.1093/bioinformatics/btl111 (2006).

147 Chavali, A. K., D'Auria, K. M., Hewlett, E. L., Pearson, R. D. & Papin, J. A. A metabolic network approach for the identification and prioritization of antimicrobial drug targets. *Trends in microbiology* **20**, 113-123, doi:10.1016/j.tim.2011.12.004 (2012).
148 Satish Kumar, V., Dasika, M. S. & Maranas, C. D. Optimization based automated curation of metabolic reconstructions. *BMC bioinformatics* **8**, 212, doi:10.1186/1471-2105-8-212 (2007).
149 Reed, J. L. *et al.* Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 17480-17484, doi:10.1073/pnas.0603364103 (2006).
150 Orth, J. D. & Palsson, B. O. Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions. *BMC systems biology* **6**, 30, doi:10.1186/1752-0509-6-30 (2012).
151 Karp, P. D., Paley, S. & Romero, P. The Pathway Tools software. *Bioinformatics* **18 Suppl 1**, S225-232 (2002).
152 Karp, P. D. *et al.* Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics* **11**, 40-79, doi:10.1093/bib/bbp043 (2010).
153 Latendresse, M., Krummenacker, M., Trupp, M. & Karp, P. D. Construction and completion of flux balance models from pathway databases. *Bioinformatics* **28**, 388-396, doi:10.1093/bioinformatics/btr681 (2012).
154 Green, M. L. & Karp, P. D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC bioinformatics* **5**, 76, doi:10.1186/1471-2105-5-76 (2004).
155 Green, M. L. & Karp, P. D. Using genome-context data to identify specific types of functional associations in pathway/genome databases. *Bioinformatics* **23**, i205-211, doi:10.1093/bioinformatics/btm213 (2007).
156 Kumar, V. S. & Maranas, C. D. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS computational biology* **5**, e1000308, doi:10.1371/journal.pcbi.1000308 (2009).
157 Herrgard, M. J., Fong, S. S. & Palsson, B. O. Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS computational biology* **2**, e72, doi:10.1371/journal.pcbi.0020072 (2006).
158 Hatzimanikatis, V. *et al.* Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603-1609, doi:10.1093/bioinformatics/bti213 (2005).
159 Orth, J. D. & Palsson, B. O. Systematizing the generation of missing metabolic knowledge. *Biotechnology and bioengineering* **107**, 403-412, doi:10.1002/bit.22844 (2010).
160 Oberhardt, M. A., Chavali, A. K. & Papin, J. A. Flux balance analysis: interrogating genome-scale metabolic networks. *Methods Mol Biol* **500**, 61-80, doi:10.1007/978-1-59745-525-1_3 (2009).
161 Joyce, A. R. *et al.* Experimental and computational assessment of conditionally essential genes in Escherichia coli. *Journal of bacteriology* **188**, 8259-8271, doi:10.1128/JB.00740-06 (2006).
162 Feist, A. M. & Palsson, B. O. The biomass objective function. *Current opinion in microbiology* **13**, 344-349, doi:10.1016/j.mib.2010.03.003 (2010).
163 Ghamsari, L. *et al.* Genome-wide functional annotation and structural verification of metabolic ORFeome of Chlamydomonas reinhardtii. *BMC genomics* **12 Suppl 1**, S4, doi:10.1186/1471-2164-12-S1-S4 (2011).
164 Zimniak, P., Holsztynska, E. J., Lester, R., Waxman, D. J. & Radominska, A. Detoxification of lithocholic acid. Elucidation of the pathways of oxidative metabolism in rat liver microsomes. *J Lipid Res* **30**, 907-918 (1989).
165 Chang, T. K., Teixeira, J., Gil, G. & Waxman, D. J. The lithocholic acid 6 beta-hydroxylase cytochrome P-450, CYP 3A10, is an active catalyst of steroid-hormone 6 beta-hydroxylation. *Biochem J* **291 (Pt 2)**, 429-433 (1993).
166 Strotkamp, D., Roos, P. H. & Hanstein, W. G. A novel CYP3 gene from female rats. *Biochim Biophys Acta* **1260**, 341-344 (1995).
167 Nagata, K. *et al.* Isolation and characterization of a new rat P450 (CYP3A18) cDNA encoding P450(6)beta-2 catalyzing testosterone 6 beta- and 16 alpha-hydroxylations. *Pharmacogenetics* **6**, 103-111 (1996).
168 Igarashi, Y. *et al.* Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res* **43**, D921-927, doi:10.1093/nar/gku955 (2015).
169 Jensen, P. A., Lutz, K. A. & Papin, J. A. TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. *BMC Syst Biol* **5**, 147, doi:10.1186/1752-0509-5-147 (2011).
170 Kolesnikov, N. *et al.* ArrayExpress update--simplifying data submissions. *Nucleic Acids Res* **43**, D1113-1116, doi:10.1093/nar/gku1057 (2015).

171 Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363-2367, doi:10.1093/bioinformatics/btq431 (2010).
172 Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**, 115-121, doi:10.1038/nmeth.3252 (2015).
173 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).
174 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).
175 Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J. & Lercher, M. J. Sybil--efficient constraint-based modelling in R. *BMC Syst Biol* **7**, 125, doi:10.1186/1752-0509-7-125 (2013).
176 De Abrew, K. N. *et al.* A novel transcriptomics based in vitro method to compare and predict hepatotoxicity based on mode of action. *Toxicology* **328**, 29-39, doi:10.1016/j.tox.2014.11.008 (2015).
177 Jensen, P. A. & Papin, J. A. MetDraw: automated visualization of genome-scale metabolic network reconstructions and high-throughput data. *Bioinformatics* **30**, 1327-1328, doi:10.1093/bioinformatics/btt758 (2014).
178 Silva, M. F. *et al.* Valproic acid metabolism and its effects on mitochondrial fatty acid oxidation: a review. *Journal of inherited metabolic disease* **31**, 205-216, doi:10.1007/s10545-008-0841-x (2008).
179 Ponsoda, X., Bort, R., Jover, R., Gomez-Lechon, M. J. & Castell, J. V. Molecular mechanism of diclofenac hepatotoxicity: Association of cell injury with oxidative metabolism and decrease in ATP levels. *Toxicol In Vitro* **9**, 439-444 (1995).
180 Felser, A. *et al.* Hepatocellular toxicity of benzbromarone: effects on mitochondrial function and structure. *Toxicology* **324**, 136-146, doi:10.1016/j.tox.2014.08.002 (2014).
181 Lewis, N. E. *et al.* Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol* **6**, 390, doi:10.1038/msb.2010.47 (2010).
182 Zielinski, D. C. *et al.* Pharmacogenomic and clinical data link non-pharmacokinetic metabolic dysregulation to drug side effect pathogenesis. *Nat Commun* **6**, 7101, doi:10.1038/ncomms8101 (2015).
183 Jorda, A., Portoles, M., Guasch, R., Bernal, D. & Saez, G. T. Effect of caffeine on urea biosynthesis and some related processes, ketone bodies, ATP and liver amino acids. *Biochemical pharmacology* **38**, 2727-2732 (1989).
184 Lee, Y. S. *et al.* Acetaminophen selectively suppresses peripheral prostaglandin E2 release and increases COX-2 gene expression in a clinical model of acute inflammation. *Pain* **129**, 279-286, doi:10.1016/j.pain.2006.10.020 (2007).
185 Kippen, I., Whitehouse, M. W. & Klinenberg, J. R. Pharmacology of uricosuric drugs. *Annals of the rheumatic diseases* **33**, 391-396 (1974).
186 Spilberg, I., Mandell, B., Mehta, J., Simchowitz, L. & Rosenberg, D. Mechanism of action of colchicine in acute urate crystal-induced arthritis. *The Journal of clinical investigation* **64**, 775-780, doi:10.1172/jci109523 (1979).
187 Bluestone, R., Kippen, I. & Klinenberg, J. R. Effect of drugs on urate binding to plasma proteins. *British medical journal* **4**, 590-593 (1969).
188 Morita, Y., Nishida, Y., Kamatani, N. & Miyamoto, T. Theophylline increases serum uric acid levels. *The Journal of allergy and clinical immunology* **74**, 707-712 (1984).
189 Choi, H. K. & Curhan, G. Coffee, tea, and caffeine consumption and serum uric acid level: the third national health and nutrition examination survey. *Arthritis and rheumatism* **57**, 816-821, doi:10.1002/art.22762 (2007).
190 Donaldson, H. H. *The Rat: Reference Tables and Data for the Albino Rat (Mus Norvegicus Albinus) and the Norway Rat (Mus Norvegicus)*. (Wistar Institute Press, 1915).
191 Parsons, H. T. THE ANTISCORBUTIC CONTENT OF CERTAIN BODY TISSUES OF THE RAT THE PERSISTENCE OF THE ANTISCORBUTIC SUBSTANCE IN THE LIVER OF THE RAT AFTER LONG INTERVALS ON A SCORBUTIC DIET. *Journal of Biological Chemistry* **44**, 587-602 (1920).
192 Donaldson, H. H. *The Rat, Data and Reference Tables for the Albino Rat (Mus Norvegicus Albinus) and the Norway Rat (Mus Norvegicus)*. (Henry H. Donaldson, 1924).
193 Berns, K. *et al.* A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431-437, doi:10.1038/nature02371 (2004).
194 Luo, B. *et al.* Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci U S A* **105**, 20380-20385, doi:10.1073/pnas.0810485105 (2008).

195 Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science (New York, N.Y.)* **343**, 84-87, doi:10.1126/science.1247005 (2014).
196 Board, P. G. The use of glutathione transferase-knockout mice as pharmacological and toxicological models. *Expert opinion on drug metabolism & toxicology* **3**, 421-433, doi:10.1517/17425255.3.3.421 (2007).
197 Hayman, G. T. *et al.* The Disease Portals, disease-gene annotation and the RGD disease ontology at the Rat Genome Database. *Database : the journal of biological databases and curation* **2016**, doi:10.1093/database/baw034 (2016).
198 Nishihara, S. *et al.* Molecular genetic analysis of the human Lewis histo-blood group system. *The Journal of biological chemistry* **269**, 29271-29278 (1994).