THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

On metabolic networks and multi-omics integration

Rasmus Ågren

Systems and Synthetic Biology
Department of Chemical and Biological Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2013

On metabolic networks and multi-omics integration

Rasmus Ågren

ISBN 978-91-7385-839-7

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 3520
ISSN 0346-718X

Department of Chemical and Biological Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 (0)31-772 1000

Cover:
Workflow for identification of transcriptionally regulated reactions. See Figure 4-4 on page 29 for details.

Printed by Chalmers Reproservice
Gothenburg, Sweden 2013

On metabolic networks and multi-omics integration

Rasmus Ågren
Systems and Synthetic Biology
Department of Chemical and Biological Engineering
Chalmers University of Technology

Abstract

Cellular metabolism is a highly complex chemical system, involving thousands of interacting metabolites and reactions. The traditional approach to understanding metabolism has been that of reductionism; by isolating and carefully measuring the involved components, the goal has been to understand the whole as the sum of its parts. This reductionist approach has successfully identified most of the components of metabolism but, unfortunately, it fails to capture the long-range and complex interactions that are essential for the functionality. Systems biology is an emerging research field which uses high-throughput data generation and mathematical modelling in order to apply a holistic, or network-centric, view on metabolism. One type of modelling framework, which is in line with this thinking, is genome-scale metabolic modelling. These models, called GEMs, represent very valuable resources, but their applications have been limited due to the large manual effort required to reconstruct them. In this project, we have developed algorithms and software for streamlining the reconstruction process, as well as for novel applications of GEMs. More specifically, we here present: the RAVEN Toolbox, a software suite for automated reconstruction and quality control; the INIT algorithm, an algorithm for inferring GEMs for human cell types; an algorithm which integrates fluxomics and transcriptomics data in order to identify transcriptionally controlled metabolic reactions.

The methods and software were used in a number of case studies to address real biological questions. These studies were: 1) Metabolic engineering of *Saccharomyces cerevisiae* for succinic acid overproduction. The predictions from the modelling were successfully validated experimentally. 2) Study of metabolic regulation in *S. cerevisiae*. This led to the identification of a small number of transcription factors and enzymes which were predicted to be controlling central parts of metabolism. 3). Penicillin production in *Penicillium chrysogenum*. This led to the reconstruction of the first GEM for *P. chrysogenum*, an important resource in itself, and to identification of metabolic engineering targets for more efficient production of penicillin. 4) Human cancer metabolism. This led to the identification of metabolic subnetworks which were predicted to be significantly more active in cancers, and to identification of potential drug targets for treatment. 5) Lipid metabolism in obesity. This led to new insights into the large-scale metabolic rearrangements associated with obesity, and to identification of possible therapeutic strategies. 6) Metabolism in non-alcoholic fatty liver disease. This led to the identification of serine deficiency as a central aspect of the disease, and to proposed therapeutic strategies for remedying it.

The work put forward in this thesis has resulted in improvements on several important aspects of genome-scale metabolic modelling, and it has shown how the framework can be applied to gain novel biological insights. As such, it can contribute to further increase the role of the framework in modelling of human health and disease.

Keywords: genome-scale metabolic model; penicillin; succinate; NASH; cancer; metabolic engineering; systems biology; metabolism; reconstruction; flux balance analysis

List of publications

This thesis is based on the following publications, referred to as Paper I to VI in the text:

I. **Agren, R.**, Otero, J.M. and Nielsen, J. (2013) Genome-scale modeling enables metabolic engineering of Saccharomyces cerevisiae for succinic acid production, *J Ind Microbiol Biot*, doi:10.1007/s10295-013-1269-3.

II. Bordel, S., **Agren, R.** and Nielsen, J. (2010) Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes, *PLoS Comput Biol*, 6(7), p. e1000859.

III. **Agren, R.**, Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I. and Nielsen, J. (2013) The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum, *PLoS Comput Biol*, 9(3), p. e1002980.

IV. **Agren, R.**[*], Bordel, S.[*], Mardinoglu, A., Pornputtapong, N., Nookaew, I. and Nielsen, J. (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT, *PLoS Comput Biol*, 8(5), p. e1002518.

V. Mardinoglu, A., **Agren, R.**, Kampf, C., Asplund, A., Nookaew, I., Jacobson, P., Walley, A.J., Froguel, P., Carlsson, L.M., Uhlen, M., Nielsen, J. (2013) Integration of clinical data with a genome-scale metabolic model of the human adipocyte, *Mol Syst Biol*, 9, p. 649.

VI. Mardinoglu, A.[*], **Agren, R.**[*], Kampf, C., Uhlen, M. and Nielsen, J. (2013) Genome-scale metabolic modeling of hepatocytes leads to identification of serine deficiency in non-alcoholic fatty liver disease, (Submitted).

Additional publications not included in this thesis:

VII. **Agren, R.**[*], Mardinoglu, A.[*], Kampf, C., Uhlen, M. and Nielsen, J. (2013) Drug discovery through the use of personalized genome-scale metabolic models for liver cancer, (Submitted).

VIII. Thiele, I., Swainston, N., Fleming, R.M., Hoppe, A., Sahoo, S., Aurich, M.K., Haraldsdottir, H., Mo, M.L., Rolfsson, O., Stobbe, M.D., Thorleifsson, S.G., **Agren, R.**, Bolling, C., Bordel, S., . . . Palsson, B.O. (2013) A community-driven global reconstruction of human metabolism, *Nat Biotechnol,* doi:10.1038/nbt.2488.

IX. Caspeta, L., Shoaie, S., **Agren, R.**, Nookaew, I. and Nielsen, J. (2012) Genome-scale metabolic reconstructions of Pichia stipitis and Pichia pastoris and in silico evaluation of their potentials, *BMC Syst Biol*, 6, p. 24.

X. Pabinger, S., Rader, R., **Agren, R.**, Nielsen, J. and Trajanoski, Z. (2011) MEMOSys: Bioinformatics platform for genome-scale metabolic models, *BMC Syst Biol*, 5, p. 20.

XI. Liu, L., **Agren, R.**, Bordel, S. and Nielsen, J. (2010) Use of genome-scale metabolic models for understanding microbial physiology, *FEBS Lett*, 584(12), pp. 2556-2564.

XII. Cvijovic, M.[*], Olivares-Hernandez, R.[*], **Agren, R.**[*], Dahr, N., Vongsangnak, W., Nookaew, I., Patil, K.R., Nielsen, J. (2010) BioMet Toolbox: genome-wide analysis of metabolism, *Nucleic Acids Res*, 38, pp. W144-149.

[*]Authors contributed equally

Contribution summary

I. Performed the modelling, assisted in the design of the fermentation experiments, analysed results, prepared and submitted the paper.

II. Formulated and implemented the sampling algorithm, participated in the analysis of results, assisted on preparation and submission of the paper.

III. Created the software, reconstructed the *Penicillium chrysogenum* model, performed the modelling, prepared and submitted the paper.

IV. Formulated and implemented the INIT algorithm, participated in the creation of the HMR database, reconstructed the cell type-specific models, participated in the analysis of results, participated in the preparation and submission of the paper.

V. Participated in the reconstruction of the adipocyte model, performed the modelling, participated in the analysis of results, assisted on preparation and submission of the paper.

VI. Participated in the reconstruction of the hepatocyte model, performed the modelling, participated in the analysis of results, participated in the preparation and submission of the paper.

VII. Formulated and implemented the tINIT algorithm, performed the modelling, participated in the analysis of results, participated in the preparation and submission of the paper.

VIII. Participated in the reconstruction of the model.

IX. Supervised the work, assisted in the analysis of results, assisted in the preparation and submission of the paper.

X. Participated in the design of the software, generated data to populate the database, assisted in the preparation and submission of the paper.

XI. Participated in the preparation and submission of the paper.

XII. Participated in the design and creation of the software, participated in the preparation and submission of the paper.

Table of contents

Lists of figures and tables

To my parents, for all that has been
To Evelina, for all that will be

“No offense, son, but that's some weak-ass thinking.
You equivocating like a motherfucker.”
Reginald “Bubbles” Cousins, The Wire (2004)

Preface

This dissertation is submitted for the partial fulfilment of the degree of doctor of philosophy. It is based on work carried out between 2009 and 2013 in the Systems and Synthetic Biology group, Department of Chemical and Biological Engineering, Chalmers University of Technology under the supervision of Professor Jens Nielsen. The research was funded by the Knut and Alice Wallenberg Foundation, Sandoz, Vetenskapsrådet, and the Chalmers Foundation.

Rasmus Ågren

May 2013

Abbreviations and symbols

A. nidulans: Aspergillus nidulans

A. niger: Aspergillus niger

A. oryzae: Aspergillus oryzae

AKG: α-ketoglutarate

BLAST: Basic local alignment search tool

BLP: Bi-level programming

BMI: Body mass index

C-mol: Carbon mol

CBM: Constraint-based model / constraint-based modelling

CE: Cholesterol ester

CS: Chondroitin sulfate

DCW: Dry cell weight

E. coli: *Escherichia coli*

EFM: Elementary flux mode

EP: Extreme pathway

FA: Fatty acid

FAD^+: Flavin adenine dinucleotide

$FADH_2$: Reduced flavin adenine dinucleotide

FBA: Flux balance analysis

FL: Fatty liver

GEM: Genome-scale metabolic model

GPR: Gene-protein-reaction

HDL: High-density lipoprotein

HMDB: Human metabolome database

HMM: Hidden Markov model

HMR: Human metabolic reaction database

HPA: Human protein atlas

HS: Heparan sulfate

HSPG: Heparan sulfate proteoglycans

IHC: Immunohistochemistry

INIT: Integrative network inference for tissues

KO: KEGG orthology

LD: Lipid droplet

LDL: Low-density lipoprotein

LP: Linear programming

LPL: Lipoprotein lipase

MCA: Metabolic control analysis

MILP: Mixed-integer linear programming

MFA: Metabolic flux analysis

MOMA: Minimisation of metabolic adjustment

NAFDL: Non-alcoholic fatty liver disease

NASH: Non-alcoholic steatohepatitis

NEFA: Non-esterified fatty acid

ORF: Open reading frame

P. chrysogenum: *Penicillium chrysogenum*

PEP: Phosphoenolpyruvate

PC: Phosphatidylcholine

PE: Phosphatidylethanolamine

PG: Proteoglycan

PS: Phosphatidylserine

QC: Quality control

QP: Quadratic programming

RAVEN: Reconstruction, analysis, and visualization of metabolic networks

ROOM: Regulatory on/off minimization

S. cerevisiae: *Saccharomyces cerevisiae*

SGR: Specific growth rate

SOP: Standard operating procedure

SSP: Serine synthesis pathway

TAG: Triacylglyceride

TCA: Tricarboxylic acid

VLDL: Very low-density lipoprotein

Nomenclature

Standard nomenclature is used for *S. cerevisiae* and *P. chrysogenum* for designating genes, proteins and gene deletions: *IDH1*, Idh1p and *Δidh1*, respectively, for isocitrate dehydrogenase 1 as an example. For *H. sapiens* it will be designated as *IDH1* and IDH1 for genes and proteins, respectively.

1 Introduction

The first genome-scale metabolic model (GEM), for the bacteria *Haemophilus influenzae*, was published in the year 2000, nine years before I started my Ph.D. studies in 2009. During these years the models grew increasingly complex, with the first model for a eukaryote published in 2003 and the first human model in 2007. Extensive method development also took place during this period, particularly for applications in metabolic engineering and strain design, where dozens of algorithms were developed. Some of these algorithms were quickly forgotten, while others proved themselves to be highly useful. The early successes in the field led to an ever larger number of GEMs being reconstructed, also for less well characterized organisms. After working with a GEM during my master's studies I had identified a number of issues that I felt should be targets for further work.

- **Model reconstruction was labour-intensive and error-prone.** Some of the published models, particularly for non-model organisms, were of rather low quality. I therefore started working on what later became the RAVEN Toolbox (**Paper III**), a software suite with focus on speeding up the reconstruction process, while at the same time ensuring a high-quality model.
- **GEMs were underused as scaffolds for omics integration.** GEMs developed when the first genome projects were finished, which allowed for incorporating gene/transcript/protein/reaction relationships in metabolic networks. A GEM can therefore be viewed as a highly structured map of metabolism in a cell, from metabolites all the way up to the genes. This would make the GEM very well suited as a scaffold for integrating and interpreting omics data of different sorts. I started working on methods for integrating transcriptomics data and flux data from fermentations into GEMs. This resulted in an algorithm for finding reactions which are likely to be transcriptionally regulated (**Paper II**).
- **There were difficulties associated with modelling of complex organisms.** My long-term goal when starting my studies was to model human metabolism and interactions between different tissues. This was associated with some issues that were not seen for simple prokaryotic organisms. Firstly, eukaryotes have their metabolism divided across subcellular compartments, and this information is not readily available. Secondly, most human cells do not actively divide, which makes the assumption of optimization of biomass yield, commonly used for microbial cells, unrealistic. Thirdly, different cell types have very different phenotypes even though they share the same genotype. It is therefore not possible to use the same GEM for all cell types, nor is it possible to reconstruct cell type-specific models only from the genome sequence. While attempting to deal with these issues I developed an algorithm for assigning sub-cellular localization in GEMs (**Paper III**) and the INIT algorithm for reconstruction of cell type-specific GEMs based on a multitude of omics types (**Paper IV**).

The software and algorithms mentioned above were then used to study succinate production in *Saccharomyces cerevisiae* (**Paper I**), regulation of metabolism in *S. cerevisiae* (**Paper II**), penicillin production in *Penicillium chrysogenum* (**Paper III**), cancer metabolism in human (**Paper IV**), adipocyte metabolism in obesity (**Paper V**), and hepatocyte metabolism in non-alcoholic fatty liver disease (**Paper VI**). Since the biological problems studied in this thesis span several areas there is no common background section. Rather, each problem is introduced in the corresponding results section. The background section represents a more

general review of the applications of genome-scale metabolic models and constraint-based modelling.

1.1 Thesis structure

This thesis represents a summary of a number of published scientific essays, a so-called compilation thesis. The thesis is divided into two parts: an extended summary and a compilation of research articles. Part one first puts the work in a larger scientific context by describing the field in which it is carried out, and the problems being studied in the field (sections 2.1 and 2.2). It then describes and explains the history and formulation of the main methodological framework underlying the work (section 2.3). After that follows an extensive examination and evaluation of the literature within areas relating to the work (section 3). This section does not address work carried out within the Ph.D. project. Part one is then concluded with a summary of the articles which form the basis for the thesis, and the results obtained in them (section 4). Lastly, some concluding remarks and future perspectives are presented (section 5). Part two contains the original research articles. The order of the articles follows the order in which the work was performed, rather than the order of publication.

2 Background

Part of the work described in this thesis deals with methods for optimizing the genetic composition of microbial organisms in order to enable production of industrially relevant metabolites. It can therefore be said to belong to the field of metabolic engineering. Other parts of the work are about utilizing a holistic view of metabolism in order to integrate and understand large-scale data sets, which would put it closer to the field of systems biology. The following two sections briefly introduce those research fields in order to outline the foundation upon which the work is based. Section 2.3 describes and explains the methodological framework underlying most of the work.

2.1 Metabolic engineering

Metabolic engineering is about analysing and modifying metabolic pathways in order to achieve some objective; normally efficient production of industrially relevant compounds. Attempts to change the metabolism of microorganisms to suit our purposes have been carried out for a long time in the biotechnology industry, for example for amino acids and antibiotics production. These early attempts utilized chemical mutagenesis to speed up the evolution process, and creative selection techniques to steer evolution in the desired direction (Stephanopoulos *et al.*, 1998). Such approaches can be very successful, as shown in the case of penicillin production, where the titre could be increased from 3-6 μM for the original strain identified by Fleming in 1928 to >75 mM in 1977 (Nielsen, 1995). However, the drawback with these techniques was that the underlying reasons for the change in phenotype remained unknown, which made it difficult to identify relevant constraints and pathways. In the late 1980s and early 1990s molecular biology tools for making genetic modifications became available; enabling targeted modifications of metabolic pathways. This became known as metabolic engineering (Bailey, 1991).

Mathematical concepts developed at about this time enabled quantification of the control each enzyme in a pathway had on the overall flux through the pathway, and for calculation of the theoretical production yields of metabolites in complex metabolic networks (see section 2.3). The general methodology in metabolic engineering, at least in its most traditional sense, is to apply mathematical modelling in order to identify constraints which could be limiting the production of the compound of interest. Such constraints could relate to for example substrate specificity of enzymes, product inhibition, medium composition, or redox balances. Once identified, genetic engineering is used to modify the genetic makeup of the organism in order to relax the constraints, leading to increased production of the compound of interest. Possible strategies include expression of heterologous enzymes, overexpression of endogenous enzymes, deletion of genes or modulation of enzymatic activity, transcriptional or enzymatic deregulation, and optimization of medium composition (Stephanopoulos *et al.*, 1998). But metabolic engineering also has a broader ambition. It tries to answer questions such as: How can the most important parameters which define the phenotype be identified? How can that information be contextualized in the control architecture of the metabolic network? How can we redirect cellular metabolism towards something that is often detrimental or even toxic to the cell? How can the network structure and dynamics be used to propose rational genetic engineering targets? In order to provide answers to questions such as these, metabolic engineering takes a holistic view of metabolism and looks at the integrated metabolic pathways rather than at individual reactions in isolation. This mind-set is closely related with

systems biology (see section 2.2). Figure 2-1 shows some types of pathway manipulations which are commonly attempted in metabolic engineering.

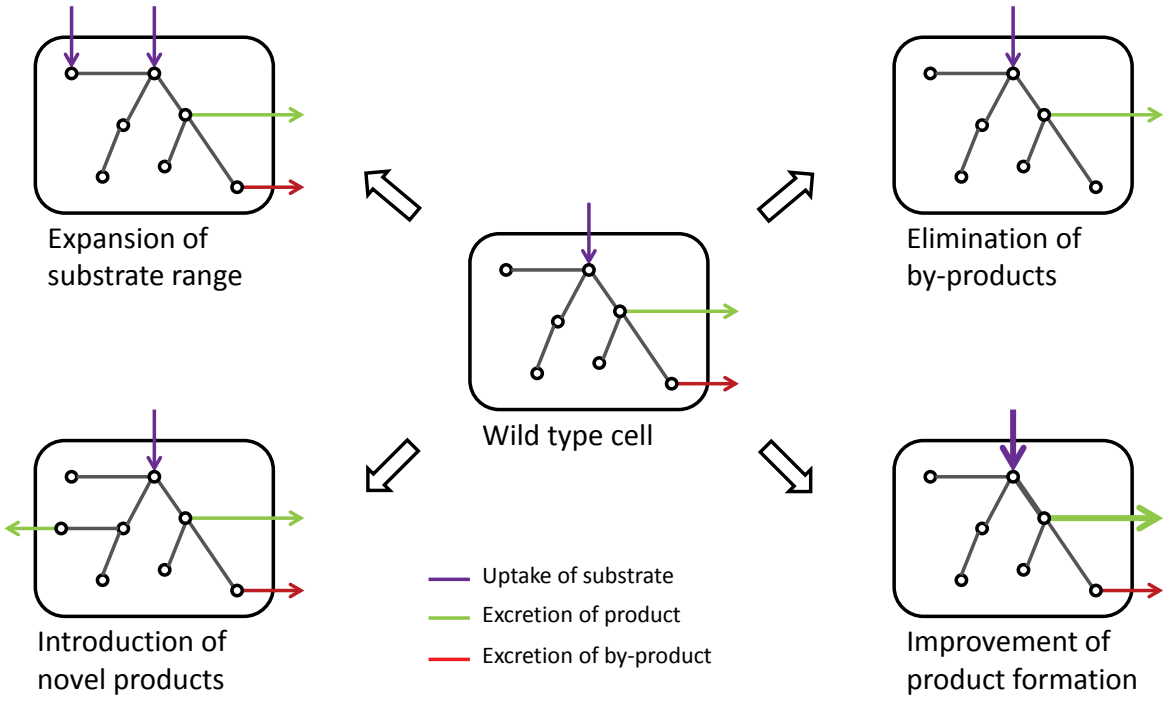


Figure 2-1. Examples of pathway modifications.

An early example of aromatic amino acid production using recombinant *Escherichia coli* can serve as a case study of the metabolic engineering workflow. *E. coli* and many other microorganisms synthesize aromatic amino acids through the condensation reaction between phosphoenolpyruvate (PEP) and erythrose 4-phosphate to form 3-deoxy-D-arabinoheptulosonate 7-phosphate (DAHP). Figure 2-2a shows the synthesis pathway of DHAP from glucose in *E. coli*. The first attempt to increase the production was made by screening of mutants which had deregulated product inhibition (Aiba *et al.*, 1980). This was followed by overexpression of the enzymes DHAP synthase (Forberg and Haggstrom, 1987) and then also transketolase (Draths *et al.*, 1992). However, the yields were still low. Forberg *et al.* (1988) then used a small metabolic model in order to identify the optimal flux distribution for DHAP production. This distribution can be seen in Figure 2-2a. It was predicted that 3 units of DHAP could be produced from 7 units of glucose. They identified the PEP phosphotransferase system, responsible for uptake of glucose, to be a suitable target. This system results in the conversion of PEP to pyruvate, which leads to a limitation in PEP for DHAP production. They further simulated the effect of introducing PEP synthase, which would regenerate the consumed PEP. The results can be seen in Figure 2-2b. The predicted yield would then rise to the double; 6 units of DHAP per 7 units of glucose. Patnaik and Liao (1994) then performed the suggested modification. The observed an almost two-fold increase in DHAP formation, from 52% yield to 90% yield, in excellent agreement with the predictions. This short example shows how powerful mathematical modelling and genetic engineering can be when used together. The field has developed tremendously since these early years, and both the dry lab and wet lab methods are now much more complex. The underlying thinking, however, remains the same.

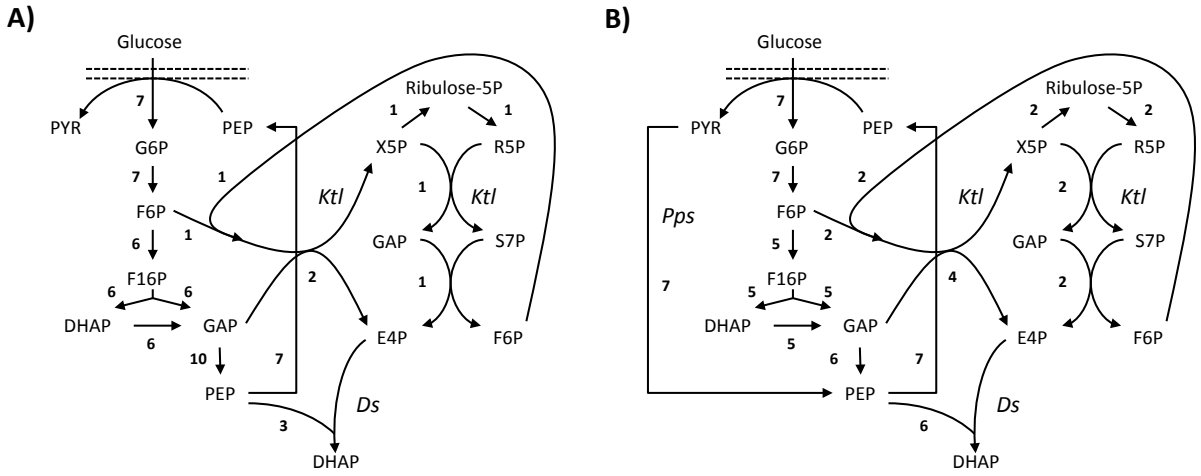


Figure 2-2. Metabolic engineering of *E. coli* for production of DHAP.A) Theoretical maximal yield of DHAP on glucose for the wild-type strain. **B)** Theoretical maximal yield of DHAP on glucose for a genetically modified strain with recombinant PEP synthase inserted. PYR: pyruvate, PEP: phosphoenolpyruvate, G6P: glucose 6-phosphate, F6P: fructose 6-phosphate, F16P: fructose 1,6-bisphosphate, DHAP: dihydroxyacetone phosphate, GAP: glyceraldehyde 3-phosphate, X5P: xylulose 5-phosphate, R5P: ribose 5-phosphate, S7P: sedoheptulose 7-phosphate, DAHP: 3-deoxy-D-arabinoheptulosonate 7-phosphate, Pps: PEP synthase, Tkt: transketolase, Ds: DAHP synthase. Adapted from Patnaik and Liao (1994).

2.2 Systems biology

The term systems biology has dual meanings. It can refer to an inter-disciplinary field which studies the interactions between components of complex biological systems. A central aspect in that interpretation is the concept of *emergent properties*. Emergence is the way complex patterns or behaviours arise from multiple relatively simple interactions. A classical example would be swarming. Each member of the swarm makes its decisions based on a few simple inputs, such as the proximity and speed relative to its neighbours, but the overall behaviour of a swarm can be awe-inspiring. Or in the words of physicist Doyne Farmer: "It's not magic...but it *feels* like magic" (Waldrop, 1992).

Systems biology claims to be particularly well posed to identify and study emergent properties, owing to its network-centric view. The underlying assumption is that biology itself has a strong link to emergence; due to the role evolution has in the growth of complexity in the natural world.

> "[In] evolutionary processes, causation is iterative; effects are also causes. And this is equally true of the synergistic effects produced by emergent systems. In other words, emergence itself... has been the underlying cause of the evolution of emergent phenomena in biological evolution; it is the synergies produced by organized systems that are the key." (Corning, 2012)

The second interpretation of the term systems biology is as a paradigm within biological sciences. It is then normally presented as the antithesis of the classical reductionist paradigm in the scientific method. The perceived limitations of the reductionist view has been described by Sauer *et al.* (2007) .

> "The reductionist approach has successfully identified most of the components and many interactions but, unfortunately, offers no convincing concepts and methods to comprehend how system properties emerge."

In practice, the systems biology approach, whether as a research field or as a scientific paradigm, often boils down to measuring multiple components simultaneously and then integrating the data with mathematical models. The field is therefore reliant on high-throughput measuring techniques such as metabolomics, transcriptomics and proteomics, as well as on methods from bioinformatics and computational biology.

As with most research fields there is no clear time point at which to put its birth. The ancestors of systems biology include the study of enzyme kinetics in the early 1900s (Michaelis *et al.*, 2011) and the application of control theory to biological systems in the 1960s and 1970s (Heinrich *et al.*, 1977). Denis Noble, who developed the first mathematical model of the working heart in 1960, is considered to be an early pioneer in the field. However, it was not until the 1990s, when the completion of the first genome projects resulted in large amounts of high quality data while at the same time computational power exploded, that the field really took off (Tomita *et al.*, 1997). Much of the work being done in systems biology is data driven rather than hypothesis driven, although this is by no means a requirement. Figure 2-3 illustrates the workflow commonly referred to as the systems biology cycle.

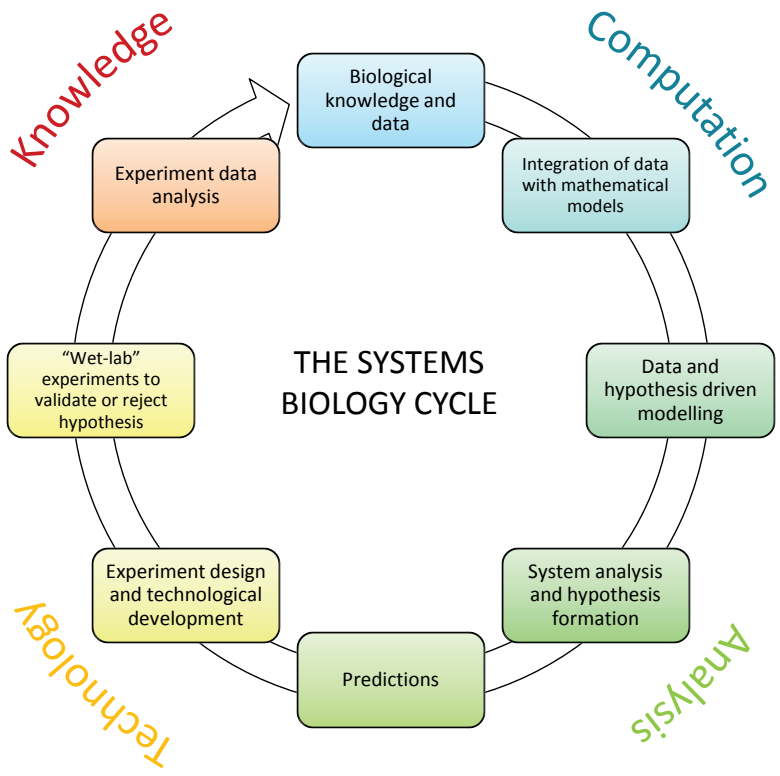


Figure 2-3. The systems biology cycle. Based on Kitano (2002a, b).

2.3 Constraint-based modelling

Mathematical modelling has been used to study metabolism for at least 100 years, since Michaelis and Menten derived their famous equation for enzyme kinetics in 1913. When the kinetic parameters for a large enough number of enzymes had been estimated, it was possible to formulate small models which could describe the basic metabolic functions of a living cell (Othmer, 1976). However, the data on kinetic parameters was fragmented, and the models

could be sensitive with respect to measurement errors (not to mention the problems associated with using *in vitro* measurements to estimate *in vivo* kinetics). There was a need to develop a mathematical framework to deal with uncertainty in data, and to quantify the control each of the enzymes had in the model. This led to the development of metabolic control analysis (MCA) (Heinrich *et al.*, 1977).

The framework was mainly applicable to small networks, and the availability of kinetic parameters continued to be limiting. If only the steady state metabolic fluxes inside the cell were of interest (rather than the dynamic change in metabolite pools) then those could be estimated in a method called metabolic flux analysis (MFA) (Aiba and Matsuoka, 1979). The method relies on measuring the rates of production/consumption of metabolites (called exchange fluxes) in the growth medium. If the set of possible enzymatic conversions is known, then the internal fluxes can be fitted from the exchange fluxes by linear regression. However, this requires that enough exchange fluxes are measured so that the resulting equation system is determined. Another issue was the determinability of fluxes in parts of the metabolism where there were cyclic or parallel reactions. By using isotope labelled substrates it was possible to track each atom, rather than each metabolite, through the metabolic network. This allowed for better resolution and more comprehensible models, but the fundamental limitations of the method remained (Wiechert, 2001). In a review paper entitled "Flux analysis of underdetermined metabolic networks: the quest for the missing constraints" Bonarius *et al.* (1997) describe how additional constraints, such as co-factor balancing or reaction reversibility, were incorporated in order to reduce the degrees of freedom and have a determined model. A large step forward was taken when the models were constrained to be optimal with respect to some cellular objective (Fell and Small, 1986). This formed the foundation for constraint-based modelling (CBM) of metabolism.

If the traditional approach to metabolic modelling is to describe the components of a model in such detail that the model correctly represents the phenotype, then the constraint-based approach is rather to impose increasingly detailed constraints on the solution space so that only relevant phenotypes are feasible. The term CBM is, at least when applied to metabolic modelling, largely synonymous to flux balance analysis (FBA), although FBA is a more narrow term. There are multiple excellent reviews describing the assumptions and mathematical formulation behind FBA (Varma and Palsson, 1994b; Edwards *et al.*, 2001; Price *et al.*, 2003). The following section will describe the methodology by using a small hypothetical metabolic network (see Figure 2-4a).

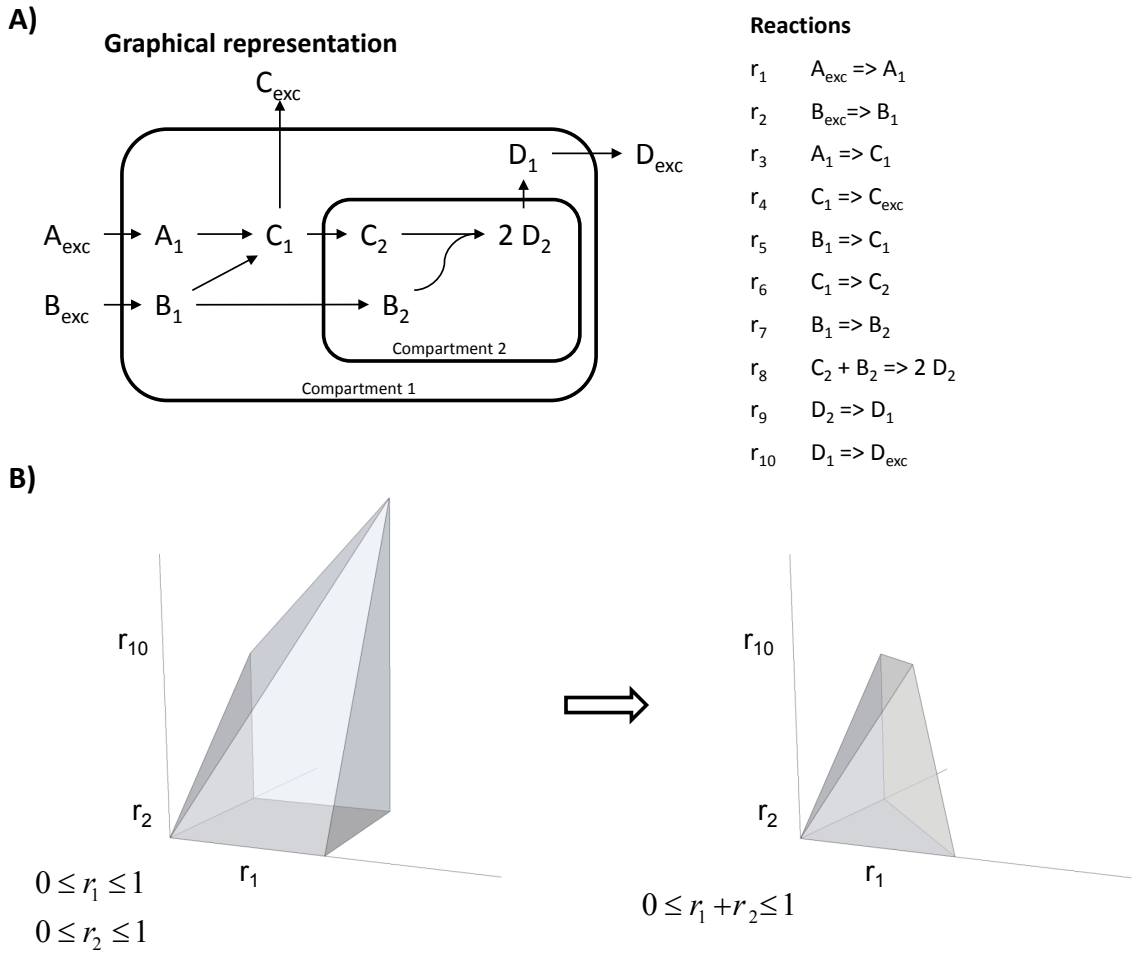


Figure 2-4. Principles of FBA.A) A small metabolic network. The network is comprised of 10 reactions (out of which 6 are internal), 12 metabolites (out of which 8 are internal) and it contains two compartments. The subscripts of the metabolites indicate which compartment they belong to. r_1, r_2, r_4 and r_{10} are exchange reactions. r_6, r_7 and r_9 are reactions which transport metabolites between compartments. Note that while, for example, C_1 and C_2 represent the same chemical compound they are regarded as unique metabolites for modelling purposes. Also note that the stoichiometries of the enzymatic conversions are expressed in the network (see r_8). **B)** The feasible solution space is shown for the reactions r_1, r_2, and r_{10}. All points in the blue cone represent feasible solutions given the constraints. As additional constraints are imposed the solution space becomes narrower.

A mass balance over a metabolite can be expressed in the general form:

$$Accumulation = Input - Output + Generation - Consumption - Dilution \quad (1)$$

Or in a more mathematical form as:

$$\frac{dx_i}{dt} = v_{in,i} - v_{out,i} + v_{generation,i} - v_{consumption,i} - \mu x_i \quad (2)$$

In (2) the rate of accumulation of metabolite x_i is defined as the rate by which it is taken up ($v_{in,i}$), minus the rate by which it is excreted ($v_{out,i}$), plus the rate by which it is generated ($v_{generation,i}$) and minus the rate by which it is consumed ($v_{consumption,i}$). The dilution term (μx_i) accounts for the decrease in concentration that comes from the fact that a cell expands as it grows. Because the intracellular concentrations of most metabolites are very low compared to the fluxes affecting them, the dilution term can generally be neglected (Stephanopoulos *et al.*, 1998). This gives:

$$\frac{dx_i}{dt} = v_{in,i} - v_{out,i} + v_{generation,i} - v_{consumption,i} \tag{3}$$

For metabolite D_2 in the figure above, for example, the equation would then read:

$$\frac{dD_2}{dt} = 2r_8 - r_9 \tag{4}$$

This relationship can be expressed in a matrix notation to represent the mass balances for all metabolites

$$\frac{d\mathbf{x}}{dt} = \mathbf{S} \cdot \mathbf{v} \tag{5}$$

In (5) **S** is a matrix which contains the stoichiometric coefficients that define the metabolic network. This matrix is referred to as the stoichiometric matrix. **v** is a vector with the rate for each reaction and **x** is a vector with the resulting changes in concentrations with respect to time for each of the internal metabolites. FBA is based on the assumption that the time scale for changes in the internal metabolite pools (typically seconds or minutes) is much faster than the time scale for growth or for changes in the environment (typically minutes or hours). It is therefore reasonable to assume that the internal metabolites are in steady state (meaning that their change in concentration is 0) (Varma and Palsson, 1994b). Equation (5) then simplifies to:

$$\mathbf{0} = \mathbf{S} \cdot \mathbf{v} \tag{6}$$

For the small network in Figure 2-4a this would look like:

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{matrix} A_1 \\ B_1 \\ C_1 \\ = C_2 \\ B_2 \\ D_2 \\ D_1 \end{matrix} \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \\ r_7 \\ r_8 \\ r_9 \\ r_{10} \end{bmatrix} \tag{7}$$

In MFA the objective would now have been to measure a sufficiently large number of fluxes to have a determined model. With 10 variables (the unknown fluxes) and 7 equations (mass balances around the internal metabolites) the system has 10-7=3 degrees of freedom (if all reactions were linearly independent). 3 fluxes would therefore have to be measured. In FBA the objective is instead to constrain the system to narrow the set of feasible flux distributions. One fundamental constraint is imposed by the thermodynamics (e.g. effective reversibility or irreversibility of reactions). In the example network all reactions are irreversible and it therefore holds that:

$$\mathbf{v} \geq 0 \tag{8}$$

For FBA to be effective three criteria have to be met: 1) the metabolic network should correctly describe the metabolic capabilities of the organism being studied, 2) the constraints should correctly describe the physiological limitations that the system operates under, 3) the objective function should correctly describe the objective which the cell strives to achieve. The first point is discussed in detail in section 3.3. The second and third points are discussed below.

Figure 2-4b shows the effect of imposing additional constraints on the model. In the left panel the uptake rates of metabolites A_1 and B_1 are constrained to be ≤ 1. This defines a feasible cone of solutions, here shown in blue. Note that not all combinations of values for r_1, r_2 and r_{10} are allowed, since their relationships are defined by the stoichiometry of the reactions. Constraints on uptake or excretion rates are the most widely used type and they are commonly based on experimentally measured fluxes. In the right panel an additional constraint has been imposed; that the sum of r_1 and r_2 should be ≤ 1. This cuts the cone and further reduces the set of allowed flux distributions. There have been many attempts to define constraints that are biologically relevant and which do not require expensive and difficult in vivo measurements of enzymatic capabilities. Examples include: physical constraints such as diffusion rates; a general upper limit on enzyme capability and molecular crowding constraints (Beg *et al.*, 2007); binary regulatory constraints on which enzymes can be active under a given condition (Covert *et al.*, 2001); energy balancing to exclude thermodynamically infeasible solutions (Beard *et al.*, 2002); thermodynamic constraints based on the standard Gibbs free energies of formation (ΔG_f^0) for metabolites (Henry *et al.*, 2006).

How does a cell adjust its intracellular fluxes given the constraints that it is under? In FBA it is assumed that cell metabolism functions according to some objective, and that such an objective can be defined as a linear combination of the reaction rates.

$$\begin{aligned} &\textit{Maximize} && \mathbf{c}^{\mathbf{T}}\mathbf{v} \\ &\textit{Subject to} && \mathbf{0} = \mathbf{S}\cdot\mathbf{v} \\ & && \textbf{lower bounds} \leq \mathbf{v} \leq \textbf{upper bounds} \end{aligned} \tag{9}$$

In (9) $\mathbf{c}$ is a vector with coefficients for each of the reactions. The expression $\mathbf{c}^T\mathbf{v}$ then becomes the product of the flux and the objective coefficient, summed over all reactions. The system defined in (9) can be efficiently solved, also for very large problems, by using linear programming (Karp, 2008). There have been many studies on what constitutes a good objective function. Some of the first suggestions were rather basic, such as to maximize the NADPH production or minimize the ATP production (Bonarius *et al.*, 1997). When the molecular composition of biomass could be quantified in sufficient detail it was possible to use maximization of growth as an objective (Varma and Palsson, 1994a). This objective proved to be a very good approximation, and still remains by far the most commonly used objective for modelling of microbial cells. More complex objectives, such as maximization of entropy production (Henry *et al.*, 2006) or combinations of several of the objectives mentioned here (Schuetz *et al.*, 2012) have been proposed since then.

The idea that a complex system such as a living cell can be modelled from a small set of physiological constraints and some general objective remains very appealing. It is therefore likely that the quest for ever more detailed constraints and more predictive objective functions will continue also in the future.

3 Applications of genome-scale metabolic models

In the previous chapter, a small metabolic model with 10 reactions was used to show the principles behind FBA and CBM. However, the models used in practice are anything but small; rather, they contain thousands of reactions and metabolites. Ever since genome sequencing took off in the 1990s is has, at least in theory, been possible to identify each enzyme that exists in an organism, and thereby infer a metabolic network which describes the full metabolic capabilities of the organism (Schilling *et al.*, 1999). These models have therefore come to be known as genome-scale metabolic models (GEMs).

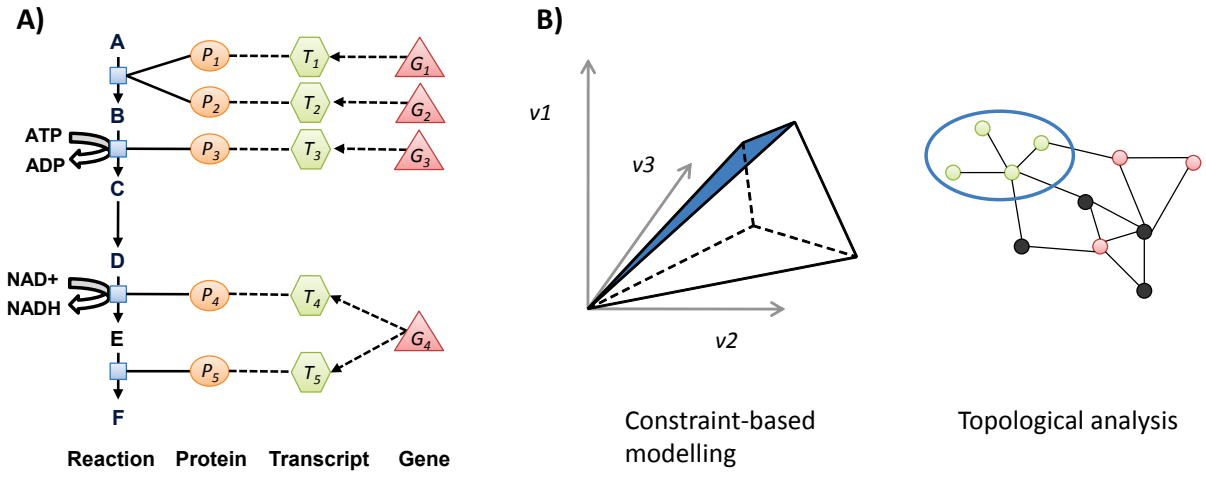


Figure 3-1. Introduction to GEMs. A) The layout of a genome-scale metabolic model (GEM). A GEM can be viewed as a highly structured map of how metabolism is controlled at different levels. At the bottom are the metabolic reactions and the metabolites which they involve. Each reaction can then be catalysed by zero or more enzymes. The enzymes are further linked to the corresponding transcripts, which in turn are linked to the corresponding genes. **B)** The two main applications of GEMs. In CBM the GEM is viewed as an equation system which describes how metabolism in a cell operates. In topological analysis the GEM is viewed as a map of how components in a cell interact with each other.

Figure 3-1a describes the general layout of GEMs. GEMs can be viewed as detailed maps of connections between the different levels of metabolism. Thereby they can provide a mechanistic description all the way from the metabolites, via reactions, enzymes and transcripts, up to the genes. Model elements at all levels can be extensively annotated so that GEMs can serve as highly structured databases. Figure 3-1b shows the two main application categories for GEMs. The following two sections will discuss how GEMs can be used for CBM (section 3.1) and for topological analysis/data integration (section 3.2). Section 3.3 deals with the reconstruction of GEMs; a very complex and time-consuming process.

3.1 Constraint-based modelling using GEMs

More than 100 algorithms for constraint-based modelling using GEMs have been published, as described in detail in an excellent review by Lewis *et al.* (2012). An in depth description of these algorithms would be outside the scope of this thesis, but has been extensively covered elsewhere (Price *et al.*, 2004; Durot *et al.*, 2009). Instead, this section will describe one algorithm each for a number of optimization frameworks, in order to illustrate the breadth of the available methods. The categories will be: 1) linear programming (LP), 2) quadratic

programming (QP), 3) mixed-integer linear programming (MILP), 4) bi-level programming (BLP), and 5) heuristic methods.

Linear programming. Linear programming represents the most fundamental of the optimization frameworks; so much that to many people CBM is synonymous to linear programming. The most widely used LP application is flux balance analysis (FBA) (see section 2.3). The approach relies on defining a linear objective function to be optimized, and then finding one solution (among the many) that is optimal with respect to the objective. An important advantage compared to the frameworks below is that LP problems can be solved to optimality very efficiently, even for large models. The objective function used for microorganisms is normally the maximization of the specific growth rate, which is consistent with the evolutionary advantage of the fastest growing species (Edwards et al., 2001). When used for strain design in metabolic engineering, the general approach is to iteratively remove enzymes from the GEM and then observe if the model produces the compound of interest as an effect of maximization of the growth rate. An example of this approach is in Lee *et al.* (2005), where the authors used FBA to suggest gene knockouts in *E. coli* with the purpose to overproduce succinic acid. The suggested modification involved a triple deletion to reduce the flux from PEP to pyruvate. When validated experimentally it resulted in a sevenfold increase in production of succinic acid.

Quadratic programming. Quadratic programming is similar to LP, but with the possibility of having quadratic terms in the objective function. This allows for minimization or maximization of the difference between fluxes, i.e. *minimize* $(v_i - v_j)^2$. The most widely used application of this optimization framework is Minimization of metabolic adjustment (MOMA) (Segre *et al.*, 2002). The underlying assumption in MOMA is that following a perturbation, such as deletion of a gene, the cell strives to minimize the distance from its flux distribution to the flux distribution of the non-perturbed cell. In a fascinating study, Wintermute and Silver (2010) used MOMA to study synthetic mutualism in auxotrophic *E. coli* mutants, and how they can complement one another's growth by cross-feeding of essential metabolites.

Mixed-integer linear programming. Mixed-integer linear programming (MILP) is based on LP, with the additional feature that variables can be constrained to only take integer values. MILP has found extensive use in algorithms for model reconstruction and gap filling (see section 3.3), since it is possible to formulate problems where a variable takes the value 1 if a reaction is included in a model and a value 0 if it is excluded. An algorithm which uses MILP for strain design is Regulatory on/off minimization (ROOM) (Shlomi *et al.*, 2005). The underlying assumption is similar to MOMA, but rather than minimization of the distance between the flux distributions, the cell is assumed to strive to minimize the difference in which reactions are active/passive. ROOM has been shown to give slightly better predictions when compared to MOMA, but at the cost of being significantly more computationally intensive (Shlomi *et al.*, 2005).

Bi-level programming. Bi-level programming represents optimization problems where one problem is embedded in another one. In this context it normally means optimization of some objective while the model is constrained to be optimal with respect to some other cellular objective. The first implementation of this optimization framework, OptKnock, has proven to be a very powerful tool for strain design (Burgard *et al.*, 2003). In OptKnock, the objective function is maximization of production of some relevant compound. In order to achieve this, the algorithm removes reactions so that production is stoichiometrically linked to optimal growth. For a small number of gene deletions this could be iteratively tested for in a brute-force approach, as described in the linear programming section. The strength of OptKnock is

that also combinations of relatively large numbers of gene deletions can be evaluated. At its foundation OptKnock is implemented as a MILP problem. OptKnock was used by Fong *et al.* (2005) in a study on lactate overproduction in *E. coli*. The algorithm suggested a triple deletion strategy resulting in 1) disabled ethanol and acetate production, 2) increased production of pyruvate and NADH, which are the precursors for lactic acid, 3) coupled uptake of glucose to the conversion of PEP to pyruvate. This deletion strategy, after a round of adaptive evolution, resulted in lactate titres of 0.87 to 1.75 g/L when the cells were grown in 2 g/L glucose.

Heuristic methods. Heuristic optimization methods are used for quickly finding an approximate solution; trading away optimality, completeness, accuracy, and/or precision. These methods also have the strength that they can make use of more general objective functions, not only the linear or quadratic forms described above. One example of a heuristic algorithm for strain design is OptGene (Patil *et al.*, 2005). It is based on randomly introducing perturbations to a population of GEMs, and then letting them compete and mate with each other based on their fitness. This approach is called evolutionary programming. OptGene has been applied to suggest metabolic engineering strategies for sesquiterpene production in *Saccharomyces cerevisiae* (Asadollahi *et al.*, 2009).

3.2 Data integration using GEMs

A number of algorithms have been developed for the purpose of using GEMs as scaffolds for data integration and interpretation. An in depth description of these algorithms would be outside the scope of this thesis, but has been extensively covered elsewhere (Joyce and Palsson, 2006; Durot *et al.*, 2009). Instead, this section will describe one or two algorithms each for three important types of omics data in order to illustrate the concept.

Fluxomics. Measured intracellular fluxes, for example using ^{13}C labelled substrates, represent a data type that is directly applicable to integration with GEMs. In order to go from measured labelling patterns in metabolites to fluxes, an atom mapping model is used. These models are similar to GEMs in that they are stoichiometric models of metabolism, but they can track each atom though the network, rather than each metabolite. They are traditionally rather small models, and only built for central carbon metabolism. This causes some problems, for example with co-factor balancing. GEMs have therefore been used to expand and complement atom mapping models, in order to also take more peripheral metabolism into account (Suthers *et al.*, 2007). As discussed in section 2.3, a driving force in CBM is the hunt for ever more precise objective functions. Burgard and Maranas (2003) developed an algorithm, ObjFind, which makes use of fluxomics data to try to infer the cellular objectives that could have given rise to the phenotype. They found that regardless of the growth condition, maximization of growth was the objective that best fitted the data. This is good news for FBA; since the methods is based on that there are simple objectives which hold for a wide range of conditions.

Metabolomics. Large-scale quantification of internal metabolites has been made possible thanks to developments in mass spectrometry and NMR technology. Since GEMs are based on the assumption that metabolite pools are in quasi-steady state, the concentrations of metabolites are not immediately possible to integrate into GEMs. Instead, the main use of metabolomics data has been to evaluate the capabilities of the GEM (can the model produce the detected metabolites?). This can then lead to directed search for the missing functions,

thereby generating new biological knowledge as well as improving on the model (Oh *et al.*, 2007). Metabolomics has also been used together with estimated Gibbs free energies of formation for metabolites in order to predict reaction directionality (Kummel *et al.*, 2006a).

Transcriptomics. As discussed in section 3.3.2, there is not a good correlation between transcript level and flux, owing to the several layers of regulation between them (Akesson *et al.*, 2004). It is therefore difficult, or impossible, to directly use the expression levels to modify the model constraints, although several attempts have been made. The most common use of transcriptomics data is therefore to classify genes in a binary fashion; either as expressed or non-expressed (see the part about GIMME in section 3.3.2). An alternative approach, developed by Patil and Nielsen (2005), uses the data in a different way. Rather than looking at the expression level, they look at the significance of differential expression between two conditions, and make use of the network topology to analyse the data. The method works by first converting the metabolic network into a bipartite graph. In a bipartite graph the metabolites are connected to genes based on the reactions in which they participate. A meta-analysis is then performed for each metabolite by testing if the genes that it is associated to are differentially expressed when taken as a group (known as gene set analysis). If so, then the metabolite is classified as a Reporter. Reporter metabolites can be said to represent "hot spots" in metabolism around which transcriptional changes occur. Reporter subnetworks, presented in the same paper, is an algorithm with a somewhat similar mind-set. Both algorithms are described in Figure 3-2. The most well-known use of metabolic network topology is elementary flux modes (EFMs) (Schuster *et al.*, 1999) and its cousin extreme pathways (EPs) (Schilling and Palsson, 2000). These are minimal sets of reactions which can operate in steady state in a metabolic network. A critical drawback is that the enumeration of EFMs or EPs is very computationally demanding, and the method is therefore only applicable for medium-sized networks. In an approach conceptually similar to Reporter subnetworks, Schwartz *et al.* (2007) used EFMs to aid in interpretation of transcriptomics data. Rather than calculating a p-value for the set of enzymes in a subnetwork, they calculated it for the set of enzymes in each EFM. The approach was then applied to study stress responses in *S. cerevisiae*.

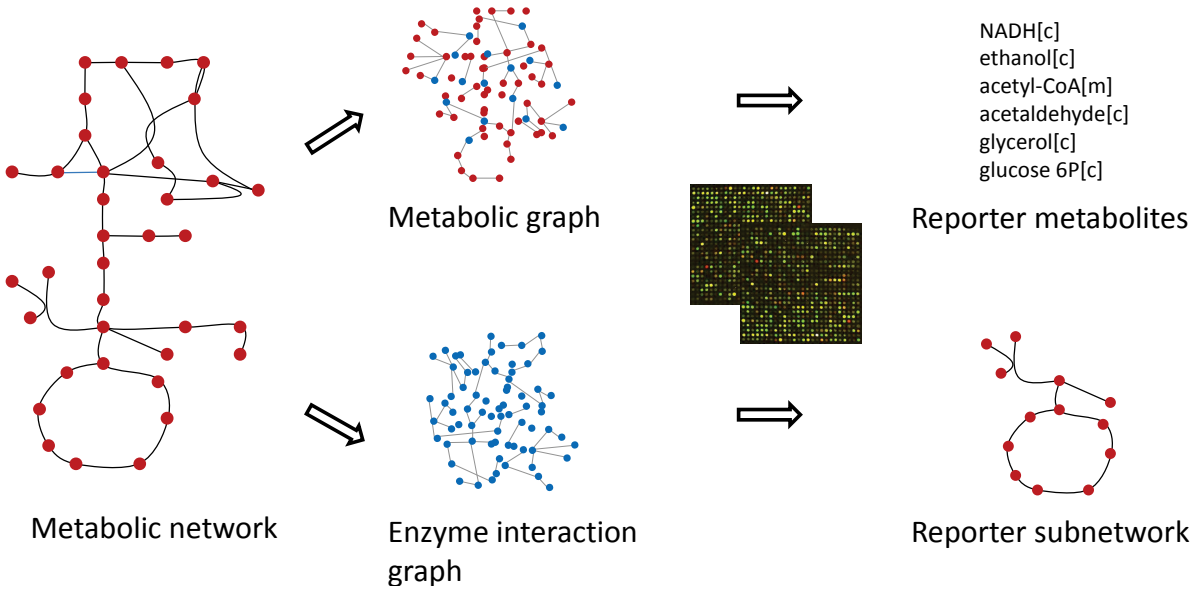


Figure 3-2. The Reporter metabolites / Reporter subnetworks algorithms. For Reporter metabolites, the metabolic network is converted into a bipartite graph, where each metabolite is connected to the genes for the reactions it participates in. A metabolite can then be scored based on the p-values for differential expression for the genes connected to it. If there is an overall significant change then the metabolite is a Reporter. For Reporter subnetworks, the metabolic network is converted into a unipartite enzyme interaction graph. A simulated annealing algorithm is then applied in order to find sets of connected enzymes which exhibit an overall significant change in expression. The metabolic network involving those enzymes can then be reconstructed from the original metabolic network. These are called Reporter subnetworks. Adapted from Patil and Nielsen (2005).

3.3 Reconstruction of GEMs

The reconstruction process of GEMs is traditionally very labour- and time-intensive, spanning from several months for a well-studied bacteria to several years for a human model (Duarte *et al.*, 2007). The very aspects that make GEMs so powerful, their scope and multi-level structure, are also what makes the reconstruction process so complex. This section gives an overview of the traditional approach, where models are manually reconstructed from genomic and bibliomic data in a bottom-up manner. The two subsections describe top-down approaches for microbial and cell type-specific models, respectively.

There has been a multitude of published descriptions of the reconstruction process, but one has had a particularly large impact on the field. In a review in Nature Protocols, Thiele and Palsson (2010) collected existing reconstruction practices and summarized them in a 96 step standard operating procedure (SOP). Figure 3-3 depicts the most important steps of the reconstruction process.

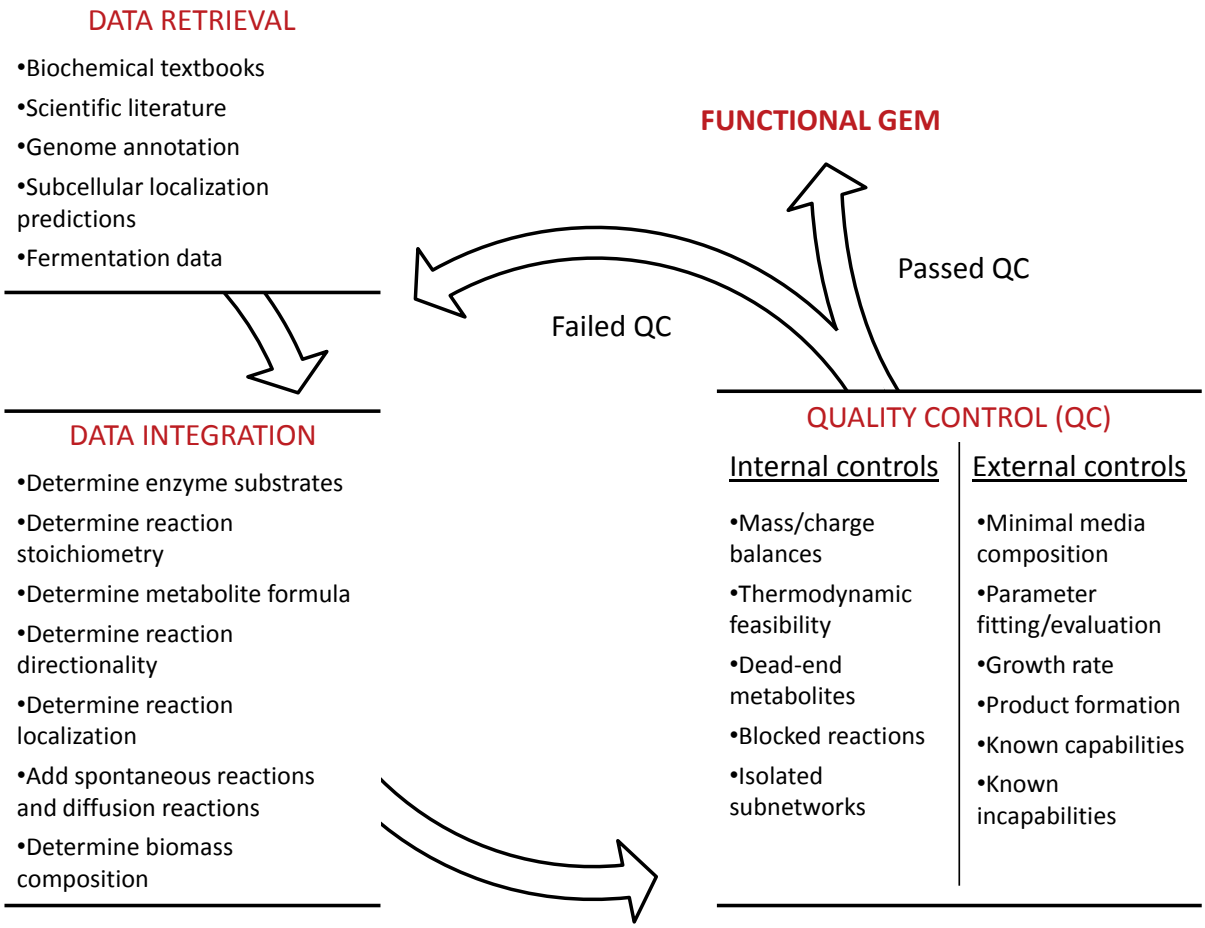


Figure 3-3. The GEM reconstruction process.

The level of detail required for reconstruction of a GEM is in a sense rather low, only the reaction stoichiometries and directionalities and the associated enzymes catalysing each reaction. This can be contrasted to the large number of kinetic parameters required for reconstruction of dynamic models. The reconstruction starts with the retrieval and organizing of the necessary input data. Depending on the organism, the availability and amount of data differs. The minimal required input can be said to be a sequenced genome and some amount of known physiological data, such as growth conditions. In general, the better the availability of physiological, biochemical and genetical data, the better the predictive ability of the model. Some data types which can be very valuable if they are available include: ^{13}C fluxomics data, measured subcellular localizations, or gene knockout libraries. The next step is the generation of a draft model. That starts by going through each of the annotated genes and deciding whether its function is within the scope of the reconstruction. Normally, only metabolic enzymes are included. There are also grey areas, such as DNA methylation, protein phosphorylation or complex glycan metabolism, which are often excluded from the GEM, even though they can be viewed as metabolic functions. The enzymes are then mapped to their respective metabolic reactions. This step is often performed via EC numbers, followed by retrieval of reactions from reaction databases such as KEGG (Ogata *et al.*, 1999) or RHEA (Alcantara *et al.*, 2012). It is important to note that EC numbers refer to the type of chemical transformation in a reaction, not the enzyme class which performs the conversion. For example, both ethanol dehydrogenase and choline dehydrogenase have EC number 1.1.1.1, since they both act on primary alcohols and use NAD^+ as a co-factor. This mapping is therefore not exact, and should be viewed as a first draft. The process described above is for

an annotated genome. If the genome for the organism of interest is not annotated, a first step would then be to use one of the many genome annotation pipelines which have been developed (Stein, 2001).

The network at this stage will most likely have a large number of issues which must be addressed. Potential issues include, but are not limited to: unclear metabolite naming, generic metabolites such as “an alcohol” or “fatty acid”, missing spontaneous conversions, wrongly assigned directionality of reactions, or generic stoichiometries such as “starch(n) + H2O => starch(n-1) + glucose”. Careful manual evaluation and modification of the reactions is used to clear up the reaction list. The reactions are then partitioned into the relevant subcellular compartments, and transport reactions are added based on literature and genomic evidence. Lastly, the chemical composition of biomass is determined based on measurements or literature.

A number of quality controls are then performed in order to validate the model. In Figure 3-3 these controls are categorized as internal or external, where internal relates to the inner workings of the network and outer to the capabilities for predicting the cellular phenotype. As GEMs are fundamentally mass balance models, it is critical that the reactions are elementally and charge balanced. This is trivial if the elemental composition is known, but that is not always the case, for example in the case of polymers. The model can then be tested for “stoichiometric inconsistencies”, meaning that there are reaction sets such as A → B and A → B + C (Gevorgyan et al., 2008). Other tests include thermodynamic and redox feasibility, so that the model cannot produce high energy compounds from low energy compounds or reduced compounds from oxidized compounds (Kummel *et al.*, 2006b). There are a number of methods for identifying reactions which cannot carry flux or metabolites which cannot be produced, and to suggest strategies for connecting them (Mahadevan and Schilling, 2003; Reed *et al.*, 2006; Satish Kumar *et al.*, 2007; Kumar and Maranas, 2009; Brooks *et al.*, 2012). This process is referred to as gap filling, and it is of central importance for the quality of the reconstruction.

The metabolic capabilities of the model must then be evaluated by comparison to the known capabilities of the organism in question. This is referred to as external controls in Figure 3-3. Such controls include that the model can grow on media that the organism in question can grow on, that it can produce products that are known to be produced by the organism, and that it can perform other known metabolic functions of the organism. It is equally important to control that the model is not too flexible, so that it can perform functions that are known to not occur in the organism, or that it can grow faster than what is seen experimentally. Lastly, there are a small number of parameters which need to be fitted from experimental data. These include ATP maintenance costs and the P/O ratio. These parameters have to be validated to be within reasonable bounds. If all quality controls pass, then the model is functional and the reconstruction complete. If any of them fails, then the metabolic network has to be further modified. This is an iterative process, where the model is modified, validated, and modified again until a functional and high-quality model is achieved. This interplay between annotation, verification, and testing is a valuable process, as it results in the refinement of both the genome annotation and the reaction network. Table 3-1 lists some databases and resources that are widely used for reconstruction of GEMs.

Table 3-1. Databases and resources for reconstruction of GEMs.

	EMBL	GenBank	SEED	BRENDA	ENZYME	UniProt	TransportDB	PSORTdb	CheBI	Pub chem	HMDB	LipidMaps	Reactome	KEGG	BioCyc	UniPathway	PubMed	RHEA	BIGG	BioModels
Biochemical activities	X	X	X	X		X	X						X	X	X	X	X	X		
Enzyme specificity				X	X	X	X										X	X		
Subcellular localization				X			X	X					X		X		X			
Reaction equation				X	X		X						X	X	X	X	X	X		
Reaction direction				X			X						X	X	X		X	X		
Metabolite annotation				X					X	X	X	X		X	X		X	X		
GPR association	X	X	X			X	X						X	X	X	X	X	X		
GEM repositories																			X	X

Based on Durot *et al.* (2009) and Mardinoglu and Nielsen (2012).

3.3.1 Automated reconstruction of microbial GEMs

There have been a number of methods published for automating parts of the reconstruction process. Most or all of these methods aim primarily at automating the annotation step, but there are also methods that integrate parts of the quality control process as well. These methods make use of reaction databases and the connections between EC numbers and genes within such databases. One of the earliest such methods was Pathway Tools; a software for generating organism-specific databases from a general database (Karp *et al.*, 2002). The input to the software is a set of genes annotated with EC numbers. Pathway Tools was not developed with GEM reconstruction in mind, but rather as a more general resource. Other methods which also matched enzymes to reactions, but without the ambition to reconstruct GEMs, were IDENTICS (Sun and Zeng, 2004), which attempted to simultaneously annotate predicted ORFs and link them to reactions by using BLAST to match known metabolic genes to a non-annotated genome, and metaSHARK (Pinney *et al.*, 2005), which used PSI-BLAST profiles rather than BLAST for the same purpose. The first software dedicated for GEM reconstruction was GEM System (Arakawa *et al.*, 2006). GEM System first identified ORFs by using GLIMMER (Delcher *et al.*, 1999), then matched the ORFs to known enzymes using BLAST. A metabolic network was then generated based on mapping to an internal database. GEM System also contained a simple algorithm for gap filling. AUTOGRAPH (Derrien *et al.*, 2007) is a software for inferring GEMs based on previously reconstructed GEMs for other species. The software that has had the highest impact by far is the Model SEED resource (DeJongh *et al.*, 2007; Henry *et al.*, 2010). Model SEED builds on the gene calling and annotation pipeline in SEED, and then uses an internal reaction database to map annotated genes to reactions. It also contains an automatic gap filling feature, where a generic biomass equation is assumed, after which the software applies a gap filling algorithm in order to ensure that the model can form biomass.

The main advantage of using software like the ones described above is that it speeds up the reconstruction process. However, it is important to note that it comes at the cost of decreased control and insight over how and why different elements are included in the model.

3.3.2 Reconstruction of cell type-specific GEMs

The cells of multicellular organisms can have very different phenotypes even though they share the same genotype. For example, the longest neural cells in a human can be more than a meter long (Fletcher and Theriot, 2004), while one of the smallest cell types, the erythrocytes, only measure 7-8 μm (Fabry *et al.*, 1981). In order to model cellular metabolism, it is therefore necessary to reconstruct a GEM specifically for the relevant cell type. In practise, this is done by starting from a generic network for the organism in question, and then manually or algorithmically select a subset of enzymes which are thought to be present in the specific cell type. In 2007 two such generic GEMs were published for human: Recon 1 (Duarte *et al.*, 2007) and EHMN (Ma *et al.*, 2007). A number of cell type-specific models have been manually reconstructed by using these models as scaffolds, including for liver (Gille *et al.*, 2010), brain (Lewis *et al.*, 2010), alveolar macrophage (Bordbar *et al.*, 2010), and a multi-tissue model for hepatocytes, adipocytes and myocytes (Bordbar *et al.*, 2011). The manual reconstruction process follows the same workflow as described in Figure 3-3. These examples are either rather small GEMs or for well-studied cell types, where there is a wealth of physiological literature available.

In parallel to this there have also been a number of algorithms developed which aim at reconstructing cell type-specific GEMs in an automated manner based on high-throughput data. Note that the problem of inferring a cell type-specific network from a generic model is closely related to the problem of inferring an organism-specific network from a generic reaction database (as described in section 3.3.1). The difference is in the input data. While the organism-specific models are reconstructed based on protein homology, the cell type-specific models have to be reconstructed from omics data. Table 3-2 lists some relevant omics types and their respective advantages and disadvantages. The first of these algorithms was GIMME (Becker and Palsson, 2008). GIMME takes transcriptome data as input and removes reactions for which the expression levels for the genes are below some threshold. It then constrains the model to perform some function, after which is uses a gap filling algorithm to reinsert the required reactions so that the model can satisfy the constraints. The state of the art algorithm, MBA (Jerby *et al.*, 2010), adds another layer of complexity by dividing the reactions with supporting evidence into two groups; one with reactions which must be included, and one with reactions that should be included. The algorithm then uses a gap filling algorithm to include as many reactions as possible from the second group while using as few reactions as possible for gap filling. These algorithms have been applied to reconstruct GEMs for liver (Jerby *et al.*, 2010), kidney (Chang *et al.*, 2010) and a generic cancer model (Folger *et al.*, 2011).

Table 3-2. Available omics types for reconstruction of cell type-specific GEMs.

	Type	Advantages	Disadvantages
Transcriptomics from DNA microarrays	Relative	Cheap, widely available, high throughput	Low correlation between gene expression and protein level
Transcriptomics from RNA-Seq	Semi-quantitative	As above, but with the added benefit that the measurements are semi-quantitative	As above
Proteomics	Semi-quantitative	Direct evidence for the presence/absence of enzymes	Expensive, not all proteins/cell types are currently covered
Metabolomics	Quantitative	Detection of a metabolite indicates that the cell must possess the metabolic capabilities to synthesize it (or be able to import it from its surroundings)	Detection of a metabolite says nothing about which pathways it was synthesized in, or about the fluxes involving it
Fluxomics	Quantitative	Direct evidence for intracellular metabolic fluxes	A metabolic model has to be used for fitting the fluxes, only available for central carbon metabolism
Bibliomics	Categorical	Can be very reliable if based on high-quality experimental data	Time- and labour-intensive to retrieve and organize the data

4 Results and discussion

In the following sections I will summarize the publications underlying the thesis and discuss their contribution to the field. The publications could broadly be divided to deal with GEMs applied to metabolic engineering of fungi (section 4.1) and GEMs applied to human health and disease (section 4.2), although there are plenty of links between them. Each section is preceded by a short review of research previously carried out in the field.

4.1 GEMs applied to metabolic engineering of fungi

Fungi have been used by humans since ancient time for production of cheese, bread, beer, wine and soy sauce. Today they are used in many industrial processes, such as the production of enzymes, vitamins, polysaccharides, alcohols, pigments, lipids, and glycolipids. Fungal secondary metabolites, in particular antibiotics, are extremely important to our health and nutrition and have tremendous economic impact (Adrio and Demain, 2003). The industrial production of β-lactam antibiotics by the mold *Penicillium chrysogenum* is one of the success stories of biotechnology, and represents one of the largest biotechnological products in terms of value, with dosage form sales of about USD 15 billion (Elander, 2003). Table 4-1 lists some important industrial applications of fungi and the species used.

Much work has gone into metabolic engineering of fungi, partly owing to their large industrial relevance and partly because several fungi are important model organisms. In terms of genome-scale metabolic modelling most of the efforts have been centred on the yeast *S. cerevisiae*. To date there are ten published *S. cerevisiae* GEMs with different scopes and applications (Osterlund *et al.*, 2012). GEMs have also been developed for the industrially relevant yeasts *Pichia pastoris* (Sohn *et al.*, 2010), *P. stipitis* (recently renamed to *Scheffersomyces stipitis*) (Caspeta *et al.*, 2012), *Candida glabrata* (Xu *et al.*, 2013), and *Yarrowia lipolytica* (Loira *et al.*, 2012) as well as for the model yeast *Schizosaccharomyces pombe* (Sohn *et al.*, 2012). Three models have been reconstructed for filamentous fungi in the *Aspergillus* genus; namely *A. niger* (Andersen *et al.*, 2008), *A. oryzae* (Vongsangnak *et al.*, 2008) and *A. nidulans* (David *et al.*, 2008).

Table 4-1. Some industrial applications of fungi.

Process	Organism
Pre-modern products	
Ang-kak	*Monascus purpurea*
Miso	*Aspergillus oryzae*
Ontjam	*Neurospora crassa*
Soy sauce	*A. oryzae, A. sojae*
Tempeh	*Rhizopus niveus*
Brewing and baking	*Saccharomyces cereviseae, S. carlbergensis*
Mold-ripened cheeses	*Penicillium roqueforti, P. camembetii*
Pharmaceuticals	
Penicillins	*P. chrysogenum*
Cephalosporin	*Cephalosporium acremonium*
Cyclosporin	*Tolypocladium inflatum*
Ergot alkaloids	*Claviceps purpurea*
Griseofulvin	*P. griseofulvin*
Mevalonin	*A. terreus*
Statins	*P. brevicompactum, P. citrinum, M. ruber, A. terreus*
Taxol	*Taxomyces andreanae*
Proteins	
α -Amylases	*A. niger, A. oryzae*
Cellulase	*Humicola insolens, P. funiculosum, Trichoderma viride*
Glucoamylases	*A. phoenicis, R. delemar, R. niveus*
Glucose oxidase	*A. niger*
Invertase	*A. niger, A. oryzae*
Laccase	*Coriolus versicolor*
Pectinase	*A. niger, A. oryzae, H. insolens*
Proteinases	*A. oryzae, A. melleus*
Recombinant enzymes	*M. miehei, M. Pusillus, Pichia pastoris, S. cerevisiae*
Organic acids	
Citric acid	*A. niger*
Itaconic acid	*A. terreus*
Solvents and fuels	
Ethanol	*S. cerevisiae*

Based on Bennett (1998); Adrio and Demain (2003); Choi *et al.* (2003).

In an excellent review Osterlund *et al.* (2012) summarized the published applications of *S. cerevisiae* GEMs since the first model was made available in 2003. They classified its applications in four categories: 1) guidance for metabolic engineering and strain improvement, 2) biological interpretation and discovery, 3) applications of novel computational frameworks, and 4) evolutionary elucidation. Figure 4-1 builds on their

classification, but also includes non-*Saccharomyces* yeasts and filamentous fungi, as well as publications from after 2010. As can be seen, the first phase was completely dominated by work on *S. cerevisiae*. Around 2007 there seems to have been an increased interest in filamentous fungi and during the last couple of years there have been a number of studies on non-*Saccharomyces* yeasts. Another trend is that up until about 2009 many of the papers are about novel computational frameworks (category 3) and about evolutionary elucidation/comparative genomics (category 4). It is not until more recently that GEMs are widely used for strain engineering (category 1). This can indicate that the field has matured, and that the mathematical methods developed in the early years are starting to prove themselves and are now being used to solve concrete problems.

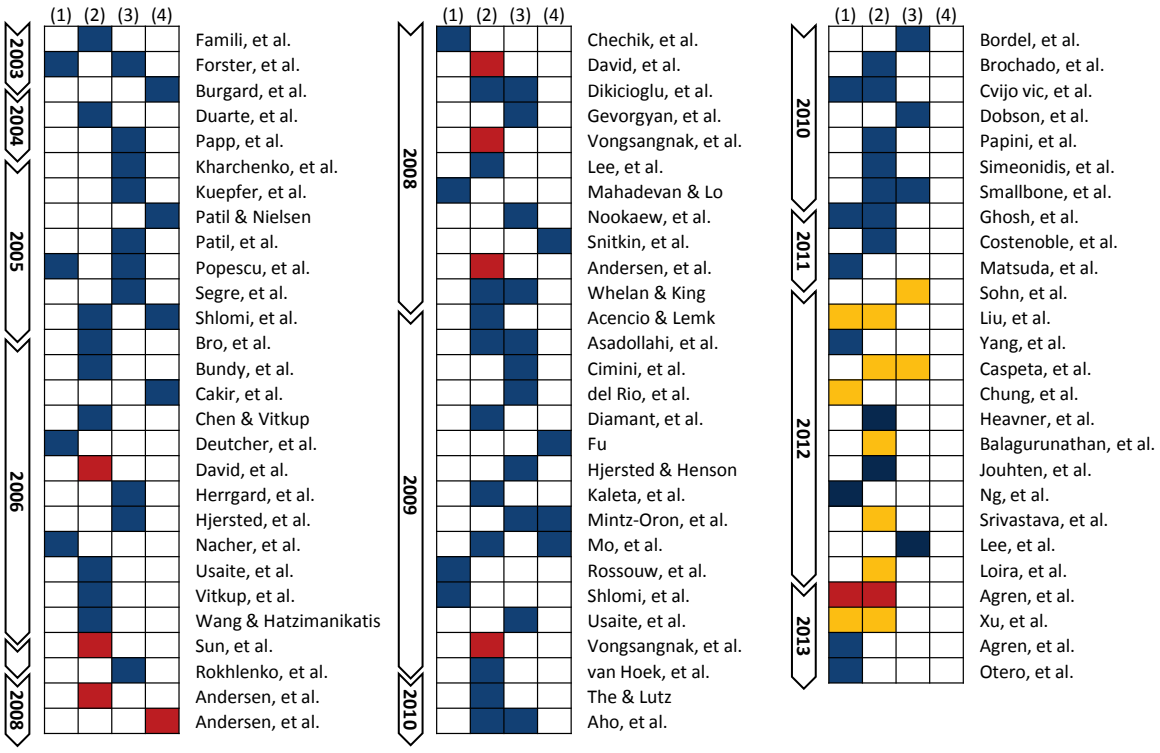


Figure 4-1. Overview of applications of fungal GEMs. The available publications which make use of GEMs to study fungal metabolism were categorized as follows: 1) guidance for metabolic engineering and strain improvement, 2) biological interpretation and discovery, 3) applications of novel computational framework, and 4) evolutionary elucidation. Blue: *S. cerevisiae*, Yellow: Non-*Saccharomyces* yeasts, Red: Filamentous fungi. Partly adapted from Osterlund *et al.* (2012).

In this section I present some of my work about genome-scale metabolic modelling applied to fungal metabolism. In **Paper I** we used genome-scale metabolic modelling to suggest knock-out targets in *S. cerevisiae* for the purpose of succinic acid production. The targets were then validated experimentally. In **Paper II** we developed an algorithm for identifying transcriptionally regulated reactions, with the aim of aiding metabolic engineering. In **Paper III** we developed a software suite for automatic reconstruction of GEMs and used this software to reconstruct a GEM for *Penicillium chrysogenum*. The model was then used to suggest metabolic engineering targets which could improve penicillin production.

4.1.1 Paper I: Genome-scale modelling enables metabolic engineering of *Saccharomyces cerevisiae* for succinic acid production

The US Department of Energy has identified succinic acid as an added-value chemical building block, with an estimated 15,000 t/year world-wide demand. The demand was predicted to grow to 270,000 t/year, representing a >2 billion USD annual market (Willke and Vorlop, 2004; McKinlay *et al.*, 2007). Succinic acid is currently used industrially in a variety of applications, such as surfactant, ion chelator, and as additive in the pharmaceutical and food industries. Currently, the only succinic acid derived from fermentation is that which is used in the food market, while the bulk is petrochemically produced from butane through maleic anhydride (McKinlay *et al.*, 2007).

Several biotechnology and metabolic engineering efforts have focused on overproduction of succinic acid in prokaryotes (Song and Lee, 2006). These bacterial hosts all grow at neutral pH, which results in secretion of the salt form, succinate, rather than the acid form. A costly acidification and precipitation step is then required in order to produce succinic acid, which is the desired product. This is a general concern when using bacterial cells for production of organic acids (Sauer *et al.*, 2008). One way to approach this issue could be to use the yeast *Saccharomyces cerevisiae* as a host. *S. cerevisiae* is a well-established, generally regarded as safe, and robust industrial production host. It is capable of growth on a variety of carbon sources, both aerobic and anaerobic, and it has a large pH operating range (3.0-6.0). Since it is capable of growth at such low pH it could be used to produce the desired acid form directly and thereby circumventing the acidification step. However, unlike the bacterial hosts described above succinate does not natively accumulate in *S. cerevisiae*.

Succinate is a TCA cycle intermediate produced from the oxidation of succinyl-CoA by succinyl-CoA synthase, or from isocitrate in a reaction catalysed by isocitrate lyase in the glyoxylate shunt. It is then further oxidized to fumarate by succinate dehydrogenase, resulting in the formation of $FADH_2$. Only limited work has been done on metabolic engineering of *S. cerevisiae* for production of succinate. The most successful work to date has been by Raab *et al.* (2010). They pursued an oxidative production route by a quadruple deletion of *SDH1*, *SDH2*, *IDP1*, and *IDH1*. This led to an interrupted TCA cycle and flux being redirected through the glyoxylate cycle instead, thereby resulting in succinate production. They could demonstrate a 0.07 C-mol/C-mol glucose succinate yield following this approach.

In **Paper I** we used FBA to propose gene deletion strategies for succinate overproduction. The main strategy was to optimize for biomass formation under constrained glucose uptake, and then observe the resulting succinate yield under a variety of conditions. Unlike the previously mentioned study, we focused primarily on anaerobic fermentation conditions, since it is a large advantage to be able to run industrial fermentations anaerobically. The three most promising single gene deletion strategies, identified under anaerobic glucose fermentation conditions, were experimentally evaluated. These strains were then physiologically and transcriptionally characterized in order to gain further knowledge into the C4 acid production by *S. cerevisiae*.

Figure 4-2 shows a comparison of the simulated specific growth rate and specific productivities compared to data generated by using the reference *S. cerevisiae* CEN.PK113-7D and BY4741 under aerobic and anaerobic glucose batch fermentations. As can be seen, the growth rate is predicted well, as is the ethanol production. However, the model is unable to capture the glycerol formation under aerobic conditions, and it underestimates the formation also under anaerobic conditions. This is due to the inability of the model to describe the Crabtree effect, as discussed earlier by Akesson *et al.* (2004).

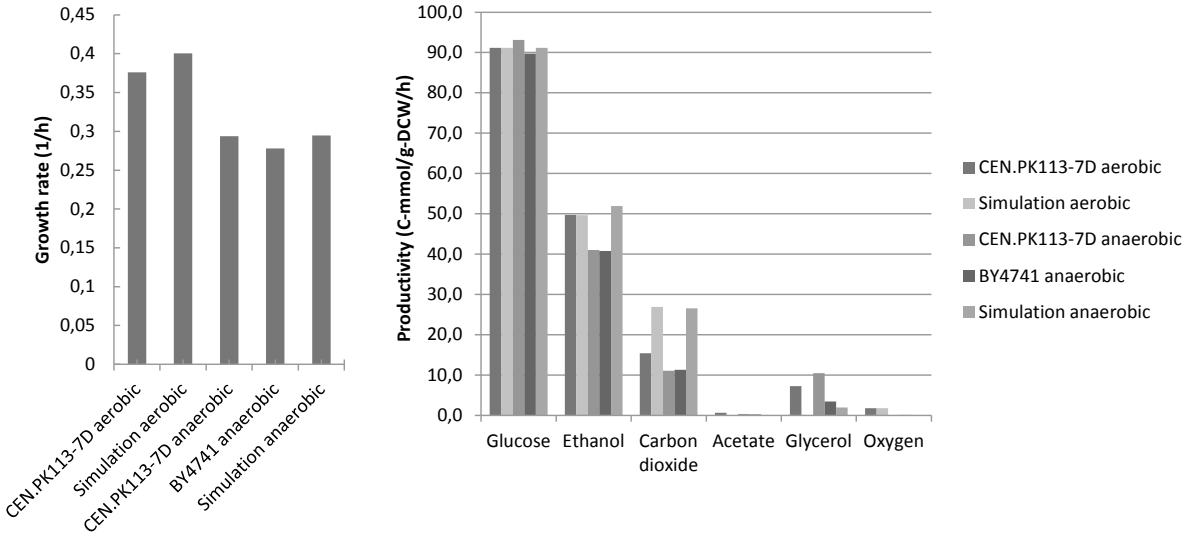


Figure 4-2. Comparison between experimental and simulated fermentation data. Comparison of the specific growth rate and specific productivities for simulated data and experimental data generated using the reference *S. cerevisiae* CEN.PK113-7D and BY4741 under aerobic and anaerobic glucose batch fermentations. For the conditions, simulation aerobic[1] and simulation anaerobic, the rO2 was constrained to 1.8 mmol-O_2/g-DCW/h and 0.016 mmol-O_2/g-DCW/h, respectively. For aerobic experimental data the specific glucose uptake rate was 91.2 C-mmol/g-DCW/h for CEN.PK113-7D. For anaerobic experimental data the specific glucose uptake rate was 93.1 C-mmol/g-DCW/h for CEN.PK113-7D and 89.7 C-mmol/g-DCW/h for BY4741. For all simulation conditions the glucose uptake rate was constrained to 91.2 C-mmol/g-DCW/h.

Overproduction of succinate was simulated using the conditions previously described. Under aerobic conditions there were no single gene deletions which resulted in succinate production. Double gene deletions resulted only in minor improvement of succinate production. Figure 4-3 presents the top single gene deletions for succinate overproduction under anaerobic conditions. There is a small but significant predicted yield on substrate for the gene deletions Δ*oac1*, Δ*mdh1*, and Δ*dic1* (0.033 C-mol/C-mol glucose). The increase in succinate yield resulted in nearly no impact on growth rate ($0.28h^{-1}$ vs. $0.30h^{-1}$, single gene deletion vs. reference case simulation, respectively). The strains *Δoac1*, *Δmdh1*, and *Δdic1* are viable null mutants, and their annotation is well known, encoding for an inner mitochondrial membrane transporter (Oac1p), malate dehydrogenase (Mdh1p), and an inner dicarboxylate mitochondrial transporter (Dic1p), respectively (Cherry et al., 1998).

In order to investigate if the single gene deletions identified in silico resulted in more succinate production, the corresponding strains were cultivated anaerobically in 2L well controlled fermenters. A comparative analysis between simulation and experimental results are presented in Figure 4-3. There is a fair agreement between model predictions and experimental data. Focusing more closely on the specific succinate productivity, the reference case, *Δmdh1*, and *Δoac1* experimentally determined yields are significantly lower than expected based on model simulations. The *Δdic1* case, however, demonstrated a significantly higher yield of succinate compared to the reference case (0.02 vs. 0.00 C-mol/C-mol glucose,

[1] The condition here referred to as "aerobic" corresponds to the condition referred to as "semi-aerobic" in **Paper I**. The original "aerobic" was for totally unconstrained oxygen uptake; a condition which badly represented the experimental data and which therefore is not discussed further here.

Δdic1 vs. reference, respectively), and was in line with the in silico prediction (0.02 vs. 0.03 C-mol/C-mol glucose, *Δdic1* experimental vs. *Δdic1* simulation, respectively). This represents a significant improvement in succinate productivity based exclusively on a novel in silico prediction.

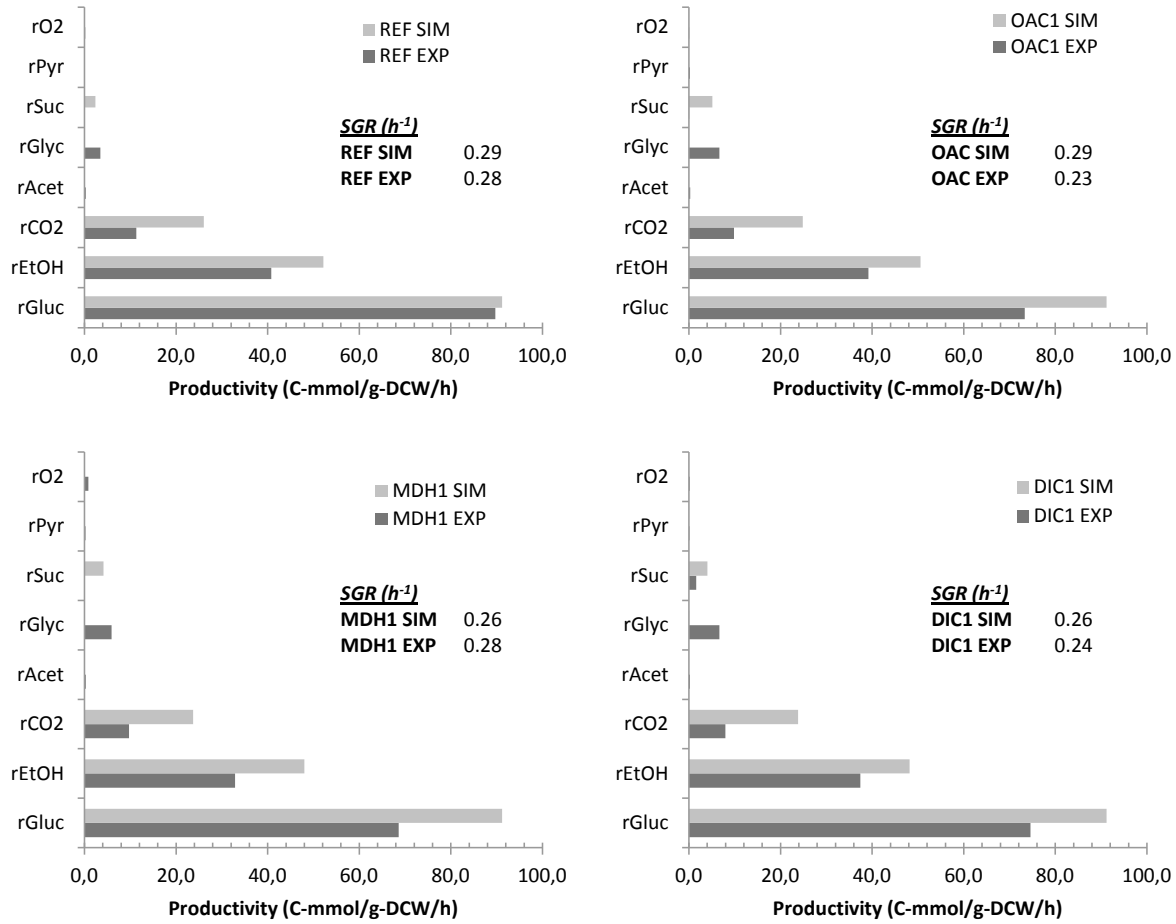


Figure 4-3. Experimental and simulated data for reference strain, *Δoac1*, *Δmdh1*, and *Δdic1* strains. Summary of the specific growth rate (SGR) and specific consumption/productivity for major products (glucose, ethanol, carbon dioxide, acetate, glycerol, succinate, pyruvate, and oxygen) for both experimentally determined data of anaerobic batch glucose fermentations and the corresponding anaerobic simulation data. The panels show the BY4741 reference strain and the single gene deletion strains *Δmdh1*, *Δdic1*, and *Δoac1*. In general, the simulated growth rates are in very good agreement with the experimental values. The *Δdic1* mutant results in some succinate production.

To gain further insight into the performance of each strain, DNA microarray profiling was performed for the anaerobic batch glucose fermentations. The complete list of differentially expressed genes[2] for *Δdic1* and *Δmdh1* were submitted for metabolic pathway annotation using the SGD Pathway Expression Viewer and Reactome databases (Paley and Karp, 2006; Matthews *et al.*, 2009). Only a rather small number of metabolic genes were identified in

[2] The number of differentially expressed genes for the *Δoac1* strain compared to the reference strain was very low, and consequently suggests that deletion of *Δoac1* causes virtually no transcriptional, and consequently, physiological differences compared to the reference BY4741 strain. No further analysis of the transcriptional data was performed for this strain.

Δdic1 and *Δmdh1* compared to the reference; a total of 10 and 20 genes respectively. Perhaps more striking is that there is an overlap of 9 metabolic genes between *Δdic1* and *Δmdh1*. The only differentially expressed gene present in the *Δdic1* condition, not present in the *Δmdh1* condition, is *DIC1*.

The identified metabolic engineering strategies through *Δdic1*, *Δmdh1*, and *Δoac1*, suggest a common mechanism. Mitochondrial redox balance must be maintained, and while respiratory metabolic activity under anaerobic conditions is reduced compared to aerobic conditions, some activity is required to support glutamate/glutamine metabolism from α-ketoglutarate (Camarasa *et al.*, 2003; Camarasa *et al.*, 2007). This leads to the accumulation of mitochondrial NADH. During anaerobic metabolism, NAD^+ regeneration occurs via the following pathways according to our simulations (where the subscript m denotes mitochondrial):

$$Oac1p : oxaloacetate \rightarrow oxaloacetate_m + H_m^+$$

$$Mdh1p : oxaloacetate_m + NADH_m \rightarrow malate_m + NAD_m^+$$

$$Dic1p : malate_m + phosphate \rightarrow malate + phosphate_m$$

$$Mir1p : H_m^+ + phosphate_m \rightarrow phosphate$$

Net reaction stoichiometry: $oxaloacetate + NADH_m \rightarrow malate + NAD_m^+$

In the cytosol, malate is then converted to oxaloacetate, and the resulting NADH is converted to NAD^+, resulting in the production of glycerol. The *Δdic1* strategy, relying on deletion of the mitochondrial dicarboxylate carrier *DIC1*, catalyses the following transport reaction:

$$Dic1p : malate + succinate_m \rightarrow malate_m + succinate$$

Assuming *DIC1* deletion, then the resulting simulated pathway is:

$$Ndip : ubiquinone_m + NADH_m \rightarrow ubiquinol_m + NAD_m^+$$

$$Sdh3p : ubiquinol_m + FAD_m \rightarrow ubiquinone_m + FADH_{2,m}$$

$$Frds1p : fumarate + FADH_{2,m} \rightarrow succinate + FAD_m$$

Net reaction stoichiometry: $fumarate + NADH_m \rightarrow succinate + NAD_m^+$

The *Δdic1* strategy relies heavily on the subcellular localization and function of Frds1p, soluble mitochondrial fumarate reductase, which continues to be poorly understood. However, recent work has suggested that a double deletion *S. cerevisiae* mutant, *Δosm1Δfrds1*, failed to grow under batch glucose anaerobic conditions. During anaerobic growth, *FRDS1* expression in the wild-type was two to eight times higher than that of *OSM1*, suggesting that formation of succinate is strictly required for the re-oxidation of $FADH_2$ and that its expression may be oxygen-regulated (Camarasa *et al.*, 2007). There was a strong upregulation of *CYC1* in both the *Δdic1* and *Δmdh1* mutants. *CYC1* facilitates electron transfer from ubiquinone cytochrome C oxidoreductase to cytochrome C oxidase. This direction, which is the normal oxidative route and ends in reduction of O_2, would not be possible under fully anaerobic conditions. The upregulation can therefore be viewed as a coping strategy to deal with the stress of redox imbalance. Deletion of *CYC1* could therefore be a way to ensure that all NAD^+ regeneration is coupled to succinate production. The strategies proposed here rely on the capacity for reductive TCA cycle activity under anaerobic conditions, and more specifically, the catalysis

of fumarate to succinate via fumarate reductase. There is evidence suggesting that *S. cerevisiae* can exhibit this metabolic state (Camarasa *et al.*, 2003; Camarasa *et al.*, 2007).

In conclusion, a GEM was used to predict single and double gene deletion strategies which could lead to increased succinate production under a variety of conditions. Three of these strategies, all utilizing anaerobic fermentation conditions, were validated in vivo and one, *Δdic1*, was identified to lead to a significant improvement in succinate yield, in close agreement with the model prediction. However, the yield was not as high as what was reported by Raab *et al.* (2010) using a quadruple gene deletion strategy and the oxidative route (0.02 C-mol/C-mol glucose vs. 0.07 C-mol/C-mol glucose). Furthermore, physiological characterization and transcriptome analysis were used to aid in interpretation of the simulations and provide a mechanistic explanation of the results. The proposed mechanisms rely heavily on compartmental transport reactions and mitochondrial redox balancing. Transcriptional profiling suggests that succinate formation is coupled to mitochondrial redox balancing, and more specifically, reductive TCA cycle activity. While far from industrial titers, this proof-of-concept suggests that in silico predictions coupled with experimental validation can be used to identify novel and non-intuitive metabolic engineering strategies.

4.1.2 Paper II: Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes

Metabolic fluxes are the result of a complex interplay involving metabolite concentrations, enzyme kinetics, gene expression, and translational regulation. Due to these multiple layers of regulation, there is generally not a clear correlation between mRNA levels, enzyme levels, and the metabolic fluxes (Akesson *et al.*, 2004). Only a rather small fraction of enzymes show a clear positive correlation between their transcription levels and the rates of the reactions that they catalyse. These reactions are said to have transcriptional regulation.

In **Paper II** we propose a method which allows for identification of enzymes that show transcriptional regulation, and therefore represent suitable targets for metabolic engineering (up- or downregulation). The method is based on integration of gene expression data with flux data by transforming quantitative flux data into a genome-scale set of statistical scores analogous to those obtained from transcriptional profiling. This works by constraining a set of experimentally determined exchange fluxes in a GEM for the organism being studied. This is then done for each of the studied conditions, or for each of the strains investigated. A random sampling algorithm is then used to generate a set of internal flux distributions which all satisfy the experimentally determined exchange fluxes. By this approach it is possible to obtain means and standard deviations for each flux in the GEM, and from there it is then possible to derive p-values for the significance of flux change between conditions (Mo *et al.*, 2009; Schellenberger and Palsson, 2009). These values can then be compared to the significance of change in gene expression for the corresponding enzymes. The comparison of flux change and gene expression allows for identification of enzymes showing a significant correlation between flux change and expression change (transcriptional regulation) as well as reactions whose flux change is likely to be driven only by changes in the metabolite concentrations (metabolic regulation). This workflow is described in more detail in Figure 4-4.

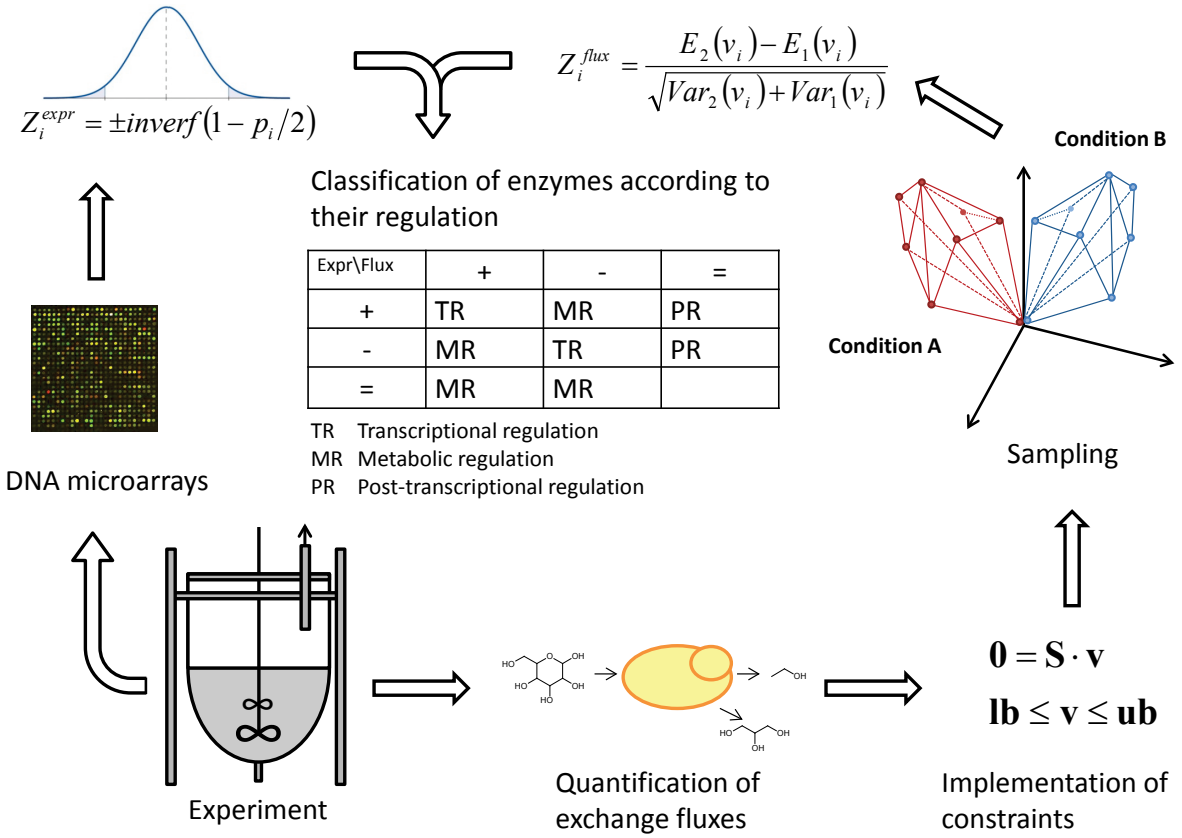


Figure 4-4. Workflow for identification of transcriptionally regulated reactions. The figure illustrates how the method can be applied to identify transcriptionally regulated reactions. Two fermentations experiments are performed; one for the test case and one reference. Gene expression levels are measured for the two cases using, for example, DNA microarrays. A statistical test is used for each of the genes to calculate the Z-score for differential expression between the cases. In parallel to this, key exchange fluxes are quantified. The corresponding exchange fluxes in two GEMs are then constrained to what was seen experimentally for each of the cases. Due to the high dimensionality of GEMs, and the redundancy in cell metabolism, there are many different internal flux distributions which all result in the measured exchange fluxes. A random sampling algorithm is applied to sample many such solutions from each of the two models. A Z-score for differential flux can then be estimated for each of the reactions from the difference in average flux divided by the square root of the sum of the variances. The Z-scores for differential expression/flux are then transformed into probabilities of change by using the cumulative Gaussian distribution. These probabilities are then used to calculate probabilities of having correlated gene expression and flux changes for the corresponding reaction. The enzymes in the network can then be classified as: 1) enzymes that have a significantly correlated change both in flux and expression level (reactions showing transcriptional regulation); 2) enzymes that show a significant change in expression but not in flux (post-transcriptional regulation); 3) enzymes that show significant changes in flux but not a change in expression (metabolic regulation).

To evaluate our method we used data for *S. cerevisiae*. Data from growth on four different carbon sources (glucose, maltose, ethanol and acetate) in chemostat cultures and from five deletion mutants grown in batch cultures were used. This summary will only focus on the different carbon sources, and not on the deletion mutants. Table 4-2 shows the top scoring enzymes in the different categories for each of the carbon source shifts.

Table 4-2. Top scoring enzymes for transcriptional, post-transcriptional and metabolic regulation for changes in carbon source. The 10 top scoring enzymes in each group are shown (or fewer when less than 10 enzymes had a probability larger than 0.95).

Carbon source shift	Enzymes showing transcriptional regulation	Enzymes showing post-transcriptional regulation	Enzymes showing metabolic regulation
Glucose/Maltose	• α-glucosidase MAL32 • Low-affinity glucose transporter HXT4	• Mevalonate kinase • Inosine-5'-monophosphate dehydrogenase IMD2 • Asparagine synthetase 1 • Glycerol-3-phosphatase 1 • Uncharacterized deaminase • Nicotinate-nucleotide pyrophosphorylase • Mevalonate kinase • Nicotinate-nucleotide pyrophosphorylase • Glycerol-3-phosphate dehydrogenase 1	• Acetate transport via proton symport
Glucose/Ethanol	• Phosphoenolpyruvate carboxykinase • Fructose-1,6-bisphosphatase • Isocitrate lyase • Malate dehydrogenase • Citrate synthase • Ribose-5-phosphate isomerise • Low-affinity glucose transporter HXT4 • External NADH-ubiquinone oxidoreductase 2 • Glucose-6-phosphate isomerase	• Formate dehydrogenase 2 • ATP-NADH kinase • Sulfate permease 1 • Formate dehydrogenase 1 • Dicarboxylate transporter • NADP-specific glutamate dehydrogenase 2 • Uncharacterized deaminase • phosphogluconolactonase 3 • 6-phosphofructo-2-kinase 2 • Nucleoside diphosphate kinase	• Fructose-bisphosphate aldolase • Triosephosphate isomerase • Pyruvate dehydrogenase E1 subunit alpha • α-ketoglutarate dehydrogenase • Succinyl-CoA ligase subunit beta • Malate synthase 2 • Glucose-6-phosphate 1-dehydrogenase • Cytochrome b-c1 subunit Rieske • Adenylate kinase
Glucose/Acetate	• Fumarate hydratase • Phosphoenolpyruvate carboxykinase • Fructose-1,6-bisphosphatase • Isocitrate dehydrogenase • Succinate-semialdehyde dehydrogenase • Citrate synthase • Isocitrate dehydrogenase subunit 1 • Pyruvate kinase 2 • Low-affinity glucose transporter HXT4	• Phospho-2-keto-3-deoxyheptonate aldolase • Ribonucleoside-diphosphate reductase large chain 1 • 6-phosphofructo-2-kinase 1 • Glutamine-dependent NAD synthetase • Ribose-phosphate pyrophosphokinase 4 • ATP-dependent permease AUS1 • Fructose-2,6-bisphosphatase • Nicotinate-nucleotide pyrophosphorylase • Squalene monooxygenase	• Fructose-bisphosphate aldolase • Triosephosphate isomerase • Ribose-5-phosphate isomerase • Inorganic pyrophosphatase • Adenylate kinase • Glutamate decarboxylase • 4-aminobutyrate aminotransferase • Tricarboxylate transport protein • Prephenate dehydrogenase

In the glucose to maltose transition, only two enzymes showed transcriptional change correlated with their flux. The α-glucosidase Mal32p, which is responsible for the breakdown of maltose into glucose, was upregulated and the glucose transporter Hxt4p was

downregulated. Only very minor adjustments could be seen in terms of fluxes, and only enzymes directly related with the substrate uptake and utilization were detected. The changes in gene expression were also few and only 11 metabolic enzymes were significantly perturbed (without significant flux changes).

The glucose to ethanol and the glucose to acetate shifts showed much more widespread changes in flux and expression. They therefore represent more interesting cases studies. In the glucose-ethanol transition 19 enzymes showed transcriptional regulation and 22 enzymes changed in expression but not in flux. For the glucose-acetate shift the same numbers were 33 and 23, respectively. Among the enzymes showing transcriptional regulation, 14 were shared between the glucose-ethanol and glucose-acetate transitions. However, there was no overlap between the sets of enzymes which do not change in flux. Metabolic regulation was observed in 21 reactions for each case, out of which 8 overlap.

The enzymes with transcriptional regulation clearly show a downregulation of enzymes involved in the uptake and utilization of glucose (e.g. glucose transporter Hxt4p or hexokinase 2) and the upregulation of enzymes involved in gluconeogenesis (e.g. fructose-1,6-biphosphatase) or the TCA cycle (e.g. succinate dehydrogenase or citrate synthase). Acetyl-CoA synthetase 2, responsible for providing acetyl-CoA to the TCA cycle, is also identified as transcriptionally upregulated, as well together with ATP synthetase and external NADH-ubiquinone oxidoreductase 2, which provide the necessary NAD^+ needed for oxidation of ethanol or acetate in the cytoplasm and thereby maintaining the redox balance in the cell. Isocitrate lyase, a key component of the glyoxylate cycle, is also transcriptionally upregulated. This allows for net formation of malate, which can then be further converted to phosphoenolpyruvate (via oxaloacetate) in order to fuel gluconeogenesis. All these changes in fluxes are consistent with what is known about the changes in metabolism between growth on glucose to C2 carbon sources like ethanol and acetate. Interestingly, not all reactions associated with this shift in flux distributions are transcriptionally regulated. Rather, the cell has selected a small number of key reactions to regulate at the transcriptional level.

We also performed an enrichment test in order to compare the transcription factors involved in the expression of the enzymes classified as having transcriptional regulation and the enzymes showing changes in expression but not in flux. We identified three transcription factors which were strongly overrepresented in the metabolic genes showing transcriptional regulation. In the glucose-ethanol transition, the transcription factors Gcr1p and Gcr2p both appeared in 11 transcriptionally regulated genes and in none of the other genes, whereas the transcription factor Hap4p appeared in 11 transcriptionally regulated genes and 5 of the other regulated genes. For the glucose-acetate transition these numbers were 15-0, 11-0 and 15-0 for the same transcription factors. This can be interpreted as if certain transcription factors are particularly involved in the transcriptional regulation of metabolic fluxes. This implies that there is global regulation of major flux alterations, which is in agreement with what has been shown experimentally for galactose metabolism (Ostergaard *et al.*, 2001).

The top scoring metabolically regulated reactions, both for the glucose-ethanol and glucose-acetate shifts, are fructose bisphosphate aldolase and triosephosphate isomerase. These reactions are known to operate close to their equilibrium and they are therefore sensitive to changes in the metabolite pools, which is consistent with metabolic regulation of the fluxes. In the two shifts the direction of these reactions is inverted. This can only be explained by a decrease in the fructose-1,6-diphosphate pool and an increase in the glyceraldehyde-3-phosphate and dihydroxyacetone pools. This hypothesis is supported by the fact that in

chemostat cultures there is not found to be any correlation between the glycolytic flux and the expression of the genes encoding for these two enzymes (Daran-Lapujade *et al.*, 2007).

In conclusion, the combined use of random sampling in GEMs and expression data allows for global identification of reactions which are either transcriptionally or metabolically regulated. The reactions exhibiting transcriptional regulation form a set of putative metabolic engineering targets, where enzyme overexpression or downregulation is likely to influence the flux through these reactions. The reactions exhibiting metabolic regulation points to parts of metabolism where the metabolite pools are possibly increasing or decreasing in connection with transcriptional changes, and thereby counteracting possible changes in enzyme concentration. This knowledge can be used to identify whether one should target changes in enzyme concentration (v_{max} changes), e.g. through overexpression, or changes in enzyme affinity (K_m changes), e.g. through expression of heterologous enzymes, in order to alter the fluxes.

4.1.3 Paper III: The RAVEN Toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*

In **Paper III** we developed a software suite named the RAVEN Toolbox (Reconstruction, Analysis, and Visualization of Metabolic Networks), which aims at automating parts of the GEM reconstruction process in order to allow for faster and easier reconstruction of high-quality GEMs. The software was then used for reconstructing a model of the ascomycetous fungi *Penicillium chrysogenum*, the organism used for industrial production of penicillin, and an important microbial cell factory. The resulting model was validated in an extensive literature survey and by comparison to fermentation data. The model was then used together with the algorithm presented in **Paper II** in order to identify transcriptionally regulated metabolic bottlenecks in penicillin fermentations. These bottlenecks can then be targets for metabolic engineering.

The RAVEN Toolbox has three main foci: 1) automatic reconstruction of GEMs based on protein homology, 2) network analysis, modelling and interpretation of simulation results, 3) visualization of GEMs using pre-drawn metabolic network maps. Figure 4-5 summarizes the capabilities of the RAVEN Toolbox.

Previously published GEMs represent a solid basis for metabolic reconstruction of models for new organisms, in particular if the organisms are closely related and therefore share many metabolic capabilities. The main advantage of using existing models compared to reaction databases, such as KEGG or BRENDA (Schomburg *et al.*, 2002), is that they contain information which can be difficult to obtain in an automated manner, in particular directionality and compartmentalization. GEMs are also typically constructed for modelling purposes, which is not the case for reaction databases. The downside is that only reactions present in the template models can be included. The RAVEN Toolbox therefore contains two approaches for automatic generation of draft models; one which relies on the metabolic functions represented in previously published models, and one method which uses the KEGG database for automatic identification of new metabolic functions that are not included in the published models.

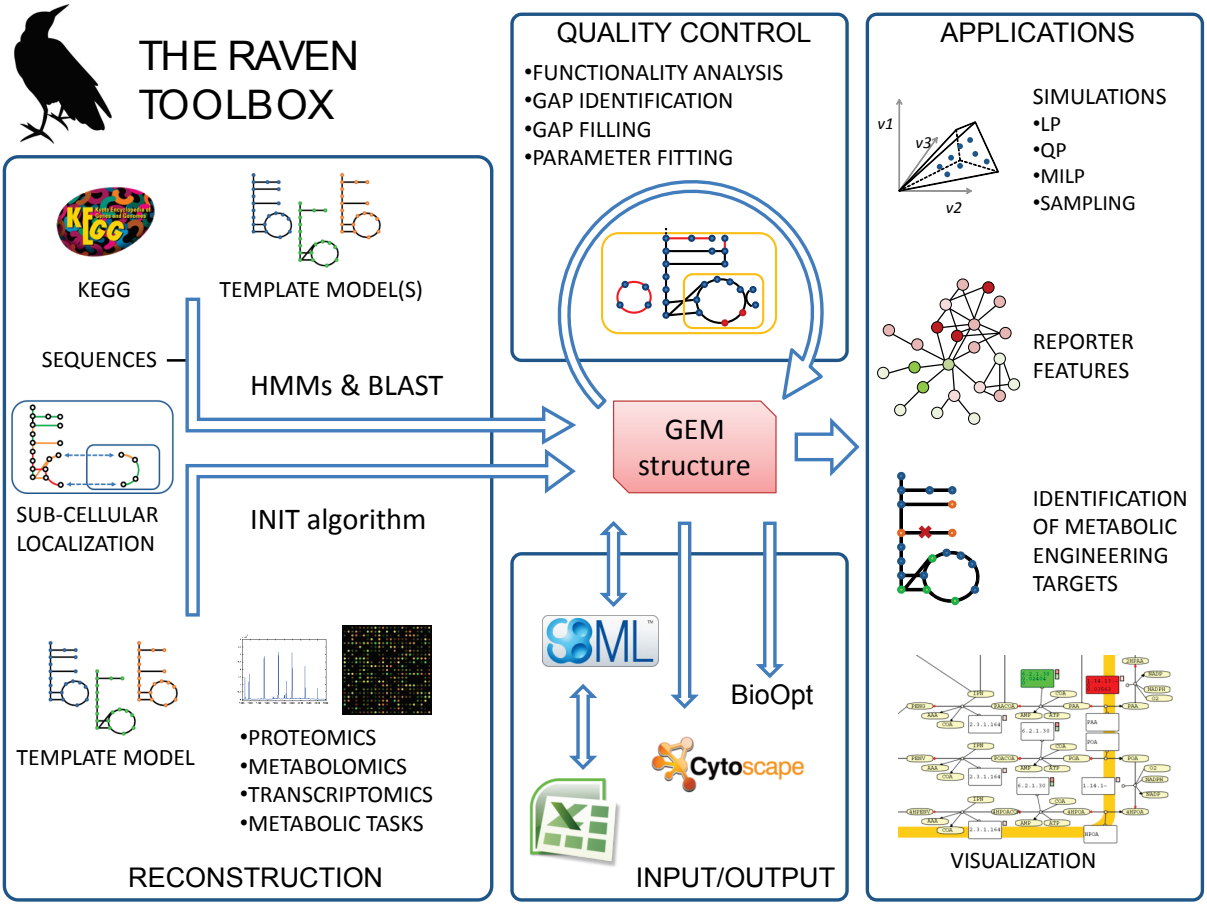


Figure 4-5. The RAVEN Toolbox. The software allows for reconstruction of GEMs based on template models or on the KEGG database. Subcellular localizations of reactions can be estimated based on signal peptides and other characteristics of the catalysing enzymes. The resulting models can be exported to a number of formats, or they can be used for various types of simulations. The RAVEN Toolbox has a strong focus on quality control. Visualization of simulation result and/or integration of other types of data can be performed by overlaying information on pre-drawn metabolic maps. The software also implements the INIT algorithm, which is a powerful approach for reconstruction of tissue-specific models (**Paper IV**). Figure taken from Agren *et al.* (2013a).

In the first approach the software generates a draft model based on protein orthology using bi-directional BLASTp (Altschul *et al.*, 1990). The second approach is also based on protein homology but requires no template models. Instead it relies on the information on protein sequences and on the assigned metabolic reactions that is available in the KEGG database. The method makes use of the KEGG Orthology (KO) IDs, which are manually annotated sets of genes that encode some specified metabolic function. Each KO is associated with a number of metabolic reactions. The aim of the method is then to assign genes to these KOs based on the consensus protein sequence. This works by generating a hidden Markov model based on the sequences for each KO using HMMER (Eddy, 1998). The final step is the querying of the set of HMMs with the protein sequences of the organism of interest. If a gene has a significant match to one KO, the reactions associated to that KO are added to the model together with the corresponding gene.

The approach proposed above will facilitate and accelerate the generation of a draft metabolic network reconstruction. However, the automated reconstruction can lead to some loss of control compared to a stricter manual, bottom-up approach. It is therefore important to identify and fill gaps in the model to ensure that the network is functioning as required. The RAVEN Toolbox therefore contains a number of novel methods to support the gap filling

process. Table 4-3 shows a comparison between the RAVEN Toolbox and some other software with similar functionalities.

Table 4-3. Comparison between the RAVEN Toolbox and some other software for automated GEM reconstruction.

	RAVEN	Model SEED[a]	AUTOGRAPH[b]	IdentiCS[c]	GEM System[d]
Includes general network	X	X		X	X
Generates functional models	X	X			
Assigns subcellular localization	X				
Can use user defined models	X		X		
Integrates gap filling	X	X			X
Offline software	X			X	
Includes visualization	X			X	X
Gene prediction		X		X	X

[a]Henry *et al.* (2010); [b]Notebaart *et al.* (2006); [c]Sun and Zeng (2004); [d]Arakawa *et al.* (2006). Taken from Agren *et al.* (2013a).

The *P. chrysogenum* metabolic network was reconstructed based on a combination of the automated reconstruction approaches in the RAVEN Toolbox, manual curation, and an extensive bibliomic survey. Three GEMs for closely related filamentous fungi, *Aspergillus nidulans* iHD666 (David *et al.*, 2006), *A. niger* iMA871 (Andersen *et al.*, 2008), and *A. oryzae* iWV1314 (Vongsangnak *et al.*, 2008), were used as template models for the reconstruction of the *P. chrysogenum* model. The model comprises 1471 unique metabolic reactions in four subcellular compartments; extracellular, cytosolic, mitochondrial, and peroxisomal. 1006 ORFs are associated to the reactions, 89 of which participate in one of 35 protein complexes. In parallel to the automatic reconstruction, an extensive literature study was performed. In total 440 cited articles provide experimental evidence for the majority of the reactions. The model was validated with respect to 76 metabolic functions known to occur in *P. chrysogenum.*

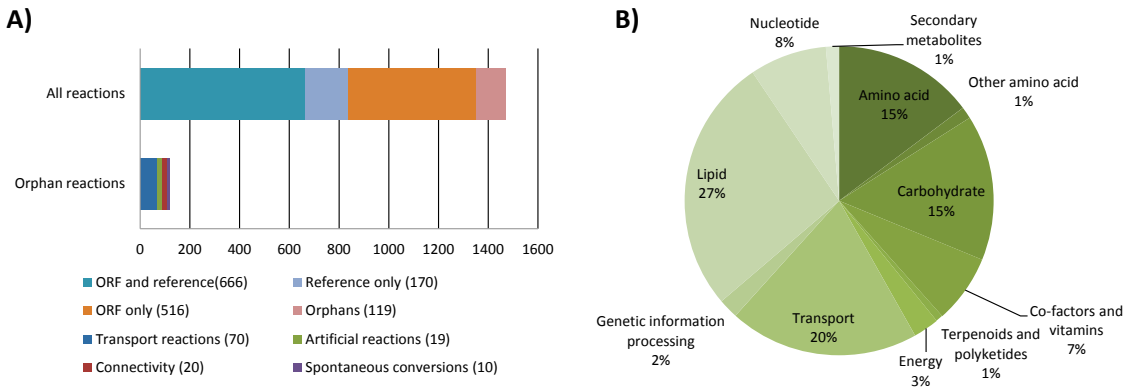


Figure 4-6. Evidence level for the *P. chrysogenum* metabolic network. A) Properties of the reconstructed network. The top bar shows the support for the 1471 unique reactions (not counting exchange reactions) sorted by the type of evidence. The bottom bar shows the orphan reactions; reactions inferred without supporting ORFs or literature references. **B**) ORF classification. The ORFs in the model are classified into broad groups based on KEGG classification.

Figure 4-6 summarises the literature support for the reactions in the model and shows a classification of the ORFs in the model based on KEGG Pathways. To illustrate the metabolic network, and to aid in interpretation of gene expression data and simulation results, a map of

the full model was drawn and annotated so as to be compatible with the visualization functions in the RAVEN Toolbox.

We then used the GEM in a study of penicillin yields and in particular the relative importance of ATP and NADPH provision during penicillin production. In a second study we showed how the model can be used to integrate fermentation data with transcriptome data using the method developed in **Paper II**.

The industrial *P. chrysogenum* strains have been subjected to 50 years of directed evolution to increase the yields and titers of penicillin, with great cost reduction and productivity gain, but the yields are still far from the theoretical maximum (Thykaer and Nielsen, 2003). Penicillin production is associated with an increased requirement of energy in the form of ATP; in the condensation of the three precursor amino acids to form the tripeptide ACV; in the reduction of sulfate; and when a side chain (the precursor molecule which is supplied to the media and which differs depending on the type of penicillin produced) is activated by ligation to coenzyme A. Penicillin production is also associated with a large requirement of NADPH, primarily needed for the reduction of sulfate, but also in the biosynthesis of valine and homoserine from α-ketobutyrate. Elucidating the impact increased ATP requirements have compared to the NADPH requirements is useful when choosing among possible metabolic engineering strategies.

The maximum theoretical yield of penicillin on glucose with sulfate as the sulfur source was calculated to be 0.42 mol penicillin/mol glucose using the reconstructed GEM. To investigate the effect of ATP, an artificial reaction was included that allowed for ATP production from ADP without any energetic costs. This resulted in a yield of 0.52 mol penicillin/mol glucose, using sulfate as the sulfur source. The conclusion is that ATP availability has a relatively small effect on the yield. The shadow prices (how much the penicillin production can increase if the availability of a metabolite were to increase by a small amount) were calculated to be 0.015 mol penicillin/mol ATP, 0.040 mol penicillin/mol NADPH, and 0.037 mol penicillin/mol NADH.

NADPH and NADH are similar when it comes to energy content, but have different roles in the metabolism, where NADPH serves primarily anabolic roles and NADH primarily catabolic roles. NADPH is mainly produced in the pentose phosphate pathway, which makes NADPH somewhat more energetically expensive to regenerate compared to NADH. In order to investigate the relative importance of NADH and NADPH an artificial reaction was included that allowed for production of NADPH from NADH to simulate a potential increase of the NADPH availability. Simulations were then carried out maximizing first for growth and then for penicillin production. The resulting flux through the artificial reaction was 8.5 times larger when maximizing for penicillin than when maximizing for growth. This demonstrates that the cells will have a much higher NADPH demand at high penicillin yields compared to normal growth conditions. Redirecting a higher flux through the pentose phosphate pathway and/or introducing NADH-dependent versions of NADPH-consuming enzymes could therefore be potential metabolic engineering strategies for achieving higher penicillin yields.

For the direct identification of possible metabolic engineering targets a gene deletion analysis was performed by searching for sets of gene deletions that resulted in an increased yield of penicillin, and which would stoichiometrically couple penicillin production to growth. This was performed using FBA, and combinations of up to three gene deletions were evaluated (MoMA was also applied and gave similar results). The only targets which could be identified were the deletion of any of the genes responsible for breakdown of phenylacetic acid

(homogentisate 1,2-dioxygenase, maleylacetoacetate isomerase, or fumarylacetoacetase). Deletion of any of these genes resulted in a predicted 21% increase in penicillin production.

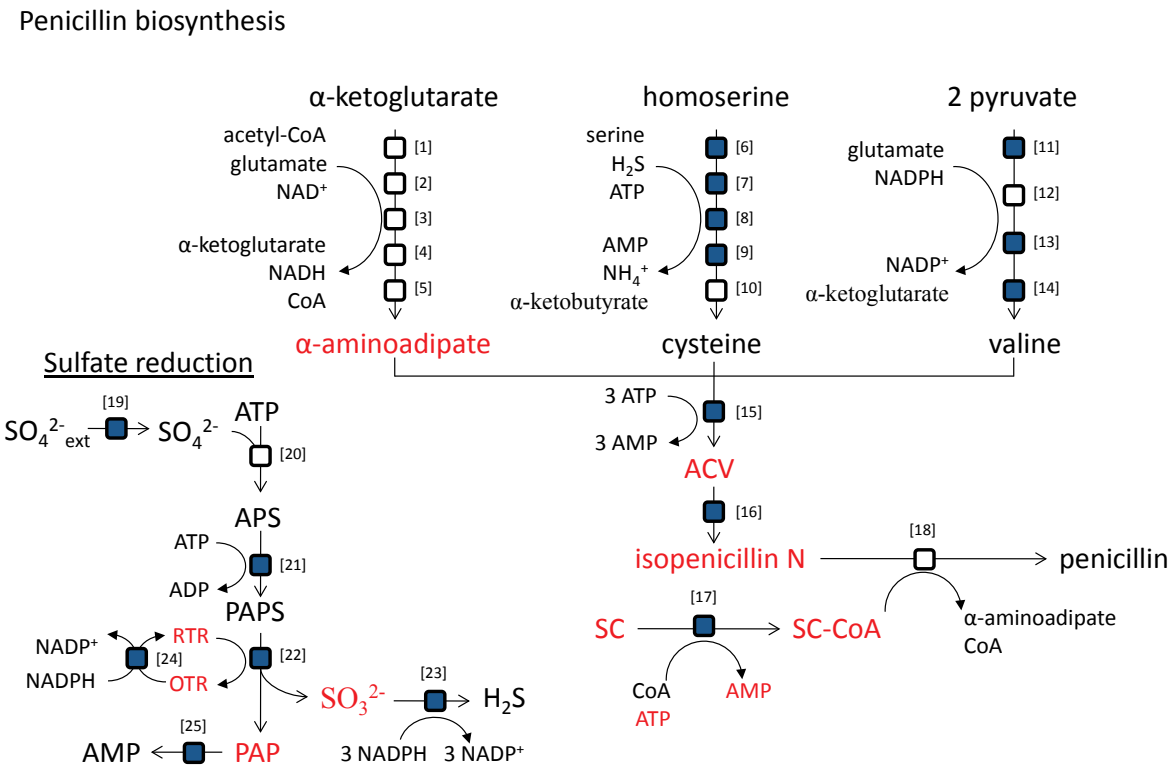


Figure 4-7. Integrative analysis of a high and a low producing strain. Depicts synthesis pathways of penicillin and important precursors. Blue boxes correspond to reactions identified as being transcriptionally controlled and upregulated by the algorithm presented in **Paper II**. Metabolites around which significant transcriptional changes occur compared to a low producing strain are coloured red. SC: side chain (e. g. the precursor molecule phenylacetic acid). The biosynthesis of penicillin starts with the condensation of the three amino acids α-aminoadipate (an intermediate in the L-lysine biosynthesis pathway), L-cystein, and L-valine to form the tripeptide ACV. ACV is further converted to isopenicillin N. For the industrially relevant types of penicillin a side-chain is supplied to the media. This side-chain is activated by ligation to coenzyme A. In the last step of penicillin biosynthesis an acyl transferase exchanges the α-aminoadipate moiety of isopenicillin N with the side-chain, thereby generating penicillin and regenerating α-aminoadipate. Since L-cystein is a sulfur-containing amino acid penicillin production is also tightly associated with sulfur metabolism. [1] homocitrate synthase; [2] homocitrate dehydrase; [3] homoaconitate hydrase; [4] homoisocitrate dehydrogenase; [5] α-aminoadipate aminotransferase; [6] homoserine transacetylase; [7] O-acetylhomoserine sulfhydrylase; [8] cystathione-β-synthase; [9] cystathione-γ-lyase; [10] acetate CoA ligase; [11] acetolactate synthase; [12] ketol-acid reductoisomerase; [13] dihydroxy acid dehydrase; [14] branched chain amino acid transferase; [15] ACV synthase; [16] isopenicillin N synthase; [17] acyl CoA ligase (side chain dependent, reaction is for phenylacetate CoA ligase); [18] isopenicillin N N-acyltransferase; [19] sulfate permease; [20] sulfate adenyl transferase; [21] adenyl sulfate kinase; [22] phosphoadenyl sulfate reductase; [23] sulfite reductase; [24] thioredoxin reductase; [25] 3'(2'),5'-bisphosphate nucleotidase.

In order to identify transcriptionally regulated metabolic bottlenecks we applied the method from **Paper II** and compared the high producing industrial strain DS17690, which has been developed by DSM, and the low producing reference strain Wis 54-1255 (Harris *et al.*, 2006). Flux data and gene expression levels for aerobic, glucose-limited chemostat fermentation of DS17690 and Wis 54-1255 were used as input to the algorithm (van den Berg *et al.*, 2008). 58 fluxes were found to be significantly changed between the high and low production strains ($p<0.05$) and 612 genes were differentially expressed ($p<0.005$). 36 reactions were identified as having significantly higher flux and upregulated genes, i.e. they are likely to have transcriptional regulation of their fluxes. Figure 4-7 shows some of the most important reactions in penicillin biosynthesis together with the responsible enzymes. Reactions that

were identified as being transcriptionally controlled and upregulated are highlighted. In addition, the Reporter Metabolites algorithm was used to identify metabolites around which significant transcriptional changes occurred (Patil and Nielsen, 2005). These metabolites are highlighted in Figure 4-7 as well.

As can be seen in Figure 4-7, a large proportion of the reactions identified as being a transcriptionally controlled are directly involved in penicillin metabolism (15 out of 38). This indicates that the capabilities of the industrial strain to produce penicillin to a large extent depend on the reactions closely related to penicillin metabolism, rather than more peripheral effects. Among these reactions are many of the reactions responsible for the synthesis of the amino acids that are precursors for ACV, as well as the two penicillin producing reactions isopenicillin N synthase and ACV synthase. This is consistent with a study on the gene copy-number effect on penicillin production (Theilgaard *et al.*, 2001). The phenylacetate:CoA ligase is high ranking but the acyl-CoA:isopenicillin N acyltransferase is absent, which is consistent with measurements of high activities of this enzyme and the low flux control estimated for this enzyme (Jorgensen *et al.*, 1995b; Nielsen and Jorgensen, 1995). Several of the reactions involved in sulfate reduction are present, as well as the sulfate permease. It is interesting to note that none of the reactions in the pentose phosphate pathway are identified even though there is an increased demand for NADPH.

We also found that the pathway from α-ketobutyrate to succinate is identified to have both increased flux and increased gene expression. α-ketobutyrate is a by-product of cysteine production via the transsulfuration pathway, and it is used for isoleucine biosynthesis. Under normal growth conditions the demand for cysteine is less than that for isoleucine, meaning that all α-ketobutyrate is converted into isoleucine. However, during high-level penicillin production the cysteine production far exceeds the need for isoleucine, requiring an alternative route for α-ketobutyrate consumption. This route involves the decarboxylation of α-ketobutyrate to yield propionyl-CoA, which then goes into the methylcitrate pathway, eventually resulting in succinate (Jorgensen *et al.*, 1995a). Several of the reactions in this pathway are identified as transcriptionally controlled by the algorithm (2-methylcitrate synthase, 2-methylcitrate dehydratase, 2-methylisocitrate dehydratase, and methylisocitrate lyase). This finding strongly supports that the transsulfuration pathway is the dominating pathway for cysteine biosynthesis, even though the enzymes for the energetically more efficient direct sulfhydrylation pathway have been identified in *P. chrysogenum* (Ostergaard *et al.*, 1998).

In conclusion, the RAVEN Toolbox enables rapid reconstruction of high-quality models, which is not possible using a traditional manual approach. It is the first software which can be used to drive the model reconstruction process and which also contains extensive functions for simulations and analysis of results. The software and workflow was validated by reconstructing the first GEM for the industrially important fungi *P. chrysogenum*. This GEM was then used to gain novel insights in penicillin biosynthesis, and for suggesting metabolic engineering targets for increased penicillin yield.

4.2 GEMs applied to human health and disease

Abnormal metabolic states are at the origin of many diseases, such as diabetes, hypertension, heart diseases and cancer. Cancer and coronary diseases are the two main causes of death in the developed countries. It is expected that by 2030 close to 200 million persons (33% of the

total population) will be obese in the EU alone, and many of these will have one or more of the following co-morbidities: diabetes, hypertension, heart disease and increased risk of cancer. The direct (medical treatment) and indirect (inability to work) costs are estimated to amount to more than €100 billion per year (Rokholm *et al.*, 2010; Caveney *et al.*, 2011). These are all complex diseases, in the sense that they are the result of a large number of interacting molecular factors. This speaks in favour of taking a holistic approach rather than a more traditional reductionist one. Genome-scale metabolic modelling can therefore be a promising methodology for study of this type of diseases, but there are still obstacles to overcome.

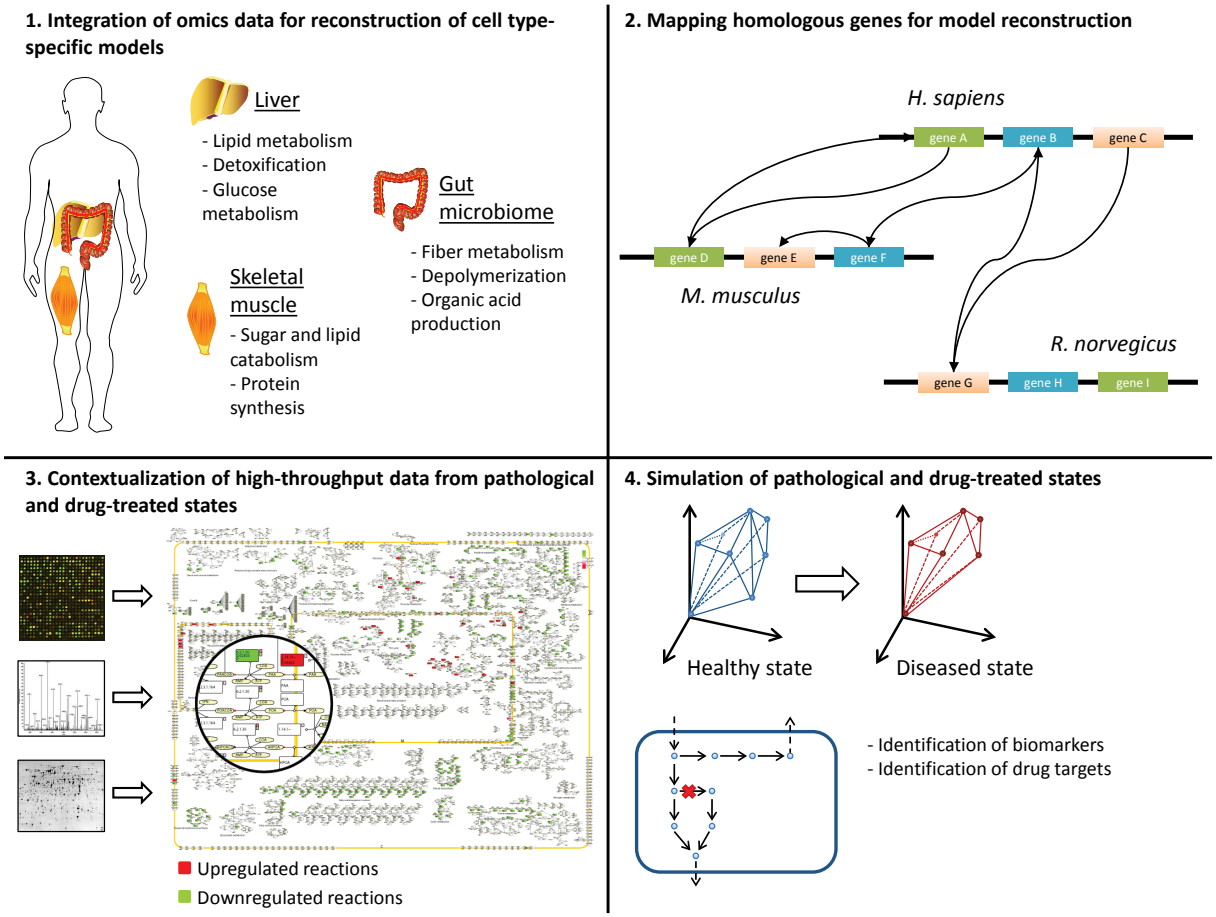


Figure 4-8. Overview of applications of GEMs in human health and disease. The available publications which make use of GEMs to study human health and disease can be categorized as follows: **1)** Integration of high-throughput data for model construction. Several algorithms have been developed which use different types of omics data to reconstruct cell type-specific models as subsets from a generic model (Becker and Palsson, 2008; Shlomi *et al.*, 2008; Wang *et al.*, 2012). There have also been attempts to integrate human GEMs with microbial GEMs, either pathogens or gut microbiota (Bordbar *et al.*, 2010; Heinken *et al.*, 2013). **2)** Mapping homologous genes for model construction. Much of medical research is carried out on mammals other than human. The high degree of homology between mammalian genes allows for reconstruction of GEMs for other mammals based on human GEMs (Sheikh *et al.*, 2005; Seo and Lewin, 2009). **3)** Contextualization of high-throughput data from pathological and drug-treated states. The work in this category uses GEMs as scaffolds for data analysis, rather than for predictive simulations. One such example is a study of the interactions between the tuberculosis bacteria and its host cell (Bordbar *et al.*, 2010). **4)** Simulation of pathological and drug-treated states. The work in the category uses GEMs for modelling, for example using FBA. Examples include simulations of hereditary metabolic diseases (Shlomi *et al.*, 2009) and simulations of the effect of potential cytostatic drugs (Folger *et al.*, 2011). Partly adapted from Bordbar and Palsson (2012).

Many of the methods developed for microorganisms, such as FBA, rely on the definition of a cellular objective. This is normally to grow as fast as possible given the available substrates. Since human cells do not grow uncontrollably those methods are not directly applicable. This

has led to that much of the work on genome-scale modelling of human cells has been focused on cancer; where the cells actually do grow uncontrollably (Folger *et al.*, 2011; Shlomi *et al.*, 2011). Another issue, which is extensively discussed in section 3.3.2, is that different cell types have different phenotypes even though they share the same genotype. There is still no clear workflow for how to generate cell type-specific models and to ensure that they are of high quality. Because of these issues, and others, the application of genome-scale metabolic modelling to human health and disease is a less mature science compared to when applied to microbial systems. Much of the work is therefore still centred on method development. Bordbar and Palsson (2012) categorized the publications which use GEMs to study human metabolism into the following four categories: 1) integration of high-throughput data for model construction, 2) mapping homologous genes for model construction, 3) contextualization of high-throughput data from pathological and drug-treated states, 4) simulation of pathological and drug-treated states. Figure 4-8 describes these categories in more detail.

In this section I present some of my work about genome-scale metabolic modelling applied to human health and disease. In **Paper IV** we developed an algorithm to reconstruct cell type-specific active metabolic networks based on different types of omics data. We then generated a large number of GEMs for cancers and their corresponding healthy cell types, and performed a statistical analysis to identify metabolic subnetworks which were more prominent in cancers. In **Paper V** we reconstructed a GEM for adipocytes and used it to study metabolic features associated to obesity. In **Paper VI** we reconstructed a GEM for hepatocytes and used it to study metabolic features associated to non-alcoholic fatty liver disease.

4.2.1 Paper IV: Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT

In **Paper IV** we describe the generation of genome-scale active metabolic networks for 69 different cell types and 16 cancer types using the INIT (Integrative Network Inference for Tissues) algorithm. We first built a generic database of human metabolism by merging and curating previously available GEMs and databases. We then developed an algorithm which integrates several types of omics data in order to generate active metabolic networks, catalogues of the metabolic reactions that are likely to be active in a given cell type, from this database. These networks represent a resource which can form the basis for simulation of metabolic interactions between organs, or act as scaffolds for interpretation of high-throughput data. Lastly, we used these networks for a comparative analysis between cancer types and healthy cell types. This allowed for identification of cancer-specific metabolic features which constitute potential drug targets for cancer treatment.

In order to provide a reliable and up to date GEM template for our tissue/cell type-specific metabolic networks, we first constructed the Human Metabolic Reaction (HMR) database, containing the elements of the previously published generic genome-scale human metabolic models (Duarte *et al.*, 2007; Ma *et al.*, 2007), as well as the HumanCyc and KEGG (Ogata *et al.*, 1999; Romero *et al.*, 2005) databases. The HMR database has a hierarchical structure in which the genes are at the top and are linked to information about their tissue-specific expression profiles reported by Su *et al.* (2004) via BioGPS (Wu *et al.*, 2009). Each gene is linked to its different splicing variants, which in turn are linked to their corresponding proteins. Each protein is linked to its tissue specific abundances in the Human Protein Atlas (HPA) (Berglund *et al.*, 2008) and to the reactions they catalyse. The reactions are linked to

metabolites, which themselves are linked to their tissue specific information collected from the Human Metabolome Database (HMDB) (Wishart *et al.*, 2007). In order to have an unambiguous characterization of metabolites and reactions, KEGG and InChI identifiers were used for standardization. Each reaction was assigned to one or several of the eight compartments included in the HMR database: nucleus, cytosol, endoplasmic reticulum, Golgi apparatus, peroxisomes, lysosomes, mitochondria and extracellular. In cases where the subcellular localization was absent from the template models it was inferred from immunohistochemical (IHC) staining in the HPA. For enzymes that were not in the HPA, Swiss-Prot and GO were used to infer localization. The HMR database was used to generate a fully connected generic human GEM, which contained 4,137 metabolites (3,397 unique), 5,535 reactions (4,144 unique), and 1,512 metabolic genes.

As previously discussed, there have been several algorithms for reconstruction of cell type specific GEMs published (see section 3.3.2). The INIT algorithm was tailored to use data from the HPA as the main evidence source for assessing the presence or absence of metabolic enzymes in each of the human cell types that are present in the HPA. In the HPA project, cell type specific high quality proteomic data is being generated based on immunohistochemistry (Uhlen *et al.*, 2005; Berglund *et al.*, 2008; Uhlen *et al.*, 2010). Tissue specific gene expression (Su *et al.*, 2004) was used as an additional source of evidence. Figure 4-9 show how INIT is used to select a subset of reactions from a generic model. The problem was formulated so that all reactions in the resulting model are able to carry flux. Instead of imposing the steady state condition for all the internal metabolites, as it is usually done, we allowed for a small positive net accumulation rate. The reason for this was that we preferred to have a network able to synthesize molecules such as NADH or NADPH, rather than only being able to use them as cofactors. If a metabolite was present in a cell type (according to the HMDB) a positive net production of this metabolite was imposed on the network in order to assure that all the reactions necessary for its synthesis are included in the cell type-specific model. A distinct advantage of the INIT algorithm compared to existing approaches such as GIMME (Becker and Palsson, 2008) or MBA (Jerby *et al.*, 2010) (see section 3.3.2) is that it makes no predetermined classification of enzymes as either present or absent in the resulting model. In order to validate the output of the algorithm, the automatically generated hepatocyte model was compared with HepatoNet1 (Gille *et al.*, 2010), a manually curated and functional model of hepatocyte metabolism.

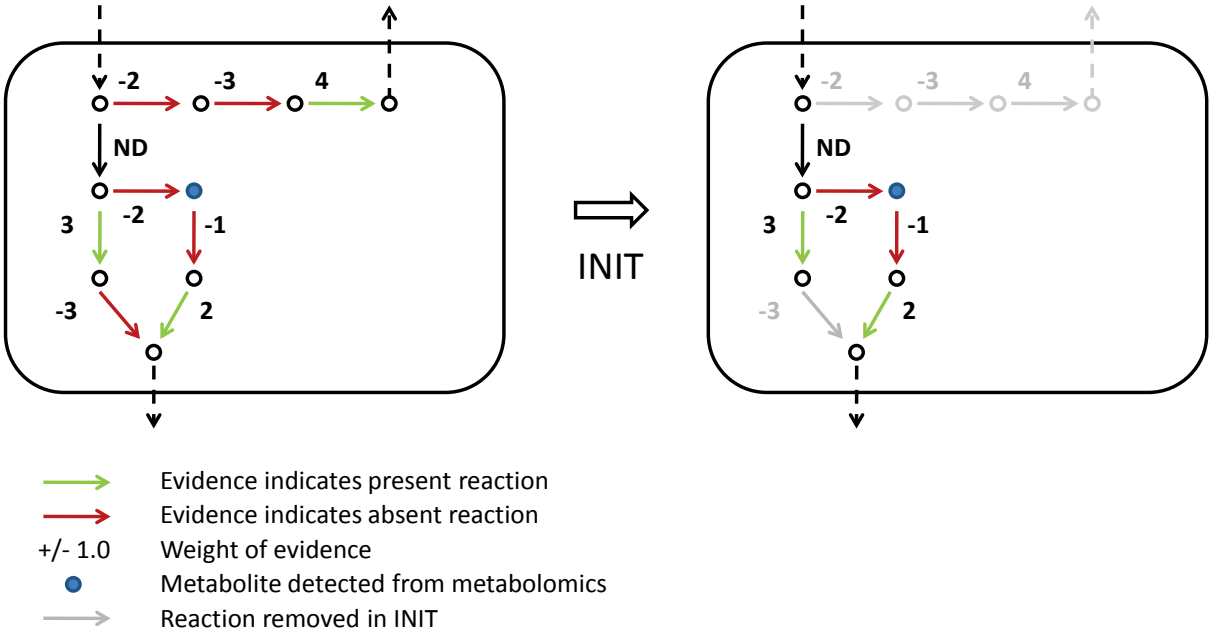


Figure 4-9. Principle of the INIT algorithm. Depending on the evidence for presence/absence of a given enzyme/gene in a cell type, a score is calculated for the reaction(s) catalysed by that enzyme. In the example above, reactions for which the evidence indicates that it should exist in the cell type are coloured green. The opposite is true for the reactions coloured red. The aim of the algorithm is to find a subnetwork in which the involved genes/proteins have strong evidence supporting their presence, but at the same time maintaining a connected and functional network. This is done by maximizing for the sum of evidence scores under the constraint that all the included reactions should be able to carry a flux, and that all the metabolites observed experimentally (metabolite coloured blue in the example above) should be synthesized from precursors that the cell is known to take up. This is then implemented as a mixed-integer linear programming problem (MILP). In the example above, the three top reactions are excluded by INIT; despite that the last of them has evidence strongly supported its presence. This is because two other reactions would have to be included in order for it to be connected, and the net score would then be negative (4-3-2<0). The path via the blue metabolite to the end product at the bottom is also negative (2 -1-2<0). However, since the blue metabolite is detected by metabolomics to exist in the cell it has to be synthesized in at least one reaction. The remaining pathway from the blue metabolite is then positive (2-1>0), and should therefore be included. The RAVEN Toolbox (**Paper III**) was used to perform the optimization and generate the cell type specific active networks.

Since the Warburg effect was observed at the beginning of the 20th century, it has been known that cancer cells show characteristic metabolic features that make them different from healthy cells (Koppenol *et al.*, 2011). This supposed metabolic similarity between cancer cells justified the development of a generic cancer GEM, which was used to identify potential drug targets against cancer proliferation (Folger *et al.*, 2011). We used INIT to infer active metabolic networks for 16 different cancer types, which can be compared with the 24 healthy cell types that they come from (there are several healthy cell types for some of the tissues associated to the cancers) in order to identify metabolic features that are characteristic of cancer. A hypergeometric test was used to identify genes and reactions that tended to be present in most of the cancer-specific active metabolic networks and absent in most of the original healthy cell types. The p-values obtained from the hypergeometric test were used to identify Reporter Metabolites (Patil and Nielsen, 2005) that are significantly more involved in the metabolism of cancer cells. These lists of genes, reactions and metabolites are cancer-specific features that are likely to be playing a specific role in proliferation of cancer cells and could be potential drug targets. Our comparative analysis between two sets of active metabolic networks can be seen as a high-throughput hypothesis generation method. These hypotheses are not based on mere correlations between cancer and the presence of a particular protein, but being based on the underlying metabolic network structure, and hereby our analysis provides a mechanistic interpretation about the possible role of each identified feature on the proliferation of cancer.

One of the most significant results from the Reporter Metabolites analysis is a much more pronounced metabolism of polyamines (PAs) such as spermidine, spermine, and putrescine in cancer cells. PAs play a variety of roles, of which several are related to oxidative stress, prevention and suppression of necrosis (Eisenberg *et al.*, 2009). PAs have long been known to be of particular importance for rapidly proliferating cells, and as such its transport and synthesis have been thoroughly investigated as anti-cancer drug targets (Seiler, 2003b). Inhibition of single enzymes in the PA synthesis pathway has proved disappointing, due to extensive regulation of the system and use of exogenous PAs by the cancer cells. Second generation drugs instead work by targeting the transport system, by structural homology to the PAs themselves, or by linking other antineoplastic drugs to the PAs (Seiler, 2003a).

Another high-ranking target is the isoprenoid biosynthesis pathway, in particular the intermediate geranylgeranyl diphosphate. This metabolite has been shown to promote oncogenic events due to its role in prenylation of important cancer proteins such as Ras and Rho GTPases (Sebti and Hamilton, 2000). Several drugs have therefore been developed to target the prenylation process (Philips and Cox, 2007) or the biosynthesis of geranylgeranyl diphosphate (Dudakovic *et al.*, 2011).

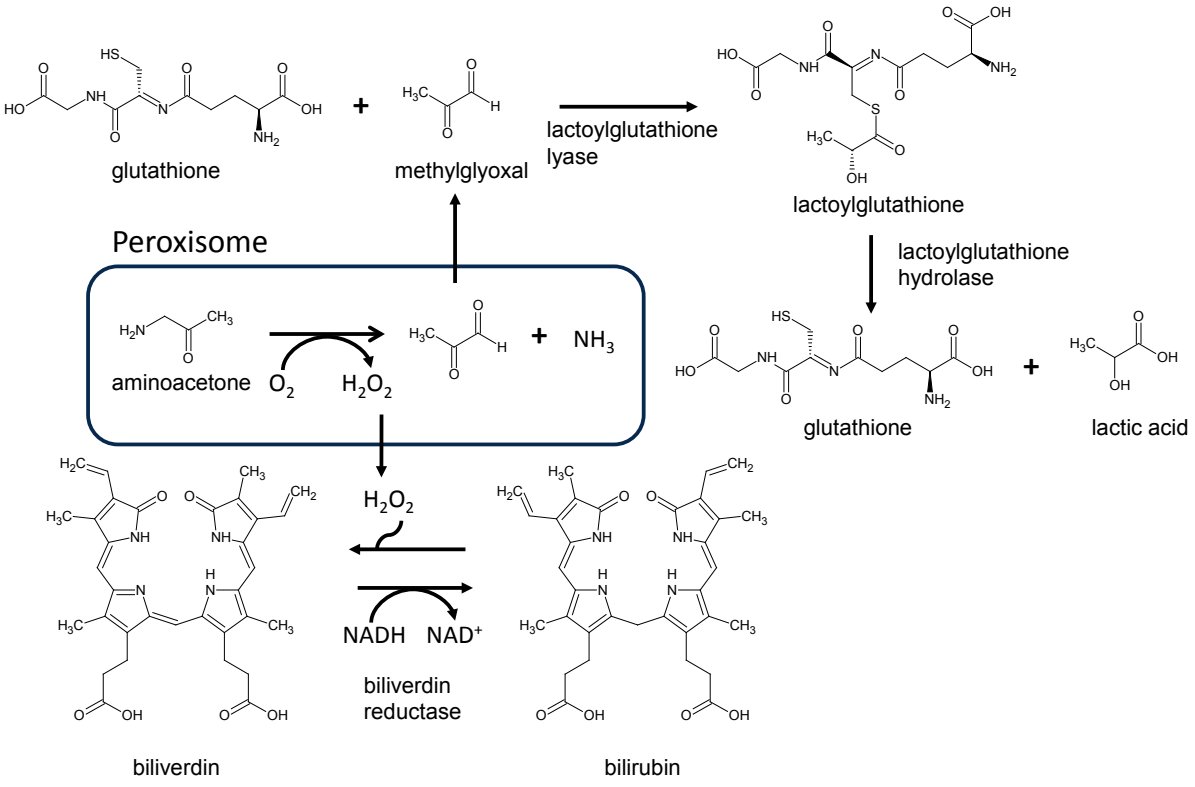


Figure 4-10. Metabolic subnetwork identified as being significantly more prominent in cancer tissues compared to their corresponding healthy tissues. Aminoacetone, which is a toxic by-product of amino acid catabolism, is converted to toxic methylglyoxal in a reaction that also result in hydrogen peroxide. The toxicity of methylglyoxal is relieved by two reaction steps involving ligation to glutathione and resulting in lactic acid. The generated hydrogen peroxide is taken care of by the enzyme biliverdin reductase. This is an example of how network-based analysis can lead to a more mechanistic interpretation of data. Figure taken from Agren *et al.* (2012).

A third prominent group among the Reporter Metabolites is prostaglandins and leukotrienes together with the intermediate HPETE. These autocrine compounds are synthesized from arachidonic acid and are elevated in connection with inflammation. They have been shown to aid in cancer progression by promoting metastasis and by influencing the immune system

(Schneider and Pozzi, 2011). Of particular interest is prostaglandin E2, where both the synthesis and degradation have been investigated as promising targets for drug development (Eruslanov *et al.*, 2009).

The fact that so many of the identified targets correspond to well known and used drug targets indicates that the method is able to generate biologically relevant hypotheses. Of particular interest are therefore the Reporter Metabolites that are currently not targeted in cancer treatment. Among the top-scoring Reporter Metabolites we identified biliverdin and bilirubin (see Figure 4-10). Biliverdin reductase and the reactions catalysed by this enzyme also appear among the genes and reaction most enriched in the cancer networks. Biliverdin reductase is known to be a major physiologic cytoprotectant against oxidative stress (Baranano *et al.*, 2002). Cancer cells are known to be exposed to high oxidative stress resulting from the hydrogen peroxide generated during the oxidation of polyamines and other products of amino acid breakdown taking place in the peroxisome. Bilirubin is oxidized to biliverdin by hydrogen peroxide and subsequently reduced back to bilirubin by biliverdin reductase. This mechanism has been proven to be a major relief system for oxidative stress and could be considered a potential target against cancer proliferation. One of the hydrogen peroxide generating reactions taking place in the peroxisomes is the transformation of aminoacetone, which is an intermediate in the degradation of glycine, into methylglyoxal. Another source of methylglyoxal in cancer cells is from gluconeogenesis (Titov *et al.*, 2010). Methylglyoxal is known to be a toxic compound (Kalapos, 1994) that has been proven to induce apoptosis in some cancer cell lines (Kang *et al.*, 1996). Methylglyoxal also appeared among our top scoring reporter metabolites and both the gene coding for lactoylglutathione lyase (an enzyme that transforms methylglyoxal and glutathione into lactoylglutathione) and its associated reactions appear among the most enriched genes and reactions in the cancer active metabolic networks. Lactoylglutathione is further transformed into glutathione and lactic acid by the enzyme lactoylglutathione hydrolase (which also shows a significant enrichment in cancer metabolic networks with a p-value of 2e-3). The mentioned two enzymes seem to be playing a relevant role in relieving the toxicity generated by methylglyoxal and could be potential drug targets against cancer proliferation. Targeting these enzymes would have the same effect on cancer cells as using methylglyoxal as a drug, but the advantage is that there would be no toxicity effects of methylglyoxal on healthy tissues.

In conclusion, the HMR database represents the most comprehensive generic human GEM (see **Paper VI**) to date and is an important resource in itself. The INIT algorithm was demonstrated to automatically generate active metabolic networks which were similar in scope compared to a high-quality manually reconstructed model. This was made possible by use of the enormous amount of proteomics data generated within the HPA project. The INIT algorithm was then applied to reconstruct GEMs for 69 cell types and 16 cancers; models which can form the basis for future work on modelling of human metabolism. The content of these models was analysed and we were able to identify a number of metabolic subnetworks which were significantly more prominent in cancers compared to their healthy counterparts. These subnetworks contained many known drug targets, but we were also able to identify a number of novel drug targets based on our analysis.

4.2.2 Paper V: Global analysis of human adipocyte metabolism in response to obesity

Adipose tissue dysfunction or overload of its lipid storage capacity can lead to wide range of diseases (e.g. immunological and inflammatory diseases), including metabolic diseases such as obesity (Lago *et al.*, 2007; Auffray *et al.*, 2009). An increased understanding of the mechanisms behind obesity and related diseases would provide valuable insights into their etiology and pathogenesis, and could lead to new treatment strategies. In **Paper V** we reconstructed a GEM for adipocytes based primarily on adipocyte specific proteome data generated within the Human Protein Atlas (HPA) project (Uhlen *et al.*, 2010).

We expanded the coverage of the HPA to include the protein profiles of adipocytes found in breast and two different soft tissues and examined the spatial distribution and the relative abundance of proteins encoded by 14,077 genes in these tissues. A total of 17,296 affinity-purified antibodies were generated and used for immunohistochemical staining of tissue micro array blocks. The proteome data was merged with previously published proteome data on adipocytes in order to increase the coverage. In total, we have proteome evidence for the presence/absence of proteins associated with 14,337 genes in adipocytes. The proteomics analysis resulted in evidence for presence of proteins associated with 7,340 genes.

As discussed in section 3.3, the subcellular localization of reactions has large implications on the functionality of GEMs, as only a portion of metabolites can be transported between compartments. Furthermore, compartments can be individually redox and/or energy balanced. The HPA includes subcellular profiling data using immunofluorescence-based confocal microscopy in three human cancer cell lines of different origin. Here, proteins were classified into eight different compartments following our HMA standard (**Paper IV**): cytosol, nucleus, endoplasmic reticulum (ER), Golgi apparatus (GA), peroxisome, lysosome, mitochondria and extracellular space. Reactions were assigned to compartments through their association with proteins in these different compartments.

In order to reconstruct a GEM for adipocytes, biochemical and genetic evidence was combined with data on protein expression and localization. HepatoNet1, a GEM for hepatocytes which is reconstructed based on the manual evaluation of the original scientific literature (Gille *et al.*, 2010), was used as a starting point for our reconstruction process and used to generate an initial candidate list of network components. Firstly, metabolism of lipids and lipoproteins in Reactome, a manually curated and peer-reviewed pathway database (Croft *et al.*, 2011), was merged into HepatoNet1. Secondly, the resulting network was combined with the evidence-based generic human models Recon1 (Duarte *et al.*, 2007) and the compartmentalized EHMN (Hao *et al.*, 2010). This combined reaction list resulted in an updated version of our Human Metabolic Reaction (HMR) database (**Paper IV**). The HMR database contains 6,049 metabolites in eight different compartments (3,162 unique metabolites), 8,107 reactions and 3,668 genes associated to those reactions. Thirdly, the existence of each protein coding gene associated to a reaction in HMR was assessed for the presence or absence in adipocytes using previously published and the here generated adipocyte-specific proteome data. This process provided us with a list of reactions that occur in adipocytes. Gaps in the resulting network were filled using the updated HMR database, public databases such as KEGG (Kanehisa *et al.*, 2010) and HumanCyc (Romero *et al.*, 2005) and manual evaluation of the literature about adipocyte metabolism. The RAVEN Toolbox (**Paper III**) was used for gap filling and quality control. This gap filling resulted in generation of iAdipocytes1809, which is a fully functional and connected GEM for adipocytes. In iAdipocytes1809, individual metabolites, rather than generic pool metabolites, for 59 fatty

acids (FAs) have been used. This allowed us to incorporate measured concentrations of different FAs in human plasma and adipocytes into the model. Figure 4-11 shows the reconstruction workflow.

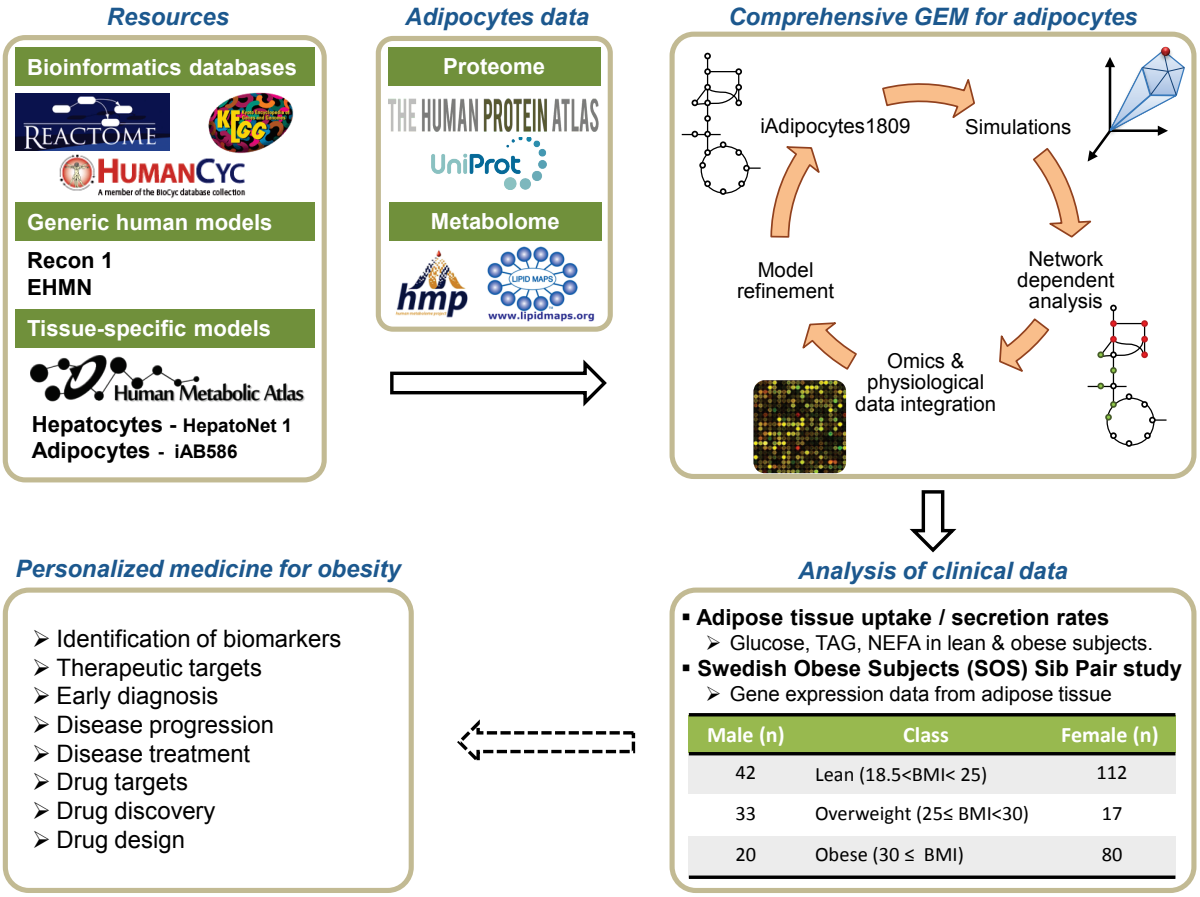


Figure 4-11. Schematic illustration of how a GEM for adipocytes may provide links between molecular processes and subject phenotypes. Here the GEM iAdipocytes1809 was reconstructed through the use of proteome, metabolome, lipidome and transcriptome data, literature based models (Recon 1, Edinburg Human Metabolic Network (EHMN) and HepatoNet1), and public resources (Reactome, HumanCyc, KEGG and the Human Metabolic Atlas). We first performed global protein profiling of adipocytes using antibodies generated within the Human Protein Atlas (HPA). We further used information on metabolome and lipidome data from the Human Metabolome Database (HMDB) and LIPID MAPS Lipidomics Gateway, respectively. The model was then used for the analysis of gene expression data obtained from subjects with different body mass indexes in the Swedish Obese Subjects (SOS) Sib Pair study and other adipose tissue relevant clinical data such as uptake/secretion rates in lean and obese subjects (see below).

In iAdipocytes1809, 59 different common long and very long chain FAs in human plasma can be taken up as NEFAs and lipoproteins. The lipid related functionality of iAdipocytes1809 is summarized in Figure 4-12. The GEM was subject to extensive quality control by using the RAVEN Toolbox (**Paper III**), following the workflow in Figure 3-3. Even a well-connected, thermodynamically correct and balanced model may not be able to perform all relevant metabolic functions, or it may be able to perform functions that it should not do (such as synthesis of essential amino acids or fatty acids). The model was therefore validated for 250 known metabolic functions of adipocytes, adapted from the definitions provided in connection with setting up HepatoNet1 (Gille *et al.*, 2010).

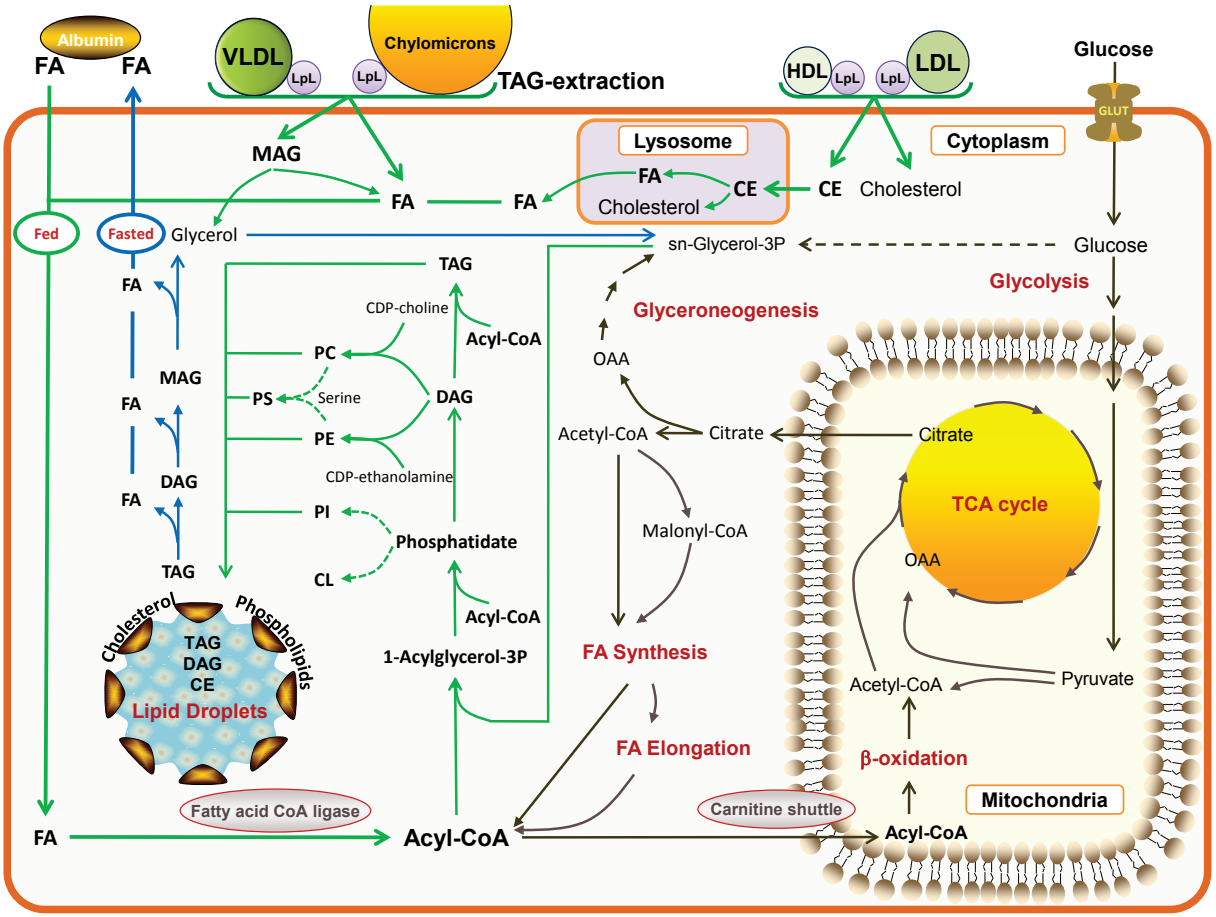


Figure 4-12. Summary of the capabilities of iAdipocytes1809. Adipocytes store lipid mainly in the form of triacylglycerols (TAGs) and cholesterol esters (CEs). They form lipid droplets (LDs) in the post-prandial state (green arrows) and release them by degrading LDs in the post-absorptive state (blue arrows) in order to provide energy for other tissues. The released fatty acids (FAs) from adipocytes are transported to other tissues by albumin. The FAs are taken up from non-esterified FAs (NEFA) and lipoproteins, including chylomicrons, very-low-density lipoprotein (VLDL) and CEs together with cholesterol are taken up with low-density lipoproteins (LDL) and high-density lipoproteins (HDL) through lipoprotein lipase (LPL). CEs taken up from lipoproteins are degraded to cholesterol and FAs in lysosomes and transported to the endoplasmic reticulum and cytosol to be stored in LDs. Adipocytes also take up glucose to be used in the de novo synthesis of FAs (black arrows) that occurs at low level in adipocytes. LDs are rich in TAGs, CEs and an unknown neutral lipid that migrated between CEs and TAGs, ether neutral lipid monoalk(en)yl diacylglycerol (MADAG). LDs also contain small amounts of free FAs, cholesterol and phospholipids including phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidylinositol (PI), ether-linked phosphatidylcholine (ePC), ether-linked phosphatidylethanolamine (ePE), lyso phosphatidylcholine (LPC), lysophosphatidylethanolamine (LPE), phosphatidylserine (PS) and sphingomyelin (SM). Formation of ePC, ePE, LPC, LPE and SM is included in iAdipocytes1809 but is not shown in figure.

The function of iAdipocytes1809 was tested by estimating the formation of LDs based on clinical data for lean and obese subjects. Recently, McQuaid *et al.* (2011) measured the delivery and transport of FAs in adipose tissue using multiple and simultaneous stable-isotope FA tracers in lean and obese subjects groups over 24 hours period. Even though abdominally obese subjects have greater adipose tissue mass than control lean subjects, the rates of delivery of NEFAs were downregulated in obese subjects. Based on measurements of the uptake of glucose and TAG and the release of NEFAs over a 24 hour period we simulated the change in LD size. We found from our simulations that lean subjects have large dynamic changes in LD formation compared with obese subjects (see Figure 4-13c). Furthermore, we predicted a lower acetyl-CoA production in obese subjects (see Figure 4-13d).

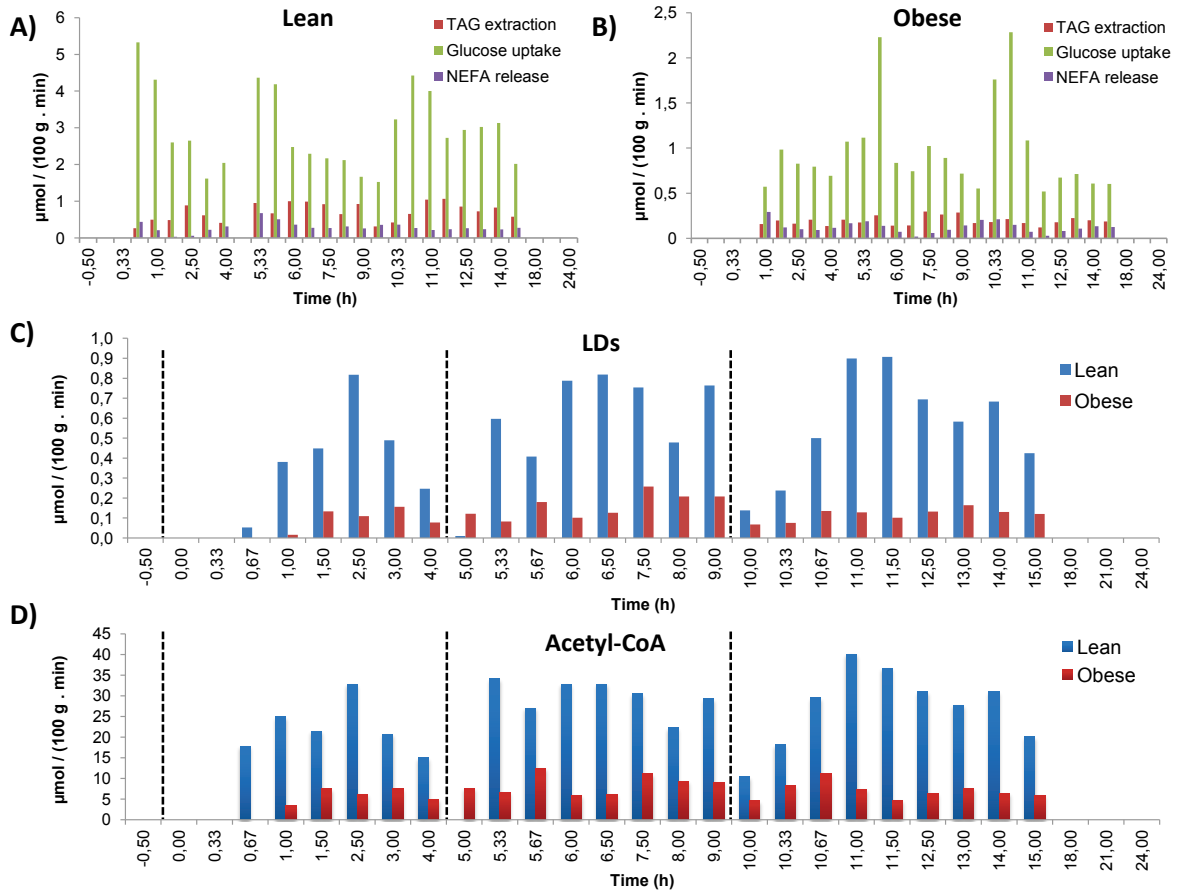


Figure 4-13. Simulated lipid droplet and acetyl-CoA production. Uptake rates for glucose and TAGs and release rates for NEFAs in adipocytes for lean (**A**) and obese (**B**) subjects were used as lower and upper bounds for input reactions (McQuaid *et al.*, 2011) together with amino acid uptake rates (Patterson *et al.*, 2002), and the amounts of LDs (**C**) and acetyl-CoA (**D**) were predicted over a 24 h period. The dashed lines at time 0, 5, 10 hours represent breakfast, lunch and dinner, respectively, for each participant of the study. Figure taken from Mardinoglu *et al.* (2013).

We then employed the iAdipose1809 GEM for the analysis of microarrays which profile the gene expression from subcutaneous adipose tissue of subjects from the Swedish Obese Subjects (SOS) Sib Pair Study. This study includes nuclear families with BMI–discordant sibling pairs (BMI difference $\geq$ 10 kg/m^2). Besides the gene expression data from the SOS Sib Pair Study, additional clinical data (e.g. plasma and WAT lipid concentrations) were also incorporated into the model. By integrating gene expression data and adipose tissue uptake/secretion rates with the reconstructed GEM, we identified metabolic differences between individuals with different BMIs by using the concept of Reporter Metabolites (Patil and Nielsen, 2005) and transcriptionally controlled reaction fluxes (**Paper II**).

The results from this analysis showed that the following pathway fluxes were transcriptionally downregulated in obese subjects: uptake of glucose, uptake of FAs, oxidative phosphorylation, mitochondrial and peroxisomal β-oxidation, FA metabolism, and TCA cycle. Furthermore, fluxes associated with beta-alanine metabolism were found to be transcriptionally downregulated in obese subjects. Previously it has been reported that blood flow, glucose uptake, release of NEFA and the extraction of TAG from plasma was significantly lower in abdominally obese subjects compared to lean subjects (McQuaid *et al.*, 2011). Most of these pathways are linked with mitochondrial dysfunction. Several therapeutic interventions, including antioxidants and chemical uncoupler treatments, have been shown to improve mitochondrial dysfunction (Kusminski and Scherer, 2012).

Mitochondrial acetyl-CoA plays a central role in different pathways in the mitochondria, where it reacts with oxaloacetate to form citrate. Citrate can then be transported from the mitochondria to the cytosol where it is participating in FA synthesis (Dean *et al.*, 2009). Acetyl-CoA derived through other principal sources, including degradation of amino acid and ketone bodies and fatty acid oxidation processes are insufficient for FA synthesis. Increasing the acetyl-CoA concentration and eventually FA synthesis in adipose tissue of obese subjects results in whole body regulation of metabolism as reported by Cao *et al.* (2008). We therefore propose to boost the metabolic activity of mitochondria in the adipocytes of obese subjects by aiming to increase the availability of mitochondrial acetyl-CoA.

As previously mentioned, beta-alanine metabolism came out as significantly changed between healthy and obese. The effect of beta-alanine as a dietary supplement was previously examined in football players and it is reported that it has effect on lean tissue accruement and body fat composition (Hoffman *et al.*, 2006). Furthermore rat studies reported that beta-alanine decreases the lipoprotein lipase (LPL) enzyme activity in adipose tissue which may help to decrease the uptake of FAs to be stored in adipocytes (Prabha *et al.*, 1988). Our results suggest that increasing the level of beta-alanine in obese subjects may help to decrease the fat composition in obese subjects.

Another high-ranking target for upregulated genes is ganglioside GM2. Gangliosides, one of the major glycosphingolipids in mammals, play major roles as mediators for cell to cell or cell to matrix recognition and regulate the transmembrane signal transducers and cell proliferation. Gangliosides in adipose tissues are also associated with insulin signalling mechanisms and it is reported that series of gangliosides GM2, GM1, and GD1a are dramatically increased in adipose tissues of obese mice (Tanabe *et al.*, 2009).

A third prominent group among the Reporter Metabolites for upregulated genes is the degradation products of heparan sulfate proteoglycans (HSPG) and keratan sulfate. These compounds are classified as glycosaminoglycans and attach to cell surface or extracellular matrix proteins. Keratan sulfate, a biomarker of proteoglycan degradation, can be expressed from stem cells in human SAT and its relevance with obesity has been reported earlier. It has been reported that catalytically active adipose tissue lipoprotein lipase (LPL) attaches to HSPG at the luminal surface of vascular endothelium (Olivecrona and Beisiegel, 1997; Lafontan, 2008) and hydrolyse the TAGs for uptake of FAs into the cell. The LPL moves between individual HSPG chains within the layer and this creates a high concentration of LPL along the surface layer of HSPG chains (Lookene *et al.*, 1996). In the presence of heparin more LPL is secreted and increased secretion was balanced by decreased degradation of LPL. There are special mechanisms that inhibit LPL and one mechanism is that LPL forms complexes with FAs (Bengtsson and Olivecrona, 1980). During the LPL hydrolysis and accumulation of FAs in the cells, the LPL is sequestered into enzyme FA complexes, lipolysis is reduced and eventually the binding of LPL to heparan sulfate is broken. If a high-affinity ligand (e.g. FAs, heparin, apoCII) is available, the LPL detaches from the cell surface to heparan sulfate chains and without ligand in the medium, the LPL recycles into the cells where it is degraded. Furthermore, several studies have reported that more sulfated polysaccharide chains increase the affinity for binding of LPL (Olivecrona and Olivecrona, 2009). One possible intervention strategy could therefore be to try to reduce the degradation rate of HSPG.

In conclusion, the first human GEM with extensive lipid metabolism was reconstructed. This was made possible by close collaboration with groups that generated large-scale proteomics data for adipocytes and transcriptomics data for healthy/obese siblings. The model could correctly capture the reduced dynamics of lipid droplet formation in obese subjects and we

could see that this was associated to mitochondrial dysfunction. The model was then used together with the algorithm developed in **Paper II** and the Reporter Metabolites algorithm to identify metabolic differences between healthy and obese siblings. This led us to hypothesize that obesity could be treated with interventions aiming a reactivating the mitochondria by increasing the availability of acetyl-CoA. An alternative approach was to target the degradation of heparan sulfate proteoglycans.

4.2.3 Paper VI: Identification of serine deficiency in non-alcoholic fatty liver disease through genome-scale metabolic modelling

Hepatocytes have a wide range of physiological functions, including production of bile and hormones, removal of toxic substances, homeostatic regulation of the plasma constituents and synthesis of most plasma proteins (Gille *et al.*, 2010). They are the most metabolically active cell types in human and play a major role in overall human metabolism. Deficiency or alterations in the metabolism of hepatocytes can lead to complicated disorders such as hepatitis, non-alcoholic fatty liver disease (NAFLD), cirrhosis and liver cancer, which are serious threats to public health (Baffy *et al.*, 2012). NAFLD is considered as the hepatic manifestation of obesity and metabolic syndrome, and encompasses a spectrum of pathological changes; ranging from simple fatty liver (FL) to non-alcoholic steatohepatitis (NASH) (Neuschwander-Tetri and Caldwell, 2003).

In **Paper VI** we reconstructed a consensus GEM for hepatocytes and applied it in order to suggest potential biomarkers and therapeutic targets for NAFLD. In parallel to this we built on the results from **Paper IV** and **Paper V** in order to reconstruct a generic human GEM.

Several generic (non-cell type-specific) GEMs for human metabolism have been previously constructed (as discussed in **Paper IV** and section 3.3.2). One such generic model is the HMR database presented in **Paper IV**. However, neither of these generic networks contain extensive lipid metabolism, which is necessary in order to study the effect of lipids on the underlying molecular mechanism of NAFLD. In **Paper V** we reconstructed as GEM for adipocytes with a strong focus on lipid metabolism. In this paper we presented HMR 2.0, in which we had integrated all published human GEMs, the original HMR database, and the lipid metabolism from the adipocyte GEM from **Paper V**. The HMR 2.0 database is the largest biochemical reaction database for human metabolism in terms of number of reactions/genes/metabolites, as well as in terms of which parts of metabolism that are covered. This represents an important step forward since lipids have major effects on the development of several important metabolic diseases (Newgard, 2012). The functionality of the model was tested using the RAVEN Toolbox (**Paper III**), in the same manner as previously described for the adipocyte GEM.

A draft hepatocyte GEM was then reconstructed from a subset of HMR 2.0 based on proteomics data from HPA and by using the tINIT algorithm (**Paper IV** and Agren *et al.* (2013b)). Previously, several GEMs for hepatocytes, including HepatoNET 1 (Gille *et al.*, 2010), iLJ1046 (Jerby *et al.*, 2010), iAB676 (Bordbar *et al.*, 2011) and iHepatocyte1154 (**Paper IV**), have been reconstructed. The draft model was then expanded to contain all of the protein coding genes and associated reactions in the previously published liver models (Figure 4-14a). In addition to the proteomics data and reactions from previously published models, protein coding genes were also included based on transcriptomics data and for connectivity reasons (Figure 4-14b). Lastly, additional clinical data for plasma and hepatocyte lipid

concentrations for individual FAs were incorporated into the model, resulting in the final iHepatocytes2260 GEM.

iHepatocytes2260 differs from previously published hepatocyte GEMs primarily in terms of coverage in lipid metabolism. Among the new lipid related functions are uptake of the remnants of lipoproteins (chylomicrons, very-low-density lipoprotein (VLDL), low-density lipoproteins (LDL) and high-density lipoproteins (HDL)), the formation and degradation of lipid droplets (LDs) and secretion of synthesized lipoproteins (VLDL, LDL, HDL) (Figure 4-14c, see also the corresponding functionality for adipocytes in Figure 4-12). The model was validating by simulating 256 different biologically defined metabolic functions (e.g. the synthesis of FAs, amino acids, cholesterol and bile acids) that is known to occur in hepatocytes. Furthermore, the ability of the model for performing gluconeogenesis was demonstrated using experimentally measured secretion rates for glucose and albumin and uptake rates for glycerol, lactate, amino acids and FAs in primary rat hepatocytes (Chan *et al.*, 2003).

In the model reconstruction process, tINIT (Agren *et al.*, 2013b) identified 61 genes (out of the 3,673 genes in the HMR 2.0 database) which had to be integrated into the model in order to maintain the functionality, even though they had been reported to be non-expressed in hepatocytes according to the HPA. We then re-analysed the immunohistochemistry (IHC) data for these 61 proteins and found that 20 (33%) of these proteins actually show presence in hepatocytes. Initial disconcordant data were due to the suboptimal titration of the antibody, misinterpretation of weak IHC staining or due to interference with other cell types besides hepatocytes present in liver (e.g. kupffer cells and sinusoids). Nine (15%) of the investigated proteins showed more concordant results to the mathematic model when re-analysed using another antibody targeting the same protein. 15 proteins (25%) with negative IHC data were kept as negative in HPA since limited literature was available, and/or concordant results were seen in subsets of the remaining panel of tissues included in the HPA high-throughput set up. The remaining 17 proteins (28%) are believed to be inaccurately assessed by IHC due to technical issues, such as antigen recognition due to antigen conformational changes, fixation or sub optimal antibody. We think this is an excellent example of how a holistic view of metabolism can lead to biological insights, in this case as a targeted way to improve on the quality of experimental data.

NAFLD, and its most severe form NASH, is progressively diagnosed worldwide (Rector *et al.*, 2008). It is tightly associated with obesity, type 2 diabetes, insulin-resistance, and hypertension and represents a severe risk for development of cirrhosis and hepatocellular carcinoma (Ascha *et al.*, 2010). Despite its severe drawbacks, liver biopsy is still the most common procedure for diagnosing NASH (Machado and Cortez-Pinto, 2012). Thus, there is a need for identifying metabolic biomarkers to diagnose NASH, as well as to subcategorize the NAFLD patients without taking biopsies. A metabolic biomarker can be defined as a metabolite which is secreted to the blood where its level differs between two different states. We used the iHepatocytes2260 GEM as a scaffold for transcriptome analysis in an attempt to identify potential such biomarkers.

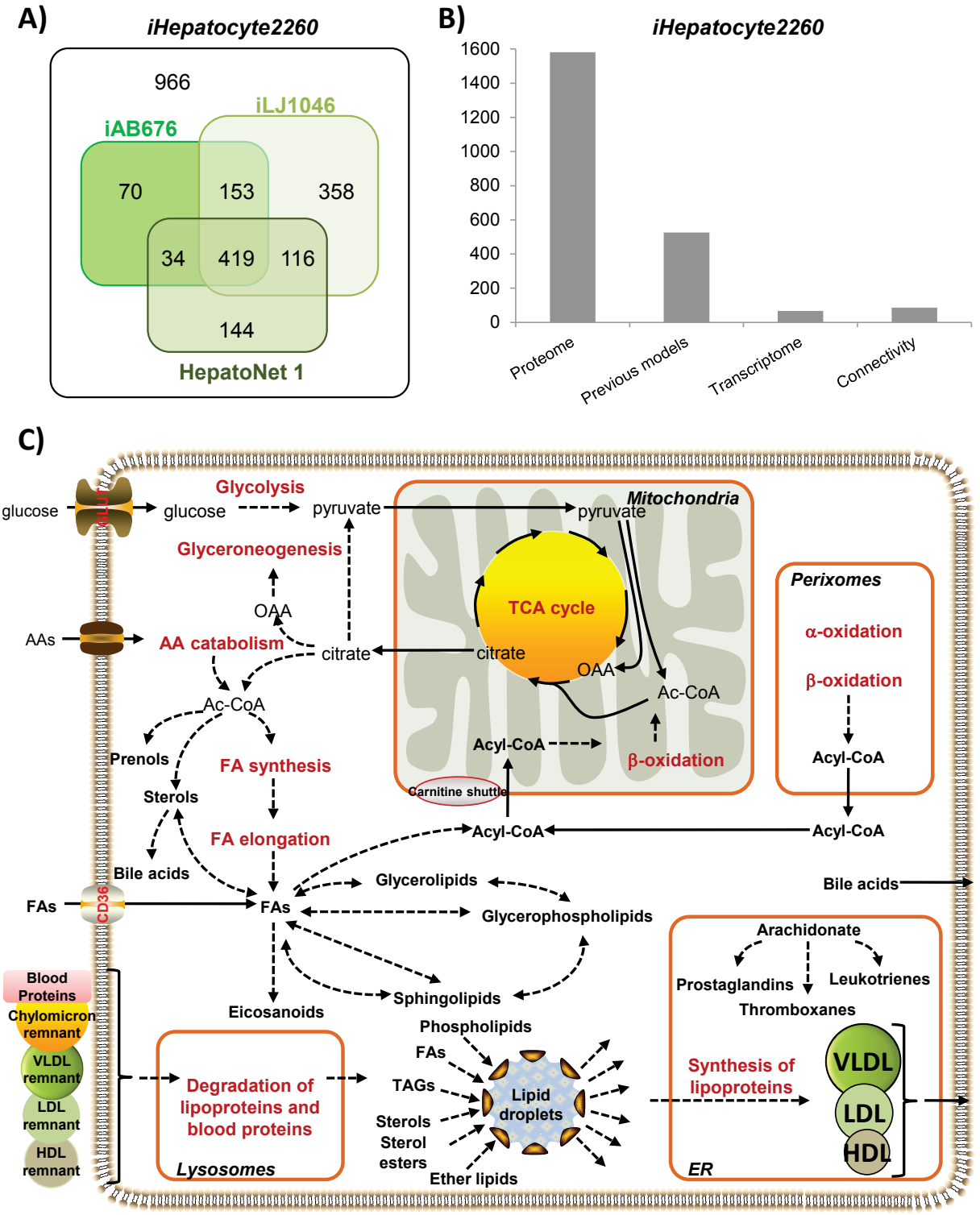


Figure 4-14. iHepatocytes2260 – a consensus GEM for hepatocytes. A) Venn diagram of the genes in iHepatocytes2260 and previously published hepatocytes GEMs. 966 new protein coding genes were included in iHepatocytes2260, primarily based on proteomics evidence provided by HPA. **B**) Genes and associated reactions were included based on proteome evidence, previously published models, transcriptome evidence, or for connectivity reasons. The number of genes included based on each category is shown. **C**) iHepatocytes2260 contains extensive lipid metabolism that is known to exist in hepatocytes, in addition to other known metabolic pathways. In the model, 59 different individual fatty acids are used, rather than generic pool names, in order to allow the integration of lipidomics data. The model can uptake the remnants of chylomicrons, very-low-density lipoprotein (VLDL), low-density lipoproteins (LDL) and high-density lipoproteins (HDL) and can form and degrade lipid droplets (LDs). Moreover the model can synthesis VLDL, LDL and HDL and secrete it to the blood. Some of the important elements of lipid metabolism are shown.

We retrieved liver gene expression data from 45 subjects out of which 19 were healthy, 10 steatotic, 9 had NASH with FL and 7 had NASH without FL (Fisher *et al.*, 2009; Lake *et al.*, 2011). We used the Piano package to compare gene expressions of NASH with and without FL to healthy subjects (Väremo *et al.*, 2013). The iHepatocyte2260 GEM was then used with the Reporter Metabolites algorithm in order to identify metabolites around which transcriptional changes occur between healthy and diseased subjects (Patil and Nielsen, 2005). A total of 50 statistically significant metabolites were such identified. In addition to several known subsystems involved in the progression of NASH, e.g. cholesterol biosynthesis, folate, vitamin B6, porphyrin, nucleotide, eicosanoid and amino acid metabolism (Greco *et al.*, 2008; Anstee and Day, 2012) we also identified several new Reporter Metabolites. These metabolites were involved in N-glycan metabolism and in the biosynthesis of the proteoglycan (PG) chondroitin sulfate (CS). PGs are composed of glycosaminoglycans, including CS and heparan sulfate (HS), and core proteins.

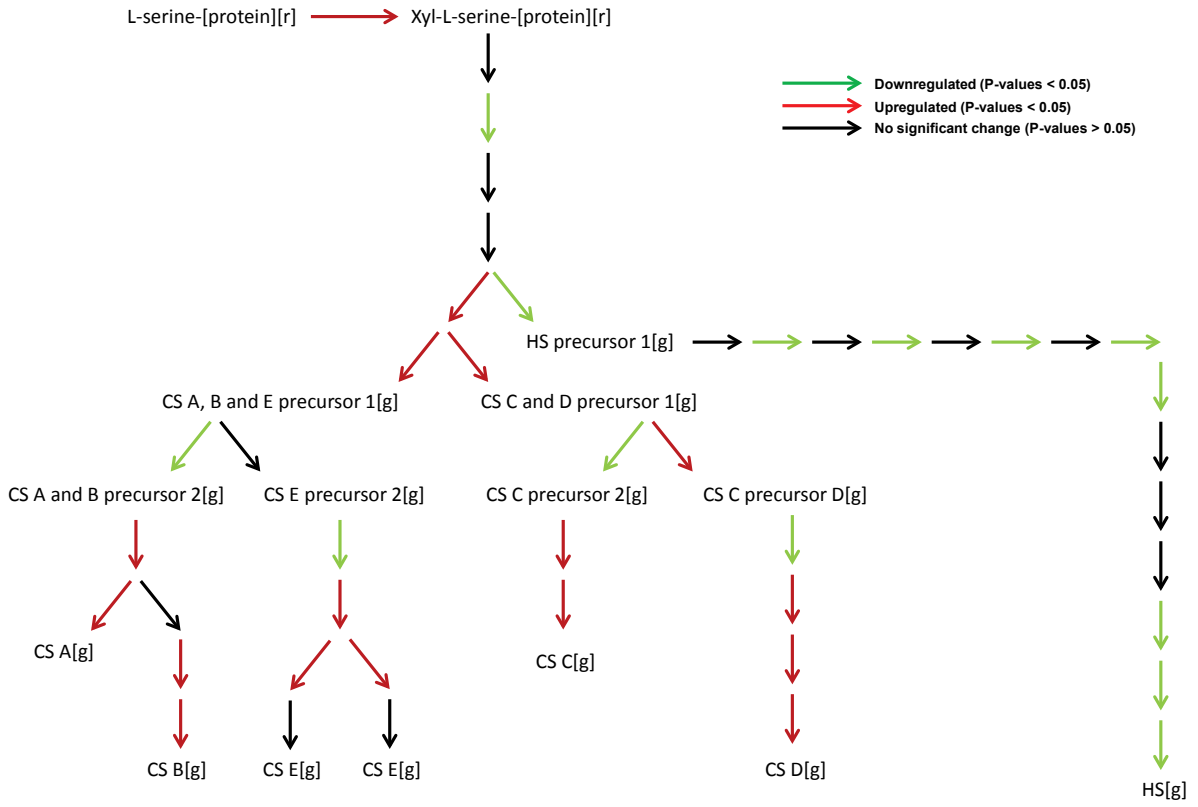


Figure 4-15. Results from Reporter Metabolites analysis. The figure shows the biosynthesis of chondroitin sulfate (CS) and heparan sulfate (HS) in Golgi apparatus, as formulated in iHepatocytes2260, together with the relative gene expression level of NASH vs. healthy samples. Red arrows indicate over-expression of a gene, whereas green arrows indicate under-expression. Non-significant changes (p-value>0.05) is indicated with black arrows.

Figure 4-15 shows the reactions in the model which involves CS or HS and the relative gene expression for the corresponding genes. As can be seen, the genes responsible for the synthesis of CS are mainly upregulated in NASH subjects while the genes responsible for the synthesis of HS are mainly downregulated. CS and HS are implicated in cancer progression (Afratis *et al.*, 2012), one of the most severe outcomes of NASH. Because of this, and because of the clear upregulation of one branch and clear downregulation of the other, we therefore suggested that the blood concentration of the metabolites associated with these pathways

might change accordingly, and that they are therefore potential biomarkers for diagnosing NASH.

The Reporter Subnetworks algorithm was then applied to identify sets of metabolic reactions which exhibit transcriptional correlation after a perturbation (in this case NASH vs. healthy) (Patil *et al.*, 2005). Figure 4-16a show the resulting subnetwork. As can be seen, amino acid metabolism has a prominent role (the non-essential amino acids serine, glycine, glutamate, glutamine, aspartate, asparagine, alanine and the essential amino acids valine and methionine are all present). Several metabolites involved in folate metabolism (e.g. tetrahydrofolate (THF), 5-methyl-THF, 5-formyl-THF and 5,10-methenyl-THF, 5,10-methylene-THF) were also identified, and these metabolites are involved in the interconversion of serine, glycine and glutamate.

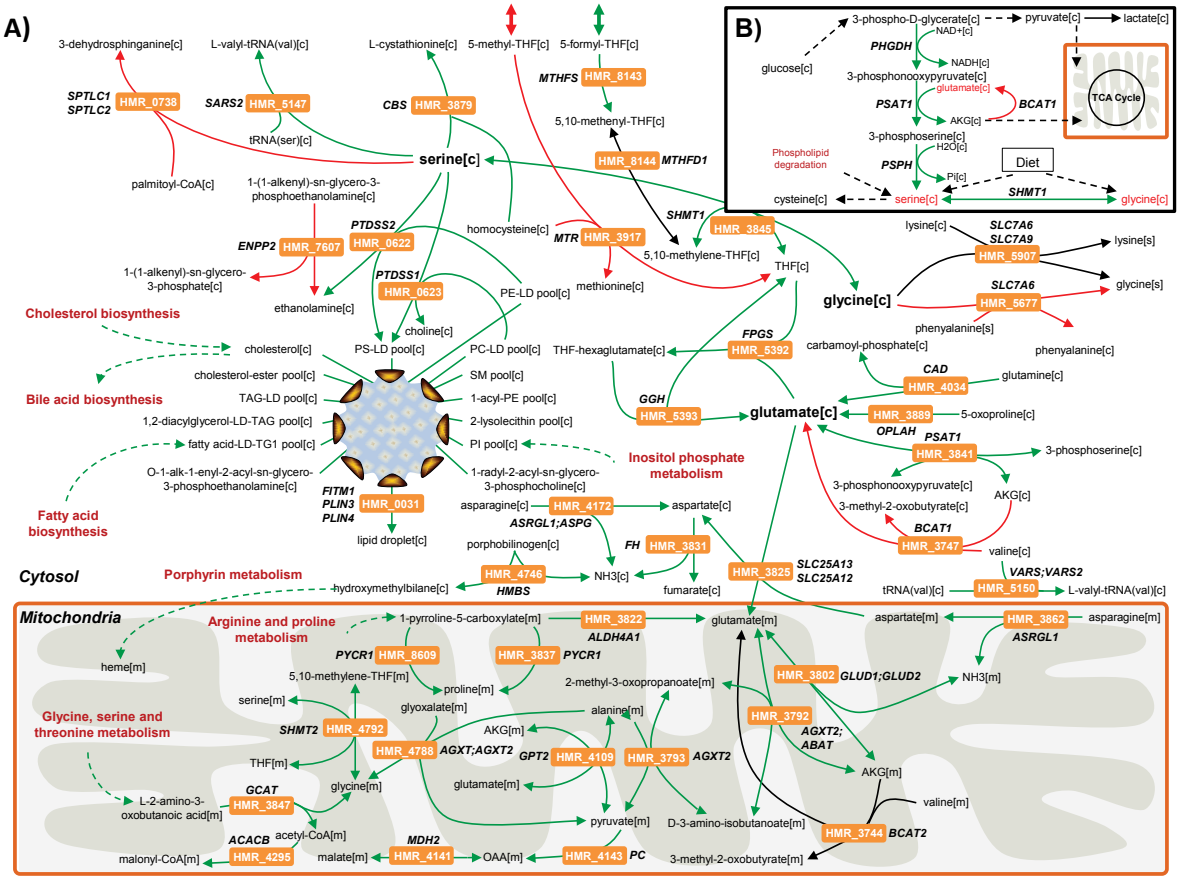


Figure 4-16. Results from Reporter Subnetworks analysis. A) The subnetwork identified using Reporter Subnetworks and gene expression data for NASH vs. healthy subjects. **B)** Some relevant reactions involved in serine biosynthesis and their corresponding change in expression. Red arrows indicate over-expression of the associated genes in NASH, whereas green arrows indicate under-expression. Non-significant changes (p-value>0.05) is indicated with black arrows.

Moreover, phosphatidylserine (PS), an essential component for formation of lipid droplets (LDs), was identified through our analysis. LDs have diverse roles in the cell, such as serving as storage for TAG and CEs or protecting the cell from excess lipids or lipophilic substances that may be toxic (Farese and Walther, 2009). The enzymes phosphatidylserine synthases (PTDSS1) and (PTDSS2) that catalyse the production of PS by condensation of phosphatidylcholine (PC) and phosphatidylethanolamine (PE), respectively, were significantly downregulated in NASH patients. The significant changes in the level of PS in

cirrhotic (severe stage of NASH) livers was previously reported in a study on changes in lipid species in subjects with cirrhotic livers compared with healthy controls (Gorden *et al.*, 2011). Given that PS is essential for hepatocytes, we hypothesize that decreased activity of these enzymes may be associated with a decrease in the endogenous level of serine, which is the second most connected node in our identified Reporter Subnetworks.

Serine is endogenously biosynthesized from a glycolytic intermediate, 3-phospho-D-glycerate. This three-step process is catalysed by phosphoglycerate dehydrogenase (*PHGDH*), phosphoserine aminotransferase 1 (*PSAT1*) and phosphoserine phosphatase (*PSPH*), as shown in Figure 4-16b. An alternative synthesis pathway is via the reversible interconversion with glycine through hydroxymethyltransferases (*SHMT1*) and (*SHMT2*). Serine can also be derived from the diet and the degradation of protein and/or phospholipids.

Through differential analysis of transcriptomics data from the NASH patients, it was also observed that gene expression of several enzymes that use serine, including *CBS* (cysteine synthesis), *SARS2* (aminoacyl-tRNA biosynthesis), *SHMT1* and *SHMT2* (glycine synthesis) were significantly downregulated (p-values < 0.05) whereas *SPTLC1* and *SPTLC2* (sphingosine synthesis) were significantly upregulated. Downregulation of *CBS* that catalyses the conversion of serine and homocysteine to L-cystathionine and upregulation of *MTR* that condensates homocysteine to methionine through the use of 5-methyl-THF indicate that there are metabolic changes around homocysteine in NASH patients. Notably, it has been earlier reported that the plasma homocysteine level can be used for diagnosing NASH and classifying steatosis and NASH patients (Gulsen *et al.*, 2005). It is not always straight forward to relate blood concentrations to gene expression levels of the involved enzymes, but our model-based analysis suggests a mechanistic explanation for this.

Taken together, the results suggest that the changes in the level of PS in liver (Gorden *et al.*, 2011) as well as the relative increase in the homocysteine blood level (Gulsen *et al.*, 2005) is caused by decreased level of endogenous serine. In order to test this hypothesis, we checked the expression level of enzymes that catalyse the biosynthesis of serine in the liver of NASH patients, and it was observed that the expression levels of *PHGDH*, *PSAT1*, *PSPH* in serine synthesis pathway (SSP) and *SHMT1* and *SHMT2* enzymes were significantly downregulated. Decreased levels of serine in NASH patients was supported by the plasma profiling of amino acids in NASH patients, and it was reported that the serine (15 % decrease, p-value=0.0568) level in the plasma is decreased (Kalhan *et al.*, 2011).

Equimolar amounts of serine and α-ketoglutarate (AKG) are synthesized in the SSP, and downregulation of reactions in SSP decrease the anaplerosis of glutamate to the TCA cycle in the form of AKG (Possemato *et al.*, 2011). Decreased level of serine also causes an accumulation of upstream glycolytic intermediates (Chaneton *et al.*, 2012), and a decreased flux of mitochondrial AKG is compensated by an increased flux of pyruvate to oxaloacetate in a healthy cell (see Figure 4-16b). In order to investigate the occurrence of this mechanism in NASH patients, we examined all mitochondrial reactions involving pyruvate as reactant in iHepatocytes2260. We found that the corresponding genes were downregulated for five out of seven such reactions. Furthermore, we investigated the expression of level of mitochondrial pyruvate carriers (*MPC1* and *MPC2*) and mitochondrial AKG/malate carrier (*SLC25A11*) and it was observed that their expression levels were downregulated in NASH patients. These indicate that the mitochondrial metabolic activity of hepatocytes is decreased in NASH patients compared to healthy subjects. This is in agreement with findings in **Paper V** where we investigated the metabolic changes in the case of fat accumulation in adipocytes in response to obesity.

Based on our analysis, increasing the serine level in hepatocytes through the uptake of serine as a dietary supplement could be beneficial for NASH patients. Activity loss of *PHGDH* in SSP in the brain, which causes low serine and glycine levels and affects neuronal function, is reversed by serine supplementation (de Koning *et al.*, 2004). The toxicity and the dosage of serine during its uptake through diet have been previously studied. Furthermore, long-term serine treatment decreased the homocysteine level in animal studies (Girard-Globa *et al.*, 1972) and in humans in a single dose situations (Verhoef *et al.*, 2004).

One other possible way to increase the serine level in order to offer the possibility for therapeutic interventions is activation of the enzymes in SSP or *SHMT1* and *SHMT2* that converts glycine to serine. Three different enzymes constitute the SSP and it is earlier reported that *PSPH* is the rate-controlling enzyme for the SSP in liver (Lund *et al.*, 1985). Activation of the SSP through the amplification of *PSPH* may also decrease the flux through pyruvate and lactate formation in cytosol since increased pyruvate and lactate levels were previously reported in NASH patients (Kalhan *et al.*, 2011).

In conclusion, the HMR 2.0 database published in this paper is the most comprehensive general human GEM, and the only one which incorporates extensive lipid metabolism. By applying the tINIT algorithm (Agren *et al.*, 2013b) in order to reconstruct a draft hepatocyte GEM, we found 61 proteins which the algorithm flagged as likely to have been misidentified in HPA. At least 51 (83%) of those proteins were indeed misidentified when the IHC stainings were re-evaluated and/or when tested for with different antibodies. This is an excellent example of how a network-centric analysis can pick up on targets which would not be possible to identify by other means. The reconstructed iHepatocyte2260 GEM was then applied to study metabolism in NASH. Our analysis suggests that it may be possible to diagnose NASH through identified metabolic biomarkers such as 5-methyl-THF, 5-formyl-THF, CS, and HS levels in blood. Furthermore, the development of therapeutics techniques based on the enhancement of endogenous serine and AKG levels may correct the underlying etiology of NASH. This could be achieved by activation (or elevated expression) of *PSPH* and *SHMT1* and inhibition of *BCAT1*.

5 Conclusions and future perspectives

5.1 Conclusions

In the introduction section I identified three key issues in need of more research. These were: 1) model reconstruction is very labour intensive and error prone, 2) GEMs are underused as scaffolds for omics integration, 3) difficulties associated with modelling of complex organisms.

In **Paper I**, genome-scale metabolic modelling was applied to succinic acid production in *S. cerevisiae*. The modelling was used to suggest single gene deletions, out of which three were validated experimentally. A central aspect in the study is the effect of oxygen on succinate production, where the simulations suggest that fully anaerobic conditions are necessary. One of the gene deletions, *Δdic1*, led to a significant succinate yield, in close agreement with the model predictions. However, the yield was not as high as what had previously been achieved by utilizing a quadruple deletion strategy and aerobic conditions (0.02 C-mol/C-mol glucose vs. 0.07 C-mol/C-mol glucose). A distinct advantage over that strategy is that anaerobic fermentations are preferred industrially. Both these strategies result in far less succinate than what is possible in bacterial hosts, and the result can therefore be seen mainly as proof of concept. The most interesting result, in my opinion, is rather that the study provides some clues on the roles of Frds1 and reductive TCA cycle in mitochondrial NAD^+ regeneration under anaerobic conditions, which is still not fully elucidated.

Paper I represents an excellent example of how powerful the systems biology cycle (see Figure 2-3) can be. Fermentation data for the wild-type was used to constrain a GEM, which was then used to study product formation under different conditions. The resulting flux distributions were visualized and analysed, and the simulation parameters were adjusted until the model correctly predicted the phenotype. Simulations were used to form hypotheses and predict the yields of product following the suggested perturbations. The predictions were then validated experimentally. Lastly, the experimental results were used to suggest further studies and identify parts of the model that might need to be revised.

One of the key issues that warranted further investigation, as identified in the introduction, was that the potential of GEMs to act as scaffolds for data integration was being underused. This gave the motivation for us to develop the algorithm in **Paper II**, which aims at integrating fermentation data with gene expression data, with the purpose of identifying transcriptionally controlled reactions. Such reactions could then form suitable targets for metabolic engineering. The algorithm was applied to study shifts in carbon sources in *S. cerevisiae* as a validation case. We identified three transcription factors which specifically regulated enzymes in transcriptionally controlled reactions. This implies that there is a global regulation of major flux alterations, which is highly relevant for metabolic engineering purposes.

Any textbook or review on metabolic engineering will have a figure on how modelling and experimental efforts interact and feed off the results from each other (as does this one, see Figure 2-3). However, this represents an ideal case and not necessarily how it works in practice. Instead, much of the modelling is based on pre-existing data from literature. A powerful aspect of the algorithm is therefore that it relies on data that is very widely available. The flexibility of the algorithm was shown in **Paper III**, **Paper V**, and **Paper VI**.

In **Paper III** we reconstructed a GEM for the filamentous fungi *Penicillium chrysogenum* and used it to study penicillin biosynthesis. The algorithm from **Paper II** was applied for comparison of an industrial high-producing strain and the wild-type strain in order to identify potential targets for increasing the penicillin yield. 36 reactions were identified as being transcriptionally controlled and upregulated in the industrial strain. They are therefore potential targets for further overexpression. We also identified three single gene deletions which were predicted to result in a 21% increase in penicillin production. In addition, we found strong evidence that in the industrial strain the transsulfuration pathway is the dominating pathway for cysteine biosynthesis, even though the enzymes for the energetically more efficient direct sulfhydrylation pathway have been identified in *P. chrysogenum*. This also represents an interesting target, as cysteine synthesis can be a limiting step in penicillin production.

The GEM for *P. chrysogenum* was reconstructed and validated using the RAVEN Toolbox. The software aims at automating parts of the GEM reconstruction process in order to allow for faster reconstruction of high-quality GEMs. It was developed to address the first issue identified in the introduction; model reconstruction is very labour intensive and error-prone. The RAVEN Toolbox has three main foci: 1) automatic reconstruction of GEMs based on protein homology and integrated quality control, 2) network analysis, modelling and interpretation of simulation results, 3) visualization of GEMs using pre-drawn metabolic network maps. It contains a number of novel approaches for gap filling, assignment of subcellular localization of reactions, and mapping of genes based on homology. This software represents by far the largest single part of the work carried out during my Ph.D. studies. As described in section 3.3, a number of other software and algorithms have been published which are partly overlapping with the RAVEN Toolbox in terms of functionality. The fundamental difference between the RAVEN Toolbox and those software is that it is not just a software for automatic reconstruction; it is a software for working with reconstruction.

In **Paper IV** we built a very comprehensive database of human metabolism, and then reconstructed cell type-specific models as subsets of this generic database. We did this by developing an algorithm which integrates different omics types, such as proteomics, transcriptomics and metabolomics, and then generates models which are in agreement with the data. The algorithm, INIT, was tailored to use large-scale proteomics data generated within the HPA project. The workflow was validated by an extensive comparison to a manually published high-quality model. We then applied the algorithm to reconstruct models for 69 cell types and 16 cancers. The models were then analysed in order to identify subnetworks which are more prominent in cancers. Such networks can be potential targets for treatment. The solutions contained several well-known targets and a few novel ones. Particularly, we found a network dealing with detoxification of aminoacetone, which we propose as a target for therapeutic intervention.

The papers described above, with the exception of **Paper I**, are primarily resource and methodology papers. The sampling algorithm, the *P. chrysogenum* GEM, the RAVEN Toolbox, the INIT algorithm, and the cell type-specific GEMs are arguably larger contributions to the scientific community than the biological interpretations drawn from applying them. The last two papers have a stronger biological component, and show how the methods and resources developed in the first set of papers can be applied.

In **Paper V** we reconstructed a GEM for adipocytes based on proteomics data generated together with our collaborators in the HPA project. The model represents a significant step forward since it is the first human GEM with extensive lipid metabolism incorporated. The

model was built on the generic database developed in **Paper IV** and validated in the RAVEN Toolbox (**Paper III**). The model could correctly capture the reduced dynamics of lipid droplet formation in obese subjects, and we could see that this was associated to mitochondrial dysfunction. The model was then used together with the algorithm developed in **Paper II** to identify metabolic differences between healthy and obese siblings. This led us to hypothesize that obesity could be treated with interventions aiming a reactivating the mitochondria by increasing the availability of acetyl-CoA. An alternative approach was to target the degradation of heparan sulfate proteoglycans.

An important lesson from this project was how important it is to have collaborators in the medical field. The reconstruction of such a high-quality model would not have been possible without the large-scale proteomics data generated specifically for this purpose. The study represents an excellent example of the usability of GEMs as scaffolds for omics integration, as is shown in the study on metabolic differences between healthy and obese siblings.

In **Paper VI** we built on the model developed in **Paper V** and adapted it to hepatocytes by using INIT (**Paper IV** and Agren *et al.* (2013b)). During this step the algorithm included 61 proteins in the model even though the proteomics data suggested that they did not exist in hepatocytes. At least 51 (83%) of those proteins were indeed misidentified when the IHC stainings were re-evaluated and/or when tested for with different antibodies. This is an excellent example of how a network-centric analysis can pick up on targets which would not be possible to identify by other means. The model was then applied to study metabolism in patients with non-alcoholic fatty liver disease. Our main findings pointed to a central role of serine, and we proposed that enhancement of endogenous serine levels may correct the underlying etiology of the disease.

A number of methods have previously been developed in order to deal with the three issues set forward in the introduction (as described in detail in section 3). Despite that, I hope to have shown that the work put forward in this thesis has contributed in some small amount to solving them, and that in doing so it has also resulted in novel biological insights.

5.2 Future perspectives

During the last couple of decades the constraint-based approach to modelling has proven to be very well-suited for metabolic engineering purposes. More recently, it has also started to prove its applicability to human health and disease. Extensive method development has been carried out in order to improve on the quality of GEMs and reduce the efforts involved in reconstructing them, part of which has been performed within this Ph.D. project. This has lowered the bar for reconstructing high-quality models, and GEMs for prokaryotes can now be routinely reconstructed with little manual input. Much effort has also gone into the development of methods for guiding metabolic engineering and strain design. Although by no means solved problems, there now exists so many algorithms for these purposes that I would suggest that the field turns its attention to some other remaining challenges while survival of the fittest sorts out the most applicable algorithms.

One issue that I think is of paramount importance is that of transport reactions. The fraction of transport reactions is often >20% for eukaryotic models and the evidence level for them is significantly worse than for the enzymatically catalysed reactions. These reactions are routinely included “for connectivity reasons”. This has large implications for modelling of eukaryotic organisms, since much of the complexity of metabolism comes from the

compartmentalization of redox, charge and energy balancing. A large-scale screening effort of transporter substrate specificity, membrane localization and transport direction for a eukaryotic model organism would be immensely valuable for the field.

For applications in human health and disease there are two issues that I think warrant special attention. The first is the absence of a standard operating procedure (SOP) for reconstruction of cell type-specific GEMs. The development of detailed SOPs for reconstruction of microbial models had a hugely positive effect on the field, and a similar push is needed for multicellular organisms as well. The second issue deals with data availability. Since constraint-based modelling has evolved in close interaction with metabolic engineering, which allows for carefully set up fermentations and quantification of internal and/or external fluxes, many of the methods are based on fluxomics data. However, this is not a common experimental setup for cultivation of mammalian cells, which are most often grown in complex media with no quantification of exchange fluxes. It would be highly relevant for the field with a medically oriented project tailored to supply the data best fitted for constraint-based modelling. This would serve as an example, both to the medical field and to the metabolic engineering field, of the capabilities of constraint-based modelling of human cells.

A third area which I'm confident will be a future focus is that of interactions between models with different objectives. Possible applications include interactions between organs, bacteria in mixed fermentations such as in the gut, cells with the same genotype but with different sets of expressed genes due to stochastic noise, or between pathogens and their hosts. The first steps towards this goal have already been taken, but there is a long way left.

Lastly, I think the field is approaching maturity when it comes to microbial systems, but that it has yet to prove itself when it comes to medical applications. A large proportion of the published papers are still based on some type of novel algorithm or method, and there are not all that many examples of applications of known methods to provide answers to concrete biological questions. The modelling is still mainly used for data analysis/mining, and not interactively and iteratively for testing hypotheses and generating new data. A few success stories would open up for more funding and collaborations.

To conclude, the field has developed immensely in the short few years that I have worked in it. All the signs point to that the coming years will be just as exciting.

Acknowledgements

First and foremost, I would like to extend my warmest gratitude to my supervisor Jens Nielsen. I have always felt like you believed in my abilities, and that you had my back no matter what. You gave me the freedom to follow my ideas whenever possible, and structure and tight reins whenever needed. I have learnt immensely from you, and your breadth of knowledge has always been an inspiration.

The following people have been fundamental for my development from a young student to something vaguely resembling a scientist. Jose Manuel Otero, without your enthusiasm and love of science I would not be where I am today. Sergio Bordel Velasco, the discussions with you during the first year of my studies were the most intellectually stimulating and rewarding ones, and really made me feel like I was actually part of academia now. Intawat Nookaew, you have been somewhat of an older brother all throughout my studies; guiding me and teaching me about the ins and outs of working in academia. My sincerest thank to all of you.

None of the work presented in this thesis has been carried out by me on my own. I would like to thank the people who have worked with me to make it possible: Wanwipa Vongsangnak, Liming Liu, Saeed Shoae, Marija Cvijovic, Natapol Pornputtapong, Stephan Pabinger, and Roberto Olivares-Hernandez. A special thanks to Adil Mardinoglu. We have worked closely together for several years now, and much of the work in this thesis is thanks to you. You have also grown to become a valued friend. I want to give a warm thank you to Erica Dahlin, Martina Butorac, Malin Nordvall, and Marie Nordquist for helping me when I was looking lost in the lab or needed help with all kinds of practicalities.

I would also like to extend my gratitude to my external collaborators at the Human Protein Atlas and at Sandoz. The data and knowledge that you have provided have enabled several of the studies presented here. A special thanks to Timo Hardiman and Rudolf Mitterbauer at Sandoz for taking care of me when in Austria.

In addition, I would like to thank all the people that I haven't collaborated with professionally, but whose minds and wits and laughs have made this whole thing the fantastic experience that it has been: Fredrik Karlsson, Leif Väremo, Siavash Partow, Amir Feizi, Verena Sievers, Kuk-ki Hong, Tobias Österlund, Christoph Knuf, Luis Caspeta-Guadarrama, Keith Tyo, Marta Papini, Goutham Vemuri, Dina Petranovic, and many more.

My deepest gratitude goes to my parents. Your unwavering love and support has made me who I am.

Evelina, you own my heart. I have no clue where life will take us now, but I know we're going to have so much fun!

References

Acencio, M.L. and Lemke, N. (2009) Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information, BMC Bioinformatics, 10, p. 290.

Adrio, J.L. and Demain, A.L. (2003) Fungal biotechnology, Int Microbiol, 6(3), pp. 191-199.

Afratis, N., et al. (2012) Glycosaminoglycans: key players in cancer cell biology and treatment, Febs Journal, 279(7), pp. 1177-1197.

Agren, R., et al. (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT, PLoS Comput Biol, 8(5), p. e1002518.

Agren, R., et al. (2013a) The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum, PLoS Comput Biol, 9(3), p. e1002980.

Agren, R., et al. (2013b) Drug discovery through the use of personalized genome-scale metabolic models for liver cancer, (Submitted).

Agren, R., et al. (2013c) Genome-scale modeling enables metabolic engineering of Saccharomyces cerevisiae for succinic acid production, J Ind Microbiol Biot, (In press).

Aho, T., et al. (2010) Reconstruction and validation of RefRec: a global model for the yeast molecular interaction network, PLoS One, 5(5), p. e10662.

Aiba, S. and Matsuoka, M. (1979) Identification of Metabolic Model - Citrate Production from Glucose by Candida-Lipolytica, Biotechnol. Bioeng., 21(8), pp. 1373-1386.

Aiba, S., et al. (1980) Enhancement of Tryptophan Production by Escherichia-Coli as an Application of Genetic-Engineering, Biotechnol Lett, 2(12), pp. 525-530.

Akesson, M., et al. (2004) Integration of gene expression data into genome-scale metabolic models, Metab Eng, 6(4), pp. 285-293.

Alcantara, R., et al. (2012) Rhea--a manually curated resource of biochemical reactions, Nucleic Acids Res, 40(Database issue), pp. D754-760.

Altschul, S.F., et al. (1990) Basic local alignment search tool, J Mol Biol, 215(3), pp. 403-410.

Andersen, M.R., et al. (2008) Metabolic model integration of the bibliome, genome, metabolome and reactome of Aspergillus niger, Mol Syst Biol, 4, p. 178.

Andersen, M.R., et al. (2009) Systemic analysis of the response of Aspergillus niger to ambient pH, Genome Biol, 10(5), p. R47.

Andersen, M.R., et al. (2011) Comparative genomics of citric-acid-producing Aspergillus niger ATCC 1015 versus enzyme-producing CBS 513.88, Genome Res, 21(6), pp. 885-897.

Anstee, Q.M. and Day, C.P. (2012) S-adenosylmethionine (SAMe) therapy in liver disease: a review of current evidence and clinical utility, J Hepatol, 57(5), pp. 1097-1109.

Arakawa, K., et al. (2006) GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes, BMC Bioinformatics, 7, p. 168.

Asadollahi, M.A., et al. (2009) Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering, Metab Eng, 11(6), pp. 328-334.

Ascha, M.S., et al. (2010) The incidence and risk factors of hepatocellular carcinoma in patients with nonalcoholic steatohepatitis, Hepatology, 51(6), pp. 1972-1978.

Auffray, C., et al. (2009) Systems medicine: the future of medical genomics and healthcare, Genome Med, 1(1), p. 2.

Baffy, G., et al. (2012) Hepatocellular carcinoma in non-alcoholic fatty liver disease: An emerging menace, J Hepatol, 56(6), pp. 1384-1391.

Bailey, J.E. (1991) Toward a science of metabolic engineering, Science, 252(5013), pp. 1668-1675.

Balagurunathan, B., et al. (2012) Reconstruction and analysis of a genome-scale metabolic model for Scheffersomyces stipitis, Microb Cell Fact, 11, p. 27.

Baranano, D.E., et al. (2002) Biliverdin reductase: a major physiologic cytoprotectant, Proc Natl Acad Sci U S A, 99(25), pp. 16093-16098.

Beard, D.A., et al. (2002) Energy balance for analysis of complex metabolic networks, Biophys J, 83(1), pp. 79-86.

Becker, S.A. and Palsson, B.O. (2008) Context-specific metabolic networks are consistent with experiments, PLoS Comput Biol, 4(5), p. e1000082.

Beg, Q.K., et al. (2007) Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity, Proc Natl Acad Sci U S A, 104(31), pp. 12663-12668.

Bengtsson, G. and Olivecrona, T. (1980) Lipoprotein lipase. Mechanism of product inhibition, Eur J Biochem, 106(2), pp. 557-562.

Bennett, J.W. (1998) Mycotechnology: the role of fungi in biotechnology, J Biotechnol, 66(2-3), pp. 101-107.

Berglund, L., et al. (2008) A genecentric Human Protein Atlas for expression profiles based on antibodies, Mol Cell Proteomics, 7(10), pp. 2019-2027.

Bonarius, H.P.J., et al. (1997) Flux analysis of underdetermined metabolic networks: The quest for the missing constraints, Trends Biotechnol., 15(8), pp. 308-314.

Bordbar, A., et al. (2010) Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions, Mol Syst Biol, 6, p. 422.

Bordbar, A., et al. (2011) A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology, BMC Syst Biol, 5, p. 180.

Bordbar, A. and Palsson, B.O. (2012) Using the reconstructed genome-scale human metabolic network to study physiology and pathology, J Intern Med, 271(2), pp. 131-141.
Bordel, S., et al. (2010) Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes, PLoS Comput Biol, 6(7), p. e1000859.
Bro, C., et al. (2006) In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production, Metab Eng, 8(2), pp. 102-111.
Brochado, A.R., et al. (2010) Improved vanillin production in baker's yeast through in silico design, Microb Cell Fact, 9, p. 84.
Brooks, J.P., et al. (2012) Gap detection for genome-scale constraint-based models, Advances in bioinformatics, 2012, p. 323472.
Bundy, J.G., et al. (2007) Evaluation of predicted network modules in yeast metabolism using NMR-based metabolite profiling, Genome Res, 17(4), pp. 510-519.
Burgard, A.P. and Maranas, C.D. (2003) Optimization-based framework for inferring and testing hypothesized metabolic objective functions, Biotechnol. Bioeng., 82(6), pp. 670-677.
Burgard, A.P., et al. (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization, Biotechnol Bioeng, 84(6), pp. 647-657.
Burgard, A.P., et al. (2004) Flux coupling analysis of genome-scale metabolic network reconstructions, Genome Res, 14(2), pp. 301-312.
Cakir, T., et al. (2006) Integration of metabolome data with metabolic networks reveals reporter reactions, Mol Syst Biol, 2, p. 50.
Camarasa, C., et al. (2003) Investigation by 13C-NMR and tricarboxylic acid (TCA) deletion mutant analysis of pathways for succinate formation in Saccharomyces cerevisiae during anaerobic fermentation, Microbiology, 149(Pt 9), pp. 2669-2678.
Camarasa, C., et al. (2007) Role in anaerobiosis of the isoenzymes for Saccharomyces cerevisiae fumarate reductase encoded by OSM1 and FRDS1, Yeast, 24(5), pp. 391-401.
Cao, H., et al. (2008) Identification of a lipokine, a lipid hormone linking adipose tissue to systemic metabolism, Cell, 134(6), pp. 933-944.
Caspeta, L., et al. (2012) Genome-scale metabolic reconstructions of Pichia stipitis and Pichia pastoris and in silico evaluation of their potentials, BMC Syst Biol, 6, p. 24.
Caveney, E., et al. (2011) Pharmaceutical interventions for obesity: a public health perspective, Diabetes Obes Metab, 13(6), pp. 490-497.
Chan, C., et al. (2003) Metabolic flux analysis of cultured hepatocytes exposed to plasma, Biotechnology and Bioengineering, 81(1), pp. 33-49.
Chaneton, B., et al. (2012) Serine is a natural ligand and allosteric activator of pyruvate kinase M2, Nature, 491(7424), pp. 458-462.
Chang, R.L., et al. (2010) Drug off-target effects predicted using structural analysis in the context of a metabolic network model, PLoS Comput Biol, 6(9), p. e1000938.
Chechik, G., et al. (2008) Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network, Nat Biotechnol, 26(11), pp. 1251-1259.
Chen, L. and Vitkup, D. (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles, Genome Biol, 7(2), p. R17.
Cherry, J.M., et al. (1998) SGD: Saccharomyces Genome Database, Nucleic Acids Res, 26(1), pp. 73-79.
Choi, B.K., et al. (2003) Use of combinatorial genetic libraries to humanize N-linked glycosylation in the yeast Pichia pastoris, Proc Natl Acad Sci U S A, 100(9), pp. 5022-5027.
Chung, B.K., et al. (2010) Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast Pichia pastoris for strain improvement, Microb Cell Fact, 9, p. 50.
Cimini, D., et al. (2009) Global transcriptional response of Saccharomyces cerevisiae to the deletion of SDH3, BMC Syst Biol, 3, p. 17.
Corning, P.A. (2012) The re-emergence of emergence, and the causal role of synergy in emergent evolution, Synthese, 185(2), pp. 295-317.
Costenoble, R., et al. (2011) Comprehensive quantitative analysis of central carbon and amino-acid metabolism in Saccharomyces cerevisiae under multiple conditions by targeted proteomics, Mol Syst Biol, 7, p. 464.
Covert, M.W., et al. (2001) Regulation of gene expression in flux balance models of metabolism, J Theor Biol, 213(1), pp. 73-88.
Croft, D., et al. (2011) Reactome: a database of reactions, pathways and biological processes, Nucleic Acids Res, 39(Database issue), pp. D691-697.
Cvijovic, M., et al. (2010) BioMet Toolbox: genome-wide analysis of metabolism, Nucleic Acids Res, 38(Web Server issue), pp. W144-149.
Daran-Lapujade, P., et al. (2007) The fluxes through glycolytic enzymes in Saccharomyces cerevisiae are predominantly regulated at posttranscriptional levels, Proc Natl Acad Sci U S A, 104(40), pp. 15753-15758.
David, H., et al. (2006) Metabolic network driven analysis of genome-wide transcription data from Aspergillus nidulans, Genome Biol, 7(11), p. R108.
David, H., et al. (2008) Analysis of Aspergillus nidulans metabolism at the genome-scale, BMC Genomics, 9, p. 163.
de Koning, T.J., et al. (2004) Prenatal and early postnatal treatment in 3-phosphoglycerate-dehydrogenase deficiency, Lancet, 364(9452), pp. 2221-2222.

Dean, J.T., et al. (2009) Resistance to diet-induced obesity in mice with synthetic glyoxylate shunt, Cell Metab, 9(6), pp. 525-536.
DeJongh, M., et al. (2007) Toward the automated generation of genome-scale metabolic networks in the SEED, BMC Bioinformatics, 8, p. 139.
del Rio, G., et al. (2009) How to identify essential genes from molecular networks?, BMC Syst Biol, 3, p. 102.
Delcher, A.L., et al. (1999) Improved microbial gene identification with GLIMMER, Nucleic Acids Res, 27(23), pp. 4636-4641.
Derrien, T., et al. (2007) AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps, Bioinformatics, 23(4), pp. 498-499.
Deutscher, D., et al. (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network, Nat Genet, 38(9), pp. 993-998.
Diamant, I., et al. (2009) A network-based method for predicting gene-nutrient interactions and its application to yeast amino-acid metabolism, Mol Biosyst, 5(12), pp. 1732-1739.
Dikicioglu, D., et al. (2008) Integration of metabolic modeling and phenotypic data in evaluation and improvement of ethanol production using respiration-deficient mutants of Saccharomyces cerevisiae, Appl Environ Microbiol, 74(18), pp. 5809-5816.
Dobson, P.D., et al. (2010) Further developments towards a genome-scale metabolic model of yeast, BMC Syst Biol, 4, p. 145.
Draths, K.M., et al. (1992) Biocatalytic Synthesis of Aromatics from D-Glucose - the Role of Transketolase, Journal of the American Chemical Society, 114(10), pp. 3956-3962.
Duarte, N.C., et al. (2004a) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model, Genome Res, 14(7), pp. 1298-1309.
Duarte, N.C., et al. (2004b) Integrated analysis of metabolic phenotypes in Saccharomyces cerevisiae, BMC Genomics, 5, p. 63.
Duarte, N.C., et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data, Proc Natl Acad Sci U S A, 104(6), pp. 1777-1782.
Dudakovic, A., et al. (2011) Geranylgeranyl diphosphate depletion inhibits breast cancer cell migration, Invest New Drugs, 29(5), pp. 912-920.
Durot, M., et al. (2009) Genome-scale models of bacterial metabolism: reconstruction and applications, FEMS Microbiol Rev, 33(1), pp. 164-190.
Eddy, S.R. (1998) Profile hidden Markov models, Bioinformatics, 14(9), pp. 755-763.
Edwards, J.S., et al. (2001) In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data, Nat Biotechnol, 19(2), pp. 125-130.
Eisenberg, T., et al. (2009) Induction of autophagy by spermidine promotes longevity, Nat Cell Biol, 11(11), pp. 1305-1314.
Elander, R.P. (2003) Industrial production of beta-lactam antibiotics, Appl Microbiol Biotechnol, 61(5-6), pp. 385-392.
Eruslanov, E., et al. (2009) Altered expression of 15-hydroxyprostaglandin dehydrogenase in tumor-infiltrated CD11b myeloid cells: a mechanism for immune evasion in cancer, J Immunol, 182(12), pp. 7548-7557.
Fabry, M.E., et al. (1981) Some aspects of the pathophysiology of homozygous Hb CC erythrocytes, J Clin Invest, 67(5), pp. 1284-1291.
Famili, I., et al. (2003) Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network, Proc Natl Acad Sci U S A, 100(23), pp. 13134-13139.
Farese, R.V., Jr. and Walther, T.C. (2009) Lipid droplets finally get a little R-E-S-P-E-C-T, Cell, 139(5), pp. 855-860.
Fell, D.A. and Small, J.R. (1986) Fat synthesis in adipose tissue. An examination of stoichiometric constraints, Biochem J, 238(3), pp. 781-786.
Fisher, C.D., et al. (2009) Hepatic cytochrome P450 enzyme alterations in humans with progressive stages of nonalcoholic fatty liver disease, Drug Metab Dispos, 37(10), pp. 2087-2094.
Fletcher, D.A. and Theriot, J.A. (2004) An introduction to cell motility for the physical scientist, Physical biology, 1(1-2), pp. T1-10.
Folger, O., et al. (2011) Predicting selective drug targets in cancer through metabolic networks, Mol Syst Biol, 7, p. 501.
Fong, S.S., et al. (2005) In silico design and adaptive evolution of Escherichia coli for production of lactic acid, Biotechnol Bioeng, 91(5), pp. 643-648.
Forberg, C. and Haggstrom, L. (1987) Effects of Cultural Conditions on the Production of Phenylalanine from a Plasmid-Harboring Escherichia-Coli Strain, Appl. Microbiol. Biotechnol., 26(2), pp. 136-140.
Forberg, C., et al. (1988) Correlation of Theoretical and Experimental Yields of Phenylalanine from Non-Growing Cells of a Rec Escherichia-Coli Strain, Journal of Biotechnology, 7(4), pp. 319-332.
Forster, J., et al. (2003a) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network, Genome Res, 13(2), pp. 244-253.
Forster, J., et al. (2003b) Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae, OMICS, 7(2), pp. 193-202.
Fu, P.C. (2009) Gene expression study of Saccharomyces cerevisiae under changing growth conditions, J Chem Technol Biot, 84(8), pp. 1163-1171.
Gevorgyan, A., et al. (2008) Detection of stoichiometric inconsistencies in biomolecular models, Bioinformatics, 24(19), pp. 2245-2251.
Ghosh, A., et al. (2011) Genome-scale consequences of cofactor balancing in engineered pentose utilization pathways in Saccharomyces cerevisiae, PLoS One, 6(11), p. e27316.

Gille, C., et al. (2010) HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology, Mol Syst Biol, 6, p. 411.
Girard-Globa, A., et al. (1972) Long-term adaptation of weanling rats to high dietary levels of methionine and serine, J Nutr, 102(2), pp. 209-217.
Gorden, D.L., et al. (2011) Increased Diacylglycerols Characterize Hepatic Lipid Changes in Progression of Human Nonalcoholic Fatty Liver Disease; Comparison to a Murine Model, Plos One, 6(8), p. e22775.
Greco, D., et al. (2008) Gene expression in human NAFLD, Am J Physiol Gastrointest Liver Physiol, 294(5), pp. G1281-1287.
Gulsen, M., et al. (2005) Elevated plasma homocysteine concentrations as a predictor of steatohepatitis in patients with non-alcoholic fatty liver disease, J Gastroenterol Hepatol, 20(9), pp. 1448-1455.
Hao, T., et al. (2010) Compartmentalization of the Edinburgh Human Metabolic Network, BMC Bioinformatics, 11, p. 393.
Harris, D.M., et al. (2006) Enzymic analysis of NADPH metabolism in beta-lactam-producing Penicillium chrysogenum: presence of a mitochondrial NADPH dehydrogenase, Metab Eng, 8(2), pp. 91-101.
Heavner, B.D., et al. (2012) Yeast 5 - an expanded reconstruction of the Saccharomyces cerevisiae metabolic network, BMC Syst Biol, 6, p. 55.
Heinken, A., et al. (2013) Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut, Gut microbes, 4(1), pp. 28-40.
Heinrich, R., et al. (1977) Metabolic regulation and mathematical models, Progress in biophysics and molecular biology, 32(1), pp. 1-82.
Henry, C.S., et al. (2006) Genome-scale thermodynamic analysis of Escherichia coli metabolism, Biophys J, 90(4), pp. 1453-1461.
Henry, C.S., et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models, Nat Biotechnol, 28(9), pp. 977-982.
Herrgard, M.J., et al. (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae, Genome Res, 16(5), pp. 627-635.
Hjersted, J.L. and Henson, M.A. (2009) Steady-state and dynamic flux balance analysis of ethanol production by Saccharomyces cerevisiae, IET systems biology, 3(3), pp. 167-179.
Hoffman, J., et al. (2006) Effect of creatine and beta-alanine supplementation on performance and endocrine responses in strength/power athletes, Int J Sport Nutr Exerc Metab, 16(4), pp. 430-446.
Jerby, L., et al. (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism, Mol Syst Biol, 6, p. 401.
Jorgensen, H., et al. (1995a) Metabolic flux distributions in Penicillium chrysogenum during fed-batch cultivations, Biotechnol Bioeng, 46(2), pp. 117-131.
Jorgensen, H., et al. (1995b) Analysis of penicillin V biosynthesis during fed-batch cultivations with a high-yielding strain of Penicillium chrysogenum, Appl Microbiol Biotechnol, 43(1), pp. 123-130.
Jouhten, P., et al. (2012) Dynamic flux balance analysis of the metabolism of Saccharomyces cerevisiae during the shift from fully respirative or respirofermentative metabolic states to anaerobiosis, FEBS J, 279(18), pp. 3338-3354.
Joyce, A.R. and Palsson, B.O. (2006) The model organism as a system: integrating 'omics' data sets, Nat Rev Mol Cell Bio, 7(3), pp. 198-210.
Kalapos, M.P. (1994) Methylglyoxal toxicity in mammals, Toxicol Lett, 73(1), pp. 3-24.
Kaleta, C., et al. (2009) Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns, Genome Res, 19(10), pp. 1872-1883.
Kalhan, S.C., et al. (2011) Plasma metabolomic profile in nonalcoholic fatty liver disease, Metabolism, 60(3), pp. 404-413.
Kanehisa, M., et al. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs, Nucleic Acids Res, 38(Database issue), pp. D355-360.
Kang, Y., et al. (1996) Effect of methylglyoxal on human leukaemia 60 cell growth: modification of DNA G1 growth arrest and induction of apoptosis, Leuk Res, 20(5), pp. 397-405.
Karp, P.D., et al. (2002) The Pathway Tools software, Bioinformatics, 18 Suppl 1, pp. S225-232.
Karp, R.M. (2008) George Dantzig's impact on the theory of computation, Discrete Optim, 5(2), pp. 174-185.
Kharchenko, P., et al. (2005) Expression dynamics of a cellular metabolic network, Mol Syst Biol, 1, p. 2005 0016.
Kitano, H. (2002a) Computational systems biology, Nature, 420(6912), pp. 206-210.
Kitano, H. (2002b) Systems biology: a brief overview, Science, 295(5560), pp. 1662-1664.
Koppenol, W.H., et al. (2011) Otto Warburg's contributions to current concepts of cancer metabolism, Nat Rev Cancer, 11(5), pp. 325-337.
Kuepfer, L., et al. (2005) Metabolic functions of duplicate genes in Saccharomyces cerevisiae, Genome Res, 15(10), pp. 1421-1430.
Kumar, V.S. and Maranas, C.D. (2009) GrowMatch: an automated method for reconciling in silico/in vivo growth predictions, PLoS Comput Biol, 5(3), p. e1000308.
Kummel, A., et al. (2006a) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data, Mol Syst Biol, 2, p. 2006 0034.
Kummel, A., et al. (2006b) Systematic assignment of thermodynamic constraints in metabolic network models, BMC Bioinformatics, 7, p. 512.
Kusminski, C.M. and Scherer, P.E. (2012) Mitochondrial dysfunction in white adipose tissue, Trends Endocrinol Metab, 23(9), pp. 435-443.
Lafontan, M. (2008) Advances in adipose tissue metabolism, Int J Obes (Lond), 32 Suppl 7, pp. S39-51.

Lago, F., et al. (2007) Adipokines as emerging mediators of immune response and inflammation, Nature clinical practice. Rheumatology, 3(12), pp. 716-724.
Lake, A.D., et al. (2011) Analysis of global and absorption, distribution, metabolism, and elimination gene expression in the progressive stages of human nonalcoholic fatty liver disease, Drug Metab Dispos, 39(10), pp. 1954-1960.
Lee, D., et al. (2012) Improving metabolic flux predictions using absolute gene expression data, BMC Syst Biol, 6, p. 73.
Lee, J.M., et al. (2008) Dynamic analysis of integrated signaling, metabolic, and regulatory networks, PLoS Comput Biol, 4(5), p. e1000086.
Lee, S.J., et al. (2005) Metabolic engineering of Escherichia coli for enhanced production of succinic acid, based on genome comparison and in silico gene knockout simulation, Appl Environ Microbiol, 71(12), pp. 7880-7887.
Lewis, N.E., et al. (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain, Nat Biotechnol, 28(12), pp. 1279-1285.
Lewis, N.E., et al. (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods, Nat Rev Microbiol, 10(4), pp. 291-305.
Liu, L., et al. (2010) Use of genome-scale metabolic models for understanding microbial physiology, FEBS Lett, 584(12), pp. 2556-2564.
Liu, T., et al. (2012) A constraint-based model of Scheffersomyces stipitis for improved ethanol production, Biotechnology for biofuels, 5(1), p. 72.
Loira, N., et al. (2012) A genome-scale metabolic model of the lipid-accumulating yeast Yarrowia lipolytica, BMC Syst Biol, 6, p. 35.
Lookene, A., et al. (1996) Interaction of lipoprotein lipase with heparin fragments and with heparan sulfate: stoichiometry, stabilization, and kinetics, Biochemistry, 35(37), pp. 12155-12163.
Lund, K., et al. (1985) The reactions of the phosphorylated pathway of L-serine biosynthesis: thermodynamic relationships in rabbit liver in vivo, Arch Biochem Biophys, 237(1), pp. 186-196.
Ma, H., et al. (2007) The Edinburgh human metabolic network reconstruction and its functional analysis, Mol Syst Biol, 3, p. 135.
Machado, M.V. and Cortez-Pinto, H. (2012) Non-invasive diagnosis of non-alcoholic fatty liver disease. A critical appraisal, J Hepatol.
Mahadevan, R. and Schilling, C.H. (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models, Metab Eng, 5(4), pp. 264-276.
Mahadevan, R. and Lovley, D.R. (2008) The degree of redundancy in metabolic genes is linked to mode of metabolism, Biophys J, 94(4), pp. 1216-1220.
Mardinoglu, A. and Nielsen, J. (2012) Systems medicine and metabolic modelling, J Intern Med, 271(2), pp. 142-154.
Mardinoglu, A., et al. (2013) Integration of clinical data with a genome-scale metabolic model of the human adipocyte, Mol Syst Biol, 9, p. 649.
Matsuda, F., et al. (2011) Engineering strategy of yeast metabolism for higher alcohol production, Microb Cell Fact, 10, p. 70.
Matthews, L., et al. (2009) Reactome knowledgebase of human biological pathways and processes, Nucleic Acids Res, 37(Database issue), pp. D619-622.
McKinlay, J.B., et al. (2007) Prospects for a bio-based succinate industry, Appl Microbiol Biotechnol, 76(4), pp. 727-740.
McQuaid, S.E., et al. (2011) Downregulation of adipose tissue fatty acid trafficking in obesity: a driver for ectopic fat deposition?, Diabetes, 60(1), pp. 47-55.
Michaelis, L., et al. (2011) The original Michaelis constant: translation of the 1913 Michaelis-Menten paper, Biochemistry, 50(39), pp. 8264-8269.
Mintz-Oron, S., et al. (2009) Network-based prediction of metabolic enzymes' subcellular localization, Bioinformatics, 25(12), pp. i247-252.
Mo, M.L., et al. (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast, BMC Syst Biol, 3, p. 37.
Nacher, J.C., et al. (2006) Identification of metabolic units induced by environmental signals, Bioinformatics, 22(14), pp. e375-383.
Neuschwander-Tetri, B.A. and Caldwell, S.H. (2003) Nonalcoholic steatohepatitis: summary of an AASLD Single Topic Conference, Hepatology, 37(5), pp. 1202-1219.
Newgard, C.B. (2012) Interplay between lipids and branched-chain amino acids in development of insulin resistance, Cell Metab, 15(5), pp. 606-614.
Ng, C.Y., et al. (2012) Production of 2,3-butanediol in Saccharomyces cerevisiae by in silico aided metabolic engineering, Microb Cell Fact, 11, p. 68.
Nielsen, J. and Jorgensen, H.S. (1995) Metabolic control analysis of the penicillin biosynthetic pathway in a high-yielding strain of Penicillium chrysogenum, Biotechnol Prog, 11(3), pp. 299-305.
Nielsen, J.H. (1995) Physiological engineering aspects of penicillium chrysogenum, Denmark, Polyteknisk forlag.
Nookaew, I., et al. (2008) The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism, BMC Syst Biol, 2, p. 71.
Notebaart, R.A., et al. (2006) Accelerating the reconstruction of genome-scale metabolic networks, BMC Bioinformatics, 7, p. 296.
Ogata, H., et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res, 27(1), pp. 29-34.
Oh, Y.K., et al. (2007) Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data, Journal of Biological Chemistry, 282(39), pp. 28791-28799.

Olivecrona, G. and Beisiegel, U. (1997) Lipid binding of apolipoprotein CII is required for stimulation of lipoprotein lipase activity against apolipoprotein CII-deficient chylomicrons, Arterioscl Throm Vas, 17(8), pp. 1545-1549.
Olivecrona, T. and Olivecrona, G. (2009) The ins and outs of adipose tissue, in Ehnholm, C. (ed), Cellular lipid metabolism, New York, Springer.
Ostergaard, S., et al. (1998) Identification and purification of O-acetyl-L-serine sulphhydrylase in Penicillium chrysogenum, Appl. Microbiol. Biotechnol., 50(6), pp. 663-668.
Ostergaard, S., et al. (2001) The impact of GAL6, GAL80, and MIG1 on glucose control of the GAL system in Saccharomyces cerevisiae, FEMS Yeast Res, 1(1), pp. 47-55.
Osterlund, T., et al. (2012) Fifteen years of large scale metabolic modeling of yeast: developments and impacts, Biotechnol Adv, 30(5), pp. 979-988.
Otero, J.M., et al. (2013) Industrial systems biology of Saccharomyces cerevisiae enables novel succinic acid cell factory, PLoS One, 8(1), p. e54144.
Othmer, H.G. (1976) The qualitative dynamics of a class of biochemical control circuits, Journal of mathematical biology, 3(1), pp. 53-78.
Pabinger, S., et al. (2011) MEMOSys: Bioinformatics platform for genome-scale metabolic models, BMC Syst Biol, 5, p. 20.
Paley, S.M. and Karp, P.D. (2006) The Pathway Tools cellular overview diagram and Omics Viewer, Nucleic Acids Res, 34(13), pp. 3771-3778.
Papini, M., et al. (2010) Phosphoglycerate mutase knock-out mutant Saccharomyces cerevisiae: physiological investigation and transcriptome analysis, Biotechnology journal, 5(10), pp. 1016-1027.
Papp, B., et al. (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast, Nature, 429(6992), pp. 661-664.
Patil, K.R. and Nielsen, J. (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology, Proc Natl Acad Sci U S A, 102(8), pp. 2685-2689.
Patil, K.R., et al. (2005) Evolutionary programming as a platform for in silico metabolic engineering, BMC Bioinformatics, 6, p. 308.
Patnaik, R. and Liao, J.C. (1994) Engineering of Escherichia coli central metabolism for aromatic metabolite production with near theoretical yield, Appl Environ Microbiol, 60(11), pp. 3903-3908.
Patterson, B.W., et al. (2002) Regional muscle and adipose tissue amino acid metabolism in lean and obese women, American journal of physiology. Endocrinology and metabolism, 282(4), pp. E931-936.
Philips, M.R. and Cox, A.D. (2007) Geranylgeranyltransferase I as a target for anti-cancer drugs, J Clin Invest, 117(5), pp. 1223-1225.
Pinney, J.W., et al. (2005) metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella, Nucleic Acids Res, 33(4), pp. 1399-1409.
Popescu, L. and Yona, G. (2005) Automation of gene assignments to metabolic pathways using high-throughput expression data, BMC Bioinformatics, 6, p. 217.
Possemato, R., et al. (2011) Functional genomics reveal that the serine synthesis pathway is essential in breast cancer, Nature, 476(7360), pp. 346-350.
Prabha, A.N.L., et al. (1988) Similar Effects of Beta-Alanine and Taurine in Cholesterol-Metabolism, J Bioscience, 13(3), pp. 263-268.
Price, N.D., et al. (2003) Genome-scale microbial in silico models: the constraints-based approach, Trends Biotechnol, 21(4), pp. 162-169.
Price, N.D., et al. (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints, Nat Rev Microbiol, 2(11), pp. 886-897.
Raab, A.M., et al. (2010) Metabolic engineering of Saccharomyces cerevisiae for the biotechnological production of succinic acid, Metab Eng, 12(6), pp. 518-525.
Rector, R.S., et al. (2008) Non-alcoholic fatty liver disease and the metabolic syndrome: an update, World journal of gastroenterology : WJG, 14(2), pp. 185-192.
Reed, J.L., et al. (2006) Systems approach to refining genome annotation, Proc Natl Acad Sci U S A, 103(46), pp. 17480-17484.
Rokhlenko, O., et al. (2007) Constraint-based functional similarity of metabolic genes: going beyond network topology, Bioinformatics, 23(16), pp. 2139-2146.
Rokholm, B., et al. (2010) The levelling off of the obesity epidemic since the year 1999--a review of evidence and perspectives, Obes Rev, 11(12), pp. 835-846.
Romero, P., et al. (2005) Computational prediction of human metabolic pathways from the complete human genome, Genome Biol, 6(1), p. R2.
Rossouw, D., et al. (2009) Comparative transcriptomic approach to investigate differences in wine yeast physiology and metabolism during fermentation, Appl Environ Microbiol, 75(20), pp. 6600-6612.
Satish Kumar, V., et al. (2007) Optimization based automated curation of metabolic reconstructions, BMC Bioinformatics, 8, p. 212.
Sauer, M., et al. (2008) Microbial production of organic acids: expanding the markets, Trends Biotechnol, 26(2), pp. 100-108.
Sauer, U., et al. (2007) Genetics. Getting closer to the whole picture, Science, 316(5824), pp. 550-551.
Schellenberger, J. and Palsson, B.O. (2009) Use of randomized sampling for analysis of metabolic networks, J Biol Chem, 284(9), pp. 5457-5461.
Schilling, C.H., et al. (1999) Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era, Biotechnol Prog, 15(3), pp. 296-303.

Schilling, C.H. and Palsson, B.O. (2000) Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis, J. Theor. Biol., 203(3), pp. 249-283.
Schneider, C. and Pozzi, A. (2011) Cyclooxygenases and lipoxygenases in cancer, Cancer Metastasis Rev, 30(3-4), pp. 277-294.
Schomburg, I., et al. (2002) BRENDA, enzyme data and metabolic information, Nucleic Acids Res, 30(1), pp. 47-49.
Schuetz, R., et al. (2012) Multidimensional optimality of microbial metabolism, Science, 336(6081), pp. 601-604.
Schuster, S., et al. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering, Trends Biotechnol, 17(2), pp. 53-60.
Schwartz, J.M., et al. (2007) Observing metabolic functions at the genome scale, Genome Biol., 8(6).
Sebti, S.M. and Hamilton, A.D. (2000) Farnesyltransferase and geranylgeranyltransferase I inhibitors and cancer therapy: lessons from mechanism and bench-to-bedside translational studies, Oncogene, 19(56), pp. 6584-6593.
Segre, D., et al. (2002) Analysis of optimality in natural and perturbed metabolic networks, Proc Natl Acad Sci U S A, 99(23), pp. 15112-15117.
Segre, D., et al. (2005) Modular epistasis in yeast metabolism, Nat Genet, 37(1), pp. 77-83.
Seiler, N. (2003a) Thirty years of polyamine-related approaches to cancer therapy. Retrospect and prospect. Part 2. Structural analogues and derivatives, Curr Drug Targets, 4(7), pp. 565-585.
Seiler, N. (2003b) Thirty years of polyamine-related approaches to cancer therapy. Retrospect and prospect. Part 1. Selective enzyme inhibitors, Curr Drug Targets, 4(7), pp. 537-564.
Seo, S. and Lewin, H.A. (2009) Reconstruction of metabolic pathways for the cattle genome, BMC Syst Biol, 3, p. 33.
Sheikh, K., et al. (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus, Biotechnol Prog, 21(1), pp. 112-121.
Shlomi, T., et al. (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations, Proc Natl Acad Sci U S A, 102(21), pp. 7695-7700.
Shlomi, T., et al. (2007) Systematic condition-dependent annotation of metabolic genes, Genome Res, 17(11), pp. 1626-1633.
Shlomi, T., et al. (2008) Network-based prediction of human tissue-specific metabolism, Nat Biotechnol, 26(9), pp. 1003-1010.
Shlomi, T., et al. (2009) Predicting metabolic biomarkers of human inborn errors of metabolism, Mol Syst Biol, 5, p. 263.
Shlomi, T., et al. (2011) Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect, PLoS Comput Biol, 7(3), p. e1002018.
Simeonidis, E., et al. (2010) Why does yeast ferment? A flux balance analysis study, Biochem Soc Trans, 38(5), pp. 1225-1229.
Smallbone, K., et al. (2010) Towards a genome-scale kinetic model of cellular metabolism, BMC Syst Biol, 4, p. 6.
Snitkin, E.S., et al. (2008) Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions, Genome Biol, 9(9), p. R140.
Sohn, S.B., et al. (2010) Genome-scale metabolic model of methylotrophic yeast Pichia pastoris and its use for in silico analysis of heterologous protein production, Biotechnology journal, 5(7), pp. 705-715.
Sohn, S.B., et al. (2012) Genome-scale metabolic model of the fission yeast Schizosaccharomyces pombe and the reconciliation of in silico/in vivo mutant growth, BMC Syst Biol, 6, p. 49.
Song, H. and Lee, S.Y. (2006) Production of succinic acid by bacterial fermentation, Enzyme and Microbial Technology, 39(3), pp. 352-361.
Srivastava, A., et al. (2012) Reconstruction and visualization of carbohydrate, N-glycosylation pathways in Pichia pastoris CBS7435 using computational and system biology approaches, Systems and Synthetic Biology.
Stein, L. (2001) Genome annotation: from sequence to biology, Nat Rev Genet, 2(7), pp. 493-503.
Stephanopoulos, G., et al. (1998) Metabolic engineering : principles and methodologies, San Diego, Academic Press.
Su, A.I., et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes, Proc Natl Acad Sci U S A, 101(16), pp. 6062-6067.
Sun, J. and Zeng, A.P. (2004) IdentiCS--identification of coding sequence and in silico reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence, BMC Bioinformatics, 5, p. 112.
Sun, J., et al. (2007) Metabolic peculiarities of Aspergillus niger disclosed by comparative metabolic genomics, Genome Biol, 8(9), p. R182.
Suthers, P.F., et al. (2007) Metabolic flux elucidation for large-scale models using C-13 labeled isotopes, Metab. Eng., 9(5-6), pp. 387-405.
Tanabe, A., et al. (2009) Obesity causes a shift in metabolic flow of gangliosides in adipose tissues, Biochem Biophys Res Commun, 379(2), pp. 547-552.
Teh, K.Y. and Lutz, A.E. (2010) Thermodynamic analysis of fermentation and anaerobic growth of baker's yeast for ethanol production, J Biotechnol, 147(2), pp. 80-87.
Theilgaard, H., et al. (2001) Quantitative analysis of Penicillium chrysogenum Wis54-1255 transformants overexpressing the penicillin biosynthetic genes, Biotechnol Bioeng, 72(4), pp. 379-388.
Thiele, I. and Palsson, B.O. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction, Nat Protoc, 5(1), pp. 93-121.
Thiele, I., et al. (2013) A community-driven global reconstruction of human metabolism, Nat Biotechnol.
Thykaer, J. and Nielsen, J. (2003) Metabolic engineering of beta-lactam production, Metab Eng, 5(1), pp. 56-69.
Titov, V.N., et al. (2010) [Methylglyoxal--test for biological dysfunctions of homeostasis and endoecology, low cytosolic glucose level, and gluconeogenesis from fatty acids], Ter Arkh, 82(10), pp. 71-77.

Tomita, M., et al. (1997) E-CELL: Software Environment for Whole Cell Simulation, Genome informatics. Workshop on Genome Informatics, 8, pp. 147-155.
Uhlen, M., et al. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics, Mol Cell Proteomics, 4(12), pp. 1920-1932.
Uhlen, M., et al. (2010) Towards a knowledge-based Human Protein Atlas, Nat Biotechnol, 28(12), pp. 1248-1250.
Usaite, R., et al. (2006) Global transcriptional and physiological responses of Saccharomyces cerevisiae to ammonium, L-alanine, or L-glutamine limitation, Appl Environ Microbiol, 72(9), pp. 6194-6203.
Waldrop, M.M. (1992) Complexity: the emerging science at the edge of order and chaos, New York, Simon & Schuster.
van den Berg, M.A., et al. (2008) Genome sequencing and analysis of the filamentous fungus Penicillium chrysogenum, Nat Biotechnol, 26(10), pp. 1161-1168.
van Hoek, M.J. and Hogeweg, P. (2009) Metabolic adaptation after whole genome duplication, Molecular biology and evolution, 26(11), pp. 2441-2453.
Wang, L. and Hatzimanikatis, V. (2006) Metabolic engineering under uncertainty--II: analysis of yeast metabolism, Metab Eng, 8(2), pp. 142-159.
Wang, Y., et al. (2012) Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE, BMC Syst Biol, 6, p. 153.
Varma, A. and Palsson, B.O. (1994a) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110, Appl Environ Microbiol, 60(10), pp. 3724-3731.
Varma, A. and Palsson, B.O. (1994b) Metabolic Flux Balancing - Basic Concepts, Scientific and Practical Use, Bio-Technology, 12(10), pp. 994-998.
Verhoef, P., et al. (2004) Dietary serine and cystine attenuate the homocysteine-raising effect of dietary methionine: a randomized crossover trial in humans, American Journal of Clinical Nutrition, 80(3), pp. 674-679.
Whelan, K.E. and King, R.D. (2008) Using a logical model to predict the growth of yeast, BMC Bioinformatics, 9, p. 97.
Wiechert, W. (2001) 13C metabolic flux analysis, Metab Eng, 3(3), pp. 195-206.
Willke, T. and Vorlop, K.D. (2004) Industrial bioconversion of renewable resources as an alternative to conventional chemistry, Appl Microbiol Biotechnol, 66(2), pp. 131-142.
Wintermute, E.H. and Silver, P.A. (2010) Emergent cooperation in microbial metabolism, Mol Syst Biol, 6, p. 407.
Wishart, D.S., et al. (2007) HMDB: the Human Metabolome Database, Nucleic Acids Res, 35(Database issue), pp. D521-526.
Vitkup, D., et al. (2006) Influence of metabolic network structure and function on enzyme evolution, Genome Biol, 7(5), p. R39.
Vongsangnak, W., et al. (2008) Improved annotation through genome-scale metabolic modeling of Aspergillus oryzae, BMC Genomics, 9, p. 245.
Vongsangnak, W., et al. (2010) Integrated analysis of the global transcriptional response to alpha-amylase over-production by Aspergillus oryzae, Biotechnol Bioeng.
Wu, C., et al. (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources, Genome Biol, 10(11), p. R130.
Väremo, L., et al. (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods., Nucleic Acids Research, p. 10.1093/nar/gkt1111.
Xu, N., et al. (2013) Reconstruction and analysis of the genome-scale metabolic network of Candida glabrata, Mol Biosyst, 9(2), pp. 205-216.
Yang, K.M., et al. (2012) Ethanol reduces mitochondrial membrane integrity and thereby impacts carbon metabolism of Saccharomyces cerevisiae, FEMS Yeast Res, 12(6), pp. 675-684.

Paper I

Genome-scale modeling enables metabolic engineering of *Saccharomyces cerevisiae* for succinic acid production

Agren, R., Otero, J.M. and Nielsen, J.

J Ind Microbiol Biot (2013), doi:10.1007/s10295-013-1269-3

Genome-scale modeling enables metabolic engineering of *Saccharomyces cerevisiae* for succinic acid production

Authors: Rasmus Agren[1], José Manuel Otero[1,2], Jens Nielsen[1*]
1. Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-41296 Gothenburg, Sweden.
2. Current Address: Vaccine & Biologics Process Development, Bioprocess Research & Development, Merck Research Labs, West Point, PA, USA.

*Corresponding author:
Jens Nielsen
Phone: +46 (0) 31 772 8633
Fax: +46 (0) 31 772 3801
nielsenj@chalmers.se
Keywords: succinic acid, metabolic engineering, genome-scale modeling, DNA microarrays, *Saccharomyces cerevisiae*

Abstract

In this work, we describe the application of a genome-scale metabolic model and flux balance analysis for the prediction of succinic acid overproduction strategies in *Saccharomyces cerevisiae*. The top three single gene deletion strategies, *Δmdh1*, *Δoac1*, and *Δdic1*, were tested using knock-out strains cultivated anaerobically on glucose, coupled with physiological and DNA microarray characterization. While *Δmdh1* and *Δoac1* strains failed to produce succinate, *Δdic1* produced 0.02 C-mol/C-mol glucose, in close agreement with model predictions (0.03 C-mol/C-mol glucose). Transcriptional profiling suggests that succinate formation is coupled to mitochondrial redox balancing, and more specifically, reductive TCA cycle activity. While far from industrial titers, this proof-of-concept suggests that in silico predictions coupled with experimental validation can be used to identify novel and non-intuitive metabolic engineering strategies.

Introduction

The chemical manufacturing industry is actively seeking cost-effective, environmentally friendly, renewable, and sustainable raw material feedstocks that will not only enable production of key chemical building blocks, but can serve as a platform for future products [28]. In 2004, the US Department of Energy identified succinic acid as an added-value chemical building block, with an estimated 15,000 t/year world-wide demand. The demand is predicted to expand to commodity chemical status with 270,000 t/year, representing a potential >2 billion USD annual market [22,40]. Within microbial metabolism succinate formation results from two routes: 1) the isocitrate lyase, Icl1p, catalyzed conversion of isocitrate to equimolar glyoxylate and succinate, and 2) from the α-keto-glutarate dehydrogenase complex, Kgd1p/Kgd2p/Lpd1p, catalyzed conversion of α-keto-glutarate to equimolar succinate, with a net production of CO2, NADH, and ATP. Succinate is subsequently depleted by the succinate dehydrogenase complex, Sdh1p/Sdh2p/Sdh3p/Sdh4p to equimolar fumarate with the net production of protonated ubiquinone [10].

Numerous industrial biotechnology efforts have focused on metabolic engineering of prokaryotes to overproduce succinic acid, including *Anaerobiospirillium succiniciproducens*, *Actinobacillus succinogenes*, *Corynebacterium glutanicum*, *Mannheimia succiniciproducens*, *Prevotella ruminocola*, *Succinivibrio dextrinosolvens*, and a metabolically engineered succinic acid over-producing *Escherichia coli*, have been presented [36]. These hosts all grow at neutral pH, which results in secretion of the salt form, succinate, rather than the acid form. This in turn requires a costly acidification and precipitation step to produce succinic acid, which is the desired product. This is a general concern when using microbial cell factories for the production of organic acids [34].

Saccharomyces cerevisiae represents a well-established, generally regarded as safe, robust, scalable industrial production host capable of growth on diverse carbon sources, chemically defined medium, both aerobic and anaerobic, and with a wide pH operating range (3.0-6.0). However, unlike the bacterial hosts described above succinate does not natively accumulate in *S. cerevisiae*. There has so far been limited work on metabolic engineering of *S. cerevisiae* for production of succinic acid for industrial applications. Succinic acid production in genetically modified sake yeast strains has been demonstrated for modification of taste profiles, primarily focusing on multi-gene deletions of citric acid cycle enzymes aconitase (Aco1p), fumarate reductase (Osm1p), α-ketoglutarate dehydrogenase (Kgdp), fumarase (Fum1), and succinate dehydrogenase (Sdh1), resulting in <0.7 g/L succinic acid on complex medium [5,6,18]. There has also been significant experimental work focused on elucidating the physiological role of cytosolic and mitochondrial fumarate reductase (Frd1p and Osm1p, respectively) in the context of facilitating anaerobic fermentation of *S. cerevisiae* [4,8,12]. Significant effort has been applied to understand succinate formation in *S. cerevisiae* by exploring SDH1 and SDH3 deletion mutants, specifically

using ^{13}C-NMR analysis of ^{13}C-labelled aspartate and glutamate supplemented anaerobic glucose fermentations, and DNA microarray analysis of aerobic and anaerobic glucose supplemented fermentations, respectively [9,11]. In both efforts, no significant succinate accumulation was observed through simple deletion of the primary succinate consuming reaction, catalyzed by the succinate dehydrogenase complex. The most successful metabolic engineering attempt to date has been by Raab et al. [33]. They pursued an oxidative production route for succinate by a quadruple deletion of *SDH1*, *SDH2*, *IDP1*, and *IDH1*. This results in an interrupted TCA cycle and flux being redirected through the glyoxylate cycle instead. Following this approach they could demonstrate a 0.07 C-mol/C-mol glucose succinate yield.

Genome-scale metabolic models (GEMs), extensively described and reviewed elsewhere [19,25,32], provide a quantitative framework for stoichiometric biochemical models annotated with gene identity, coupled with mass-balance boundary conditions, to enable simulations of how the metabolic network operates under different conditions. For *S. cerevisiae*, the most well characterized eukaryote in systems biology, a number of GEMs have been developed. The models differ in scope, compartmentalization and intended applications [26]. GEMs have been widely used to identify metabolic engineering targets *in silico*, e.g. by using an evolutionary programming method and flux balance analysis (FBA) for identification of multiple knockout targets [31]. Otero et al. recently applied this method for the identification of sets of gene deletions which would link succinate production to growth in *S. cerevisiae* [27]. The proposed strategy relies on a triple deletion of *SDH1*, *SER3*, and *SER33*, which leads to disabled serine synthesis from glycolysis. Since serine is required for growth it must then be synthesized from glycine. Glycine production, in turn, is coupled to succinate production through the glyoxylate shunt. Following this strategy the authors reported a succinate yield of 0.02 C-mol/C-mol glucose, but the strain relied on glycine supplementation. Adaptive evolution followed by additional metabolic engineering steps resulted in a succinate yield of 0.05 C-mol/C-mol glucose without the need for glycine supplementation.

Here we use FBA to explore succinic acid overproduction strategies based on single and double gene deletions. Unlike the previously mentioned studies we focus primarily on anaerobic fermentation conditions, since it is a significant advantage from an industrial viewpoint to be able to run fermentations anaerobically. The top three single gene deletion strategies, identified under anaerobic glucose fermentation conditions, were experimentally evaluated. Furthermore, these three strains were physiologically and transcriptionally characterized with the objective of gaining further knowledge into the C4 acid production by *S. cerevisiae*.

Materials and methods

Modeling

The iFF708 GEM was used for all simulations [14]. We chose to use iFF708 even though there are more recent GEMs available. This was because we believed that the relative small number of subcellular compartments and the focus on central carbon metabolism made it the most suited model for studying succinate production. The following compounds are necessary for growth in iFF708 and were unconstrained in all simulations: ammonia, phosphate, and sulfate. Ergosterol and zymosterol are necessary for growth under anaerobic conditions but were unconstrained in all simulations. A maintenance ATP requirement of 1 mmol/g-DCW/h was used, following the calculations in Forster et al. [14]. Unless otherwise stated, all simulation conditions shared an identical objective function: maximizing growth under a limiting glucose uptake rate. The glucose uptake rate, based on experimentally determined glucose uptake rates of the *S. cerevisiae* CEN.PK113-7D under batch aerobic glucose fermentation conditions (see Table 2), was fixed to 15.2 C-mmol/g-DCW/h (91.2 mmol/g-DCW/h). For simulations referred to as aerobic or semi-aerobic the oxygen uptake rate, rO_2, was unconstrained or constrained to 1.8 mmol O_2/g-DCW/L, respectively. For simulations referred to as anaerobic, the rO_2 was constrained to 0.016 mmol O_2/g-DCW/L, rather than strictly zero. This was because a small succinate production, 0.003 C-mol/C-mol glucose, is predicted under anaerobic conditions. If succinate production is constrained to zero the model predicts no growth. However, this behavior is not seen experimentally. In the model this is because the production of orotate from dihydryoortate, catalyzed by dihydroorotate dehydrogenase (encoded by *URA1*) and required for pyrimidine synthesis, is coupled to the reduction of ubiquinone to ubiquinol. Under aerobic conditions oxygen serves as the final electron acceptor and enables ubiquinone regeneration, while under anaerobic conditions flavin adenine dinucleutoide (FAD) serves as the electron acceptor for ubiquinone regeneration. FAD must then be regenerated by the transfer of electrons to fumarate, producing succinate. Given that experimentally it would be difficult to ensure 0 mmol O_2, potential gene deletions were therefore screened for micro aerobic conditions, where rO_2 was constrained to 0.016 mmol O_2/g-DCW/h, which was the minimum rO_2 required for sustaining cell growth at the same rate regardless of whether succinate production is constrained to zero or unconstrained.

The gene deletions were tested for using a brute force approach where all combinations of single or double deletions were evaluated, rather than by using a faster algorithm like OptKnock [7]. This was because any deletion strategy was to be evaluated experimentally, and only single or double deletions were within the scope of this project. All simulations were carried out using the RAVEN Toolbox [1].

Strains

The reference strain Saccharomyces cerevisiae BY4741 (*MATa; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0*) and the single deletion strains were all received from the European *Saccharomyces Cerevisiae* Archive for Functional Analysis (Frankfurt, Germany). The reference strain *Saccharomyces cerevisiae* CEN.PK113-7D (*MATa; URA3; HIS3 LEU2; TRP1; MAL2-8C; SUC2*) was received from the Scientific Research and Development GmbH (Oberursel, Germany) [37]. The single gene-deletion knock-out strains used throughout this study and their corresponding genotype are presented in Table 1.

Table 1. *Saccharomyces cerevisiae* strain description and genotype.

Strain name	Strain genotype	Source
CEN.PK113-7D	MATa URA3 HIS3 LEU2 TRP1 SUC2 MAL2-8[3]	SRD GmbH[1]
Reference (REF)	BY4741: MATa; his3Δ1; leu2Δ0' met15Δ0; ura3Δ0	EURO-SCARF[2]
ΔMDH1	BY4741: MATa; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0; YKL085w:kanMX4	
ΔOAC1	BY4741: MATa; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0; YKL120w:kanMX4	
ΔDIC1	BY4741: MATa; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0; YLR348c:kanMX4	
ΔSDH3[2]	BY4743: MATa/α; his3Δ1/ his3Δ1; leu2Δ0/ leu2Δ0; met15Δ0/ met15Δ0; ura3Δ0/ ura3Δ0; YKL141w:kanMX4/YKL141w	

[1]Scientific Research and Development GmbH (Oberursel, Germany)
[2]European Saccharomyces Cerevisiae Archive for Functional Analysis (Frankfurt, Germany)
[3]All strains were haplioid of mating typ a, with the exception of the *ΔSDH3* strain, which is diploid and mating type a/α. A haploid strain of *ΔSDH3* was reported as not viable

Medium formulation

A chemically defined minimal medium of composition 5.0 g L^{-1} $(NH_4)_2SO_4$, 3.0 g L^{-1} KH_2PO_4, 0.5 g L^{-1} $MgSO_4 \cdot 7H_2O$, 1.0 mL L^{-1} trace metal solution, 300 mg L^{-1} uracil, 800 mg L^{-1} lysine, 200 mg L^{-1} histidine, 200 mg L^{-1} methionine, 0.05 g L^{-1} antifoam 204 (Sigma-Aldrich A-8311), and 1.0 mL L^{-1} vitamin solution was used for all shake flask and 2L well-controlled fermentations [39]. The trace element solution included 15 g L^{-1} EDTA, 0.45 g L^{-1} $CaCl_2 \cdot 2H_2O$, 0.45 g L^{-1} $ZnSO_4 \cdot 7H_2O$, 0.3 g L^{-1} $FeSO_4 \cdot 7H_2O$, 100 mg L^{-1} H_3BO_4, 1 g L^{-1} $MnCl_2 \cdot 2H_2O$, 0.3 g L^{-1} $CoCl_2 \cdot 6H_2O$, 0.3 g L^{-1} $CuSO_4 \cdot 5H_2O$, 0.4 g L^{-1} $NaMoO_4 \cdot 2H_2O$. The pH of the trace metal solution was adjusted to 4.00 with 2M NaOH and heat sterilized. The vitamin solution included 50 mg L^{-1} d-biotin, 200 mg L^{-1} para-amino benzoic acid, 1 g L^{-1} nicotinic acid, 1 g L^{-1} Ca•pantothenate, 1 g L^{-1} pyridoxine HCl, 1 g L^{-1} thiamine HCl, and 25 mg L^{-1} m•inositol. The pH of the vitamin solution was adjusted to 6.5 with 2M NaOH, sterile-filtered and the solution was stored at 4°C. The final formulated medium, excluding glucose and vitamin solution supplementation, is adjusted to pH 5.0 with 2M NaOH and heat sterilized. For carbon-limited cultivations the sterilized medium is supplemented with 20 g L^{-1} glucose, heat sterilized separately, and 1.0 mL L^{-1} vitamin solution is added by sterile filtration (0.20 µm pore size Ministart®-Plus Sartorius AG, Goettingen, Germany). For anaerobic fermentations a total of 4 g L^{-1} ergosterol and 168 g L^{-1} Tween 80 dissolved in pure ethanol was supplemented.

Shake flask cultivations and stirred tank fermentations

Shake flask cultivations were completed in 500 mL Erlenmeyer flasks with two diametrically opposed baffles and two side-necks with septums for sampling by syringe. Flasks were heat sterilized with 100 mL of medium, inoculated with a single colony, and incubated at 30°C with orbital shaking at 150 RPM. Stirred tank fermentations were completed in well-controlled, aerobic or anaerobic, 2.2L Braun Biotech Biostat B fermentation systems with a working volume of 2L (Sartorius AG, Goettingen, Germany). The temperature was controlled at 30°C. The fermenters were outfitted with two disk-turbine impellers rotating at 600 RPM. Dissolved oxygen was monitored with an autoclavable polarographic oxygen electrode (Mettler-Toledo, Columbus, OH). During aerobic cultivation the air sparging flow rate was 1 vvm. During anaerobic cultivation nitrogen containing less than 5 ppm O_2 was used for sparging at a constant flow rate of 2 vvm, with less than 1% air saturated oxygen in the fermenter as confirmed by dissolved oxygen and off-gas analysis. The higher flow rate of 2 vvm was employed to ensure anaerobic conditions; however, it is acknowledged that ethanol stripping was likely to increase. The pH was kept constant at 5.0 by automatic addition of 2M KOH. Off-gas passed through a condenser to minimize the evaporation from the fermenter. The fermenters were inoculated from shake flask precultures to an initial OD_{600} 0.005.

Fermentation analysis

Off-gas analysis: The effluent fermentation gas was measured every 30 seconds for determination of $O_{2(g)}$ and $CO_{2(g)}$ concentrations by the off-gas analyzer Brüel and Kjær 1308 (Brüel & Kjær, Nærum, Denmark).

Biomass determination: The optical density (OD) was determined at 600 nm using a Shimadzu UV mini 1240 spectrophotometer (Shidmazu Europe GmbH, Duisberg, Germany). Duplicate samples were diluted with deionized water to obtain OD_{600} measurements in the linear range of 0-0.4 OD_{600}. Samples were always maintained at 4°C post-sampling until OD_{600} and dry cell weight (DCW) measurements were performed. DCW measurements were determined through the exponential phase, until stationary phase was confirmed according to OD_{600} and off-gas analysis. Nitrocellulose filters (0.45 µm Sartorius AG, Goettingen, Germany) were used. The filters were pre-dried in a microwave oven at 150W for 10 min., and cooled in a desiccator for 10 min. 5.0 mL of fermentation broth were filtered, followed by 10 mL DI water. Filters were then dried in a microwave oven for 20 min. at 150W, cooled for 15 minutes in a desiccator, and the mass was determined.

Metabolite concentration determination: All fermentation samples were immediately filtered using a 0.45 µm syringe-filter (Sartorius AG, Goettingen, Germany) and stored at -20°C until further analysis. Glucose, ethanol, glycerol, acetate, succinate, pyruvate, fumarate, citrate, oxalate, and malate were determined by HPLC analysis using an Aminex HPX-87H ion-exclusion column (Bio-Rad Laboratories, Hercules, CA). The column was maintained at 65°C and elution performed using 5 mM H_2SO_4 as the mobile phase at a flow rate of 0.6 mL min^{-1}. Glucose, ethanol, glycerol, acetate, succinate, citrate, fumarate, malate, oxalate were detected on a Waters 410 differential refractometer detector (Shodex, Kawasaki, Japan), and acetate and pyruvate were detected on a Waters 468 absorbance detector set at 210 nm.

Transcriptomics

RNA sampling and isolation: Samples for RNA isolation from the late-exponential phase of glucose-limited batch cultivations were taken by rapidly sampling 25 mL of culture into a 50 mL sterile Falcon tube with 40 mL of crushed ice in order to decrease the sample temperature to below 2°C in less than 10 seconds. Cells were immediately centrifuged (4000 RPM at 0°C for 2.5 min.), the supernatant discarded, and the pellet frozen in liquid nitrogen and it was stored at -80°C until total RNA extraction. Total RNA was extracted using the FastRNA Pro RED kit (QBiogene, Carlsbad, USA) according to manufacturer's instructions after partially thawing the samples on ice. RNA sample integrity and quality was determined prior to hybridization with an Agilent 2100 Bioanalyzer and RNA 6000 Nano LabChip kit according to the manufacturer's instruction (Agilent, Santa Clara, CA).

Probe preparation and hybridization to DNA microarrays: mRNA extraction, cDNA synthesis, labeling, and array hybridization to Affymetrix Yeast Genome Y2.0 arrays were performed according to the manufacturer's recommendations (Affymetrix GeneChip® Expression Analysis Technical Manual, 2005-2006 Rev. 2.0). Washing and staining of arrays were performed using the GeneChip Fluidics Station 450 and scanning with the Affymetrix GeneArray Scanner (Affymetrix, Santa Clara, CA).

Microarray gene transcription analysis: Affymetrix Microarray Suite v5.0 was used to generate CEL files of the scanned DNA microarrays. These CEL files were then processed using the statistical language and environment R v5.3 (R Development Core Team, 2007, www.r-project.org), supplemented with Bioconductor v2.3 (Biconductor Development Core Team, 2008, www.bioconductor.org) packages Biobase, affy, gcrma, and limma. The probe intensities were normalized for background using the robust multi-array average (RMA) method only using perfect match (PM) probes after the raw image file of the DNA microarray was visually inspected for acceptable quality. Normalization was performed using the qspline method and gene expression values were calculated from PM probes with the median polish summary. Statistical analysis was applied to determine differentially expressed genes using the limma statistical package. Moderated t-tests between the sets of experiments were used for pair-wise comparisons. Empirical Bayesian statistics were used to moderate the standard errors within each gene and Benjamini-Hochberg's method was used to adjust for multi-testing. A cut-off value of adjusted $p<0.05$ was used for statistical significance, unless otherwise specified [35]. Gene Ontology process annotation was performed by submitting differentially expressed gene (adjusted $p<0.05$) lists to the Saccharomyces Genome Database GO Term Finder resource and maintaining a cut-off value of $p<0.01$ [10].

Table 2. Physiological characterization.

Strain	CEN.PK113-7D Aerobic		CEN.PK113-7D		Reference (BY4741)		*Δmdh1*		*Δdic1*		*Δoac1*		*Δsdh3*
	Value	± σ	Value	± σ	Value	± σ	Value	± σ	Value	± σ	Value	± σ	Value
Specific growth rate (h^{-1})	0.38	0.00	0.29	0.01	0.28	0.00	0.28	0.01	0.24	0.02	0.23	<0.00	0.26
Product Yields (C-mol/C-mol substrate)													
Y_{SX}	0.17	0.02	0.13	0.01	0.12	0.02	0.17	0.06	0.12	0.00	0.13	0.00	0.18
Y_{SEtOH}	0.54	0.04	0.44	0.03	0.45	0.01	0.48	0.00	0.50	0.01	0.54	0.01	0.53
Y_{SCO2}	0.16	0.00	0.12	0.00	0.13	0.03	0.14	0.01	0.11	0.00	0.14	0.00	0.18
Y_{SAcet}	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Y_{SGlyc}	0.08	0.03	0.11	0.30	0.08	0.01	0.09	0.00	0.09	0.00	0.09	0.00	0.11
Y_{SSuc}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00
Y_{SPyr}	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Productivities (C-mmol/g-DCW/h)													
r_{Gluc}	91.2	6.0	93.1	4.0	89.7	2.8	68.6	18.2	74.5	2.0	73.3	3.0	68.5
r_{EtOH}	49.7	6.6	41.0	5.5	40.8	0.8	32.9	8.7	37.4	1.9	39.2	3.4	36.0
r_{CO2}	15.4	0.0	11.1	0.0	11.3	3.3	9.7	2.1	7.9	0.1	9.9	0.1	12.0
r_{Acet}	0.7	0.3	0.4	0.1	0.3	0.4	0.3	0.1	0.2	0.1	0.3	0.1	0.4
r_{Glyc}	7.3	3.4	10.5	2.2	3.5	4.8	5.9	1.6	6.7	0.3	6.7	0.3	7.3
r_{Suc}	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	1.6	0.1	0.1	0.1	0.3
r_{Pyr}	0.4	0.0	0.3	0.0	0.1	0.1	0.2	0.1	0.1	0.1	0.2	0.1	0.2
r_{O2}	1.8	0.3	0.0	0.0	0.1	0.2	0.9	0.0	0.2	0.2	0.0	0.2	0.4
Carbon Recovery (%)	96.3	4.0	81.30	3.2	79.0	6.2	88.6	6.7	84.7	1.0	89.6	1.0	100.3
Succinate Titer (g/L)	0.02	0.01	0.00	0.00	0.03	0.00	0.03	0.00	0.23	0.03	0.02	0.03	0.04
Biomass Titer (g/L)	1.96	0.08	2.00	0.12	2.05	0.13	2.07	0.198	1.97	0.00	2.08	0.00	2.23
Y_{XSuc} (g/g-biomass)	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.11	0.02	0.01	0.02	0.02

Results

Model validation and comparison to experimental data

The predictive power of the model was evaluated using fermentation data of *S. cerevisiae* CEN.PK113-7D (see Table 2). Batch aerobic and anaerobic glucose fermentations were performed in well-controlled 2L fermentations, and compared to corresponding simulation conditions. The objective function, growth, was maximized while constraining glucose uptake rate, and for anaerobic conditions, constraining the oxygen uptake rate (rO_2) to 0.016 mmol O_2/g-DCW/h as discussed in Methods. Table 2 demonstrates that under aerobic conditions, 96.3 ± 4.0% of all carbon is recovered, and distributed across ethanol (54%), acetate (1%), glycerol (8%), carbon dioxide (16%), and biomass (17%) formation. Fig 1 shows the results of simulated carbon distributions and the specific growth rate when oxygen was unconstrained. As can be seen there is a poor agreement with batch glucose aerobic experimental data due to the inability of the model to describe the Crabtree effect, as discussed earlier by Akesson et al. [2]. When rO_2 was constrained to experimentally determined fermentation values of 1.8 mmol O_2/g-DCW/h, referred to as semi-aerobic, the simulation accurately predicted the specific growth rate (0.38 vs. 0.40h^{-1}, experimental vs. simulation, respectively), ethanol yield (0.54 vs. 0.54 C-mol/C-mol glucose), and biomass yield (0.17 vs. 0.18 C-mol/C-mol glucose). However, carbon dioxide (0.16 vs. 0.30 C-mol/C-mol glucose) and glycerol (0.08 vs. 0.0 C-mol/C-mol glucose) yields were in poor agreement. While the relatively high carbon recovery observed experimentally in aerobic batch glucose fermentation suggests carbon dioxide measurements were accurate, it should be noted the theoretical ratio of carbon dioxide to ethanol production under purely fermentative glucose metabolism is 1:2, and experimentally under both aerobic, and anaerobic conditions in CEN.PK113-7D and BY4741 the ratio observed is 1:3 [23]. The original iFF708 model's ability to predict carbon dioxide production rates was validated experimentally with aerobic glucose-limited continuous cultivation data, and demonstrated excellent fit between dilution rates 0.1 and 0.38h^{-1}, representing a broad span of respiratory quotients

[13]. This therefore suggests that carbon dioxide metabolism in CEN.PK113-7D and BY4741 under batch glucose fermentation conditions deviates from theoretical expectations, or when considered in the context of a highly interconnected network, not fully described by the stoichiometry of iFF708. It is not expected that the discrepancy in carbon dioxide predictive power would significantly alter the succinate overproduction strategies identified. It is further interesting to note that in the same work, the only data point not predicted by the original iFF708 was the glycerol production rate at the higher dilution rate, $0.38h^{-1}$, most representative of batch conditions [13].

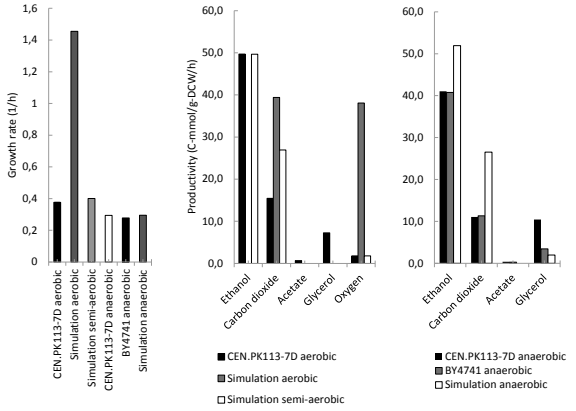


Fig 1. Comparison between experimental and simulated fermentation data. Comparison of the specific growth rate and specific productivities for simulated data and experimental data generated using the reference *S. cerevisiae* CEN.PK113-7D and BY4741 under aerobic and anaerobic glucose batch fermentations. For the condition, simulation aerobic, simulation semi-aerobic, simulation anaerobic, the rO_2 was unconstrained, constrained to 1.8 mmol-O_2/g-DCW/h, and constrained to 0.016 mmol-O_2/g-DCW/h, respectively. For aerobic experimental data the specific glucose uptake rate was 91.2 C-mmol/g-DCW/h for CEN.PK113-7D. For anaerobic experimental data the specific glucose uptake rate was 93.1 C-mmol/g-DCW/h for CEN.PK113-7D and 89.7 C-mmol/g-DCW/h for BY4741. For all simulation conditions the glucose uptake rate was constrained to 91.2 C-mmol/g-DCW/h.

Biomass formation as a result of glucose respiro-fermentative metabolism, with a high dependence on oxygen availability and glucose concentration, results in the formation of excess NADH [24]. Excess NADH, both cytosolic and mitochondrial, is a direct result of biomass required ATP generation, and compartmental redox balance is possible through cytosolic NADH dehydrogenases, the glycerol-3-phosphate shuttle, and mitochondrial redox shuttles [15,17,20,29]. Glycerol formation results from redox balancing. NADH regeneration to NAD^+ in the cytosol, with subsequent glycerol production, can be reduced through expression of a cytosolic NADH oxidase [38].

Model validation was initially performed using *S. cerevisiae* CEN.PK113-7D batch glucose aerobic fermentation data; however, realizing that succinate metabolic engineering strategies would likely require exploration of anaerobic metabolism, similar comparative analysis for anaerobic fermentations was performed. More specifically, the reference *S. cerevisiae* BY4741 was also included, noting that gene deletion strategies to be identified in silico could rapidly be evaluated in vivo using the systematic Yeast Knock-Out (YKO) library available from the *Saccharomyces* Gene Deletion Project [41]. Under anaerobic conditions, the carbon recovery for both strains CEN.PK113-7D and BY4741 are significantly less compared to aerobic conditions (Table 2); however, when evaluating experimental and simulation values for specific growth rate and specific productivities there is reasonable agreement. Specifically, for CEN.PK113-7D, BY4741, and anaerobic simulations the specific growth rate was 0.29, 0.27, and 0.29 h^{-1}, respectively. For ethanol (41.0, 40.8, 51.9 C-mmol/g-DCW/h), glycerol (10.5, 3.5, 2.0 C-mmol/g-DCW/h), and carbon dioxide (11.1, 11.3, 26.5 C-mmol/g-DCW/h) specific productivities the agreement between experimental and model simulations were fair, but indicating that the lack of carbon recovery is likely a result of ethanol stripping and evaporation from the bioreactor. Emphasis was therefore placed on ensuring simulation conditions and constraints captured experimentally observed metabolite production, with less focus on matching exact flux values.

Gene deletion strategies for succinate overproduction

Overproduction of succinic acid was evaluated in silico using the various simulation conditions previously described. Prior to investigating those results, the maximum theoretical yield of succinic acid was determined in silico. Assuming 1 mmol ATP/g-DCW/h maintenance cost and a 10 mmol glucose/g-DCW/h uptake rate, the maximum succinate yield is 0.51 g/g-glucose. This maximum yield is based on FBA when $[H^+]$ was balanced. The exact mechanism by which succinate is transported across the cytosolic membrane has yet to be clearly elucidated, with literature suggesting both dicarboxylic acid proton-coupling, and the absence of such coupling [3]. If $[H^+]$ is treated as an external metabolite (e.g., unconstrained), the maximum yield of succinate is 0.98 g/g-glucose. Furthermore, if carbon dioxide uptake is permitted, enabling carboxylation reactions, the maximum theoretical yield is 1.124 g/g-glucose. Given the lack of physiological characterization of succinate transport, and the relatively high impact of assumptions surrounding $[H^+]$ balancing, external $[H^+]$ was balanced throughout all simulations, and the maximum succinate yield was assumed to be 0.51 g/g-glucose (0.52 C-mol/C-mol glucose). This represents a worst case scenario in terms of the

theoretical potential for *S. cerevisiae* to stoichiometrically overproduce succinate.

Under aerobic conditions there are no single gene deletions that result in increased succinate production (see Online Resource Table 1). Interestingly, the reference case simulation under aerobic conditions with no gene deletions produces a small amount of succinate (0.003 C-mol/C-mol glucose), which is not observed experimentally. If succinate excretion is constrained to zero then optimization of growth results in a similar growth rate but while producing glycerol, under minimal amounts of oxygen, and then acetate under increasing amounts of oxygen. However, experimentally, both glycerol and acetate production are observed while succinate production is absent. Under aerobic conditions there is a strong sensitivity of succinate yield on substrate to rO_2 and for $rO_2 > 2$ mmol O_2/g-DCW/h the succinate yield on substrate is zero (to be discussed later).

Under aerobic conditions double gene deletions only resulted in minor improvement of succinate production (data not shown). Nearly all of the predictions required the deletion of the succinate dehydrogenase complex (Sdh3p), which catalyzes the conversion of succinate to fumarate in the TCA cycle, and represents the primary succinate consumption reaction in *S. cerevisiae* central carbon metabolism. In addition to previous work suggesting that succinate dehydrogenase complex interruption does not lead to succinate accumulation [9,11], Table 2 confirms that deletion *Δsdh3* in the BY4741 strain also fails to accumulate succinate.

Table 3 presents the top single gene deletions for succinate overproduction under anaerobic conditions. It shows that a significant increase in the succinate yield, by a factor of approximately 10-fold from the simulated reference case, can be obtained for the single gene deletions *Δoac1, Δmdh1*, and *Δdic1* (0.033 C-mol/C-mol glucose vs. 0.003 C-mol/C-mol glucose, single gene deletion vs. reference case simulation, respectively). Furthermore the significant increase in succinate yield on substrate resulted in nearly no impact on growth rate ($0.28h^{-1}$ vs. $0.30h^{-1}$, single gene deletion vs. reference case simulation, respectively). Physiologically, it was confirmed that *Δoac1*, *Δmdh1*, and *Δdic1* are viable null mutants, and their annotation is well known, encoding for an inner mitochondrial membrane transporter (OAC1p), malate dehydrogenase (MDH1p), and an inner dicarboxylate mitochondrial transporter (DIC1p), respectively [10]. Interestingly, further simulations of the best double gene deletions resulted in the same order of magnitude succinate yields on substrate compared to the aforementioned single gene deletions.

Table 3. Top gene deletions under anaerobic constraints for succinate yield.

Simulation conditions	Genotype	Specific growth rate [h^{-1}]	Y_{SSuc} [C-mol/C-mol glucose]
TOP SINGLE GENE DELETIONS ANAEROBIC	No deletions	0.30	0.003
	Δ*oac1*	0.28	0.033
	Δ*mdh1*	0.28	0.033
	Δ*dic1*	0.28	0.032
	Δ*fum1*	0.30	0.005
	Δ*met22*	0.30	0.004
TOP DOUBLE GENE DELETIONS ANAEROBIC	No deletions	0.29	0.003
	Δ*mdh1*Δ*yat1*	0.25	0.061
	Δ*mdh1*Δ*cat2*	0.25	0.061
	Δ*dic1*Δ*yat1*	0.25	0.060
	Δ*dic1*Δ*cat2*	0.25	0.060
	Δ*dic1*Δ*cit2*	0.25	0.056
	Δ*mdh1*Δ*put2*	0.26	0.051
	Δ*mdh1*Δ*kgd1*	0.26	0.050
	Δ*mdh1*Δ*lsc2*	0.26	0.050
	Δ*dic1*Δ*oac1*	0.25	0.051
	Δ*dic1*Δ*lsc2*	0.26	0.049

Physiological characterization of gene deletion strains

In order to explore and validate if the single gene deletions identified in silico result in more succinate production, the corresponding strains of the BY4741 background (see Table 1) were cultivated anaerobically in 2L well controlled fermenters. Fermentation results are presented in Table 2, and comparative analysis between simulation and experimental results are presented in Fig 2. It is seen that there is a fair agreement between model predictions and experimental data. Focusing more closely on the specific succinate productivity, the reference case, *Δmdh1*, and *Δoac1* experimentally determined yields are significantly lower than expected based on model simulations. The *Δdic1* case, however, demonstrated a significantly higher yield of succinate compared to the reference case (0.02 vs. 0.00 C-mol/C-mol glucose, *Δdic1* vs. reference, respectively), and was in-line with the in silico prediction (0.02 vs. 0.03 C-mol/C-mol glucose, *Δdic1* experimental vs. *Δdic1* anaerobic simulation, respectively). This represents a significant improvement in succinate productivity based exclusively on a novel in silico prediction.

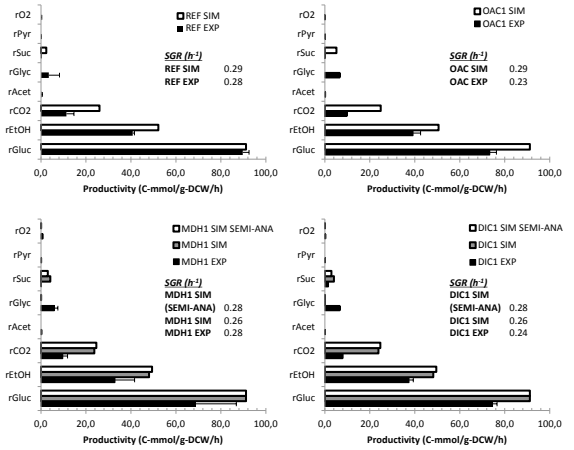


Fig 2. Experimental and simulation comparative data for reference, *Δoac1*, *Δmdh1*, and *Δdic1* strains. Summary of the specific growth rate (SGR) and specific consumption/productivity values for major carbon products (glucose, ethanol, carbon dioxide, acetate, glycerol, succinate, pyruvate, and oxygen) for both experimentally determined data of anaerobic batch glucose fermentations, and corresponding anaerobic simulation data of the BY4741 reference strain, and single gene deletion strains *Δmdh1*, *Δdic1*, and *Δoac1*. In general, the experimental data suggests a lower specific growth rate compared to the predicted growth rate, whether anaerobic simulations (referred to as SIM) or semi-anaerobic simulations (referred to as SIM SEMI-ANA) are considered. The simulation data for *Δmdh1* and *Δdic1* conditions attempt to highlight the significant sensitivity to relatively small changes in rO_2, where the SIM SEMI-ANA simulation constrains rO_2 to 0.02 mmol O_2/g-DCW/L compared to 0 mmol O_2/g-DCWL/L, while impacting growth rate significantly. Both glucose and oxygen are consumed; however, are presented as positive values. Succinate production was noted under all simulated conditions. However, it was only observed under the *Δdic1* experimental condition.

Transcriptome characterization of gene deletion strains

To gain further insight into the physiological performance of each strain identified via simulation results, genome-wide DNA microarray profiling was completed under anaerobic batch glucose fermentations. Table 4 provides an overall summary of the comparative transcriptome of differentially expressed genes between *Δdic1*, *Δmdh1*, and *Δoac1* strains, each compared to the reference strain. The number of differentially expressed genes for the *Δoac1* strain compared to the reference strain was very low, and consequently suggests that deletion of *Δoac1* causes virtually no transcriptional, and consequently, physiological differences compared to the reference BY4741 strain. The *Δdic1* and *Δmdh1* strains, compared to the reference strain, had 117 and 209 differentially expressed genes, respectively. Of these genes a total of 33% and 23% were up-regulated genes and 66% and 76% were down-regulated genes, for the *Δdic1* and *Δmdh1* strains, respectively. The average fold change of differentially expressed genes for the *Δdic1* strain, both up- and down-regulated, was ≈2.5-fold greater than *Δmdh1*. Given the relatively low differential expression for the *Δoac1* strain, no further analysis of the transcriptional data was performed for this strain.

The differentially expressed genes sets for *Δdic1* and *Δmdh1* were submitted for gene ontology (GO) process annotation. Table 5 presents the statistically significant GO process annotation terms, showing a high degree of similarity for the two strains, with changes mainly in genes involved in energy metabolism and electron transport. It's particularly interesting to note that there is a large overlap for the two strains and there were only four GO process categories that were unique to *Δmdh1* as compared to *Δdic1*, and these are involved in sterol transport, lipid transport, generation of precursor metabolites and energy, and energy derivation by oxidation of organic compounds.

Table 4. Summary of differentially expressed genes.

Comparative transcriptome	*Δdic1* vs. REF (n=2)	*Δmdh1* vs. REF (n=2)	*Δoac1* vs. REF[1] (n=2)
No. differentially expressed genes ($p\text{-value}_{B\text{-}H}$<0.01)	117	209	5
Up-regulated	39	49	3
Down-regulated	78	160	2
Average log-fold change (± std. dev)			
Up-regulated	1.45 (1.61)	1.09 (0.60)	-
Down-regulated	-1.98 (1.53)	-1.55 (1.23)	-

[1]For the comparison of *Δoac1* vs. REF, the $p\text{-value}_{B\text{-}H}$<0.1 criteria was applied and resulted in only 5 differentially expressed genes. Given the low number of differentially expressed genes no average log-fold change is reported.

Given the high degree of similarity in the GO process annotation for both the *Δdic1* and *Δmdh1* conditions, the complete list of differentially expressed genes were submitted for metabolic pathway annotation using the SGD Pathway Expression Viewer and Reactome databases [21,30]. The results are presented in Online Resource Table 2 with color coding of genes according to their log-fold change and direction of expression relative to the reference case. Only a relatively small number of metabolic genes are identified in *Δdic1* and *Δmdh1* compared to the reference; a total of 10 and 20 genes respectively. Perhaps more striking is that there is an overlap of 9 metabolic pathway genes between both *Δdic1* and *Δmdh1*. The only differentially

expressed gene present in the *Δdic1* condition, not present in the *Δmdh1* condition, is *Δdic1*.

Discussion

Succinic acid overproduction metabolic engineering strategies in *S. cerevisiae* are limited. The most successful metabolic engineering strategy to date relies on oxidative production of succinate and is based on a quadruple gene deletion, with the aim of redirection of flux from the TCA cycle to the glyoxylate cycle [33]. Here we evaluated specifically anaerobic growth conditions, and the top single gene deletion targets identified resulted in significantly higher succinate yields on substrate. However, the yield was not as high as what was reported using the oxidative route (0.02 C-mol/C-mol glucose vs. 0.07 C-mol/C-mol glucose).

Table 5. Process Gene Ontology annotation of differentially expressed genes of *ΔDIC1*:REF and *ΔMDH1*:REF.

Gene Ontology	Genes annotated in *Δdic1*	Genes annotated in *Δmdh1*	p-value *Δdic1*	p-value *Δmdh1*
mitochondrial electron transport, cytochrome C to oxygen	*COX4, COX6, CYC1, COX7, COX5A*	*COX4, COX6, CYC1, COX7, COX5A*	1.20E-04	4.75E-03
electron transport chain	*COX4, QCR10, COX6, CYC1, COX7, COX5A*	*COX4, QCR10, COX6, CYC1, COX7, COX5A, QCR2*	5.80E-04	3.04E-03
respiratory electron transport chain	*COX4, QCR10, COX6, CYC1, COX7, COX5A*	*COX4, QCR10, COX6, CYC1, COX7, COX5A, QCR2*	5.80E-04	3.04E-03
ATP synthesis coupled electron transport	*COX4, QCR10, COX6, CYC1, COX7, COX5A*	*COX4, QCR10, COX6, CYC1, COX7, COX5A, QCR2*	5.80E-04	3.04E-03
mitochondrial ATP synthesis coupled electron transport	*COX4, QCR10, COX6, CYC1, COX7, COX5A*	*COX4, QCR10, COX6, CYC1, COX7, COX5A, QCR2*	5.80E-04	3.04E-03
oxidation reduction	*COX4, QCR10, COX6, CYC1, COX7, COX5A*	*COX4, QCR10, COX6, CYC1, COX7, COX5A, QCR2*	5.80E-04	3.04E-03
sterol transport	-	*SWH1, SUT1, PDR11, DAN1, AUS1, HES1, SUT2*	-	5.14E-05
energy derivation by oxidation of organic compounds	-	*PET9, HOR2, BMH1, COX4, COX13, QCR10, COX6, PIG2, CYC1, MDH1, PET10, PUF3, NDE1, COX7, COX5A, QCR2*	-	1.84E-03
generation of precursor metabolites and energy	-	*PET9, HOR2, BMH1, HXK1, COX4, COX13, QCR10, COX6, PIG2, CYC1, MDH1, PET10, PUF3, NDE1, COX7, COX5A, PFK27, QCR2*	-	2.65E-03
lipid transport	-	*SWH1, SUT1, PDR11, DAN1, AUS1, HES1, FAA1, SUT2*	-	2.94E-03

As discussed in Introduction, a recent paper by Otero et al. also makes use of flux balance analysis for the purpose of succinate production in S. cerevisiae [27]. Attempts to reproduce those simulations resulted in significantly reduced succinate yield, and could only be obtained if a constraint preventing acetaldehyde secretion was imposed (data not shown). The underlying reason for this was that threonine aldolase was erroneously assigned to be irreversible in the direction of glycine and acetaldehyde to threonine in the original version of iFF708. In later models, and in the iFF708 version which we used, this has been corrected to be in the opposite direction [16]. This in turn provides a metabolic route for synthesis of serine from threonine, and therefore uncouples serine synthesis and succinate production.

The metabolic engineering strategies identified through *Δdic1*, *Δmdh1*, and *Δoac1*, suggest a common mechanism. Mitochondrial redox balance must be maintained, and while respiratory metabolic activity under anaerobic conditions is reduced compared to aerobic conditions, some activity is required to support glutamate/glutamine metabolism from α-keto-glutarate [8,9]. This results in the production of NADH. During anaerobic metabolism, NAD^+ regeneration occurs via the following pathways according to our simulations (where the subscript m denotes mitochondrial):

OAC1p: oxaloacetate → $oxaloacetate_m + H^+_m$

MDH1p: $oxaloacetate_m + NADH_m \rightarrow malate_m + NAD^+_m$

DIC1p: $malate_m$ + phosphate → malate + $phosphate_m$

MIR1p: $H^+_m + phosphate_m \rightarrow$ phosphate

Net reaction stoichiometry: oxaloacetate + $NADH_m \rightarrow$ malate + NAD^+_m

In the cytosol malate is then converted to oxaloacetate, and the resulting NADH is converted to NAD^+ with the production of glycerol. If we consequently assume that *Δmdh1* were deleted, then NAD^+ regeneration occurs in the following manner according to the simulations:

FUM1p: $malate_m \rightarrow fumarate_m$

NDIp: $ubiquinone_m + NADH_m \rightarrow ubiquinol_m + NAD^+_m$

SDH3p: $ubiquinol_m + FAD_m \rightarrow ubiquinone_m + FADH2_m$

OSM1p: $fumarate_m + FADH_{2,m} \rightarrow succinate_m + FAD_m$

DIC1p: malate + $succinate_m \rightarrow malate_m$ + succinate

Net reaction stoichiometry: malate + $NADH_m \rightarrow$ succinate + NAD^+_m

The above mechanism is highly dependent on several metabolic pathway assumptions, particularly that there are no other mitochondrial reactions capable of NAD^+_m regeneration. Also, the *Δmdh1* strategy is highly sensitive to rO_2, as shown in Fig 3, because of the succinate production driven requirement for electron donation from ubiquinone to succinate and not oxygen. Even small values of rO_2 (<0.1 mmol O_2/g-DCW/h) result in no succinate production. If *COX1* (encoding subunit I of cytochrome C oxidase) and *RIP1* (encoding ubiquinol cytochrome C reductase) are deleted in combination with *MDH1*, to eliminate oxygen reactivity, the rO_2 range across which succinate yield is observed for the *Δmdh1* strategy is extended to 0.6 mmol O_2/g-DCW/h (see Fig 3). Furthermore, it should be noted that additional multi-gene deletion strategies leveraging the general *Δmdh1* strategy could be expanded. A simple triple gene deletion strategy of *Δmdh1 Δcat2 Δcit2* was simulated (data not shown), and resulted in further improved succinate yield on glucose (0.08 C-mol/C-mol glucose vs. 0.03 C-mol/C-mol glucose for only *Δmdh1*). CAT2 and CIT2 encode carnitine acetyl-CoA transferase and citrate synthase, respectively.

The *Δdic1* strategy, relying on deletion of the mitochondrial dicarboxylate carrier DIC1p, catalyzes the following transport reaction, noting the intermediate transport of orthophosphate:

Dic1: malate + $succinate_m \rightarrow malate_m$ + succinate

(malate + $orthophosphate_m \rightarrow malate_m$ + orthophosphate)

(succinate + $orthophosphate_m \rightarrow succinate_m$ + orthophosphate)

Assuming *DIC1* deletion, then the resulting simulated pathway is:

NDIp: $ubiquinone_m + NADH_m \rightarrow ubiquinol_m + NAD^+_m$

SDH3p: $ubiquinol_m + FAD_m \rightarrow ubiquinone_m + FADH_{2,m}$

FRDS1p: fumarate + $FADH_{2,m} \rightarrow$ succinate + FAD_m

Net reaction stoichiometry: fumarate + $NADH_m \rightarrow$ succinate + NAD^+_m

The *Δdic1* strategy relies heavily on the sub-cellular localization and function of FRDS1p, soluble mitochondrial fumarate reductase, which continues to be poorly understood. However, recent work has suggested that a double deletion *S. cerevisiae* mutant, *Δosm1 Δfrds1*, failed to grow under batch glucose anaerobic conditions. Furthermore, during anaerobic growth, *FRDS1* expression in the wild-type was two to eight times higher than that of *OSM1*, suggesting that formation of succinate is strictly required for the re-oxidation of $FADH_2$ and its expression may be oxygen-regulated [8]. While neither *FRDS1* nor *OSM1* were significantly differentially expressed in the *Δmdh1* or *Δdic1* mutants compared to the reference strain, *FRDS1* was slightly up-regulated in the *Δdic1* mutant compared to the *Δmdh1* mutant (log10 fold change 0.11 vs. -0.10, respectively).

There was a strong upregulation of *CYC1* in both the *Δdic1* and *Δmdh1* mutants. *CYC1* facilitates electron transfer from ubiquinone cytochrome C oxidoreductase to cytochrome C oxidase. This direction, which is the normal oxidative route and ends in reduction of O_2, would not be possible under fully anaerobic conditions. The upregulation can therefore be viewed as a coping strategy to deal with the stress of redox imbalance. Deletion of *CYC1* could therefore be a way to ensure that all NAD^+ regeneration is coupled to succinate production.

Lastly, as shown (see Online Resource Table 2) there was strong up-regulation of *CYC1* in both the *Δdic1* and *Δmdh1* mutants. *CYC1* facilitates electron transfer from ubiquinone cytochrome C oxidoreductase to cytochrome C oxidase. This direction, which is the normal oxidative route and ends in reduction of O_2, would not be possible under fully anaerobic conditions. The upregulation can therefore be viewed as a coping strategy to deal with the stress of redox imbalance. Deletion of *CYC1* could therefore be a way to ensure that all NAD^+ regeneration is coupled to succinate production.

However, this does not explain the lack of succinate production observed in the *Δmdh1* mutant. It has been suggested that mitochondrial $FADH_2$ could be oxidized in the cytosol, which may provide an explanation for the failure of the *Δmdh1* and *Δoac1* mutants to produce any succinate [12]. In any event, the strategies proposed here rely on the capacity for reductive TCA cycle activity under anaerobic conditions, and more specifically, the catalysis of fumarate to succinate via fumarate reductase. There is data suggesting that *S. cerevisiae* can exhibit this metabolic state [8,9].

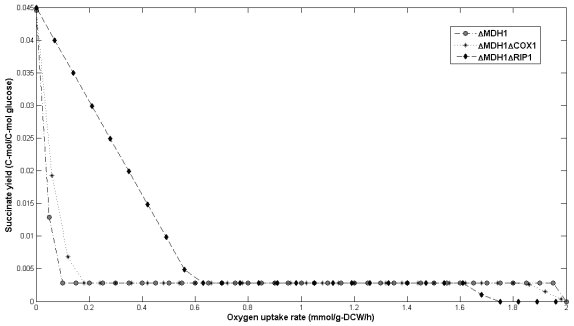


Fig 3. Oxygen sensitivity of succinate yield on glucose. Succinate yield on glucose when rO_2 is constrained between 0 and 2 mmol O_2/g-DCW/h, while maximizing for growth under constrained glucose uptake rate. Data is shown for the single gene deletion of *MDH1*, the double gene deletion of *MDH1* and *COX1*, and the double gene deletion of *MDH1* and *RIP1*. The reference model has a succinate yield of approximately 0.003 C-mol/C-mol glucose (similar to the baseline seen for the three deletion strategies) up until 1.8 mmol O_2/g-DCW/h, after which is drops to zero. *MDH1* encodes malate dehydrogenase, *RIP1* encodes ubiquinol cytochrome C reductase, and *COX1* encodes subunit I of the cytochrome C oxidase.

In conclusion, a genome-scale metabolic model was used to predict single deletion strategies that could lead to increased succinate production and that were physiologically feasible during anaerobic growth. Three of these strategies were validated in vivo and one, *Δdic1*, was identified to lead to a significant improvement in succinate yield on substrate, in close agreement with the model prediction. Furthermore, physiological characterization and transcriptome analysis were used to propose biological mechanisms. The mechanisms proposed rely heavily on inter-compartmental transport reactions as well as redox balancing, both identified as the dominant GO process categories in the *Δdic1* succinate overproducing mutant. Further in vivo characterization of the transport reactions, and subsequent corresponding modifications to the genome-scale network reconstruction would be required for further improvements and understanding of metabolic engineering strategies.

Acknowledgements

José Manuel Otero was a Merck Doctoral Fellow and acknowledges financial support from Merck Research Labs, Merck & Co., Inc. We also acknowledge funding from Knut and Alice Wallenberg Foundation and the Chalmers Foundation.

References

1. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, et al. (2013) The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum. PLoS Comput Biol 9: e1002980.
2. Akesson M, Forster J, Nielsen J (2004) Integration of gene expression data into genome-scale metabolic models. Metab Eng 6: 285-293.
3. Aliverdieva DA, Mamaev DV, Bondarenko DI, Sholtz KF (2006) Properties of yeast Saccharomyces cerevisiae plasma membrane dicarboxylate transporter. Biochemistry (Mosc) 71: 1161-1169.
4. Arakawa K, Yamada Y, Shinoda K, Nakayama Y, Tomita M (2006) GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. BMC Bioinformatics 7: 168.
5. Arikawa Y, Kobayashi M, Kodaira R, Shimosaka M, Muratsubaki H, et al. (1999) Isolation of sake yeast strains possessing various levels of succinate- and/or malate-producing abilities by gene disruption or mutation. J Biosci Bioeng 87: 333-339.
6. Arikawa Y, Kuroyanagi T, Shimosaka M, Muratsubaki H, Enomoto K, et al. (1999) Effect of gene disruptions of the TCA cycle on production of succinic acid in Saccharomyces cerevisiae. J Biosci Bioeng 87: 28-36.
7. Burgard AP, Pharkya P, Maranas CD (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84: 647-657.
8. Camarasa C, Faucet V, Dequin S (2007) Role in anaerobiosis of the isoenzymes for Saccharomyces cerevisiae fumarate reductase encoded by OSM1 and FRDS1. Yeast 24: 391-401.
9. Camarasa C, Grivet JP, Dequin S (2003) Investigation by 13C-NMR and tricarboxylic acid (TCA) deletion mutant analysis of pathways for succinate formation in Saccharomyces cerevisiae during anaerobic fermentation. Microbiology 149: 2669-2678.
10. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, et al. (1998) SGD: Saccharomyces Genome Database. Nucleic Acids Res 26: 73-79.
11. Cimini D, Patil KR, Schiraldi C, Nielsen J (2009) Global transcriptional response of Saccharomyces cerevisiae to the deletion of SDH3. BMC Syst Biol 3: 17.
12. Enomoto K, Arikawa Y, Muratsubaki H (2002) Physiological role of soluble fumarate reductase in redox balancing during anaerobiosis in Saccharomyces cerevisiae. FEMS Microbiol Lett 215: 103-108.
13. Famili I, Forster J, Nielsen J, Palsson BO (2003) Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proc Natl Acad Sci U S A 100: 13134-13139.
14. Forster J, Famili I, Fu P, Palsson BO, Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res 13: 244-253.
15. Geertman JM, van Maris AJ, van Dijken JP, Pronk JT (2006) Physiological and genetic engineering of cytosolic redox metabolism in Saccharomyces cerevisiae for improved glycerol production. Metab Eng 8: 532-542.
16. Herrgard MJ, Swainston N, Dobson P, Dunn WB, Arga KY, et al. (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol 26: 1155-1160.

17. Jagow G, Klingenberg M (1969) [Hydrogen pathways in the mitochondrion of Saccharomyces carlsbergensis]. Hoppe Seylers Z Physiol Chem 350: 1155.
18. Kubo Y, Takagi H, Nakamori S (2000) Effect of gene disruption of succinate dehydrogenase on succinate production in a sake yeast strain. J Biosci Bioeng 90: 619-624.
19. Liu L, Agren R, Bordel S, Nielsen J (2010) Use of genome-scale metabolic models for understanding microbial physiology. FEBS Lett 584: 2556-2564.
20. Luttik MA, Overkamp KM, Kotter P, de Vries S, van Dijken JP, et al. (1998) The Saccharomyces cerevisiae NDE1 and NDE2 genes encode separate mitochondrial NADH dehydrogenases catalyzing the oxidation of cytosolic NADH. J Biol Chem 273: 24529-24534.
21. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37: D619-622.
22. McKinlay JB, Vieille C, Zeikus JG (2007) Prospects for a bio-based succinate industry. Appl Microbiol Biotechnol 76: 727-740.
23. Nielsen JH, Villadsen J, Lidén G (2003) Bioreaction engineering principles. New York: Kluwer Academic/Plenum Publishers. xv, 528 p. p.
24. Nissen TL, Schulze U, Nielsen J, Villadsen J (1997) Flux distributions in anaerobic, glucose-limited continuous cultures of Saccharomyces cerevisiae. Microbiology 143 (Pt 1): 203-218.
25. Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5: 320.
26. Osterlund T, Nookaew I, Nielsen J (2012) Fifteen years of large scale metabolic modeling of yeast: developments and impacts. Biotechnol Adv 30: 979-988.
27. Otero JM, Cimini D, Patil KR, Poulsen SG, Olsson L, et al. (2013) Industrial systems biology of Saccharomyces cerevisiae enables novel succinic acid cell factory. PLoS One 8: e54144.
28. Otero JM, Panagiotou G, Olsson L (2007) Fueling industrial biotechnology growth with bioethanol. Adv Biochem Eng Biotechnol 108: 1-40.
29. Overkamp KM, Bakker BM, Kotter P, van Tuijl A, de Vries S, et al. (2000) In vivo analysis of the mechanisms for oxidation of cytosolic NADH by Saccharomyces cerevisiae mitochondria. J Bacteriol 182: 2823-2830.
30. Paley SM, Karp PD (2006) The Pathway Tools cellular overview diagram and Omics Viewer. Nucleic Acids Res 34: 3771-3778.
31. Patil KR, Rocha I, Forster J, Nielsen J (2005) Evolutionary programming as a platform for in silico metabolic engineering. BMC Bioinformatics 6: 308.
32. Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2: 886-897.
33. Raab AM, Gebhardt G, Bolotina N, Weuster-Botz D, Lang C (2010) Metabolic engineering of Saccharomyces cerevisiae for the biotechnological production of succinic acid. Metab Eng 12: 518-525.
34. Sauer M, Porro D, Mattanovich D, Branduardi P (2008) Microbial production of organic acids: expanding the markets. Trends Biotechnol 26: 100-108.
35. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article3.
36. Song H, Lee SY (2006) Production of succinic acid by bacterial fermentation. Enzyme and Microbial Technology 39: 352-361.
37. van Dijken JP, Bauer J, Brambilla L, Duboc P, Francois JM, et al. (2000) An interlaboratory comparison of physiological and genetic properties of four Saccharomyces cerevisiae strains. Enzyme Microb Technol 26: 706-714.
38. Vemuri GN, Eiteman MA, McEwen JE, Olsson L, Nielsen J (2007) Increasing NADH oxidation reduces overflow metabolism in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 104: 2402-2407.
39. Verduyn C, Postma E, Scheffers WA, Van Dijken JP (1992) Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation. Yeast 8: 501-517.
40. Willke T, Vorlop KD (2004) Industrial bioconversion of renewable resources as an alternative to conventional chemistry. Appl Microbiol Biotechnol 66: 131-142.
41. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, et al. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285: 901-906.

Paper II

Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes

Bordel, S., **Agren, R.** and Nielsen, J.

PLoS Comput Biol (2010), 6(7), p. e1000859

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Sampling the Solution Space in Genome-Scale Metabolic Networks Reveals Transcriptional Regulation in Key Enzymes

Sergio Bordel, Rasmus Agren, Jens Nielsen*

Systems Biology, Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

Abstract

Genome-scale metabolic models are available for an increasing number of organisms and can be used to define the region of feasible metabolic flux distributions. In this work we use as constraints a small set of experimental metabolic fluxes, which reduces the region of feasible metabolic states. Once the region of feasible flux distributions has been defined, a set of possible flux distributions is obtained by random sampling and the averages and standard deviations for each of the metabolic fluxes in the genome-scale model are calculated. These values allow estimation of the significance of change for each reaction rate between different conditions and comparison of it with the significance of change in gene transcription for the corresponding enzymes. The comparison of flux change and gene expression allows identification of enzymes showing a significant correlation between flux change and expression change (transcriptional regulation) as well as reactions whose flux change is likely to be driven only by changes in the metabolite concentrations (metabolic regulation). The changes due to growth on four different carbon sources and as a consequence of five gene deletions were analyzed for *Saccharomyces cerevisiae*. The enzymes with transcriptional regulation showed enrichment in certain transcription factors. This has not been previously reported. The information provided by the presented method could guide the discovery of new metabolic engineering strategies or the identification of drug targets for treatment of metabolic diseases.

Citation: Bordel S, Agren R, Nielsen J (2010) Sampling the Solution Space in Genome-Scale Metabolic Networks Reveals Transcriptional Regulation in Key Enzymes. PLoS Comput Biol 6(7): e1000859. doi:10.1371/journal.pcbi.1000859

Editor: Jennifer L. Reed, University of Wisconsin-Madison, United States of America

Received March 18, 2010; **Accepted** June 14, 2010; **Published** July 15, 2010

Funding: This work was funded by the European Research Council. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nielsenj@chalmers.se

Introduction

Systems Biology aims to use mathematical models to integrate different kinds of data in order to achieve a global understanding of cellular functions. The data to be integrated differ both in their nature and measurability. The availability of DNA microarrays allows for the comparative analysis or mRNA levels between different strains and conditions. These data provide genome-wide information, and changes in expression at different conditions are expressed in statistical terms such as p-values or Z-scores that quantify the level of significance in transcriptional changes. The availability of annotated genome-scale metabolic networks allowed mapping of the transcriptional changes in metabolic genes on to their corresponding metabolic pathways and defining significantly up or down regulated sub-networks [1]. Even though this allows for identification of transcriptional hot-spots in metabolism, this does still not provide information about whether there are any changes in metabolic fluxes in these pathways, as it has been shown that in general there is no clear correlation between gene expression and protein concentration [2] or metabolic flux [3,4].

Metabolic fluxes are the result of a complex interplay between enzyme kinetics, metabolite concentrations, gene expression and translational regulation. Metabolic fluxes can be directly measured using ^{13}C labeling experiments [5]. However, flux data obtained using this approach differ from gene expression data in two main features: 1) their determination is only possible for a relatively small subset of all the reactions in a genome-scale metabolic network and 2) they are indirect data in the sense that the fluxes are quantified obtained by fitting measured labeling patterns using a simple metabolic model. The complexity of the mRNA-flux dependence and the disparity in the nature of both kinds of data make their integration an important challenge.

In this paper we propose a method to integrate gene expression data with flux data by transforming a limited amount of quantitative flux data into a genome-scale set of statistical scores similar to the one obtained from DNA microarrays. In order to do that, a set of experimental exchange fluxes are fixed for each of the studied conditions or for each of the strains investigated, and a sampling algorithm is then used to obtain a set of flux distributions satisfying the experimental values. This approach allows for obtaining means and standard deviations for each flux in the genome-scale network. From the mean and standard deviation it is possible to derive statistical scores for the significance of flux change between conditions [6,7]. Random sampling in the region of feasible flux distributions has been previously used to study the statistical distribution of flux values and determine a flux backbone of reactions carrying high fluxes [8] as well as to define modules of reactions whose fluxes are positively correlated [9,10]. Also mitochondria related diseases have been analyzed using random sampling [11]. All the works published so far used the Hit and Run algorithm to perform the sampling [7].

Author Summary

The sequencing of full genomes and the development of high-throughput analysis technologies have made available both genome-scale metabolic networks and simultaneous transcription data for all the genes of an organism. Genome-scale metabolic models, with the assumption of steady state for the internal metabolites, allow the definition of a region of feasible metabolic flux distributions. This space of solutions can be further constrained using experimental flux measurements (normally production or uptake rates of external compounds). Here a random sampling method was used to obtain average values and standard deviations for all the reaction rates in a genome-scale model. These values were used to quantify the significance of changes in metabolic fluxes between different conditions. The significance in flux changes can be compared to the changes in gene transcription of the corresponding enzymes. Our method allowed for identification of specific reactions that are transcriptionally regulated, and we further identified that these reactions can be ascribed to a few key transcription factors. This suggests that the regulation of metabolism has evolved to contain a few flux-regulating transcription factors that could be the target for genetic manipulations in order to redirect fluxes.

By dividing the average difference among two conditions (e.g. carbon sources or mutant strains) by its standard deviation, it is possible to obtain Z scores for each metabolic flux. These scores can be transformed into p-values that measure the significance of change of each flux (see methods). By comparing these p-values with the p-values derived from gene-expression arrays, the enzymes in the network can be classified as: 1) enzymes that have a significantly correlated change both in flux and expression level (reactions showing transcriptional regulation), 2) enzymes that show a significant change in expression but not in flux (we will refer to them as showing post-transcriptional regulation) and 3) enzymes that show significant changes in flux but not a change in expression (metabolic regulation). Hereby we provide a framework that allows for global classification of reaction fluxes into those that are transcriptionally regulated, post-transcriptionally regulated and metabolically regulated (see Fig. 1). This will have substantial impact on the field of metabolic engineering where changes in gene-expression are often used as the key means to alter metabolic fluxes. In the paper we show the use of the presented framework for the analysis of the yeast *Saccharomyces cerevisiae* grown at different growth conditions and for the analysis of different deletion mutants.

The combined use of random sampling of genome-scale metabolic networks and expression data allows for global mapping of reactions that are either transcriptionally or metabolically regulated. This information can be used to guide the engineering of microbial strains or as a diagnosis tool for studying metabolic diseases in humans. In particular we should highlight that reactions in which there is no relation between gene transcription level and metabolic flux are not suitable targets for flux increase via gene over-expression. Through analysis of different data sets the method revealed that many changes in gene expression are not correlated with a corresponding change in metabolic fluxes. The use of gene-expression data alone can therefore be misleading. However, our method allowed for identification of many specific reactions that are indeed transcriptionally regulated, and we further identified that the expression of these enzymes is regulated a few key transcription factors. This fact suggests that the

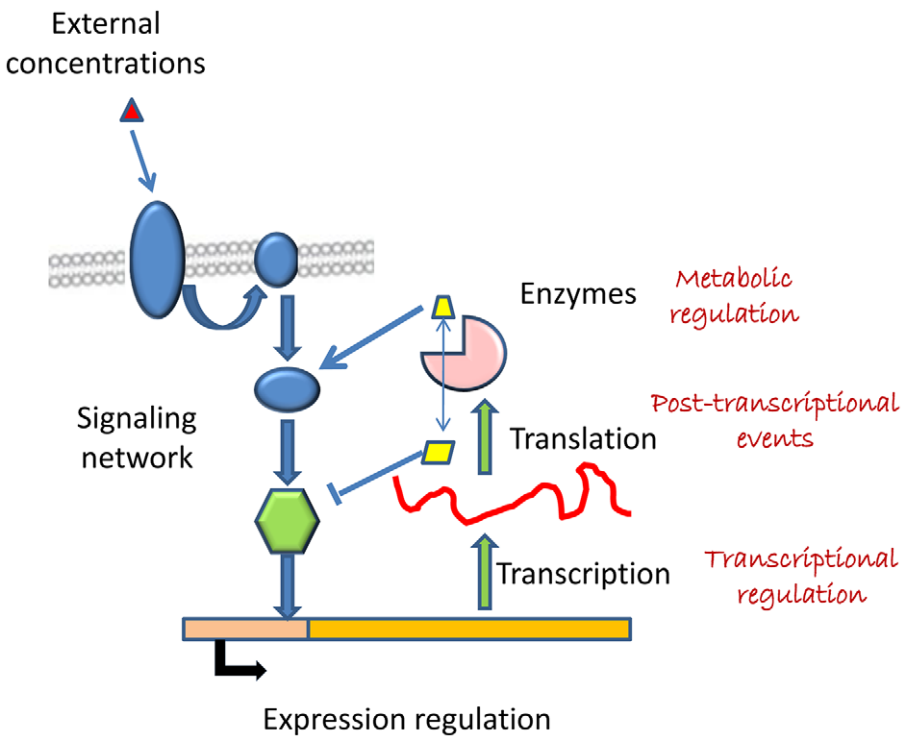


Figure 1. Illustration of the regulatory mechanisms of cellular metabolism. The fluxes can be regulated at the level of mRNA transcription, by the concentrations of the metabolites or by intermediate steps such as translation or activation of the enzymes.
doi:10.1371/journal.pcbi.1000859.g001

regulation of metabolism has evolved to contain a few flux-regulating transcription factors that could be the target for genetic manipulations in order to redirect fluxes.

Results/Discussion

Comparisons between different sampling methods

Here we propose a sampling method that finds extreme solutions among the feasible flux distributions of the metabolic network. These solutions correspond to the corners in the region of allowed flux distributions, and in mathematical terms they are elements of the convex basis of the region of feasible solutions (which is a convex set). The COBRA Toolbox [12] includes a random sampling option that uses the Hit and Run algorithm [13] to obtain points uniformly distributed in the region of allowed solutions. The difference between the two sampling methods is illustrated in Fig. 2.

In order to assess the accuracy of our sampling method to estimate the average fluxes and their standard deviations, we compared a set of internal fluxes measured with ^{13}C labeling [14] with predictions using 500 sampling points obtained using the sampling method in the convex basis and 500 sampling points obtained using the sampling algorithm implemented in the COBRA Toolbox. The results are summarized in Table 1 where our method is labeled Convex Basis (CB), because it samples elements of the convex basis of the region of allowed solutions (see above), and the method from the COBRA Toolbox is labeled Hit and Run (HR). The Z values in the table are the number of standard deviations that the real value is deviating from the calculated mean.

The means obtained by the two sampling methods are very similar for most of the reactions; however the standard deviations found using the HR algorithm are significantly smaller. With the HR method the real values for the fluxes in many cases deviate several standard deviations from the mean, A high value of Z indicates that the real value has a very low chance of being obtained using the considered sampling method (or in other words: the real value does not belong to the family of solutions that is generated by the sampling method). The number of samples with the HR algorithm was increased up to 5000 to check possible effects of the sample size on the standard deviation. Only small increases were observed for the standard deviations of the studied fluxes.

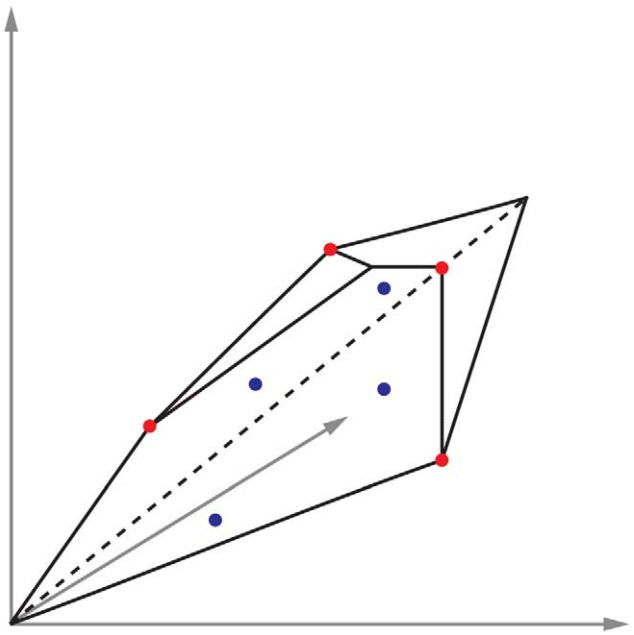


Figure 2. The red points illustrate the sampling in the corners of the region of allowed solutions. The blue points illustrate the uniform random sampling inside the space of allowed solutions.
doi:10.1371/journal.pcbi.1000859.g002

Using the CB algorithm we obtain higher standard deviations and the real flux is for most reactions less than one standard deviation away from the mean flux. We can therefore conclude that the CB sampling method gives more realistic standard deviations for the fluxes. This is important if we want to compare the significance of flux changes between conditions. An under-estimated standard deviation would make some flux changes appear as being significant even though they may not be in reality, and our method therefore gives a more conservative list of significantly changed reaction fluxes than the HR algorithm.

Comparisons between different carbon sources and mutant strains

To evaluate our method we used data for the yeast *S. cerevisiae*. Data from growth on four different carbon sources (glucose, maltose, ethanol and acetate) in chemostat cultures and five deletion mutants (grr1Δ, hxk2Δ, mig1Δ, mig1Δmig2Δ and gdh1Δ) grown in batch cultures were used. The exchange fluxes and gene expression data for the mentioned conditions have been published earlier [15–17].

Our method obtains probability scores for each enzyme in the metabolic network (see methods) and this allowed us to classify the enzymes as transcriptionally regulated (correlation between flux and gene expression), post-transcriptionally regulated (changes in gene expression don't cause changes in flux) and metabolically regulated (changes in flux are not caused by changes in gene-expression). The cut-off chosen for this classification was a probability score above 0.95. Tables 2 and 3 show the 10 top scoring enzymes in each group (or fewer when less than 10 enzymes had a score exceeding 0.95). The method is illustrated in Fig. 3.

The method to identify the significance of flux changes relies on a set of measured external fluxes, and in some cases strains that don't show significant changes in external fluxes have changes in internal fluxes [18]. These changes cannot be identified with our method, and our estimations of the significance of flux changes can therefore be seen as conservative estimates. The lists of transcriptionally and metabolically regulated reactions are therefore more reliable than the list of post-transcriptionally regulated reactions (in which some fluxes may be changed in reality but their change pass undetected).

The reactions showing transcriptional regulation form a set of putative targets where enzyme over-expression or down regulation will influence the flux through these reactions. The reactions showing metabolic regulation points to parts of the metabolism where the pools of metabolites are possibly increasing or decreasing in connection with transcriptional changes and hereby counteracting possible changes in enzyme concentration as a result of transcriptional changes. This knowledge can be used to identify whether one should target changes in enzyme concentration (v_{max} changes), e.g. through over-expression, or changes in enzyme affinity (K_m changes), e.g. through expression of heterologous enzymes, in order to alter the fluxes.

Effects of different carbon sources. In the glucose to maltose transition, only two enzymes showed transcriptional change correlated with their flux. The α-glucosidase MAL32,

Table 1. Real and estimated fluxes in *S. cerevisiae* at aerobic and anaerobic growth conditions.

Aerobic conditions	Flux	Mean (HR)	Variance (HR)	Mean (CB)	Variance (CB)	Z (HR)	Z (CB)
Fructose-bisphosphate aldolase	0.70	0.76	0.0001	0.73	0.0098	3.38	0.19
Pyruvate kinase	1.50	1.49	0.0351	1.43	0.344	0.024	0.12
Pyruvate dehydrogenase	0.47	1.03	0.0004	0.98	0.015	28	4.14
Citrate synthase	0.71	0.99	0.0178	0.94	0.1868	2.08	0.53
Pyruvate carboxylase	0.32	0.23	0.0014	0.22	0.0124	2.44	0.96
Glucose-6-phosphate 1-dehydrogenase	0.52	0.37	0.0012	0.41	0.0881	4.25	0.37
Pyruvate decarboxylase	0.53	0.07	0.0004	0.08	0.0092	23	4.7
Anaerobic conditions							
Fructose-bisphosphate aldolase	5.82	5.62	0.0053	5.67	0.30	2.78	0.28
Pyruvate kinase	10.37	10.39	1.93	10.49	25.62	0.019	0.025
Pyruvate dehydrogenase	0.26	0.59	0.0114	0.87	0.45	3.16	0.91
Citrate synthase	0.19	0.96	0.0271	1.03	1.03	4.7	0.83
Pyruvate carboxylase	0.32	0.32	0.0261	0.88	2.82	0.028	0.33
Glucose-6-phosphate 1-dehydrogenase	0.32	1.35	0.0441	1.20	2.40	4.9	0.56
Pyruvate decarboxylase	9.54	9.82	1.71	9.21	2.20	0.22	0.22

HR refers to the Hit and Run algorithm and CB to the Convex Basis algorithm.
doi:10.1371/journal.pcbi.1000859.t001

responsible for the breakdown of maltose into glucose was up-regulated and the glucose transporter HXT4 was down-regulated. The metabolic adjustment in terms of fluxes was minimal and only enzymes directly related with substrate uptake and utilization were detected. The changes in gene expression were also low and only 11 metabolic enzymes were significantly perturbed (without significant flux changes).

The glucose to ethanol and the glucose to acetate transitions showed widespread flux and expression changes, and they are therefore more interesting study cases. In the glucose-ethanol transition 19 enzymes showed transcriptional regulation and 22 other enzymes changed in expression but not in flux. For the glucose-acetate transition the same numbers were 33 and 23 respectively. We can see that about one half of the genes that changed in transcription level also showed significant flux changes, this proportion is higher than in the case of deletion mutants (as it will be discussed later). Among the enzymes showing transcriptional regulation, 14 were shared between the glucose-ethanol and glucose-acetate transitions. Interestingly, no overlap was found between the sets of enzymes that don't change in flux. Metabolic regulation was observed in 21 reactions for each case, among which 8 overlap.

The enzymes showing transcriptional regulation clearly show a down regulation of enzymes involved in glucose uptake and utilization (e.g. Glucose transporter HXT4 or Hexokinase 2) and the up-regulation of enzymes involved in the gluconeogenesis (e.g. Fructose-1,6-biphosphatase) and the TCA cycle (e.g. Succinate dehydrogenase or Citrate synthase). The AcCoA synthetase 2, responsible for supplying AcCoA to the TCA cycle is also transcriptionally up-regulated as well as the ATP synthetase (involved in the respiratory chain) and the external NADH-ubiquinone oxidoreductase 2, which supplies the necessary NAD^+ to oxidize ethanol or acetate in the cytoplasm and maintain the redox balance in the cell. Isocitrate lyase, a key component of the glyoxylate cycle, is also transcriptionally up regulated and this allows for net formation of malic acid that can further be converted to phopshoenolpyruvate (via oxaloacetate) that fuels the gluconeogenesis. All these changes in fluxes are consistent with general knowledge about the changes in metabolism from glucose to C2 carbon sources like ethanol and acetate, but what is interesting to see is that not all the reactions associated with these flux changes are transcriptionally regulated, but the cell have selected a few key reactions to regulate at the transcriptional level and these are identified using our method.

In order to make a deeper analysis, we performed an enrichment test to compare the transcription factors involved in the expression of the enzymes showing transcriptional regulation and the enzymes showing changes in expression but not in flux. We found three transcription factors that were strongly over-represented in the metabolic genes showing transcriptional regulation. In the glucose-ethanol transition, the transcription factors Gcr1 and Gcr2 both appeared in 11 transcriptionally regulated genes and in none of the other genes, whereas the transcription factor Hap4 appeared in 11 transcriptionally regulated genes and 5 of the other regulated genes. For the glucose-acetate transition these numbers were 15-0, 11-0 and 15-0 for the same transcription factors. This means that certain transcription factors are especially involved in the transcriptional regulation of metabolic fluxes (the same kind of enrichment was observed in the deletion mutants, as will be discussed later), and to our knowledge this has not been previously reported. It basically implies that there is global regulation of major flux alterations, which is similar to what has experimentally been shown to be the case for galactose metabolism [19].

The top scoring metabolically regulated reactions, both for glucose-ethanol and glucose-acetate, are the Fructose biphosphate aldolase and the Triosephosphate isomerase. These reactions are known to operate close to the equilibrium and are therefore very sensitive to changes in the metabolic pools, which is consistent with metabolic regulation of the fluxes. In the considered cases the direction of these reactions is inverted. This can only be explained by a decrease in the fructose-1,6-diphosphate pool and an increase in the glyceraldehyde-3-phosphate and dihydroxyacetone pools. This hypothesis is supported by the fact that in chemostat cultures, there was not found any correlation between the glycolytic flux and the expression of the genes encoding these two enzymes [20].

Table 2. Top scoring enzymes for transcriptional, post-transcriptional and metabolic regulation for changes in carbon source.

Carbon source shift	Enzymes showing transcriptional regulation	Enzymes showing post-transcriptional regulation	Enzymes showing metabolic regulation
Glucose-Maltose	• α-glucosidase MAL32	• Mevalonate kinase	• Acetate transport via proton symport
	• Low-affinity glucose transporter HXT4	• Inosine-5′-monophosphate dehydrogenase IMD2	
		• Asparagine synthetase 1	
		• (DL)-glycerol-3-phosphatase 1	
		• Uncharacterized deaminase	
		• Nicotinate-nucleotide pyrophosphorylase	
		• Mevalonate kinase	
		• Mevalonate kinase	
		• Glycerol-3-phosphate dehydrogenase [NAD+] 1	
Glucose-Ethanol	• Phosphoenolpyruvate carboxykinase	• Formate dehydrogenase 2	• Fructose-bisphosphate aldolase
	• Fructose-1,6-bisphosphatase	• ATP-NADH kinase	• Triosephosphate isomerase
	• Isocitrate lyase	• Sulfate permease 1	• Pyruvate dehydrogenase E1 component subunit alpha [m]
	• Malate dehydrogenase [c]	• Formate dehydrogenase 1	• Alpha-ketoglutarate dehydrogenase
	• Citrate synthase [p]	• Dicarboxylate transporter [m]	• Succinyl-CoA ligase [ADP-forming] subunit beta [m]
	• Ribose-5-phosphate isomerise	• NADP-specific glutamate dehydrogenase 2	• Malate synthase 2, glyoxysomal
	• Low-affinity glucose transporter HXT4	• Uncharacterized deaminase	• Glucose-6-phosphate 1-dehydrogenase
	• External NADH-ubiquinone oxidoreductase 2 [m]	• Probable 6-phosphogluconolactonase 3	• Cytochrome b-c1 complex subunit Rieske [m]
	• Glucose-6-phosphate isomerase	• 6-phosphofructo-2-kinase 2	• Adenylate kinase [c]
		• Nucleoside diphosphate kinase	
Glucose-Acetate	• Fumarate hydratase [m]	• Ribonucleoside-diphosphate reductase large chain 1	• Fructose-bisphosphate aldolase
	• Phosphoenolpyruvate carboxykinase [ATP]	• Phospho-2-keto-3-deoxyheptonate aldolase	• Triosephosphate isomerase
	• Fructose-1,6-bisphosphatase	• 6-phosphofructo-2-kinase 1	• Ribose-5-phosphate isomerase
	• Isocitrate dehydrogenase [NADP] [c]	• Glutamine-dependent NAD(+) synthetase	• Inorganic pyrophosphatase
	• Succinate-semialdehyde dehydrogenase [NADP+]	• Ribose-phosphate pyrophosphokinase 4	• Adenylate kinase [c]
	• Citrate synthase [p]	• ATP-dependent permease AUS1	• Glutamate decarboxylase
	• Isocitrate dehydrogenase [NAD] subunit 1 [m]	• Fructose-2,6-bisphosphatase	• 4-aminobutyrate aminotransferase
	• Pyruvate kinase 2	• Nicotinate-nucleotide pyrophosphorylase [carboxylating]	• Tricarboxylate transport protein
	• Low-affinity glucose transporter HXT4′	• Squalene monooxygenase	• Prephenate dehydrogenase [NADP+]
			• Tricarboxylate transport protein

doi:10.1371/journal.pcbi.1000859.t002

The results for the glucose-ethanol change are summarized in Fig. 4.

Effects of gene deletions. As mentioned above several different deletion strains were evaluated and the number of enzymatic reactions showing transcriptional regulation was 26 for the grr1Δ strain, 25 for the hxk2Δ strain, 11 for the mig1Δ strain, 8 for the mig1Δmig2Δ strain and 0 for the gdh1Δ strain. The reactions showing post-transcriptional regulation were 73, 70, 46, 36 and 89 for the same strains, respectively. These numbers clearly show, that in contrast to growth on different carbon sources, most of the transcriptional changes do not result in correlated changes in metabolic fluxes. This indicates that many transcriptional changes are indeed happening in order to minimize the metabolic adjustment resulting from a gene deletion, and the most extreme case is the gdh1Δ strain, where no transcriptional changes seem to be correlated to flux changes. This behaviour is consistent with the MOMA (Minimization of Metabolic Adjustment) hypothesis [21].

There is a substantial overlap between the strains grr1Δ and hxk2Δ, 15 transcriptionally regulated reactions were shared between both strains and 28 post-transcriptionally regulated reactions were also shared. An enrichment test was performed in order to find transcription factors regulating the enzymes showing transcriptional regulation. For the grr1Δ, the most significantly enriched transcription factors were Pho2 (which regulates the expression of 10 of the 26 enzymes with transcriptional regulation and only 6 of the 73 enzymes showing post-transcriptional regulation) and Bas1 (which regulates 10 out of 26 and 7 out of

Table 3. Top scoring enzymes for transcriptional, post-transcriptional and metabolic regulation upon deletion of specific genes.

Mutants	Enzymes showing transcriptional regulation	Enzymes showing post-transcriptional regulation	Enzymes showing metabolic regulation
Wild type-grr1Δ	• GMP synthase [glutamine-hydrolyzing]	• ADP, ATP carrier protein 2	• Acetyl-CoA carboxylase malonyltransferase
	• Phosphoribosylformylglycinamidine synthase	• Glycogen [starch] synthase isoform 1	• [Acyl-carrier-protein]
	• Dihydroorotase	• AMP deaminase'	• [Acyl-carrier-protein] acetyltransferase
	• Imidazole glycerol phosphate synthase hisHF	• Sugar transporter STL1	• Grouped fatty acid synthesis [c]
	• Adenylosuccinate lyase	• Phosphofructokinase 2	
	• Pantoate–beta-alanine ligase	• D-3-phosphoglycerate dehydrogenase 2	
	• Inorganic phosphate transporter	• Probable 6-phosphogluconolactonase 4	
	• Threonine dehydratase [m]	• Glycerol-3-phosphate dehydrogenase [NAD+] 1	
	• ATP phosphoribosyltransferase	• Alcohol dehydrogenase 4	
	• Histidinol-phosphate aminotransferase	• Xanthine phosphoribosyltransferase 1	
Wild type-hxk2Δ	• Phosphoribosylformylglycinamidine synthase	• Nucleoside diphosphate kinase	• Mannose-6-phosphate isomerase
	• Pyruvate decarboxylase isozyme 3	• Cytochrome b-c1 complex subunit Rieske [m]	• Acetyl-CoA carboxylase
	• Phosphoglycerate mutase 2	• Homocysteine S-methyltransferase 1	• [Acyl-carrier-protein] malonyltransferase
	• Alcohol dehydrogenase 5	• ATP-NADH kinase	• [Acyl-carrier-protein] acetyltransferase
	• Imidazole glycerol phosphate synthase hisHF	• 1,4-alpha-glucan-branching enzyme	• Grouped fatty acid synthesis [c]
	• Dihydroorotase'	• NAD-dependent malic enzyme [m]	
	• Hexokinase-2	• Pyrroline-5-carboxylate reductase	
	• Phosphoglycerate mutase 3	• Galactose-1-phosphate uridylyltransferase	
	• Phosphofructokinase 1		
	• Adenylosuccinate lyase		
Wild Type-mig1Δ	• Probable 6-phosphogluconolactonase 1	• ADP-sulfurylase	• Mannose-6-phosphate isomerase
	• Orotidine 5'-phosphate decarboxylase	• Purine nucleoside phosphorylase	• Mannose-1-phosphate guanyltransferase
	• Transketolase 2	• 3',5'-cyclic-nucleotide phosphodiesterase 2	• Acetyl-CoA carboxylase
	• Chorismate mutase	• Phosphoserine phosphatase	• Grouped fatty acid synthesis [c]
	• Prephenate dehydratase'	• Cytidine deaminase	
	• Amidophosphoribosyltransferase	• 1,3-beta-D-glucan-UDP glucosyltransferase	
	• Phosphatidate cytidylyltransferase	• Vacuolar acid trehalase	
	• Diacylglycerol pyrophosphate phosphatase 1	• Low-affinity glucose transporter HXT1	
	• Inositol-3-phosphate synthase	• Trehalose-phosphatase	
Wild Type-mig1Δmig2Δ	• Branched-chain-amino-acid aminotransferase [c]	• Cystathionine beta-synthase	• Phosphomannomutase
	• Phosphoribosylformylglycinamidine synthase	• Isocitrate lyase	• Mannose-1-phosphate guanyltransferase
	• Pantothenate kinase	• Mitochondrial dicarboxylate transporter	• Acetyl-CoA carboxylase
	• GMP synthase [glutamine-hydrolyzing]	• High-affinity glucose transporter HXT2	• Grouped fatty acid synthesis [c]
	• Imidazole glycerol phosphate synthase hisHF	• Uridylate kinase	
	• 2-isopropylmalate synthase	• Nucleoside diphosphate kinase	
		• Cytidine deaminase	
		• Potassium-activated aldehyde dehydrogenase [m]	
		• Acetyl-coenzyme A synthetase 2	
		• Pyrroline-5-carboxylate reductase	
Wild Type-gdh1Δ		• Malate synthase 2 [g]	• Glycerol uptake/efflux facilitator protein
		• Sugar transporter STL1	
		• 1,3-beta-glucan synthase component FKS3	
		• High-affinity glucose transporter HXT2	
		• S-(hydroxymethyl)glutathione dehydrogenase	

Table 3. Cont.

Mutants	Enzymes showing transcriptional regulation	Enzymes showing post-transcriptional regulation	Enzymes showing metabolic regulation
		• Methylenetetrahydrofolate dehydrogenase [NAD+]	
		• Acetate transport via proton symport	
		• Homoaconitase [m]	
		• Homoisocitrate dehydrogenase [m]	
		• Glutamate 5-kinase	

doi:10.1371/journal.pcbi.1000859.t003

73 enzymes in the mentioned groups). For the hxk2Δ strain the most significant transcription factor enrichment was found for the factors Pho2 (with 7 and 5 enzymes in each of the groups) and Bas1 (with 7 and 6 enzymes in each of the groups).

Pho2 and Bas1 are partner proteins that regulate the transcription of genes involved in purine and histidine biosynthesis [22]. It is possible that the slower growth rate observed in grr1Δ and hxk2Δ with respect to the wild type is due to a down-

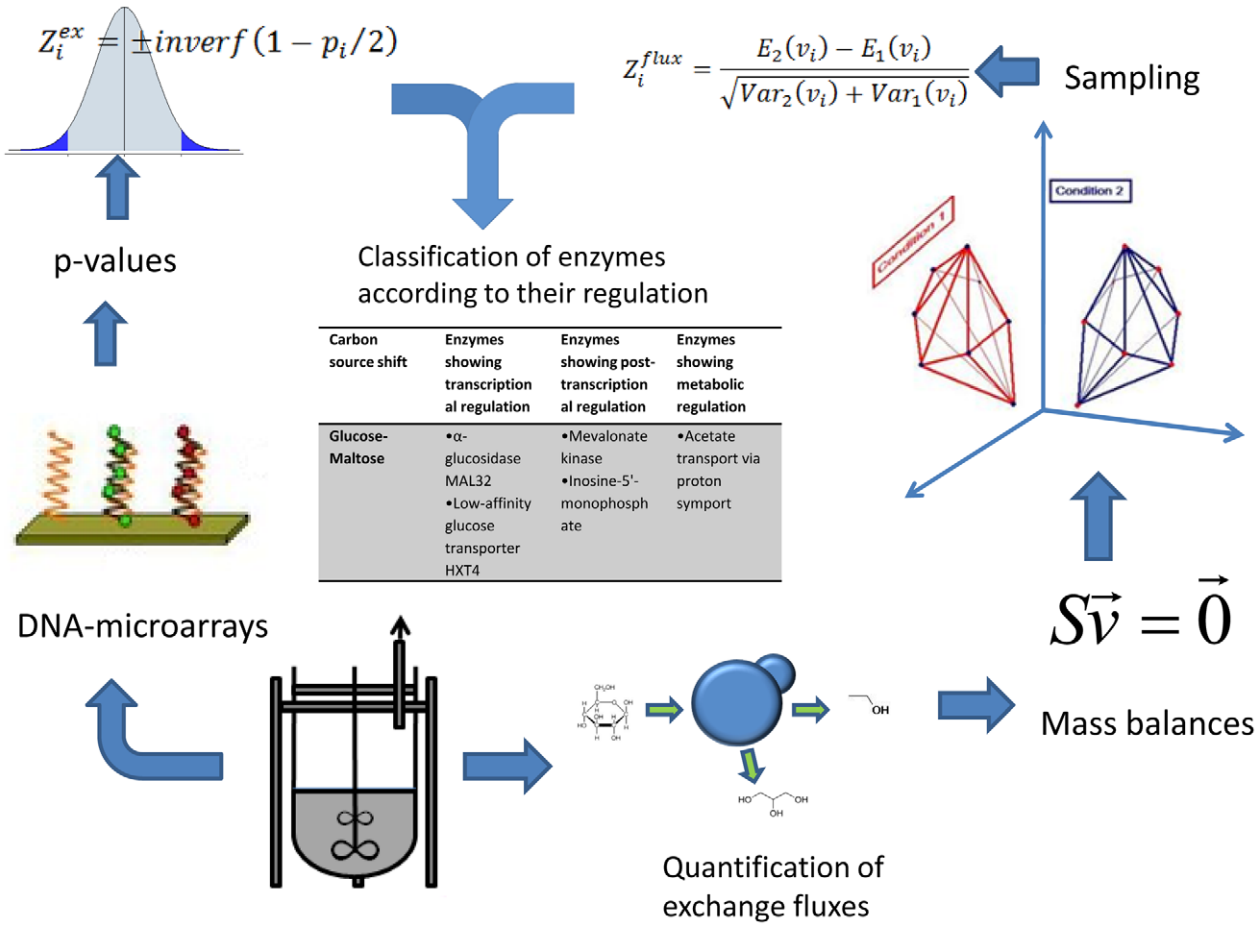


Figure 3. This figure illustrates the different steps of our method. Two kinds of data are extracted from fermentations, gene expression data and production and consumption rates of different metabolites. The gene expression data are transformed into significance scores and p-values for the expression change of the metabolic genes. The measured fluxes are used to constrain the solution spaces corresponding to different conditions. A sampling among the allowed solutions gives averages and standard deviations for each reaction rate in the metabolic network. These values can be obtained to obtain significance scores and p-values for the changes in reaction rates. The p-values for changes in expression and in reaction rates can be combined to obtain the probabilities for a correlated change between both values (transcriptional regulation), changes in rate not correlated to transcriptional changes (metabolic regulation) and changes in transcription that are not correlated to changes in rate (which we refer to as posttranscriptional regulation).
doi:10.1371/journal.pcbi.1000859.g003

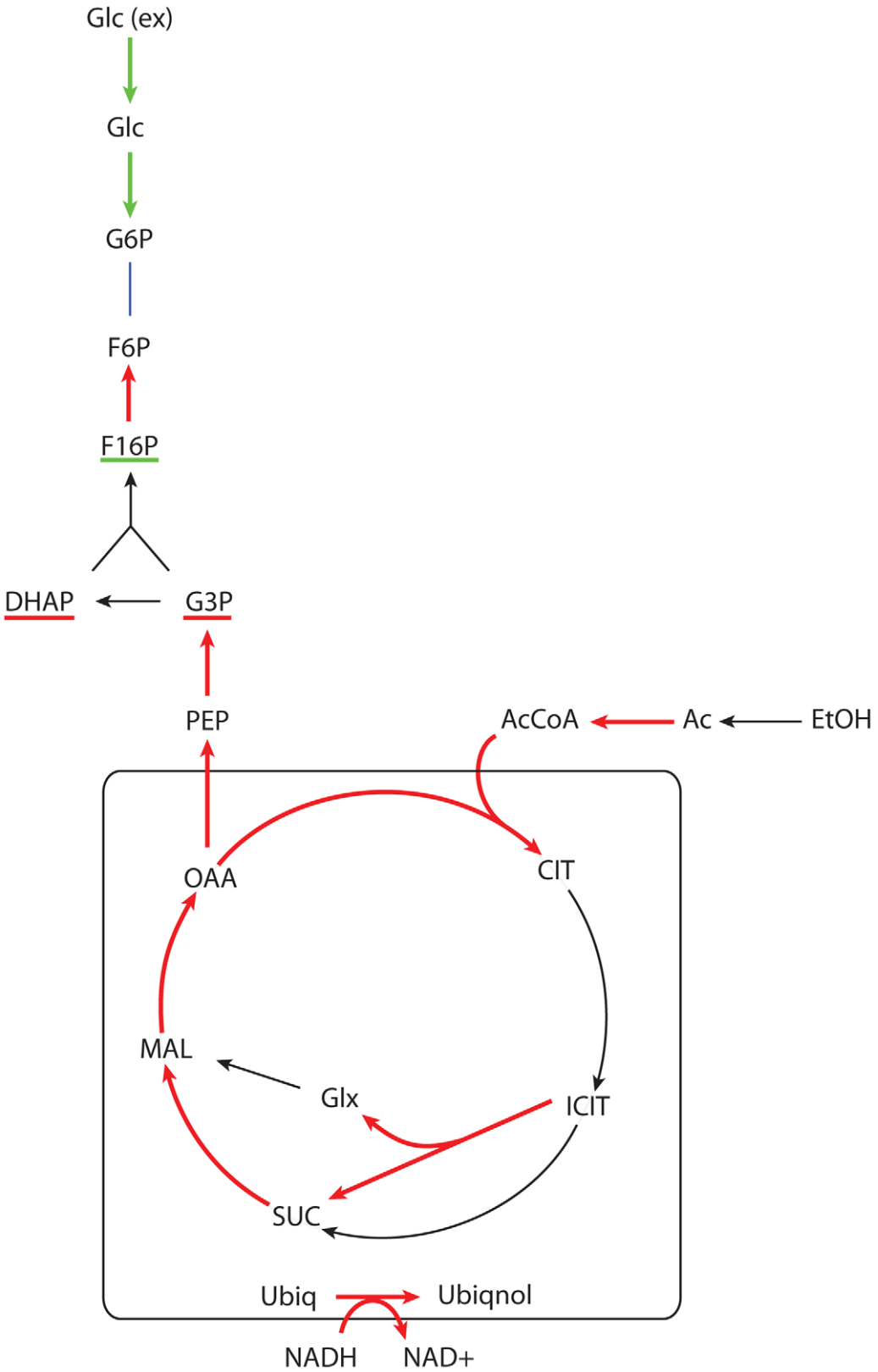


Figure 4. Main reactions showing transcriptional up (red) or down (green) regulation associated with the glucose-ethanol shift. The underlined metabolite pools are those that are expected to increase (red) or decrease (green) according to the observed metabolic regulation.
doi:10.1371/journal.pcbi.1000859.g004

regulation of purine and histidine biosynthesis resulting from lower activities of Pho2 or Bas1.

Strains grr1Δ and hxk2Δ show specific growth rates of 0.23 and 0.22 h^{-1} respectively [16], however the biomass yields were 0.09 g-DW g^{-1} and 0.2 g-DW g^{-1}. The specific glucose uptake rate for the hxk2Δ strain is significantly lower than for the grr1Δ strain as well as for the reference strain. This is associated with the observation that the glycolytic flux in the hxk2Δ strain shows transcriptional down-regulation of five enzymes. In the upper glycolysis, the Hexokinase 2 has been deleted and the Phosphofructokinase 1 is down-regulated. The Phosphofructokinase 1 was also strongly down-regulated in the grr1Δ strain but the decrease in flux was not as large as in the hxk2Δ strain. In the lower glycolysis of the hxk2Δ, all the three iso-enzymes of Phosphoglycerate mutase were down-regulated as well as the phosphoglycerate kinase. No down-regulation for these enzymes was seen in the grr1Δ strain. The glycerol-3-phosphate dehydrogenase has two iso-enzymes. The first of those isoenzymes was up-regulated in the grr1Δ and the hxk2Δ strains; however its expected flux decreased in both cases (more significantly in hxk2Δ). The second iso-enzyme did not show important changes in grr1Δ but was down-regulated in hxk2Δ. This is not the only case in which different iso-enzymes show different regulatory patterns, and in these cases our method for having flux estimations independent from transcriptome analysis is particularly useful.

The strain hxk2Δ showed a strong decrease in ethanol production compared to grr1Δ. All the alcohol dehydrogenase iso-enzymes were down-regulated in a similar way in both strains. The explanation of the differences in flux towards ethanol should be found in the pyruvate decarboxylase. Pyruvate decarboxylase 3 was strongly down-regulated in both strains, however, pyruvate decarboxylase 2 was up-regulated. This up-regulation was much more significant in grr1Δ, which could explain a higher flux from pyruvate to AcCoA in this strain. The results for the hxk2Δ mutant are summarized in Fig. 5.

The mig1Δ mutant shows a higher specific growth rate than the wild type. In general it is transcriptionally very similar to the wild type [16]. An enrichment test in transcription factors between transcriptional regulated and post-transcriptional regulated reactions was performed and the factor Sfp1 was found. This factor is known to regulate ribosome production and is nutrient sensitive [23]. This could mean that the deletion of *MIG1* activates a response against starvation that results in an increased specific growth rate. Among the transcriptionally regulated reactions, a slight down-regulation of the PP pathway is observed together with an up-regulation of several amino-acid production pathways.

In the mig1Δmig2Δ mutant there is a slight decrease in the specific growth rate. All the 8 transcriptionally regulated reactions were down-regulated and belonged to amino-acid biosynthesis pathways. The enrichment test found the factors Cbf1 and Gcn4 (represented in 4 and 1 out of 8 reactions and 5 and 6 out of 36 reactions). Gcn4 is known to regulate amino-acid biosynthetic genes [24] and it seems that the up-regulation of amino-acid biosynthesis due to the deletion of *MIG1* is cancelled by an opposite effect due to *MIG2*.

In all the mutants discussed above, the AcCoA carboxylase and the fatty acid synthesis showed metabolic regulation. This could indicate that in the studied cases the AcCoA pool was the main parameter responsible for adjusting the rate of lipid biosynthesis to match the changes in specific growth rates.

The experiments for the gdh1Δ mutant were performed in chemostat cultures [17] with the same dilution rate. The only observed change in exchange fluxes was a small decrease in glycerol production and only a few significant changes were identified in the metabolic fluxes. However, there were significant transcriptional changes in many metabolic pathways. This again points to the hypothesis that changes in transcription mainly results in altered metabolite levels such that metabolic homeostasis can be maintained. This is supported by metabolome analysis of this mutant, which showed that there were many changes in the metabolite levels [25] and in fact many of these changes were associated with changes also in the transcription of associated enzymes [26]. It is possible that the chemostat conditions, by imposing the same specific growth rate, forced the mutant strain to important transcriptional changes in order to keep the fluxes unchanged.

In Fig. 6 we aim to provide a global visualization of the changes for all the studied metabolic conditions.

Methods

Sampling in the region of feasible solutions

The steady state condition and the irreversibility of some reactions impose limitations on the flux distributions attainable by the cell [18]. The set of feasible solutions can be further constrained by fixing some fluxes to their experimental values. In general, the fluxes most accessible to experimental determination are those corresponding to uptake or secretion rates. After fixing a subset of fluxes, genome scale models still have a large number of degrees of freedom. In this study we used the genome scale model iFF708 for *S. cerevisiae* [27]. Random sampling has previously been performed [7] by enclosing the region of allowed solutions in a parallelepiped with the same dimensions as solution space (the null space of the stoichiometric matrix) and generating random points inside this parallelepiped. The points that lie inside the region of possible solutions are then selected. The COBRA Toolbox [12] uses a Hit and Run algorithm to generate random points in this way. In this work instead of sampling inside the region of allowed solutions we sampled at its corners.

In order to obtain corners in the space of allowed solutions we used the simplex method with a random set of objective functions to be maximized. The maximization of each of these objective functions will give a corner in the space of solutions. The constraints imposed upon each optimization are:

$$S\vec{v}=\vec{0} \tag{1}$$

$$v_i \geq 0 \forall i \in \{irreversible\} \tag{2}$$

$$v_j = v_j^{\exp} \forall j \in \{measured\} \tag{3}$$

The values of the measured fluxes ($\boldsymbol{v^{exp}}$) are different between conditions. This fact changes the shape of the region of feasible solutions between different conditions. S is the stoichiometric matrix of the network.

In order to reduce the effects of internal loops we first identified all the reactions that can get involved in loops using the FVA (Flux Variability Analysis) option in the COBRA Toolbox. The reactions that can be involved in loops are unbounded and show the default maximal or minimal value set in the COBRA Toolbox (1000 or −1000). If these bounds were kept, the means and standard deviations for these reactions would be unrealistic [6] and cannot be used for further analysis. In order to reduce the effect of loops, the default maximal and minimal fluxes for the reactions involved in loops, were set to a smaller value in order to reduce the loop effect. In order to select an appropriate value the bounds were increased from 0 in steps of 0.1 until the minimal value that allows obtaining flux distributions consistent with the experimental

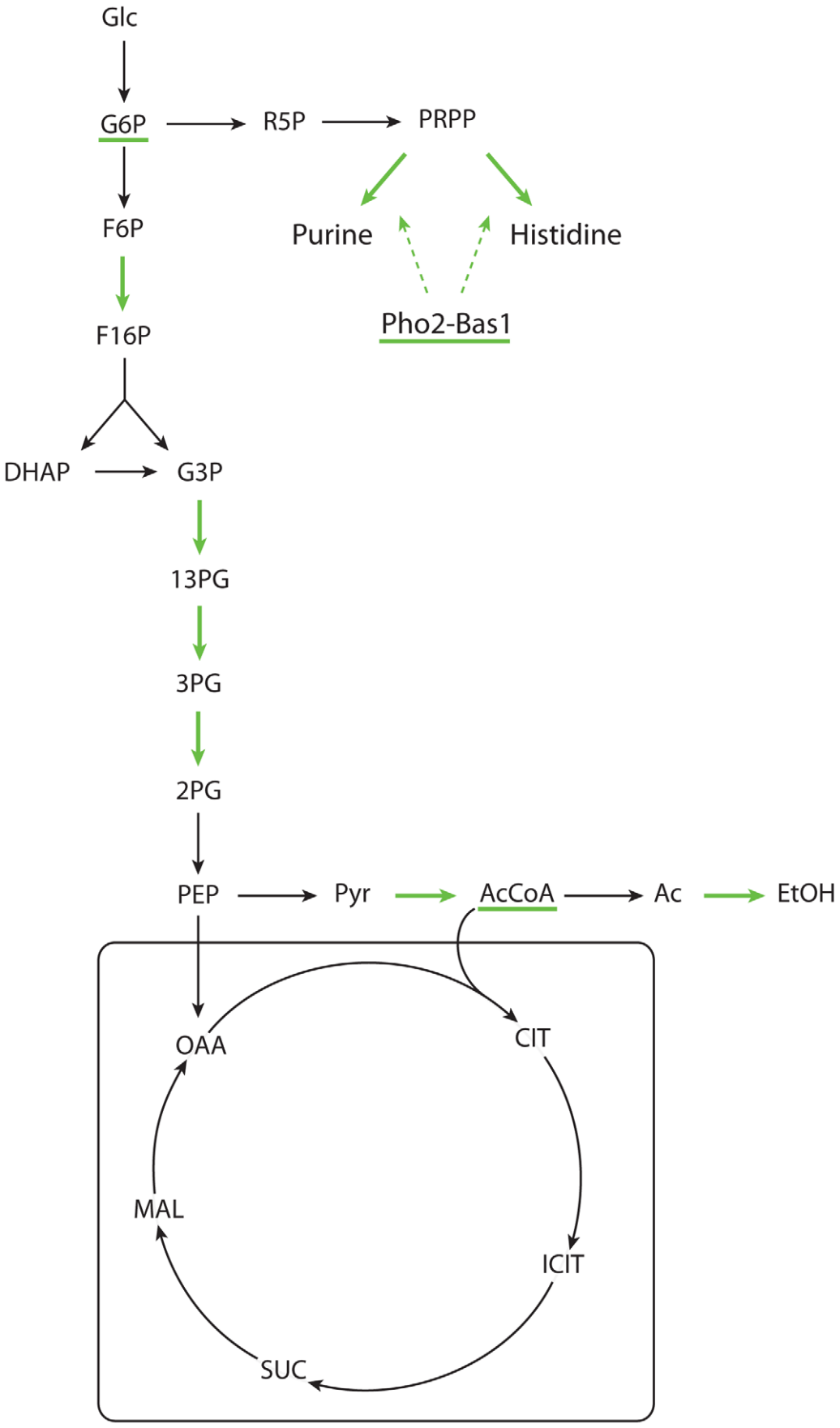

Glc
G6P
R5P
PRPP
F6P
Purine
Histidine
Pho2-Bas1
F16P
DHAP
G3P
13PG
3PG
2PG
PEP
Pyr
AcCoA
Ac
EtOH
OAA
CIT
MAL
ICIT
SUC

Figure 5. Main reactions showing transcriptional up (red) or down (green) regulation associated with the deletion of HXK2. The underlined metabolite pools are expected to increase (red) or decrease (green) according to the observed metabolic regulation. The transcription factors controlling the down-regulated pathways are also underlined in green.
doi:10.1371/journal.pcbi.1000859.g005

fluxes is found. These values went from 1 to 15 mmol h^{-1}g-DW^{-1} depending on each condition. Also no weights (eq. 4) were assigned to the reactions involved in loops in order to avoid objective functions that maximize the activity of loops.

Random objective functions were generated by selecting random pairs of reactions and assigning them random weights (the reactions involved in loops were excluded from these choices). The weights (w_i) assigned to each reaction were generated by dividing a random number between 0 and 1 by the maximal flux for this reaction obtained using FVA. This normalization was made to account for the different size orders of the different reactions. The objective functions take the form:

$$F = w_i v_i + w_j v_j \quad (4)$$

One solution is obtained for each of the objective functions generated.

Our objective is to obtain means and standard deviations for each flux in each of the compared conditions and use them to get a Z-score quantifying the significance of change in each flux between the considered conditions. This score is equal to the difference between the means in each of the conditions divided by the standard deviation of this difference (note that the variance of the difference is the sum of the two variances and the standard deviation its square root).

$$Z_i^{flux} = \frac{E_2(v_i) - E_1(v_i)}{\sqrt{Var_2(v_i) + Var_1(v_i)}} \quad (5)$$

The difference between averages in the numerator follows a normal distribution (according to the central limit theorem) with a standard deviation equal to the standard deviation of the flux (the denominator in eq. (5)) divided by the square root of the number of samples. Therefore, Z itself follows a normal distribution with a standard deviation equal to the inverse of the square root of the number of samples.

The Z score measures the significance of change in terms of standard deviations. If the error in the Z score is lower than 0.15, no information would be lost in terms of classifying a reaction as significantly changed or not. The order of size of a genome-scale model is about 1000 reactions. A reasonable accuracy for the Z-scores would be to expect errors higher than 0.15 on the Z score only for 1 reaction in the whole model. This means a p-value of 0.001. If we want to keep the error on the Z score under 0.15 with a probability of 0.999 we need 500 samples, and this was therefore selected as the sampling number.

Classification of enzymes according to their changes in flux and expression level

The Z-scores can be transformed into probabilities of change by using the cumulative Gaussian distribution. Once we have Z-

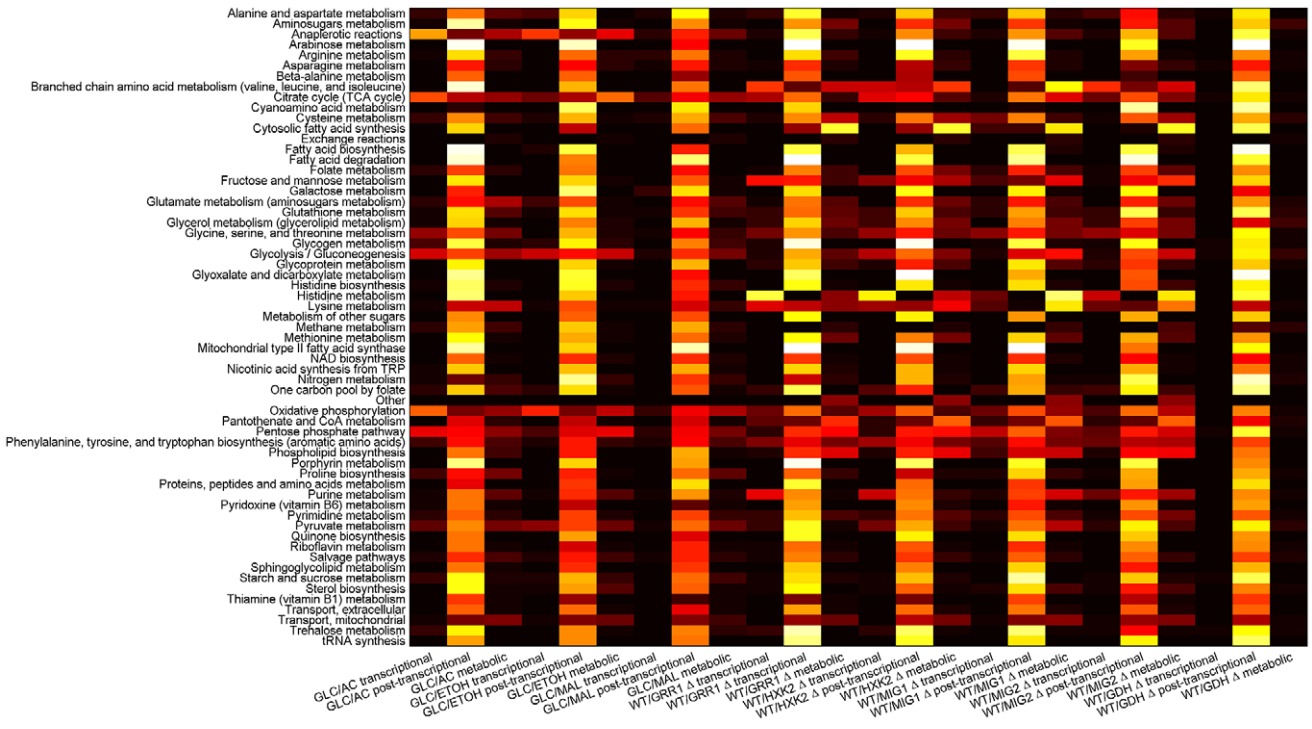


Figure 6. This figure illustrates the extent of transcriptional, post-transcriptional and metabolic regulation observed in different metabolic processes for each of the studied cases. The metabolic processes are defined in the same way as in the iFF708 model. The brightness of the color is proportional to the probability of a reaction in the corresponding process to show transcriptional, post-transcriptional and metabolic regulation respectively. The black correspond to 0 and the white to 1.
doi:10.1371/journal.pcbi.1000859.g006

Table 4. Decision table to assign transcriptional regulation (TR), post transcriptional regulation (PR) or metabolic regulation (MR) to the different enzymes depending on the observed up-regulation (+), down-regulation (−) or lack of change (=) of flux and gene expression.

Exp\Flux	+	−	=
+	TR	MR	PR
−	MR	TR	PR
=	MR	MR	

doi:10.1371/journal.pcbi.1000859.t004

scores for the significance of flux changes and Z-scores for the significance of gene-expression changes we can obtain probabilities of having correlated expression and flux changes for each enzyme.

An increase in enzyme expression can result in an increase of flux (transcriptional regulation). In order to evaluate the probability for a reaction of being transcriptionally regulated we multiply the probability of its enzyme level changing by the probability of its flux changing in the same direction (obtained using the cumulative normal distribution).

$$P_{tri} = \Phi\left(Z_i^{flux}\right)\Phi\left(Z_i^{\exp}\right) \tag{6}$$

$$\Phi(Z) = \int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \tag{7}$$

If there is a decrease in expression and a decrease in flux, both Z-scores are negative and we will use the absolute values of the Zs in eq. (6). If there is an increase in expression and a negative flux becomes more negative, we will use the absolute value of the Z-score for the flux change. If the direction of the flux changes between conditions, this change must be driven by the metabolic concentrations and no by transcriptional regulation, therefore a P_{tri} of zero is assigned by default.

In the same way as in eq. (6) we can define probabilities for the expression level changing and for the flux not changing (post transcriptional regulation).

$$P_{pri} = erf\left(Z_i^{\exp}\right)\left(1 - erf\left(Z_i^{flux}\right)\right) \tag{8}$$

$$erf(Z) = \int_{-Z}^{Z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \tag{9}$$

Now we use the error function because we want to evaluate the probability of change in any direction. The absolute value of Z is used in all the cases.

The probability of a change in flux but not in transcription (metabolic regulation) can be obtained for each reaction as follows:

$$P_{mri} = erf\left(Z_i^{flux}\right)\left(1 - erf\left(Z_i^{\exp}\right)\right) \tag{10}$$

Each of these three probabilities can be associated to each enzyme in the metabolic network.

Table 4 summarizes the criteria to assign each type of regulation.

Author Contributions

Conceived and designed the experiments: SB JN. Analyzed the data: SB RA. Wrote the paper: SB.

References

1. Patil KR, Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci U S A 102: 2685–2689.
2. Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19: 1720–1730.
3. Yang C, Hua Q, Shimizu K (2002) Integration of the information from gene expression and metabolic fluxes for the analysis of the regulatory mechanisms in Synechocystis. Appl Microbiol Biotechnol 58: 813–822.
4. Moxley JF, Jewett MC, Antoniewicz MR, Villas-Boas SG, Alper H, et al. (2009) Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p. Proc Natl Acad Sci U S A 106: 6477–6482.
5. Fong SS, Nanchen A, Palsson BO, Sauer U (2006) Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes. J Biol Chem 281: 8024–8033.
6. Mo ML, Palsson BO, Herrgård MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC systems biology 3: 37.
7. Schellenberger J, Palsson BØ (2008) Use of randomized sampling for analysis of metabolic networks. J Biol Chem 284: 5457–5461.
8. Almaas E, Kovács B, Vicsek T, Oltvai ZN, Barabási AL (2004) Global organization of the metabolic fluxes in the bacterium *Escherichia coli*. Nature 427: 839–843.
9. Papin JA, Reed JL, Palsson BØ (2004) Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. Trends Biochem Sci 29: 641–647.
10. Jamshidi N, Palsson BO (2007) Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC Syst Biol 1: 26.
11. Thiele I, Price ND, Vo TD, Palsson BØ (2005) Candidate metabolic network states in human mitochondria: Impact of diabetes, ischemia, and diet. Journal of Biological Chemistry 280: 11683–11695.
12. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, et al. (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. Nat Prot 2: 727–738.
13. Lovasz L (1999) Hit and Run mixes fast. Math Program 86: 443–461.
14. Jouhten P, Rintala E, Huuskonen A, Tamminen A, Toivari M, et al. (2008) Oxygen dependence of metabolic fluxes and energy generation of Saccharomyces cerevisiae CEN.PKI I3-IA. BMC Sys Biol 2: 60–79.
15. Daran-Lapujade P, Jansen MLA, Daran JM, van Gulik W, de Winde JH, et al. (2004) Role of transcriptional regulation in controlling fluxes in central carbon metabolism of Saccharomyces cerevisiae. J Biol Chem 279: 9125–9138.
16. Westergaard SL, Oliveira AP, Bro C, Olsson L, Nielsen J (2007) A systems biology approach to study glucose repression in the yeast Saccharomyces cerevisiae. Biotechnol Bioeng 96: 134–145.
17. Bro C, Regenberg B, Nielsen J (2004) Genome-wide transcriptional response of a Saccharomyces cerevisiae strain with an altered redox metabolism. Biotechnol Bioeng 85: 269–276.
18. Price ND, Reed JL, Palsson BØ (2004) Genome scale models of microbian cells: Evaluating the consequences of constraints. Nat Rev Microbiol 2: 886–897.
19. Ostergaard S, Walløe KO, Gomes CSG, Olsson L, Nielsen J (2001) The impact of GAL6, GAL80 and MIG1 on glucose control of the GAL system in Saccharomyces cerevisiae. FEMS Yeast Res 1: 47–55.
20. Daran-Lapujade P, Rossell S, van Gulik WM, Luttik MAH, de Groot MJL, et al. (2007) The fluxes through glycolytic enzymes in Saccharomyces cerevisiae are predominantly regulated at posttranscriptional levels. Proc Natl Acad Sci U S A 104: 15753–15758.
21. Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A 99: 1801–1806.
22. Bhoite LT, Allen JM, Garcia E, Thomas LR, Gregory ID, et al. (2002) Mutations in the Pho2 (Bas2) transcription factor that differentially affect activation with its partner proteins Bas1, Pho4, and Swi5. J Biol Chem 227: 37612–37618.

23. Marion RM, Regev A, Segal E, Barash Y, Koller D, et al. (2004) Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. Proc Natl Acad Sci U S A101: 14315–14322.
24. Hope IA, Struhl K (1987) Gcn4, a eukaryotic transcriptional activator protein, binds as a dimer to target DNA. EMBO J 6: 2781–2784.
25. Villas-Bôas SG, Moxley JF, Åkesson M, Stephanopoulos G, Nielsen J (2005) High throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. Biochem J 388: 669–677.
26. Cakir T, Patil KR, Önsan ZI, Ülgen KÖ, Kirdar B, et al. (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. Mol Syst Biol 2: 50.
27. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res 13: 244–253.

Paper III

The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum*

Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I. and Nielsen, J.

PLoS Comput Biol (2013), 9(3), p. e1002980

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum*

Rasmus Agren[1], Liming Liu[1], Saeed Shoaie[1], Wanwipa Vongsangnak[2], Intawat Nookaew[1], Jens Nielsen[1]*

1 Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden, **2** Center for Systems Biology, Soochow University, Suzhou, China

Abstract

We present the RAVEN (Reconstruction, Analysis and Visualization of Metabolic Networks) Toolbox: a software suite that allows for semi-automated reconstruction of genome-scale models. It makes use of published models and/or the KEGG database, coupled with extensive gap-filling and quality control features. The software suite also contains methods for visualizing simulation results and omics data, as well as a range of methods for performing simulations and analyzing the results. The software is a useful tool for system-wide data analysis in a metabolic context and for streamlined reconstruction of metabolic networks based on protein homology. The RAVEN Toolbox workflow was applied in order to reconstruct a genome-scale metabolic model for the important microbial cell factory *Penicillium chrysogenum* Wisconsin54-1255. The model was validated in a bibliomic study of in total 440 references, and it comprises 1471 unique biochemical reactions and 1006 ORFs. It was then used to study the roles of ATP and NADPH in the biosynthesis of penicillin, and to identify potential metabolic engineering targets for maximization of penicillin production.

Citation: Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, et al. (2013) The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum*. PLoS Comput Biol 9(3): e1002980. doi:10.1371/journal.pcbi.1002980

Editor: Costas D. Maranas, The Pennsylvania State University, United States of America

Received September 11, 2012; **Accepted** January 24, 2013; **Published** March 21, 2013

Funding: This project has been financed by European Research Council (Grant 247013), the EU-funded project SYSINBIO and Sandoz. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nielsenj@chalmers.se

Introduction

Genome sequencing projects have in recent years contributed enormously to our understanding of the metabolic capabilities of cellular systems. Functional annotation of the gene products allow for reconstruction of genome-scale metabolic models (GEMs) that summarize these metabolic capabilities in a consistent and compact way [1,2]. A number of mathematical tools, including sampling of available metabolic states [3,4] and methods borrowed from computational geometry [5], have been developed to analyze the resulting networks and to gain insight into the complex interactions that give rise to the metabolic capabilities. GEMs have also been used extensively for simulation of metabolism, particularly for metabolic engineering purposes [6,7]. Since these models connect metabolites, proteins, and genes they are particularly well suited for the integration of metabolomics, proteomics, and genomics which is, in a sense, the goal of systems biology [8,9].

The foundation of a GEM is the functional annotation of the genes. The first GEMs were primarily for model organisms for which direct evidence exists in the literature for a large proportion of the genomically encoded functions [10,11]. However, as the number of genome sequencing projects increases there is a growing demand of GEMs for less well known organisms. These models must by necessity be built largely relying on protein homology to more well-characterized organisms [12]. This, together with the large amount of manual work that is involved in a strict bottom-up reconstruction, has sparked interest in more automated approaches to model reconstruction. There are now a number of tools available for automated annotation of genes [13,14]. However, the annotated genes must be linked to metabolic reactions in a way so as to generate a functional metabolic model. This includes the addition of spontaneous reactions and non-carrier mediated transport across membranes as well as sub-cellular localization of enzymes. Most importantly, the model must also be constructed in a way so that all reactions are balanced and well-connected [15]. This tends to become a problem if the gene-reaction relationship is automatically inferred from databases, partly due to differences in metabolite naming, but mainly because of how complex carbohydrates and complex lipids are represented. Because of the aforementioned issues it makes sense to use previously reconstructed models as templates for new GEMs. Here we present the RAVEN Toolbox, which allows the user to input GEM(s) for one or more template organisms, their corresponding protein sequences, and the protein sequences of the target organism. A GEM for the target organism is then constructed based on orthology between the protein sequences of the target organism and the organisms of the template models. Metabolic functions not present in the template models can obviously not appear in the new model, and to account for these missing reactions the RAVEN Toolbox also includes a functionality that matches proteins to KEGG Orthology (KO) categories [16] by using Hidden Markov models to capture the representative amino acid pattern in each KO. The resulting metabolic network can be used for automatic or manual gap filling, or it can be used on its own as a draft network.

Author Summary

Genome-scale models (GEMs) are large stoichiometric models of cell metabolism, where the goal is to incorporate every metabolic transformation that an organism can perform. Such models have been extensively used for the study of bacterial metabolism, in particular for metabolic engineering purposes. More recently, the use of GEMs for eukaryotic organisms has become increasingly widespread. Since these models typically involve thousands of metabolic reactions, the reconstruction and validation of them can be a very complex task. We have developed a software suite, RAVEN Toolbox, which aims at automating parts of the reconstruction process in order to allow for faster reconstruction of high-quality GEMs. The software is particularly well suited for reconstruction of models for eukaryotic organisms, due to how it deals with sub-cellular localization of reactions. We used the software for reconstructing a model of the filamentous fungi *Penicillium chrysogenum*, the organism used in penicillin production and an important microbial cell factory. The resulting model was validated through an extensive literature survey and by comparison with published fermentation data. The model was used for the identification of transcriptionally regulated metabolic bottlenecks in order to increase the yield in penicillin fermentations. In this paper we present the RAVEN Toolbox and the GEM for *P. chrysogenum*.

Several approaches which also aim at generating GEMs from either a template model or from a general database have been published [17–20]. Table 1 summarizes the capabilities of the RAVEN Toolbox compared to some other published approaches when it comes to automatic reconstruction. However, the largest difference is maybe not in the approaches taken, but in that the RAVEN Toolbox is a complete software for all tasks involving reconstruction and simulation of GEMs. In this aspect the RAVEN Toolbox is more similar to the COBRA Toolbox, but with extensive reconstruction capabilities [21]. Even though the RAVEN Toolbox can be used for fully automated reconstruction, in a manner similar to Model SEED, the intended purpose is to make use of the extensive quality control and gap identification/gap filling features for increasing the quality of reconstructions, as well as for decreasing the time needed for reconstructing high-quality models.

The RAVEN Toolbox was evaluated for its ability to reconstruct a GEM for the well studied yeast *Saccharomyces cerevisiae*, and the resulting GEM was compared with a manually reconstructed model. Thereafter the RAVEN toolbox was used for the reconstruction of a GEM of the industrially important mold *Penicillium chrysogenum*. The *Penicillium* genus encompasses species of great economical, medical, and environmental importance [22]. Members of the *Penicillium* genus serve important roles in the food industry, both as some of the main spoilers of fresh vegetables and as essential actors in the production of blue cheeses. Most importantly though, they are sources of major antibiotics, particularly penicillin and griseofulvin.

The industrial production of β-lactam antibiotics, such as penicillins and cephalosporins, is one of the success stories of biotechnology. Today the β-lactams represent one of the largest biotechnological products in terms of value, with sales of about USD 15 billion [23]. The industrial *P. chrysogenum* strains have been subjected to 50 years of directed evolution to increase the yields and titers of penicillin, with great cost reduction and productivity gain, but the yields are still far from the theoretical maximum [24]. A GEM of *P. chrysogenum* could aid in identification of metabolic bottlenecks as well as in elucidating the underlying reason for the significantly better performance of industrial strains compared to low producing strains.

Results

The RAVEN Toolbox

A software suite named the RAVEN Toolbox (Reconstruction, Analysis, and Visualization of Metabolic Networks) was developed. The toolbox is a complete environment for reconstruction, analysis, simulation, and visualization of GEMs and runs within MATLAB. The software imports and exports models in two formats: the widely used Systems Biology Markup Language (SBML) format [25] and a Microsoft Excel model representation. Both these formats allow for extensive annotation of model components, such as International Chemical Identifier strings (InChI) [26] for metabolites or database identifiers for reactions and genes. The native model format for the RAVEN Toolbox follows the format of the yeast consensus metabolic network [27], but models in the COBRA Toolbox format can also be imported [21]. The Microsoft Excel representation enables the user to set simulation parameters such as bounds and objective function coefficients directly in the spread sheets. This simplifies the modeling process for users not comfortable with working within a scripting environment, as well as providing a simpler, but less rigorous, model format compared to SBML. The software, together with a manual, a set of tutorials and a detailed description of the supported file formats is available through the BioMet Toolbox [28] (http://www.sysbio.se/BioMet). Figure 1 summarizes the capabilities of the RAVEN Toolbox.

Table 1. Comparison between the RAVEN Toolbox and other software for automatic GEM reconstruction.

	RAVEN	*Model SEED* [*20*]	*AUTOGRAPH* [*18*]	*IdentiCS* [*60*]	*GEM System* [*17*]
Includes general network	X	X		X	X
Generates functional models	X	X			
Assigns sub-cellular localization	X				
Can use user defined models	X		X		
Integrates gap filling	X	X			X
Offline software	X			X	
Includes visualization	X			X	X
Gene prediction				X	X

doi:10.1371/journal.pcbi.1002980.t001

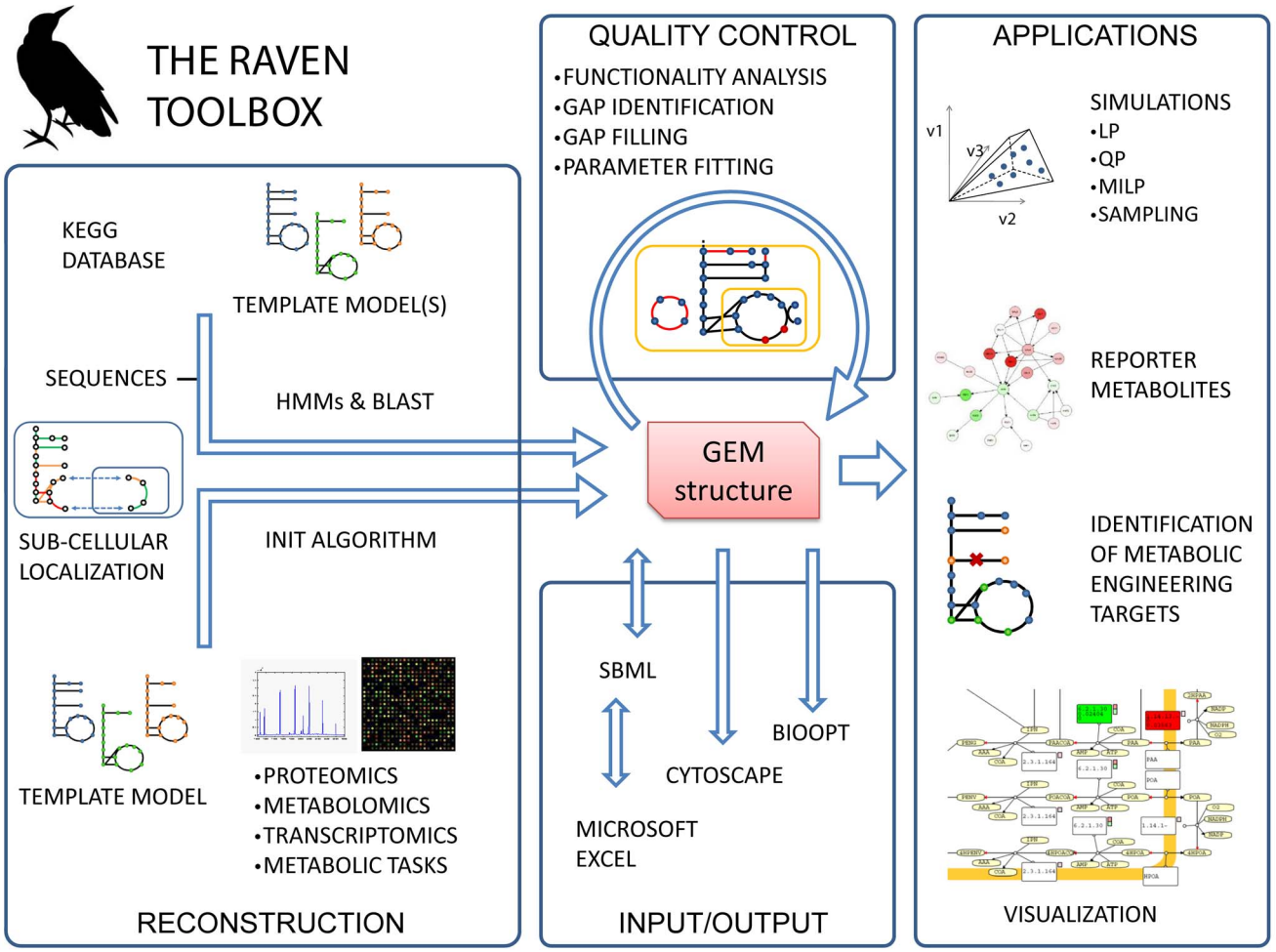


Figure 1. The RAVEN Toolbox. The software allows for reconstruction of GEMs based on template models or on the KEGG database. The resulting models can be exported to a number of formats, or they can be used for various types of simulations. The RAVEN Toolbox has a strong focus on quality control. Visualization of simulation result and/or integration of other types of data can be performed by overlaying information on pre-drawn metabolic maps. The software also implements the INIT algorithm, which is a powerful approach for reconstruction of tissue-specific models [59]. HMM: Hidden Markov model, LP: Linear programming, QP: Quadratic programming, MILP: Mixed-integer linear programming.
doi:10.1371/journal.pcbi.1002980.g001

The software has three main foci: 1) automatic reconstruction of GEMs based on protein homology, 2) network analysis, modeling and interpretation of simulation results, 3) visualization of GEMs using pre-drawn metabolic network maps.

Automated reconstruction of GEMs based on protein homology. Previously published GEMs represent a solid basis for metabolic reconstruction of models for new organisms, in particular if the organisms are closely related and therefore share many metabolic capabilities. The main advantage of using existing models compared to reaction databases, such as KEGG or BRENDA [29], is that they contain information that can be difficult to obtain in an automated manner, in particular directionality and compartmentalization. There have been attempts to predict the directionality of reactions based on the estimates of the standard Gibbs energies of formation for the involved metabolites [30]. However, we believe that manually reconstructed networks for related species can be a more reliable source of directionality information. The same is true for compartmentalization. Even though the RAVEN Toolbox contains methods for inferring subcellular localization based on predictors, it is to be viewed as an aid rather than an exact method. GEMs are also typically constructed for modeling purposes, which is not the case for reaction databases. The downside is that only reactions present in the template models can be included. The RAVEN Toolbox therefore contains two approaches for automatic generation of draft models; while the method mentioned above relies on the metabolic functions represented in previously published models, the complementary method uses the KEGG database for automatic identification of new metabolic functions that are not included in the published models.

The first approach lets the user supply a number of existing GEMs and FASTA files with protein sequences for the template models and for the organism of interest. The software then generates a draft model based on protein orthology. The default implementation uses bi-directional BLASTp [31] for evaluation of protein homology, but the software also supports other homology measurements as long as a score can be assigned to each pair wise protein comparison. The resulting model can be exported as a SBML file or be used in MATLAB for simulation and further analysis.

The second approach is also based on protein homology but requires no template models. Instead it relies on the information

on protein sequences and on the assigned metabolic reactions that is available in the KEGG database. The method makes use of the KEGG Orthology (KO) IDs, which are manually annotated sets of genes that encode some specified metabolic function. Each KO is associated with a number of metabolic reactions. The aim of the present method is to assign genes to these KOs based on the consensus protein sequence. The tool first downloads all relevant parts of the KEGG database to a local directory and parses these files to generate a GEM representing a metabolic network across all of the species annotated in KEGG, i.e. this would lead to a network comprising 7029 metabolites, 8398 reactions and 843369 genes, when using the most current version of the KEGG database. A GEM for the organism of interest is then constructed by choosing a subset of this larger model and linking the reactions with the corresponding genes. The protein sequences for each KO are retrieved and aligned using MUSCLE [32]. The user has the option to only use genes from organisms of a given phylogenetic distance from the target organism, e.g. only fungal genes or only eukaryotic genes. A hidden Markov model is then generated based on the sequences for each KO using HMMER [33]. The final step is the querying of the set of HMMs with the protein sequences of the organism of interest. If a gene has a significant match to one KO, the reactions associated to that KO are added to the model together with the corresponding gene. This process is fully automated, and the user only needs to supply a FASTA file with protein sequences. Users who do not subscribe to KEGG can download pre-trained HMMs for eukaryotes and prokaryotes through the BioMet Toolbox (http://www.sysbio.se/BioMet). These HMMs are based on the last open version of KEGG. More advanced users can set parameters that affect how genes are mapped to KOs and how general, unbalanced, or otherwise problematic reactions from KEGG should be dealt with.

Model analysis and simulation. The approach proposed above will facilitate and accelerate the generation of a draft metabolic network reconstruction. The automated reconstruction can lead to some loss of control compared to a stricter manual, bottom-up approach. It is therefore important to identify and fill gaps in the model to ensure that the network is functioning as required. In a high quality model all reactions should be able to have a flux if all uptake and excretion reactions are allowed and net synthesis of most metabolites should be possible (the exception would normally be some co-factors). The second criterion is important, since the large degree of freedom in GEMs allow for internal loops where reactions can carry flux but where no net consumption or synthesis of metabolites occurs. The RAVEN Toolbox contains a number of methods to support the gap filling process. The following section describes the suggested workflow for gap identification and filling when starting from a draft network.

- Gap filling traditionally centers on adding reactions in order to enable production of all precursors needed for biomass production. However, it is equally important to ensure that the model cannot produce anything when there is no uptake of metabolites. The reactions which enable this type of behavior are typically those which involve polymers, metabolite pools, or other abstract metabolites but they can also simply be erroneous reactions. A brute force solution would be to exclude all reactions which are not elementally balanced, but this could result in a large fraction of the network being deleted, as many metabolites typically lack information about elemental composition. The *makeSomething* and *consumeSomething* functions identifies such reactions by solving the mixed integer linear programming (MILP) problem of finding the smallest set of reactions which results in the net synthesis or consumption of any metabolite. The solutions can then be cross-referenced to balancing information from *getElementalBalance* in order to identify reactions which are both active and have wrong/lacking composition. This process can also be done automatically using *removeBadRxns*.
- After the user has added relevant exchange reactions *canProduce/canConsume* can be used to generate a list of the metabolites that can have net synthesis or consumption. Early on in the reconstruction process it is likely that not all biomass precursors can be synthesized. The function *checkProduction* can be very useful in this situation. It calculates the smallest set of metabolites which must have net synthesis in order to enable net synthesis of all other metabolites. This gives the user information such as "in order to synthesize biomass, you must enable synthesis of valine and coenzyme A" or "if synthesis of choline is enabled, the following set of metabolites could also be synthesized". The function also allows the user to set rules about merging compartments, since it can be easier to first make sure that the model is functional with merged compartments and deal with transport and sub-cellular localization afterwards.
- Ideally all reactions should be able to carry flux if all relevant exchange reactions are available. The function *haveFlux* can be used to identify reactions which cannot carry flux, and also to distinguish between reactions which cannot carry flux because some substrate cannot be synthesized and those which cannot carry flux because some product cannot be further consumed. However, because of the many internal loops in GEMs it is common that reactions can carry flux and appear well-connected even if they are not connected to the rest of the metabolic network. *getAllSubGraphs* can be used to identify such subnetworks using Tarjan's algorithm [34].
- The function *fillGaps* can be used to retrieve reactions from a set of template models or from KEGG in order to generate a functional network. The user can set constraints on their model, such as that it should be able to produce biomass from minimal media, and *fillGaps* will then solve the MILP problem of including the minimal set of reactions from a set of template models in order to satisfy the constraints. The same function can be used to enable net synthesis of all metabolites or to enable flux through all reactions. This approach is similar to that taken in Model SEED, and enables fully automatic model reconstruction. However, we suggest that GEM reconstruction should be done iteratively and with manual input and that the results from these algorithms are to be viewed as suggestions to point the user in the right direction.
- In eukaryotes the enzymatic reactions are distributed between different organelles. To determine which reactions occur where is a difficult task, and one of the more time-consuming steps in the reconstruction process. The RAVEN Toolbox takes a first step towards speeding up this step by including a method for assigning subcellular localization to enzymatic reactions in an automated fashion. The algorithm aims at assigning localization in a manner that is consistent with signal peptide composition and physiochemical protein properties, while at the same time maintaining a well-connected and functional network. The default predictor is WoLF PSORT, which is distributed with the RAVEN Toolbox [35]. A parser for other predictors, such as CELLO is also included [36]. In short, the algorithm works by generating fully connected solutions, which are then scored based on the agreement to the predicted localization and the number of transport reactions which had to be included in order to have a connected

network. The problem is solved using simulated annealing. A more detailed description is available in Figure 2.

The RAVEN Toolbox also contains a number of methods for performing simulations using GEMs. In this aspect it is similar to the COBRA Toolbox [21]. Most of the features of COBRA Toolbox are also present in the RAVEN Toolbox, with the exception of dynamic FBA. This includes linear programming such as FBA, quadratic programming such as MoMA, mixed integer linear programming applications, and random sampling. Utility functions such as setting constraints and objectives, adding or removing model elements, presenting simulation outputs, sensitivity analysis, screening for gene deletions, and fitting model parameters such as maintenance ATP consumption are also included. For a full description of all functions of the toolbox, see the supplied manual. The RAVEN Toolbox uses MOSEK (MOSEK ApS, Copenhagen, Denmark) for solving the underlying optimization problems. MOSEK is proprietary software but a full featured license is freely available for academic use.

Validation of the workflow. The RAVEN Toolbox pipeline was validated by constructing a model for *Saccharomyces cerevisiae*, a model organism for which several GEMs have been constructed. To compare the quality of the automatically generated model to a manually curated one, some kind of reference was needed. As all models contain errors it would not be very relevant to simply compare the similarity between the RAVEN Toolbox generated model and a previously published model. Saccharomyces Genome Database (SGD) was therefore used as a reference with respect to the enzymes present in *S. cerevisiae* and their subcellular localization. A model was generated from KEGG in a fully automatic manner and then compared to the iIN800 model, a model which has been shown to have excellent simulation capabilities [37]. It should be noted that this fully automatic reconstruction is not how the RAVEN Toolbox is intended to be used for reconstruction. We suggest that the user view the output of each step as suggestions, and manually fill gaps or fix problematic reactions. However, we wanted to perform an evaluation of the overall reconstruction feature of the RAVEN Toolbox.

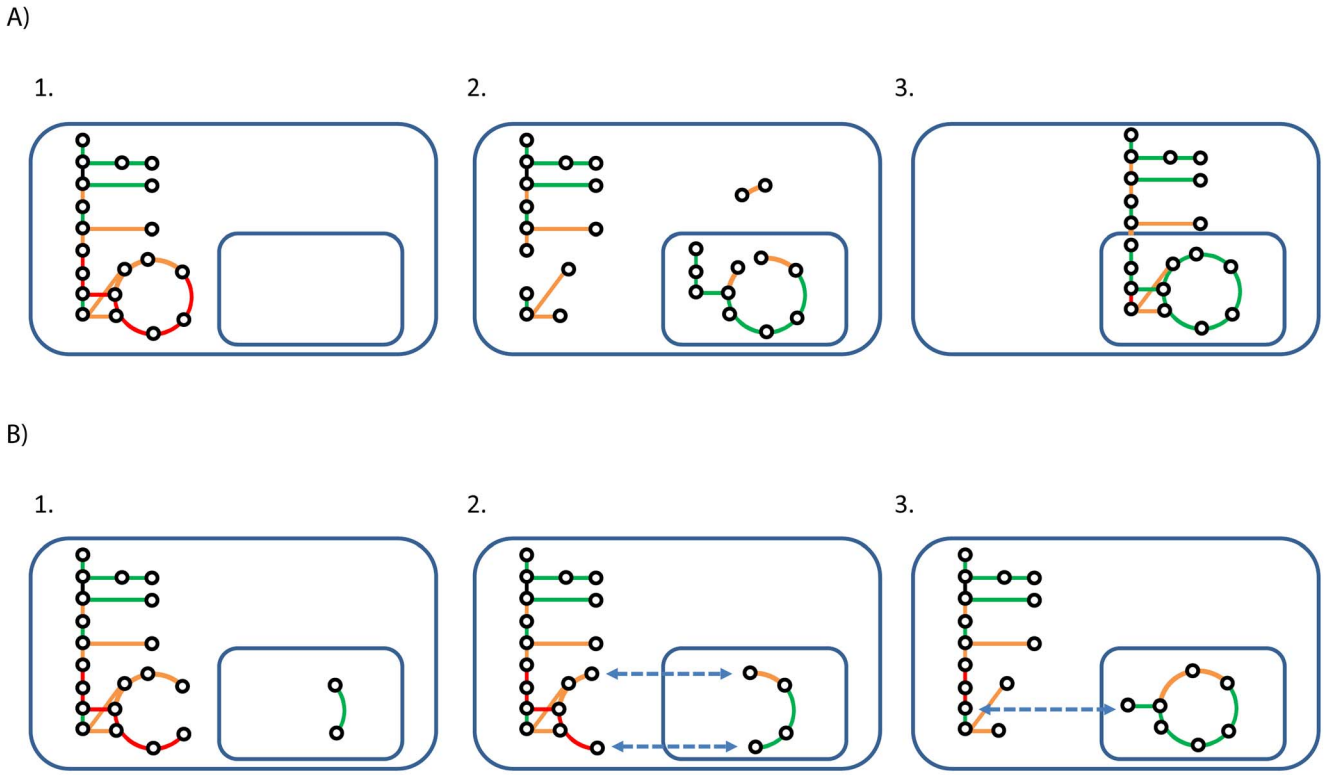


Figure 2. Prediction of subcellular localization of reactions. Circles correspond to metabolites and lines correspond to reactions. Green lines are reactions which are in their correct compartment according to the predictions. Red are reactions which are in an incorrect compartment and orange are reactions where there is no strong indication for either compartment. A) There is a tradeoff between connectivity and agreement with predicted localization. Network 1 represents the extreme case where connectivity is much more important than predicted localization scores. All reactions are then localized to the cytosol. Network 2 represents the other extreme case where the reactions are localized only based on localization scores and with no regard for connectivity. This would result in an unconnected network. Network 3 represents the case where the network is connected, while still being in good agreement with the localization scores. The underlying assumption in the algorithm is that a good network is characterized by being fully connected, in the sense that all metabolites are synthesized in at least one reaction and consumed in at least one reaction, while still being in good agreement with the localization scores and relying on the smallest possible number of transport reactions to achieve this. B) Summary of the localization algorithm. 1. The algorithm first randomly moves one gene product and its associated reaction(s) to another compartment. The probabilities depend on the scores for the gene products in their respective compartments. 2. This may result in an unconnected network. The algorithm then tries to find a small set of reactions which, when moved, reconnects the network. If moving these reactions would result in a large decrease of fitness, then the network is connected by including transport reactions for some metabolites instead. 3. The connected network is then scored as the sum of scores for all genes in their assigned compartment, minus the cost of all transport reactions that had to be included in order to keep the network connected. The user can set the relative weight given to transport compared to gene localization. The overall problem is solved using simulated annealing.
doi:10.1371/journal.pcbi.1002980.g002

The model was generated using *getKEGGModelForOrganism* with the settings to only use eukaryotic genes when training the HMMs, a cutoff of 1e-30 when matching genes to the HMMs, and to exclude reactions labeled as general or incomplete in KEGG. *S. cerevisiae* genes were excluded in the training of the HMMs to simulate reconstruction of an organism for which there is little previous gene annotation. Not all unbalanced or erroneous reactions were labeled as such, and this resulted in that the KEGG model could produce some metabolites without any uptakes. *removeBadRxns* identified 79 reactions which enabled such production (see Table S1). Out of these 72 were unbalanced, general or polymer reactions and as such were correctly removed. 7 reactions were correct, but lacked composition about the metabolites (it is a setting in *removeBadRxns* whether it is allowed to remove such reactions).

Based on experimental minimal media the model was allowed uptake of glucose, phosphate, sulfate, NH3, oxygen and the essential nutrients 4-aminobenzoate, riboflavin, thiamine, biotin, folate, and nicotinate [38]. Uptake of the carriers carnitine and acyl-carrier protein was allowed for modeling purposes (many compounds are bound to them and therefore net synthesis of these compounds is not possible without them). This was the only manual step in the reconstruction of the yeast model.

The resulting model contained 1126 reactions, 1144 metabolites and 713 genes (before compartmentalization). 521 (73%) of those genes were shared with iIN800. 192 genes were unique to the automatically reconstructed model and 91 genes were unique to the iIN800 model (since there are no transport reactions in KEGG, all transporters were excluded from iIN800 for the purpose of this comparison). Figure 3 shows a classification of the genes that are unique to either the automatically reconstructed model or to iIN800 (see Table S2 for details). As can be seen, the automatically reconstructed model has a significantly larger proportion of enzymes compared to the published model.

Given the inputs the model could have net-synthesis of 476 (42%) of the 1144 metabolites (see Table S3 for details). As a comparison, 456 (66%) out of 683 unique metabolites in iIN800 could be synthesized given the same inputs. Among the 476 metabolites were 19 out of the 22 standard amino acids (leucine, methionine, and taurine could not be synthesized), the nucleotides needed for RNA and DNA synthesis (ATP, GTP, CTP, UTP, dATP, dGTP, dCTP, and dTTP), fatty acids, sterols such as lanosterol and ergosterol, important co-factors such as NADH, NADPH, FADH2 and CoA, and the building blocks needed for cell wall assembly (UDP-glucose, UDP-N-acetyl-D-glucosamine and mannose). The only major biomass constituents that could not be synthesized were complex lipid compounds such as phospholipids and sphingolipids. This is because of the combinatorial nature of fatty acid metabolism (given $\sim$20 fatty acids there are $20!/2!(20-2)=190$ possible versions of phosphatidylcholine) and how it is represented in KEGG.

As the next step of the fully automatic reconstruction, *fillGaps* was used to automatically fill gaps in the yeast network using the full KEGG database as a template. This resulted in 45 reactions being added, which in turn enabled the synthesis of 91 metabolites that could previously not be synthesized (see Table S4). Among them were the three amino acids that were previously missing. A closer investigation of the reactions which were added (see Table S5) showed that out of the 45 added reactions, 17 had evidence to support that they should be included in the model, 9 had inconclusive or missing evidence, and 19 reactions should not have been included in the model. 5 of the 91 genes that were previously unique to iIN800 were also added in this process.

The RAVEN Toolbox also contains a method for partitioning enzymatic reactions to compartments in a manner that keeps the network connected, but at the same time in agreement with the results from predictors of protein localization (see Figure 2 for details). The default predictor, WoLF PSORT, was used to predict the protein localization of all ORFs in the FASTA file. *predictLocalization* was then used to partition the network between mitochondria and cytosol. The transport cost was set to 0.1. Table S6 lists the genes for which the corresponding reactions were assigned to the mitochondria. 119 gene products were assigned to the mitochondria and the remaining 594 gene products were assigned to the cytosol. Out of the 119 predicted mitochondrial gene products, 72% were listed as mitochondrial in the SGD based on experimental evidence. The same calculations for iIN800 give that 91 gene products are mitochondrial and that 83% were listed as mitochondrial in SGD. Localization predictions based only on primary protein sequences are not very exact, and the resulting model from *predictLocalization* will not be totally biologically correct. The main issue is that all transport reactions are formulated as passive diffusion, while in reality other types of transport are also taking place. However, the method is able to quickly generate a connected model where the enzyme localizations are in almost as good agreement to SGD as a published model. This could be useful for many applications, such as when using metabolic networks for integrating omics data, and it constitutes a first step towards fully automated reconstruction of eukaryote GEMs.

These results show both strengths and weaknesses of using a fully automatic approach to reconstruction. A model capable of producing all the needed building blocks for synthesis of protein, RNA, DNA, and the cell wall was generated solely from a FASTA file and with almost no manual input. The automated gap filling identified 17 new reactions, out of which 8 were not present in the published *S. cerevisiae* model. As was shown in Figure 3, the quality in terms of included genes was as good or better compared to the published model. On the other hand, the gap filling included 19 reactions which did not belong in the model, and complex lipids could not be synthesized. The sub-cellular localization of enzymes was up to par with the published model, but with the drawback that all transport reactions were formulated as passive diffusion. In a real situation a reconstruction should therefore be done in an iterative manner, with manual input after each iteration (the user would, for example, remove the 19 bad reactions from the template model and then run *fillGaps* again).

Visualization of GEMs. Stoichiometric metabolic models have been proven to generate remarkably good predictions when it comes to the central carbon metabolism in microorganisms. However, the lack of kinetic and regulatory information is a rather large simplification and it is possible to get simulation results that have little biological meaning (such as thermodynamically disallowed loops). It is therefore imperative to understand the underlying reasons for a change in predicted phenotype after a perturbation such as a gene deletion. Due to the large dimensionality of GEMs interpretation of flux distributions is a rather daunting task. Visualization of fluxes can aid with interpretation, as well as provide an instant overview of how the system functions. Software that aims at network visualization based purely on connectivity, such as Cytoscape [39] or CellDesigner [40], cannot provide a comprehensible or well organized image of GEMs due to their size. The RAVEN Toolbox allows for visualization of simulation results based on manually drawn maps. The maps are drawn in CellDesigner and each reaction is labeled with the corresponding reaction identifier in the model. The map can then be imported to MATLAB and the

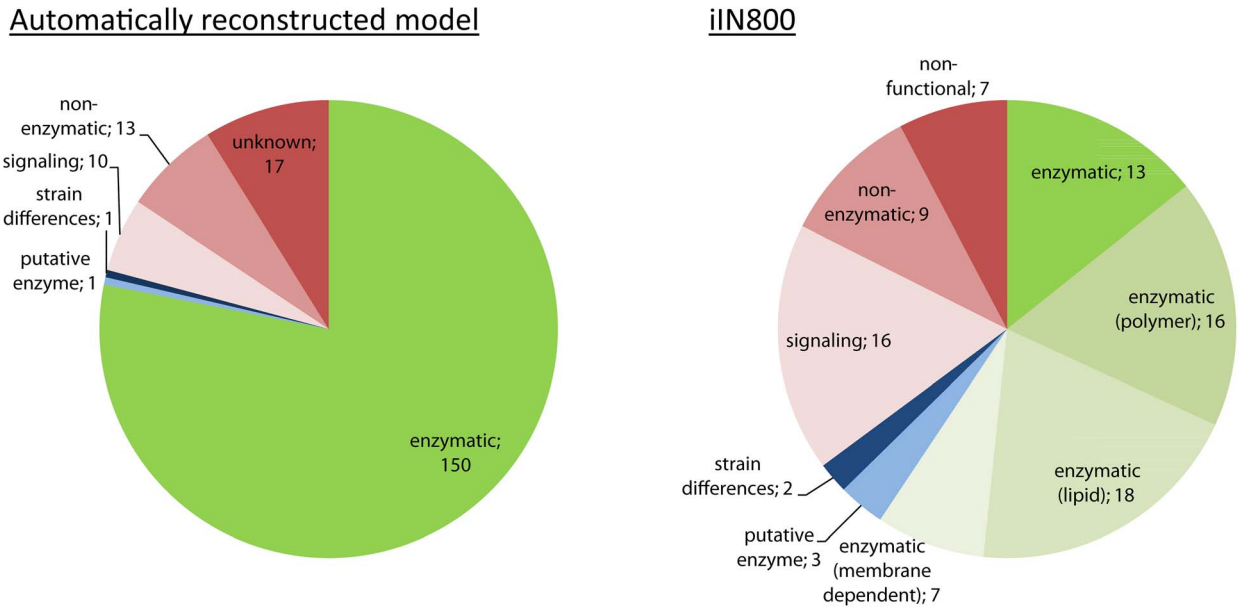


Figure 3. Overview of the genes which are unique to the automatically reconstructed model and iIN800, respectively. Saccharomyces Genome Database was used to classify the genes. Green corresponds to genes where the function is well-defined and suited for GEMs, basically enzymes involved in metabolism. Red corresponds to genes where the function is unknown, where the corresponding protein is not an enzyme or where the function is in signaling rather than metabolism. These genes should normally not be present in a GEM. Blue corresponds to genes that are putative enzymes or where the ORF is a functional enzyme in some strains but not in others. As can be seen, the automatically reconstructed model has both a larger number of unique genes and a larger proportion of enzymes compared to the published model. For iIN800 some enzymatic genes are further classified as "polymer", "lipid" or "membrane". These are parts of metabolism where an automatically generated model from KEGG would have particular drawbacks compared to a manually reconstructed model. "Polymer" corresponds mainly to genes involved in sugar polymer metabolism, which is an area that contains many unbalanced reactions in KEGG. Such reactions were excluded when the validation model was generated, so the corresponding genes could not be included. The same holds for "lipid", where the reactions contain many general metabolites. This also results in excluded reactions. "Membrane" corresponds to reactions which depend on one metabolite but in two different compartments. This compartmentalization information is absent in KEGG so the equation becomes incorrect and it is therefore excluded.
doi:10.1371/journal.pcbi.1002980.g003

reactions are colored according to the change in the corresponding flux between simulation conditions. Gene expression data can be incorporated in the map to illustrate the correlation between flux and gene expression. This will extraordinarily facilitate the comparison and interpretation of flux distributions found for different environmental conditions. The resulting map is exported as a pdf-file. Figure 4 show a close up on penicillin metabolism in the peroxisome overlaid on the full *P. chrysogenum* map (see following section for details).

Comparative genomics of template species

In order to assign metabolic functions to the genes present in the *P. chrysogenum* genome, sequence alignment analysis was performed. Three fungi from the *Aspergillus* genus (*A. oryzae*, *A. niger* and *A. nidulans*) were selected for sequence comparison based on being closely related fungi outside of the *Penicillium* genus and on having previously reconstructed GEMs (see Figure S1). Table 2 shows some genome characteristics of the *Aspergilli* in comparison with *P. chrysogenum*. Initially pairwise comparison was done by similarity searching of the protein sequences of *P. chrysogenum* against the protein sequences known to be involved in the metabolism of the three *Aspergillus* species as described in the Methods section. With a chosen threshold of the E-value, identity, and alignment length, a list of inferred metabolic functions was generated. The results are summarized in Table 2. Pairwise comparison shows that *A. oryzae* has the highest number of sequence homologues of proteins with metabolic functions with *P. chrysogenum* (915 sequences). This result suggests that metabolism of *A. oryzae* is probably closer related to *P. chrysogenum* than *A. nidulans* and *A. niger* which have less sequence homologues of 576 and 563, respectively. Upon completion of the similarity searching, the results suggest that 1143 genes in *P. chrysogenum* could be assigned as orthologous metabolic genes from the three *Aspergillus* species used for comparison. The large number of metabolic orthologues indicates that the existing GEMs for closely related species could be a sound foundation upon which to reconstruct the new model.

Reconstruction and comparative analysis of the *P. chrysogenum* metabolic network

Using the RAVEN Toolbox the metabolic network of *P. chrysogenum* metabolic network was reconstructed. The metabolic network comprises 1471 unique metabolic reactions in four subcellular compartments; extracellular, cytosolic, mitochondrial, and peroxisomal (Table 3). 1006 ORFs are associated to the reactions, 89 of which participate in one of 35 protein complexes. In parallel to the automatic reconstruction, an extensive literature study was performed. In total 440 cited articles provide experimental evidence for the majority of the reactions. All model components were extensively annotated to adhere to the MIRIAM standard for biological models [41]. The model was validated with respect to 76 important metabolic functions (the supplementary file simulations.xls is an input file to *checkTasks*, which was used to performed the validation). There are 30 reactions in the model which cannot carry flux if all uptakes are allowed, i.e. dead-end reactions (see Table S7). According to the naming conventions for metabolic networks the presented model is denoted as *iAL1006* [42]. Table 3 shows the division of model elements between the four compartments.

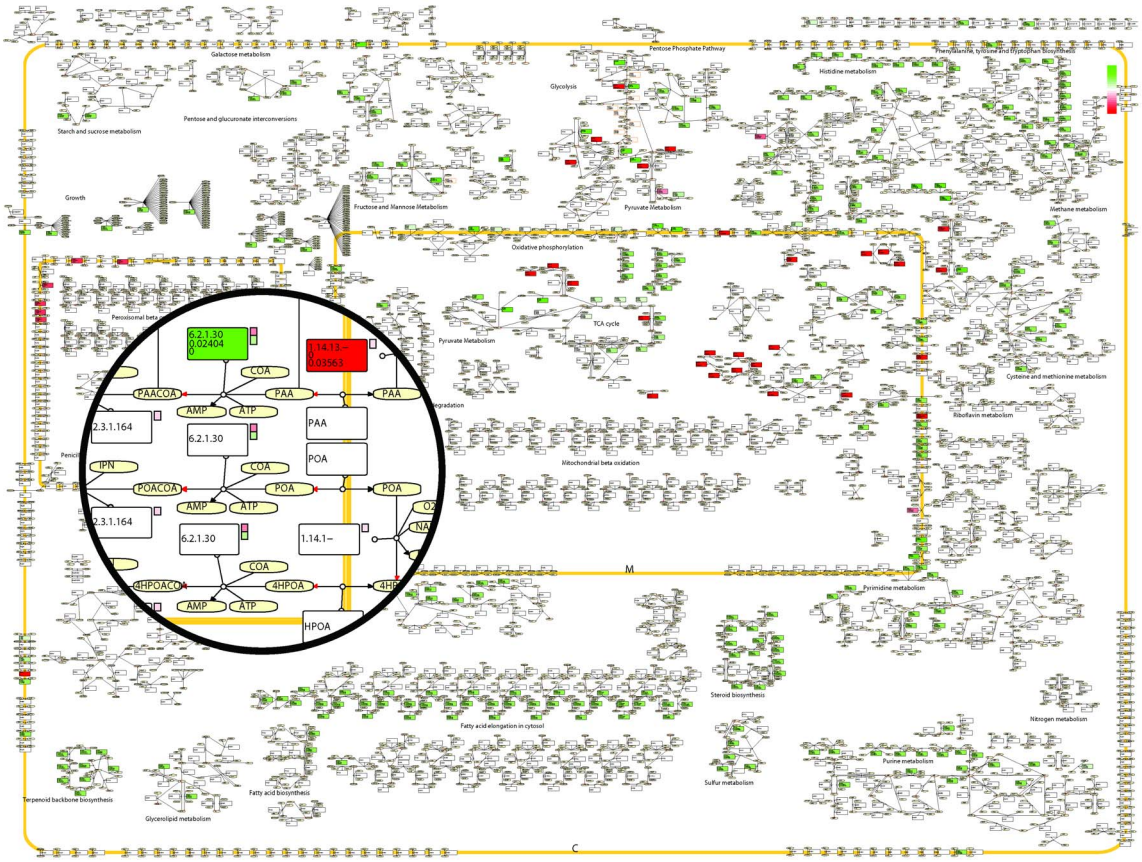


Figure 4. Example of the visualization capabilities of the RAVEN Toolbox. The figure shows a small section of the *Penicillium* metabolic map, depicting peroxisomal penicillin metabolism, superimposed on the full map. Rectangles correspond to reactions and ellipses correspond to metabolites. The broad yellow line represents the peroxisomal membrane. Reactions are colored based on the log-fold change in flux between a reference and a test case, where green represents a higher flux in the test case and red a lower flux. The positive direction of reversible reactions (defined as from left to right in the model equations) is indicated by a red arrow head. For reactions carrying flux in any of the simulated cases, the flux values are printed in the reaction box. The small squares to the right of some of the reactions correspond to the log-fold change of transcript levels of the genes associated to that reaction. The gene-reaction relation is retrieved from the model structure and not implicitly specified in the CellDesigner map.
doi:10.1371/journal.pcbi.1002980.g004

Table 2. Comparison of genome characteristics and metabolic function assignment between *P. chrysogenum* and three *Aspergillus* species.

Features	***ANi***	***AO***	***AN***	***PC***
Genome size (Mb)	30.1	37.2	34.9	32.2
Number of chromosomes/supercontigs	8	8	8	49
Number of total protein sequences	10 560	12 074	11 197	12 811
Functional assignments				
Pairwise comparison	***ANi* and *PC***	***AO* and *PC***	***AN* and *PC***	***PC***
Number of protein sequence orthologues[a]	5749	5614	5632	-
Number of metabolic orthologues based on COG[b]	1316	1471	1313	2330
Number of metabolic orthologues based on GEMs[c]	576	915	563	1143

ANi: *A. nidulans*, *AO*: *A. oryzae*, *AN*: *A. niger*, *PC*: *P. chrysogenum*.
[a]Proteins were regarded as orthologues if E-value <1e-30, identity >40%, sequence coverage >50%, and alignment length >200 amino acids.
[b]Described as being present in the functional category of metabolism based on the COG database [61].
[c]Described as being present in the corresponding previously published genome-scale metabolic model; *A. nidulans iHD666* [52]; *A. niger iMA871* [53]; *A. oryzae iWV1314* [54].
doi:10.1371/journal.pcbi.1002980.t002

Table 3. Network characteristics of the reconstructed metabolic network of *P. chrysogenum*.

ORFs	1006	
EC-numbers	627	
Metabolites[a]	1235	
Extracellular metabolites		160
Cytosolic metabolites		728
Mitochondrial metabolites		242
Peroxisomal metabolites		105
Reactions[b]	1471	
Extracellular reactions		175
Cytosolic reactions		835
Mitochondrial reactions		324
Peroxisomal reactions		137

[a]Exchange metabolites are not included.
[b]Exchange reactions are not included. Transport reactions from the cytosol to any other compartment are included in the count for that compartment, i.e. mitochondrial transport reactions are regarded as mitochondrial reactions.
doi:10.1371/journal.pcbi.1002980.t003

Figure 5 summarizes the literature support for the reactions in the model and shows a classification of the ORFs in the model based on the KEGG pathways. The full list of reactions, metabolites, and genes are supplied in Microsoft Excel format and SBML format in Dataset S1. Both these formats are compatible with the RAVEN Toolbox. To illustrate the metabolic network, and to aid in interpretation of gene expression data and simulation results, a map of the full model was drawn in CellDesigner and annotated so as to be compatible with the visualization functions in the RAVEN Toolbox. The CellDesigner file is available in Dataset S1. Even though the map is drawn for *Penicillium* metabolism it can be used as a template for generation of maps for other organisms as well. Lastly, the *P. chrysogenum* GEM has also been added to the model repository in the BioMet Toolbox, which allows for a variety of analyses and simulations to be carried out.

Comparison of fungal metabolic networks

To evaluate the similarity between the reconstructed network and the template networks, the networks were compared with respect to identical reactions and involved metabolites. Only *A. oryzae iWV1314* and *A. niger iMA871* were used in the comparison since only a small number of reaction were inferred from *A. nidulans iHD666*. Figure 6 illustrates the results. 534 reactions were unique to *iAL1006*. The large discrepancy between the models is primarily because of differences in how lipid metabolism is formulated and due to differences in localization. There are also differences in how reactions catalyzed by protein complexes are described, where one reaction for each subunit is formulated rather than lumping the reactions. The difference in metabolic capabilities between the models is therefore smaller than what is indicated by the Venn diagrams (Figure 6). The unique capabilities of the *P. chrysogenum* model are mainly in penicillin metabolism and transport (data not shown). In general, the reactions that were inferred from *A. oryzae* but not from *A. niger* are predominantly involved in co-factor synthesis and in sugar polymer metabolism. The reactions inferred from *A. niger iMA871* but not from *A. oryzae iWV1314* are mainly involved in lipid metabolism. The key statistics of the reconstructed *P. chrysogenum* network compared to those of other fungal networks is available in Table S8.

Biomass composition and parameter fitting

Growth is described as production of biomass, which in turn is regarded as drain of the macromolecules and building blocks that constitute the cellular components. The demand of each component is estimated based on published data on the biomass composition. The main components and their content within the biomass are listed in Table 4 (see also Table S9 for a detailed description). The cost of biomass production does not only include synthesis of precursors and polymerization of macromolecules, but also factors such as maintaining turgor pressure, transport costs, protein turnover, and membrane leakage. These costs are summarized as an ATP requirement for non-growth associated maintenance, m_{ATP}, and for growth associated maintenance, K_{xATP}, i.e. ATP costs not directly associated with biomass synthesis but associated with cell growth (can be maintenance of membrane potentials across an expanding cell). These parameters were determined by linear regression to glucose-limited chemostat experiments in the presence of phenoxyacetic acid (POA) [43]. The values were hereby estimated to be 4.14 mmol ATP/g DW/h for m_{ATP} and 104 mmol ATP/g DW for K_{xATP}. The growth-associated ATP cost is significantly higher than for the template organisms (64 mmol ATP/g DW/h in *A. niger iMA871*). This could possibly be an effect of the presence of phenoxyacetic acid, which is added to the fermentation medium under industrial penicillin producing conditions. It is believed that phenoxyacetic acid, being a lipophilic weak acid, acts as a proton un-coupler which would manifest itself as a high ATP maintenance cost [44]. The P/O ratio is fitted by assigning the number of cytosolic protons needed to synthesize one ATP by the F_0F_1-ATPase. This is a small simplification since the number of protons pumped across the mitochondrial membrane might also differ between organisms. This parameter was estimated to 3.75 (3.88 in *A. niger iMA871*). Figure S2 shows the agreement of model simulations with experimental fermentation data after parameter fitting.

Simulations and integrative data analysis of penicillin biosynthesis

A genome-scale metabolic model is a powerful tool that can be used for exploring the metabolic capabilities of the cell, as well as being used as a scaffold for integrative data analysis. Here we present two case studies to illustrate the use of the reconstructed *P. chrysogenum* model. The first case is a study of penicillin yields and in particular the relative importance of ATP and NADPH provision during penicillin production. In the second study we show how the model can be used to integrate fermentation data with transcriptome data using a recently published sampling algorithm to aid in the interpretation of high-throughput data [4].

Penicillin yields. Penicillin production is associated with an increased requirement of energy in the form of ATP; in the condensation of the three precursor amino acids to form the tripeptide ACV; in the reduction of sulfate; and when a side chain (the precursor molecule which is supplied to the media and which differs depending on the type of penicillin produced) is activated by ligation to coenzyme A. Penicillin production is also associated with a large requirement of NADPH; primarily needed for the reduction of sulfate but also in the biosynthesis of valine and homoserine from α-ketobutyrate. Elucidating the impact increased ATP requirements have compared to the NADPH requirements is useful when choosing among possible metabolic engineering strategies.

Different types of penicillin can be produced by changing the side chain that is supplied to the medium (e.g. supplementation of

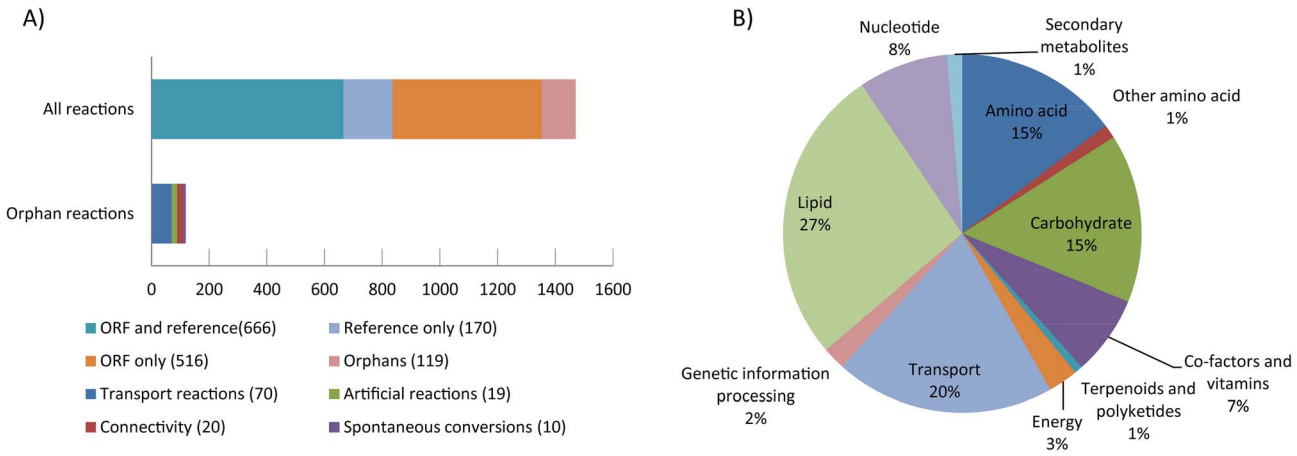


Figure 5. Evidence level for the *P. chrysogenum* metabolic network. A) Properties of the reconstructed network. The top bar shows the support for the 1471 unique reactions (not counting exchange reactions) sorted by the type of evidence. The bottom bar shows the orphan reactions; reactions inferred without supporting ORFs or literature references. B) ORF classification. The ORFs in the model are classified into broad groups based on KEGG classification.
doi:10.1371/journal.pcbi.1002980.g005

phenylacetic acid result in penicillin G and supplementation of phenoxyacetic acid result in penicillin V). However, this has no impact on the yield and it is therefore not necessary to specify the type of penicillin being produced for theoretical evaluations. The maximum theoretical yield of penicillin on glucose with sulfate as the sulfur source was calculated to be 0.42 mol penicillin/mol glucose using the reconstructed genome-scale metabolic model. This is in agreement with what has previously been published [45]. If the sulfur source is sulfite the maximal theoretical yield is found to be 0.45 mol penicillin/mol glucose and if it is hydrogen sulfide it is 0.51 mol penicillin/mol glucose. The difference between using sulfite and hydrogen sulfide is relatively large and can be attributed to the differences in NADPH cost (3 NADPH are consumed in the sulfite reduction to hydrogen sulfide). This points to the importance of NADPH availability for penicillin production. To investigate the effect of ATP an artificial reaction was included that allowed for ATP production from ADP without any energetic costs. This resulted in a yield of 0.52 mol penicillin/mol glucose, using sulfate as the sulfur source. The conclusion is that ATP availability has a relatively small effect on the yield, comparable to that of NADPH consumption in the sulfate reduction. The shadow prices (how much the penicillin production can increase if the availability of a metabolite were to increase by a small amount) were calculated to be 0.015 mol penicillin/mol ATP, 0.040 mol penicillin/mol NADPH, and 0.037 mol penicillin/mol NADH.

NADPH and NADH are similar when it comes to energy content, but have different roles in the metabolism, where NADPH serves primarily anabolic roles and NADH primarily catabolic roles. NADPH is mainly produced in the pentose phosphate pathway, which makes NADPH somewhat more energetically expensive to regenerate compared to NADH. In order to investigate the relative importance of NADH and NADPH an artificial reaction was included that allowed for production of NADPH from NADH to simulate a potential increase of the NADPH availability. Simulations were then carried out maximizing first for growth and then for penicillin production. The resulting flux through the artificial reaction was 8.5 times larger when maximizing for penicillin than when maximizing for growth. This demonstrates that the cells will have a much higher NADPH demand at high penicillin yields compared to normal

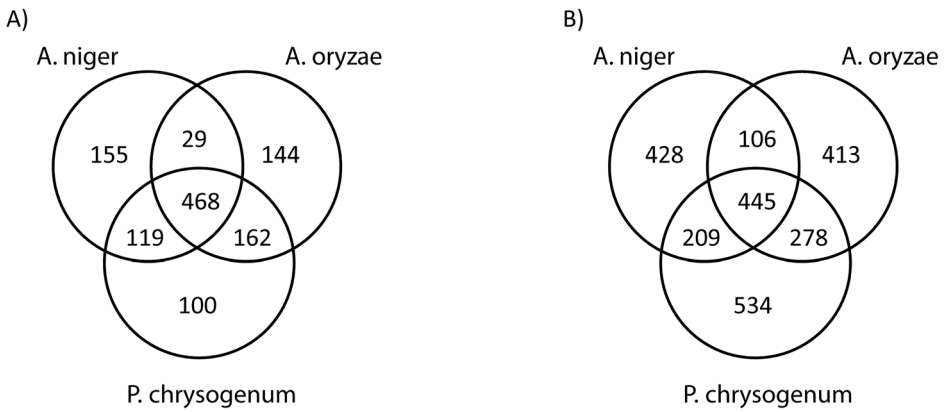


Figure 6. Venn diagrams of model statistics for the template models *A. oryzae* iWV1314 and *A. niger* iMA871 and the *P. chrysogenum* iAL1006 model. A) The number of chemically distinct metabolites shared and specific for the three models, not counting presence in multiple compartments. B) The number of unique reactions shared and specific for the three models. The overlap with *A. nidulans iHD666* is not shown here.
doi:10.1371/journal.pcbi.1002980.g006

Table 4. Biomass composition of *P. chrysogenum*.

Components	Content (g/g DW)
Protein	0.45
RNA	0.08
DNA	0.01
Lipids	0.05
Phospholipids	0.035
Sterolesters	0.010
Triacylglycerides	0.005
Carbohydrates	0.25
Cell wall	0.22
Glycogen	0.03
Soluble pool	0.08
Amino acids	0.04
Nucleotides	0.02
Total[a]	0.90

[a]8% of the dry weight is constituted by ash [43]. The remaining 2% are other soluble metabolites.

doi:10.1371/journal.pcbi.1002980.t004

growth conditions. Redirecting a higher flux through the pentose phosphate pathway and/or introducing NADH-dependent versions of NADPH-consuming enzymes could therefore be potential metabolic engineering strategies for achieving higher penicillin yields.

For the direct identification of possible metabolic engineering targets a gene deletion analysis was performed by searching for sets of gene deletions that result in an increased yield of penicillin, and which would stoichiometrically couple penicillin production to growth. This was performed using FBA, and combinations of up to three gene deletions were evaluated (MoMA was also applied and gave similar results). The only targets which could be identified were the deletion of any of the genes responsible for breakdown of phenylacetic acid (homogentisate 1,2-dioxygenase, maleylacetoacetate isomerase, or fumarylacetoacetase). Deletion of any of these genes resulted in a 21% increase in penicillin production when maximizing for growth.

Identification of transcriptionally regulated metabolic bottlenecks. The metabolism of cells is redundant in the sense that different sets of metabolic reactions can be used to generate the same net phenotype. A recently developed method aims at finding potential metabolic engineering targets by identifying genes that are differentially expressed between different cultivation conditions and where the corresponding reactions exhibit significantly changed fluxes for the same conditions [4]. Changes in expression level of such genes are then assumed to be likely to result in altered fluxes. The algorithm finds these transcriptionally regulated reactions by random sampling of the solution space, after which it compares the statistics of the sampling with the statistic of the mRNA expression. Here we applied this method to compare the high producing industrial strain DS17690, which has been developed by DSM, and the low producing reference strain Wis 54-1255 (see Methods for details) [46].

A total of 58 fluxes were found to be significantly changed between the high and low production strains ($p<0.05$) and 612 genes were differentially expressed ($p<0.005$). Out of those, 36 reactions were identified as having significantly higher flux and up-regulated genes (see Table S10), i.e. they are likely to have transcriptional regulation of their fluxes. Figure 7 shows some of the most important reactions in penicillin biosynthesis together with the responsible enzymes and the corresponding model IDs. Reactions that were identified as probably being transcriptionally controlled and up-regulated are highlighted. In addition, the Reporter Metabolites algorithm was used to identify metabolites around which significant transcriptional changes occurred [47]. These metabolites are highlighted in Figure 7 as well (see Table S11 for a full list of reporter metabolites).

As can be seen in Figure 7, a large proportion of the reactions identified as being a transcriptionally controlled are directly involved in penicillin metabolism (15 out of 38). This indicates that the capabilities of the industrial strain to produce penicillin to a large extent depend on the reactions closely related to penicillin metabolism, rather than more peripheral effects. Among these reactions are many of the reactions responsible for the synthesis of the amino acids that are precursors for ACV as well as the two penicillin producing reactions isopenicillin N synthase and ACV synthase, which is consistent with a study on the gene copy-number effect on penicillin production [48]. The phenylacetate:-CoA ligase is high ranking but the acyl-CoA:isopenicillin N acyltransferase is absent, which is consistent with measurements of high activities of this enzyme and the low flux control estimated for this enzyme [49,50]. Several of the reactions involved in sulfate reduction are present as well as the sulfate permease. It is interesting to note that none of the reactions in the pentose phosphate pathway are identified even though there is an increased demand for NADPH.

We also found that the pathway from α-ketobutyrate to succinate is identified to have both increased flux and increased gene expression. α-ketobutyrate is a by-product of cysteine production via the transsulfuration pathway, and it is used for isoleucine biosynthesis. Under normal growth conditions the demand for cysteine is less than that for isoleucine, meaning that all α-ketobutyrate is converted into isoleucine. However, during high-level penicillin production the cysteine production far exceeds the need for isoleucine, requiring an alternative route for α-ketobutyrate consumption. This route involves the decarboxylation of α-ketobutyrate to yield propionyl-CoA, which then goes into the methylcitrate pathway, eventually resulting in succinate [45]. Several of the reactions in this pathway are identified as transcriptionally controlled by the algorithm (2-methylcitrate synthase, 2-methylcitrate dehydratase, 2-methylisocitrate dehydratase, and methylisocitrate lyase). This finding strongly supports that the transsulfuration pathway is the dominating pathway for cysteine biosynthesis, even though the enzymes for the energetically more efficient direct sulfhydrylation pathway have been identified in *P. chrysogenum* [51].

Discussion

The RAVEN Toolbox, a software suite for semi-automated reconstruction and simulation of genome-scale metabolic models was developed. The RAVEN Toolbox is the first software that contains methods both for model reconstruction and for a wide variety of simulation approaches. A visualization feature for simulation results and a feature that allows the user to manipulate metabolic models and set simulation parameters via Microsoft Excel are provided in order to make the software easy to use. The RAVEN Toolbox was evaluated for its ability to reconstruct GEMs by generating a model for *S. cerevisiae*. The reconstructed model compares well with a manually reconstructed model. This demonstrates that the RAVEN Toolbox is useful for reconstruction of novel models, in particular eukaryotic models, due to its feature for automatic assignment of sub-cellular localization. We used the RAVEN Toolbox to reconstruct a GEM for *P. chrysogenum* by using three models for closely related fungal species. Extensive manual

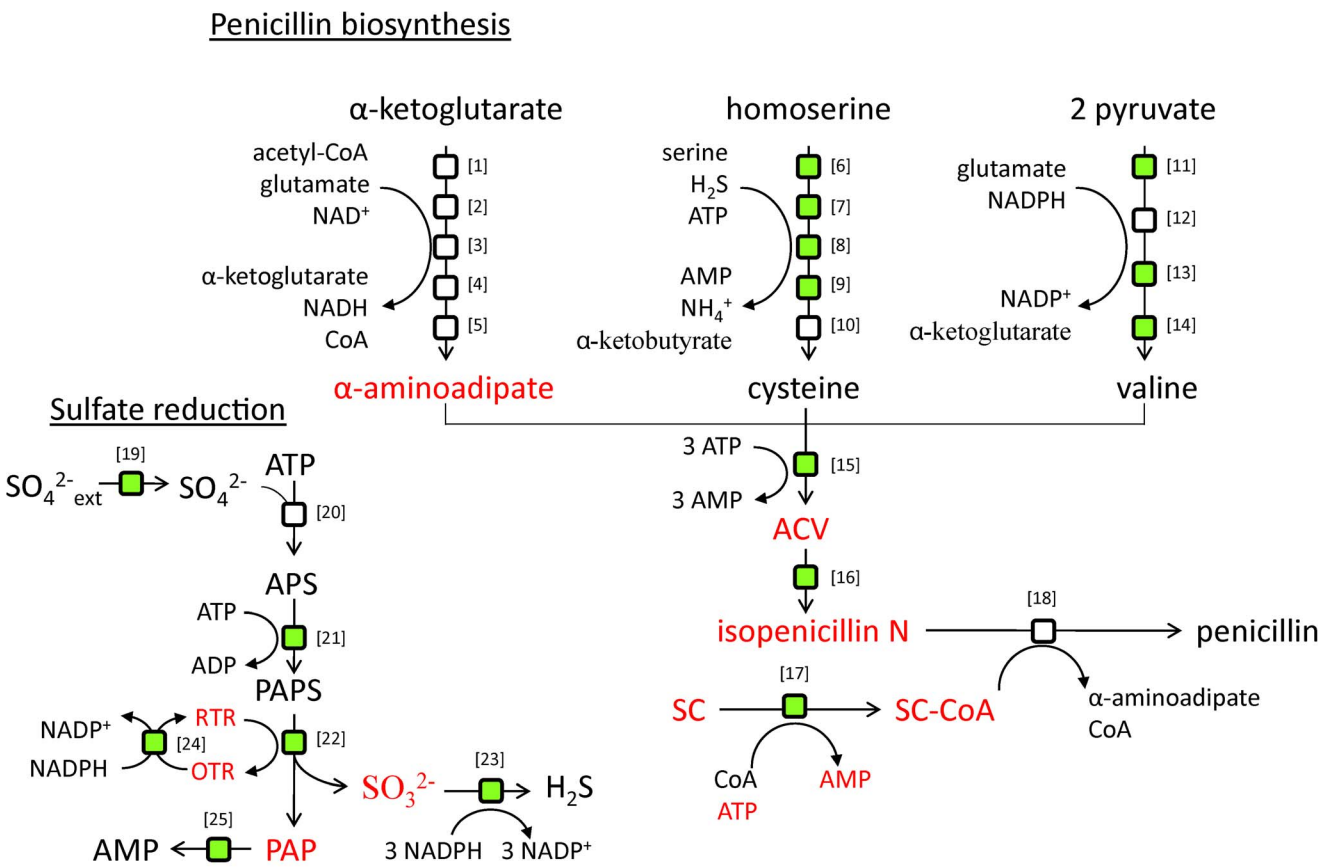


Figure 7. Integrative analysis of a high and a low producing strain. Depicts synthesis pathways of penicillin and important precursors. Green boxes correspond to reactions identified as being transcriptionally controlled and up-regulated by the algorithm (see text). Metabolites around which significant transcriptional changes occur compared to a low producing strain are colored red. SC: side chain (e. g. the precursor molecule phenylacetic acid). The biosynthesis of penicillin starts with the condensation of the three amino acids α-aminoadipate (an intermediate in the L-lysine biosynthesis pathway), L-cystein, and L-valine to form the tripeptide ACV. ACV is further converted to isopenicillin N. For the industrially relevant types of penicillin a side-chain is supplied to the media. This side-chain is activated by ligation to coenzyme A. In the last step of penicillin biosynthesis an acyl transferase exchanges the α-aminoadipate moiety of isopenicillin N with the side-chain, thereby generating penicillin and regenerating α-aminoadipate. Since L-cystein is a sulfur-containing amino acid penicillin production is also tightly associated with sulfur metabolism. The corresponding model IDs for the enzymes are indicated within parentheses. [1] homocitrate synthase (r0683); [2] homocitrate dehydrase (r0684); [3] homoaconitate hydrase (r0685); [4] homoisocitrate dehydrogenase (r0688); [5] α-aminoadipate aminotransferase (r0689); [6] homoserine transacetylase (r0600); [7] O-acetylhomoserine sulfhydrylase (r0601); [8] cystathione-β-synthase (r0632); [9] cystathione-γ-lyase (r0606); [10] acetate CoA ligase (r0025); [11] acetolactate synthase (r0465); [12] ketol-acid reductoisomerase (r0653); [13] dihydroxy acid dehydrase (r0656); [14] branched chain amino acid transferase (r0648); [15] ACV synthase (r0814); [16] isopenicillin N synthase (r0812); [17] acyl CoA ligase (side chain dependent, reaction is for phenylacetate CoA ligase) (r0747); [18] isopenicillin N N-acyltransferase (r0813); [19] sulfate permease (r1408); [20] sulfate adenyl transferase (r1151); [21] adenyl sulfate kinase (r1147); [22] phosphoadenyl sulfate reductase (r1148); [23] sulfite reductase (r1149); [24] thioredoxin reductase (r0419); [25] 3′(2′),5′-bisphosphate nucleotidase (r1150).
doi:10.1371/journal.pcbi.1002980.g007

validation of the model was performed; both to validate the reconstruction method and to ensure a high-quality model. The resulting *P. chrysogenum* model consists of 1471 reactions, 1235 metabolites and 1006 genes. 440 cited articles provide experimental evidence for the majority of the reactions. Considerable efforts were spent on standardizing and annotating the template models in order to adhere to MIRIAM standards. The standardized template models and the reconstructed *P. chrysogenum* model are available through the BioMet Toolbox. This collection of fungal models, together with the complementary method of generating metabolic networks based on the KEGG database, constitutes an excellent platform for the reconstruction of metabolic networks for other eukaryotic organisms.

Methods

The *P. chrysogenum* metabolic network was reconstructed based on a combination of automated reconstruction approaches, manual curation, and an extensive bibliomic survey. Figure 8 gives an overview of the whole reconstruction process.

Inferring reactions based on protein homology

Three GEMs for other filamentous fungi, *A. nidulans* iHD666 [52], *A. niger iMA871* [53], and *A. oryzae iWV1314* [54], were used as template models for the reconstruction of a *P. chrysogenum* model. Efforts were taken in order to standardize the template models to facilitate the automatic reconstruction. This standardization primarily involved metabolite naming, but also how to represent more complex aspects of metabolism such as polymers and lipids. As part of the standardization effort a large majority of the metabolites were assigned database identifiers and chemical structure information. This annotation step allowed for verification that all reactions were elementally balanced, which in turn led to a number of inconsistencies in the template models being corrected. The revised

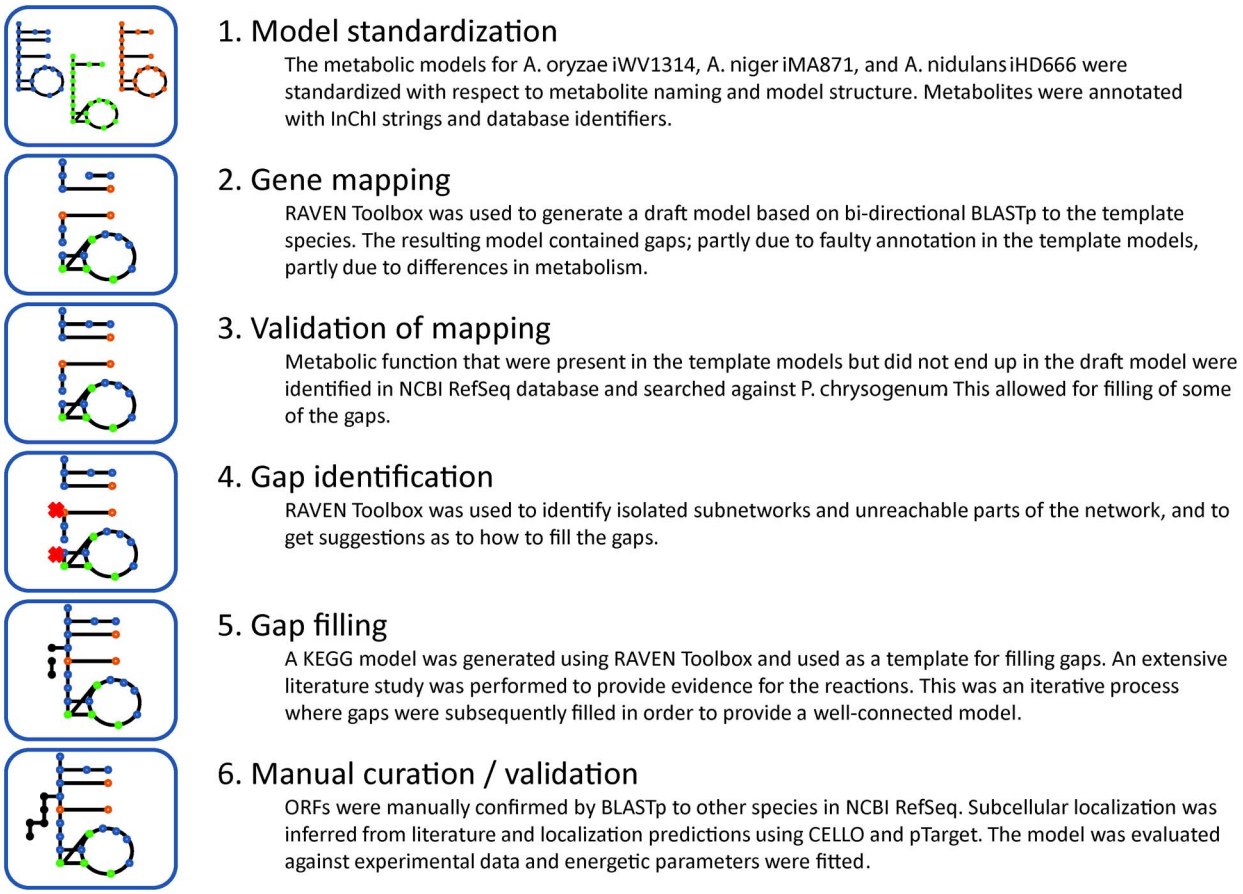


Figure 8. Overview of the iAL1006 reconstruction process.
doi:10.1371/journal.pcbi.1002980.g008

models for the three *Aspergillus* species are available as up-dates in the BioMet Toolbox (http://www.sysbio.se/BioMet) [28].

A draft GEM for *P. chrysogenum* was then constructed based on bidirectional best hits of BLASTp between the template model proteins and their orthologues in *P. chrysogenum* using the RAVEN Toolbox. Proteins were regarded as orthologues if E-value <1e-30, identity >40%, sequence coverage (>50%) and alignment length (>200 amino acids).

The protein sequences of *P. chrysogenum* Wisconsin 54-1255 (annotation, version 1) were obtained from the EMBL database (http://www.ebi.ac.uk/embl/). The protein sequences of *A. nidulans* FGSC A4 (annotation, version 4) were taken from the Broad Institute database (http://www.broadinstitute.org/annotation/genome/aspergillus_group). The protein sequences of *A. oryzae* RIB40 (annotation, version 1) were taken from the DOGAN database (http://www.bio.nite.go.jp/dogan/project/view/AO). The protein sequences of *A. niger* ATCC1015 (annotation, version 3) were taken from the JGI database (http://genome.jgi-psf.org/Aspni5/Aspni5.home.html).

Gap filling

The first draft model based on homology to template models contained gaps due to incorrect annotation in the template models and lacked reactions in parts of metabolism that were unique to *P. chrysogenum* (Figure 8). Therefore, another draft model was generated from KEGG using the RAVEN Toolbox. An E-value <1e-50 was used as cut off in the gene assignment. This model was used for filling gaps in the draft network and for suggesting metabolic pathways that were not included based on the template models. No reactions were included based solely on presence in the KEGG model, and gene assignments were only included after careful manual validation against the NCBI RefSeq database.

Gaps in the draft metabolic network were identified using the gap finding capabilities of the RAVEN Toolbox (Figure 8). The initial network was rather disconnected and biomass production from glucose was not possible. Firstly, the software was used to identify which metabolites had to be connected in order to produce biomass. This resulted in the addition of some spontaneous reactions from the template models. The second step was to ensure that as many metabolites as possible could be produced. When non-connected metabolites were identified, the KEGG model was queried for candidate reactions which could connect that metabolite. A targeted literature search was then conducted to find evidence for the presence of such a reaction. In only few cases this resulted in the addition of transport reactions based solely on connectivity issues (<2%, see Figure 5). The final step was to use the RAVEN Toolbox to identify reactions that could not carry a flux during growth on any of the available carbon sources. There are, however, sets of reactions that are included in the model even though they cannot currently carry a flux. One such example is the synthesis and loading of tRNA, which is included to allow for possible future extension of the model to cover protein synthesis.

Compartmentalization and transport

The model has four compartments: extracellular space, cytosol, mitochondrion, and peroxisome. Reactions with unknown

localization, or where the real localization is not represented by one of the compartments in the model, were assigned to the cytosol. Enzymes reported to be present in the cell wall were assigned to the extracellular space, and those present in the mitochondrial membrane were mainly assigned to the mitochondria. The peroxisome was included primarily because of its role in penicillin metabolism and β-oxidation of fatty acids. Transport reactions between compartments were inferred mainly from the template fungal models and backed up with literature evidence. However, there were situations where a transport reaction had to be included in order to have a functional network, even when no literature evidence could be found. For enzymes reported to have isoenzymes in several compartments, the ORF assignments to each compartment were based on localization predictions from CELLO [36] and pTarget [55].

Simulations

Unless otherwise specified, simulations were carried out using unlimited uptake of oxygen, phosphate, sulfate, NH3, thiamin and pimelate. The carbon source was glucose. Excretion of all exchange metabolites was allowed. Biosynthesis of L-cysteine was only allowed through the transsulfuration pathway. Since the energy content of NADH and NADPH is similar it is possible that cycles convert one into the other. These cycles normally take the form of reversible reactions that can utilize either NADH or NADPH. Since such futile cycles make it difficult to study NADPH/NADH metabolism using GEMs, and since they are probably not active in the cell, they were identified and blocked in analogy to what has been done in previous models [10]. The way by which they were identified was minimizing/maximizing for the flux through an artificial reaction NADH+NADP(+)< = > NAD(+)+NADPH while not allowing for uptake of any carbon source. Reactions were then identified and deleted until this reaction could no longer carry a flux. The following reactions were deleted: D-mannitol:NAD+ 2-oxidoreductase (r0181); ethanol:-NADP+ oxidoreductase (r0019); ethanol:NAD+ oxidoreductase (r0020); (S)-3-hydroxybutanoyl-CoA:NADP+ oxidoreductase (r0081). None of these reactions can be expected to be active during the studied conditions. Deletion of D-mannitol:NAD+ 2-oxidoreductase does not inactivate the NADPH regenerating mannitol cycle, which has a role in NADPH regeneration in many fungal species [56].

The penicillin yields were calculated by setting the glucose uptake rate to 1.0 mmol/gDW/h and maximizing for penicillin production. Free ATP was simulated by including an artificial reaction in the form ADP+Pi = >ATP+H2O.

Integrative analysis

A random sampling algorithm was applied in order to identify transcriptionally regulated metabolic bottlenecks [4]. Flux data and gene expression levels for aerobic, glucose-limited chemostat fermentation of DS17690 and Wis 54-1255 were used as input to the algorithm [57]. The exchange fluxes were fitted to the reported values using a quadratic fitting. 5000 sampling iterations were performed for each of the two strains. The expression data set of the study were retrieved from GEO database (GSE9825) as CEL format then normalized together with PLIER workflow (http://media.affymetrix.com/support/technical/technotes/plier_technote.pdf). Two way ANOVA were employed to evaluate the differentially expressed genes with respect to the strain (DS17690 and Wis 54-1255) with multiple correction following [58]. The Reporter algorithm [47] was employed to integrate the transcriptome data with the reconstructed GEM to identify key metabolites in the network.

Supporting Information

Dataset S1 The iAL1006 genome-scale model of *P. chrysogenum* in SBML and Excel formats, together with a metabolic map for visualization and a task list for model validation.
(ZIP)

Figure S1 Proteome comparison of genomes in *Fungi*. ALR (the ratio of alignment length to query sequence length): >0.50, identity: >0.40. The red shades refer to protein homology that can found within a genome (paralog). The green shades refer to protein homology that can found between two genomes (ortholog).
(PDF)

Figure S2 Agreement of model simulations with experimental fermentation data. Data from glucose-limited chemostat with defined medium containing glucose, inorganic salts and phenoxyacetate.
(PDF)

Table S1 Reactions which were excluded from the general KEGG model after running *removeBadRxns*. 72 reactions were unbalanced, general or polymer reactions and were therefore correctly removed. 7 reactions were correct in KEGG, but were removed because they lacked metabolite composition (it is a setting in *removeBadRxns* whether it is allowed to remove such reactions).
(PDF)

Table S2 Comparison of an automatically reconstructed model for *S. cerevisiae* to a published model of the same organism (iIN800) in terms of included genes. The table shows the genes that are unique to either the automatically reconstructed or the manually reconstructed model, and a classification of the genes into groups that reflect how well suited they are for being included in a GEM. Genes labeled as "enzymatic" should be included, while all other groups should probably be excluded. For iIN800 some enzymatic genes are further classified as "polymer", "lipid" or "membrane". These are parts of metabolism where an automatically generated model from KEGG would have particular drawbacks compared to a manually reconstructed model. "Polymer" corresponds mainly to genes involved in sugar polymer metabolism, which is an area that contains many unbalanced reactions in KEGG. Such reactions were excluded in the validation, so the corresponding genes could not be included. The same is true for "lipid", where the reactions contain many general metabolites, which also results in excluded reactions. "Membrane" corresponds to reactions which depend on any one metabolite in different compartments. This compartmentalization information is absent in KEGG so such a reaction would read, for example, A+B = >A+C. "A" here might mean "A(cytosolic)" and "A(mitochondrial)", but since that information is missing, the equation becomes incorrect and it is therefore excluded. "Signaling" corresponds to proteins which are primarily involved in signaling, even though they might have an enzymatic capability.
(PDF)

Table S3 Metabolites which could be synthesized in the automatically reconstructed *S. cerevisiae* model from minimal media (glucose, phosphate, sulfate, NH3, oxygen, 4-aminobenzoate, riboflavin, thiamine, biotin, folate, and nicotinate). Uptake of the carriers carnitine and acyl-carrier protein was allowed for modeling purposes (many compounds are bound to them and therefore net synthesis of these compounds is not possible without them).
(PDF)

Table S4 New metabolites which could be synthesized in the automatically reconstructed *S. cerevisiae* model from minimal media after gap-filling. These metabolites were all present in the model before the addition of new reactions.
(PDF)

Table S5 Reactions which were added to the automatically reconstructed *S. cerevisiae* model by *fillGaps*. Out of the 45 added reactions 17 has evidence to support that they should be included in the model, 9 has inconclusive of missing evidence, and 19 should not have been included in the model.
(PDF)

Table S6 Genes where their corresponding reactions were localized to the mitochondria after running *predictLocalization* (transport cost = 0.1). The color indicates whether the gene product is mitochondrial in SGD, where green means that it does, yellow that it is unclear, and red that it does not.
(PDF)

Table S7 Reactions which cannot carry flux even when all uptake reactions are unconstrained.
(PDF)

Table S8 Comparison of metabolic models.
(PDF)

Table S9 Biomass composition calculations for *P. chrysogenum*.
(PDF)

Table S10 Reactions with significantly higher flux in DS17690 compared to Wis 54-1255 where the corresponding genes are also up-regulated. Ranked by significance ($p<0.05$).
(PDF)

Table S11 Reporter metabolites when comparing the DS17690 and Wis 54-1255 strains. Ranked by significance. Top 40 best scoring metabolites are shown.
(PDF)

Acknowledgments

We thank Timo Hardiman and Rudolf Mitterbauer at Sandoz for valuable feedback throughout the project, and Sinisa Bratulic for helping with the generation of the CellDesigner map.

Author Contributions

Performed the reconstruction: RA LL. Bioinformatics analyses: WV IN. Developed the sub-cellular localization algorithm: RA SS. Wrote the software and performed the simulations: RA. Supervised the work: IN JN. Conceived and designed the experiments: RA JN. Performed the experiments: RA LL SS. Analyzed the data: RA WV IN. Wrote the paper: RA.

References

1. Liu L, Agren R, Bordel S, Nielsen J (2010) Use of genome-scale metabolic models for understanding microbial physiology. FEBS Lett 584: 2556–2564.
2. Price ND, Papin JA, Schilling CH, Palsson BO (2003) Genome-scale microbial in silico models: the constraints-based approach. Trends Biotechnol 21: 162–169.
3. Price ND, Schellenberger J, Palsson BO (2004) Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. Biophys J 87: 2172–2186.
4. Bordel S, Agren R, Nielsen J (2010) Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. PLoS Comput Biol 6: e1000859.
5. Schuster S, Dandekar T, Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends Biotechnol 17: 53–60.
6. Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, et al. (2005) In silico design and adaptive evolution of Escherichia coli for production of lactic acid. Biotechnol Bioeng 91: 643–648.
7. Alper H, Jin YS, Moxley JF, Stephanopoulos G (2005) Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. Metab Eng 7: 155–164.
8. Cakir T, Patil KR, Onsan Z, Ulgen KO, Kirdar B, et al. (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. Mol Syst Biol 2: 50.
9. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7: 129–143.
10. Forster J, Famili I, Fu P, Palsson BO, Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res 13: 244–253.
11. Edwards JS, Palsson BO (2000) The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A 97: 5528–5533.
12. Otero JM, Nielsen J (2009) Industrial Systems Biology. Biotechnology and Bioengineering 105: 439–460.
13. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, et al. (2000) InterPro–an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16: 1145–1150.
14. Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28: 405–420.
15. Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols 5: 93–121.
16. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29–34.
17. Arakawa K, Yamada Y, Shinoda K, Nakayama Y, Tomita M (2006) GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. BMC Bioinformatics 7: 168.
18. Notebaart RA, van Enckevort FH, Francke C, Siezen RJ, Teusink B (2006) Accelerating the reconstruction of genome-scale metabolic networks. BMC Bioinformatics 7: 296.
19. Pinney JW, Shirley MW, McConkey GA, Westhead DR (2005) metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. Nucleic Acids Res 33: 1399–1409.
20. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28: 977–982.
21. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, et al. (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nature Protocols 2: 727–738.
22. Pitt JI (1979) The genus Penicillium and its teleomorphic states Eupenicillium and Talaromyces. New York: Academic Press. 634 p.
23. Elander RP (2003) Industrial production of beta-lactam antibiotics. Appl Microbiol Biotechnol 61: 385–392.
24. Thykaer J, Nielsen J (2003) Metabolic engineering of beta-lactam production. Metab Eng 5: 56–69.
25. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19: 524–531.
26. Stein SE, Heller SR, Tchekhovski D (2003) An open standard for chemical structure representation: The IUPAC Chemical Identifier. Proceedings of the 2003 International Chemical Information Conference: 131–143
27. Herrgard MJ, Swainston N, Dobson P, Dunn WB, Arga KY, et al. (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol 26: 1155–1160.
28. Cvijovic M, Olivares-Hernandez R, Agren R, Dahr N, Vongsangnak W, et al. (2010) BioMet Toolbox: genome-wide analysis of metabolism. Nucleic Acids Res 38 Suppl: W144–149.
29. Schomburg I, Chang A, Schomburg D (2002) BRENDA, enzyme data and metabolic information. Nucleic Acids Res 30: 47–49.
30. Fleming RM, Thiele I (2011) von Bertalanffy 1.0: a COBRA toolbox extension to thermodynamically constrain metabolic models. Bioinformatics 27: 142–143.
31. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
32. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
33. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755–763.
34. Tarjan RE (1983) Space-Efficient Implementations of Graph Search Methods. Acm Transactions on Mathematical Software 9: 326–339.
35. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, et al. (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res 35: W585–587.
36. Yu CS, Chen YC, Lu CH, Hwang JK (2006) Prediction of protein subcellular localization. Proteins 64: 643–651.
37. Nookaew I, Jewett MC, Meechai A, Thammarongtham C, Laoteng K, et al. (2008) The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism. BMC Syst Biol 2: 71.
38. McDonald PN (2001) Two-hybrid systems. Methods and protocols. Introduction. Methods Mol Biol 177: v-viii.
39. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504.
40. Funahashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, et al. (2008) CellDesigner 3.5: A versatile modeling tool for biochemical networks. Proceedings of the Ieee 96: 1254–1265.

41. Le Novere N, Finney A, Hucka M, Bhalla US, Campagne F, et al. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). Nat Biotechnol 23: 1509–1515.
42. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol 4: R54.
43. Nielsen J (1997) Physiological engineering aspects of Penicillium chrysogenum. Singapore: World Scientific. 269 p.
44. Henriksen CM, Nielsen J, Villadsen J (1998) Modelling of the protonophoric uncoupling by phenoxyacetic acid of the plasma membrane potential of Penicillium chrysogenum. Biotechnol Bioeng 60: 761–767.
45. Jorgensen H, Nielsen J, Villadsen J, Mollgaard H (1995) Metabolic flux distributions in Penicillium chrysogenum during fed-batch cultivations. Biotechnol Bioeng 46: 117–131.
46. Harris DM, Diderich JA, van der Krogt ZA, Luttik MA, Raamsdonk LM, et al. (2006) Enzymic analysis of NADPH metabolism in beta-lactam-producing Penicillium chrysogenum: presence of a mitochondrial NADPH dehydrogenase. Metab Eng 8: 91–101.
47. Patil KR, Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci U S A 102: 2685–2689.
48. Theilgaard H, van Den Berg M, Mulder C, Bovenberg R, Nielsen J (2001) Quantitative analysis of Penicillium chrysogenum Wis54-1255 transformants overexpressing the penicillin biosynthetic genes. Biotechnol Bioeng 72: 379–388.
49. Jorgensen H, Nielsen J, Villadsen J, Mollgaard H (1995) Analysis of penicillin V biosynthesis during fed-batch cultivations with a high-yielding strain of Penicillium chrysogenum. Appl Microbiol Biotechnol 43: 123–130.
50. Nielsen J, Jorgensen HS (1995) Metabolic control analysis of the penicillin biosynthetic pathway in a high-yielding strain of Penicillium chrysogenum. Biotechnol Prog 11: 299–305.
51. Ostergaard S, Theilgaard HBA, Nielsen J (1998) Identification and purification of O-acetyl-L-serine sulphhydrylase in Penicillium chrysogenum. Applied Microbiology and Biotechnology 50: 663–668.
52. David H, Hofmann G, Oliveira AP, Jarmer H, Nielsen J (2006) Metabolic network driven analysis of genome-wide transcription data from Aspergillus nidulans. Genome Biol 7: R108.
53. Andersen MR, Nielsen ML, Nielsen J (2008) Metabolic model integration of the bibliome, genome, metabolome and reactome of Aspergillus niger. Mol Syst Biol 4: 178.
54. Vongsangnak W, Olsen P, Hansen K, Krogsgaard S, Nielsen J (2008) Improved annotation through genome-scale metabolic modeling of Aspergillus oryzae. BMC Genomics 9: 245.
55. Guda C, Subramaniam S (2005) pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes. Bioinformatics 21: 3963–3969.
56. Hult K, Veide A, Gatenbeck S (1980) The Distribution of the Nadph-Regenerating Mannitol Cycle among Fungal Species. Archives of Microbiology 128: 253–255.
57. van den Berg MA, Albang R, Albermann K, Badger JH, Daran JM, et al. (2008) Genome sequencing and analysis of the filamentous fungus Penicillium chrysogenum. Nat Biotechnol 26: 1161–1168.
58. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125: 279–284.
59. Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, et al. (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS Comput Biol 8: e1002518.
60. Sun J, Zeng AP (2004) IdentiCS–identification of coding sequence and in silico reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence. BMC Bioinformatics 5: 112.
61. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.

Paper IV

Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT

Agren, R.[*], Bordel, S.[*], Mardinoglu, A., Pornputtapong, N., Nookaew, I. and Nielsen, J.

PLoS Comput Biol (2012), 8(5), p. e1002518

[*]Authors contributed equally

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT

Rasmus Agren[☯], Sergio Bordel[☯], Adil Mardinoglu[¶], Natapol Pornputtapong[¶], Intawat Nookaew, Jens Nielsen*

Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

Abstract

Development of high throughput analytical methods has given physicians the potential access to extensive and patient-specific data sets, such as gene sequences, gene expression profiles or metabolite footprints. This opens for a new approach in health care, which is both personalized and based on system-level analysis. Genome-scale metabolic networks provide a mechanistic description of the relationships between different genes, which is valuable for the analysis and interpretation of large experimental data-sets. Here we describe the generation of genome-scale active metabolic networks for 69 different cell types and 16 cancer types using the INIT (Integrative Network Inference for Tissues) algorithm. The INIT algorithm uses cell type specific information about protein abundances contained in the Human Proteome Atlas as the main source of evidence. The generated models constitute the first step towards establishing a Human Metabolic Atlas, which will be a comprehensive description (accessible online) of the metabolism of different human cell types, and will allow for tissue-level and organism-level simulations in order to achieve a better understanding of complex diseases. A comparative analysis between the active metabolic networks of cancer types and healthy cell types allowed for identification of cancer-specific metabolic features that constitute generic potential drug targets for cancer treatment.

Citation: Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, et al. (2012) Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT. PLoS Comput Biol 8(5): e1002518. doi:10.1371/journal.pcbi.1002518

Editor: Costas D. Maranas, The Pennsylvania State University, United States of America

Received January 10, 2012; **Accepted** March 30, 2012; **Published** May 17, 2012

Funding: We acknowledge the Knut and Alice Wallenberg Foundation and the Chalmers Foundation for financial contribution to this work. The Lars Hierta Memorial Foundation supported the work on the comparative analysis of cancer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nielsenj@chalmers.se

☯ These authors contributed equally to this work.

¶ These authors also contributed equally to this work.

Introduction

Abnormal metabolic states are at the origin of many diseases such as diabetes, hypertension, hearth diseases and cancer, which can be seen in many aspects as a metabolic disease. Cancer and coronary diseases are the two main causes of death in the developed countries. It is expected that by 2030 close to 200 million persons (33% of the total population) will be obese in the EU alone, and many of these will have one or more of the following co-morbidities: diabetes, hypertension, heart disease and increased risk of cancer, and the direct (medical treatment) and indirect (inability to work) costs are estimated to amount to more than €100 billion per year [1,2]. The molecular mechanisms involved in these kinds of diseases are complex and in many cases different underlying molecular causes lead to the same disease phenotypes. A good understanding of human metabolism in different human cell types, whole tissues, and the interactions between them is therefore a necessary step towards efficient diagnosis and treatment of these diseases. Metabolism is, however, complex and involves a very large number of individual reactions that are highly interconnected through the sharing of common metabolites [3]. Understanding the function of metabolism therefore requires analysis of the complete metabolic network, and this is best done through the use of so-called genome-scale metabolic models (GEMs) [4,5,6].

There are three generic genome-scale human metabolic networks currently available, namely Recon1 [7], the Edinburgh Human Metabolic Network (EHMN) [8] and HumanCyc [9]. These reconstructions, however, are not tissue specific, which prevents their applicability to the study of particular human cell types or diseases. Tissue specific transcription profiles were used to generate tissue specific models for 10 different human tissues [10], which are subsets of Recon1, but these networks were not sufficiently flexible to explore the metabolic states of the tissues under various genetic and physiological conditions [11]. The same group later proposed a different algorithm that combines transcriptomic and proteomic data to generate a more flexible liver specific metabolic model [11], also using Recon1 as a template. Besides the mentioned automatically generated models, an extensive effort led to the publication of a manually reconstructed and annotated liver specific metabolic model referred as HepatoNet1 [12]. Models have also been developed for kidney [13], brain [14], erythrocytes [15] and alveolar macrophages [16]. Computational methods used to construct cell

Author Summary

Many serious diseases have a strong metabolic component. The abnormal metabolic states of diseased cells could therefore be targets for treatment. However, metabolism is a highly complex and interconnected system in which thousands of metabolic reactions occur simultaneously in any given cell type. In order to understand how metabolism of a diseased cell differs from its healthy counterpart we must therefore study the system as a whole. We have developed an algorithm that integrates several types of data in order to generate active metabolic networks; catalogues of the metabolic reactions that are likely to be active in a given cell type. We applied this algorithm to data for 69 healthy cell types and 16 cancer cell types. These metabolic networks can form the basis for simulation of metabolic interactions between organs or as scaffolds for interpretation of high-throughput data. We used these networks to perform an analysis between cancer and healthy cell types in order to identify cancer specific metabolic features that constitute potential drug targets. Several of the resulting targets were already known and used clinically, but we also found high-ranking reactions and metabolites which have not yet been investigated as drug targets.

type specific metabolic models aim to integrate the evidence about the presence or absence of metabolic enzymes in a particular cell type, while at the same time maintaining a well-connected network (e.g. metabolites consumed in one reaction should be able to be produced in another reaction or to be taken up from the cell environment). Transcriptome data are often noisy and differences in mRNA expression are not absolute but relative to a reference condition, and in most cases do not correlate well with enzyme levels [17]. In the frame of the Human Protein Atlas (HPA) [18,19,20] cell type specific high quality proteomic data are being generated based on specific antibodies, and this represents an essential source for protein evidence in different human cell types.

Here we present a pipeline for automatic identification of expressed cell type specific genome-scale metabolic networks (Figure 1). A key element of the pipeline is the INIT (Integrative Network Inference for Tissues) algorithm (Figure 2), which relies on the HPA as the main evidence source for assessing the presence or absence of metabolic enzymes in each of the human cell types that are present in the HPA. Tissue specific gene expression [21] was used as an extra source of evidence in INIT. Metabolomic data from the Human Metabolome Database (HMDB) [22] are also used as constraints in such a way that if a metabolite has been found in a particular tissue the resulting network should be able to produce this metabolite from simple precursors. More details can be found in the description of the method.

The output of our analysis is a cell type specific metabolic network for each of the cell types profiled in the HPA. As we are using HPA and gene expression data our networks do not represent the complete metabolic network that may be expressed in each cell type, but solely the part of the metabolic network that is expressed and hence the part of the network that is likely to be active. In order to provide a reliable and up to date genome-scale model template for our tissue/cell type specific metabolic networks, we first constructed the Human Metabolic Reaction (HMR) database containing the elements of previously published generic genome-scale human metabolic models [7,8,9] as well as the KEGG [23] database. This HMR database, which is publically available at www.metabolicatlas.com, will be periodically updated as new reactions are added to KEGG or MetaCyc or expression profiles for more proteins become available in HPA or other databases.

In order to evaluate the capability of our pipeline to generate reliable tissue specific metabolic networks, the metabolic model generated for hepatocytes was compared to HepatoNet1 [12], which is an extensively manually curated and annotated model of high quality. The availability of active metabolic networks corresponding to a broad set of healthy human cell types and cancers allows for a comparative analysis between cancer and healthy cell types in order to identify cancer specific metabolic features that constitute potential drug targets.

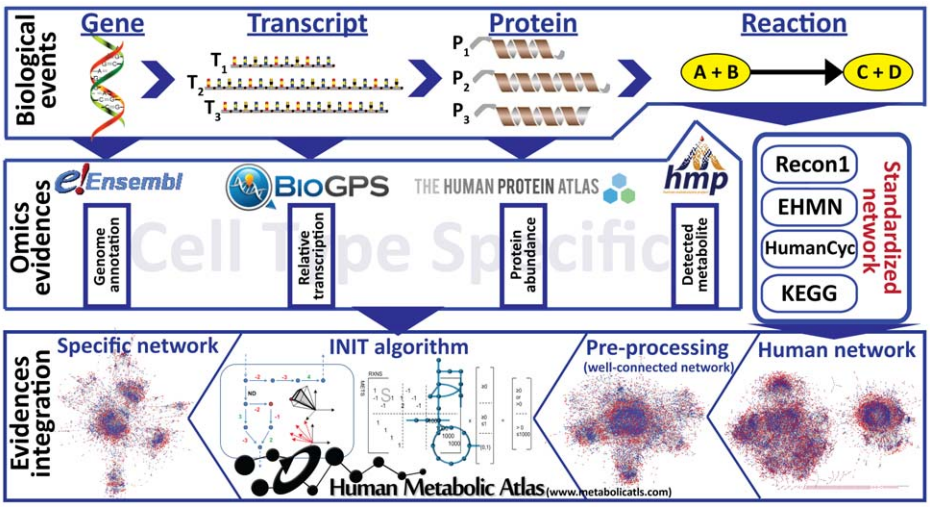


Figure 1. General pipeline used in the reconstruction of cell specific genome-scale metabolic networks. Biological information at the genome, transcriptome, proteome and metabolome levels contained in publicly available databases and generic human GEMs (Recon1, EHMN, HumanCyc) is integrated to form a generic human metabolic network, which is processed in order to obtain the connected *iHuman1512* network. Subsequently, the cell type specific evidence is used to generate cell type specific subnetworks using the INIT algorithm.
doi:10.1371/journal.pcbi.1002518.g001

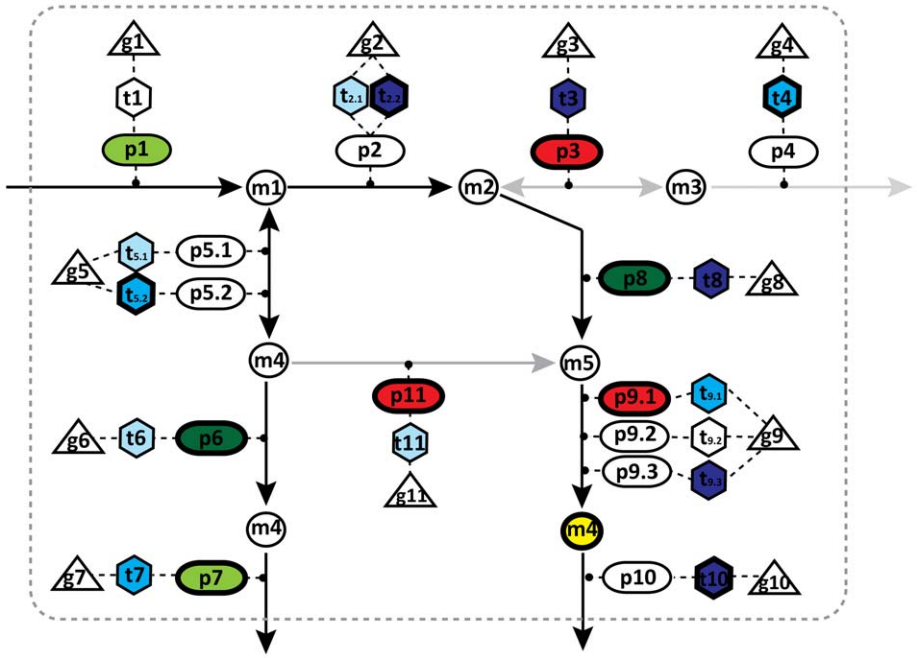


Figure 2. Illustration of the principles of the INIT algorithm. The hierarchical structure of GEMs is characterized by its gene-transcript-protein-reaction (GTPR) associations. In GEMs, each metabolic reaction is associated to one or more enzymes, which in turn are associated to transcripts and genes. Depending on the evidence for presence/absence of a given enzyme/gene in a cell type, a score can be calculated for the reaction(s) catalyzed by that enzyme. The HPA evidence scores are illustrated as red, light, medium and dark green representing negative, weak, moderate and strong evidence, respectively. The transcriptome evidence scores (GeneX), which are illustrated as red, light, medium, and dark blue representing low, medium and high expression, respectively. No evidence is present as white object. For some metabolites (yellow filled circle), metabolomic data are available to prove that they are present in the considered cell type. The aim of the algorithm is to find a sub-network in which the involved genes/proteins have strong evidence supporting their presence in the cell type under consideration. This is done by maximizing the sum of evidence scores. All the included reactions should be able to carry a flux and all the metabolites observed experimentally should be synthesized from precursors that the cell is known to take up. The bold lines represent the resulting network after optimization.
doi:10.1371/journal.pcbi.1002518.g002

Results/Discussion

Database construction

The existing methods for the inference of tissue specific active metabolic networks have only used Recon1 as a scaffold. In order to integrate other sources of information we constructed the Human Metabolic Reaction database (HMR), containing the two existing genome-scale metabolic models, Recon1 and EHMN, as well as incorporating information from HumanCyc and KEGG.

The HMR database has a hierarchical structure in which the genes are at the top and are linked to information about their tissue specific expression profiles reported by Su et al [21] via BioGPS [24]. Each gene is linked to its different splicing variants and those to their corresponding proteins. Each protein is linked to its tissue specific abundances in the HPA database [18] and to the reactions they catalyze. The reactions are linked to metabolites that themselves are linked to their tissue specific information collected from the HMDB [22]. The HMR database will be regularly updated with new reactions contained in future genome-scale human metabolic reconstructions, as well as with new evidence included in future versions of the HPA, HMDB and newly published specific transcriptome data. Details regarding the construction and curation of the HMR database are available in the Methods section. The INIT algorithm requires a connected template human metabolic model as input, and this template model was generated from HMR. The template model contains 4,137 metabolites (3,397 unique) and 5,535 reactions (4,144 unique), which are associated to 1,512 metabolic genes. This template model is referred to as *iHuman1512*.

Generation of 69 tissue specific and 16 cancer type specific genome-scale active metabolic networks

Using the INIT algorithm (see supplementary material for a detailed description), genome-scale active networks for 69 different cell types and 16 cancers were automatically generated. The resulting active metabolic networks are provided in SBML [25] format and are available at www.metabolicatlas.com.

The tissue specific models generated were compared with the BRENDA [26] collection of detected enzymes in various tissues. A hypergeometric test was carried out using the R statistical software. The reported p-values are the probabilities of obtaining an overlap higher than the observed with a random set of metabolic genes of the same size as the corresponding BRENDA entry. As it is shown in Table S1, all the comparisons between the models generated by our algorithm and BRENDA showed overlaps with p-values lower than 5e-4. Our computational liver model (*iHepatocyte1154*) shows a p-value of 1e-200, which is similar to the value obtained by comparing the manually reconstructed HepatoNet1 to BRENDA. 55% of the genes in *iHepatocyte1154* are also in BRENDA, while only 43% of the genes in HepatoNet1 are in BRENDA. The comparatively high p-values are for tissues for which there are very few annotated enzymes in BRENDA.

In order to validate the output of our algorithm, our automatically generated hepatocyte model was compared with HepatoNet1 [12], a manually curated and functional model of hepatocyte metabolism. The comparison was carried out at the gene level to avoid ambiguous decisions about reaction similarity. The overlap between the lists of genes included in each of the

models is showed in Figure 3. Our hepatocyte model (*iHepatocyte1154*) contains 1,154 genes, of which 452 are also included in HepatoNet1 and 702 are absent. The evidence for the expression and translation of the 702 absent genes is as good as the evidence for the 452 genes that are in both networks, and we are therefore confident that the presence of most of the 702 extra genes has been correctly inferred by our algorithm. The HepatoNet1 network contains 261 genes not included in *iHepatocyte1154*, of which 156 were absent from our initial connected human network. Our algorithm could therefore not have assigned these genes to the hepatocyte sub-network and their existence reveals just a limitation of the data that were used as an input and not a limitation of our algorithm. 80 of these genes were not in HMR (see Table S2), but closer examination revealed that the majority (62 genes) of these genes corresponded to reactions that were actually present in HMR, but with different or absent gene associations. 13 out of the 18 remaining genes encode for transporters to the sinusoidal space; a type of blood vessels in the liver and therefore not a part of hepatocytes. The other 76 genes that were absent from *iHuman1512*, and their corresponding reactions, were removed because of being unbalanced, unconnected or otherwise problematic (see Table S3). 105 genes included in HepatoNet1, and present in *iHuman1512*, were not assigned by our algorithm to the hepatocyte-specific network. These genes correspond to 237 reactions, 132 of them still exist in *iHepatocyte1154* associated to different isoenzymes. The experimental evidences for the presence of these 105 genes (see Table S4) in the hepatocytes is mostly weak or negative, even slightly worse than the evidence for the 253 genes that were both rejected by our algorithm and absent in HepatoNet1, and we are therefore confident that these 105 genes were correctly rejected. This shows the importance of using cell type specific data when reconstructing GEMs, as enzyme isoforms can be differentially expressed in different cell types. Based on the above we can conclude that the

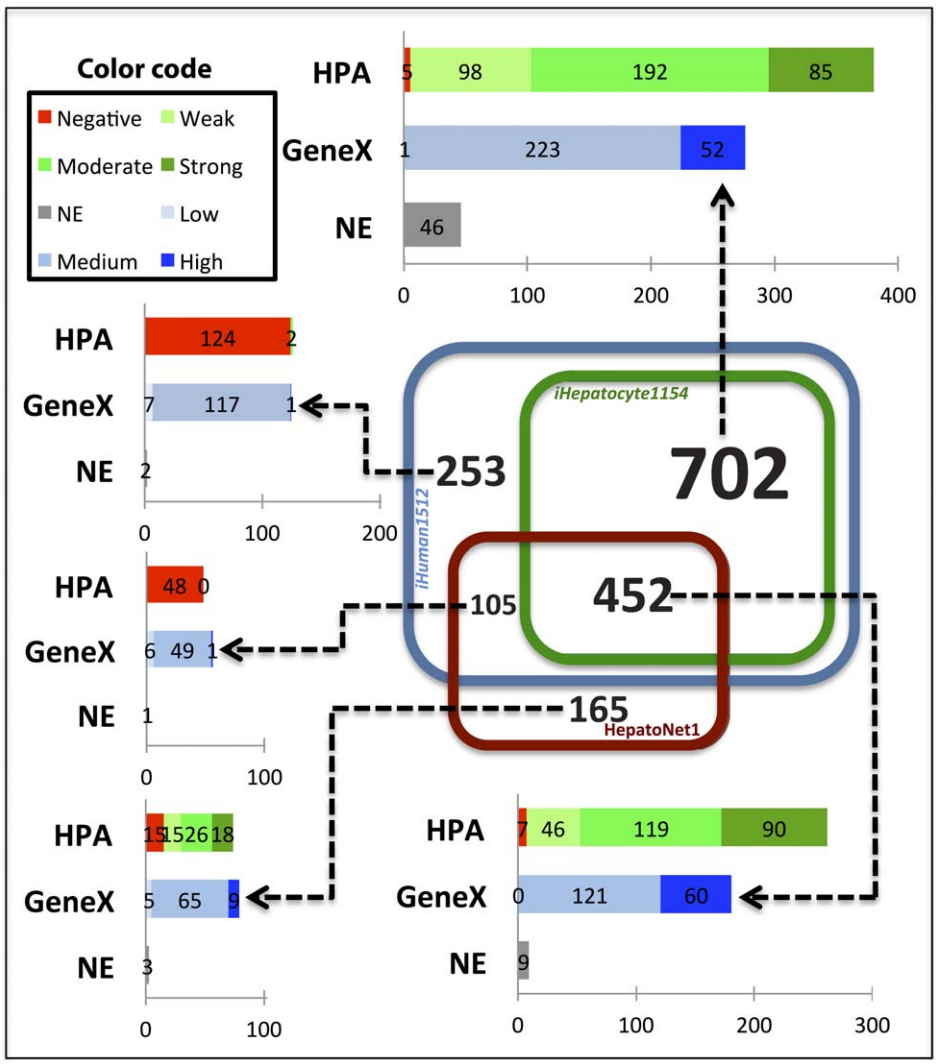


Figure 3. Gene content comparison between our hepatocyte model and HepatoNet1. The Venn diagram shows the overlap in terms of included genes between three models. The blue, green and red squares represent *iHuman1512*, our hepatocyte model *iHepatocyte1154* and HepatoNet1, respectively. The distribution of evidence scores of each section of the Venn diagram is plotted. The HPA evidence scores are illustrated as red, light, medium and dark green represent negative, weak moderate and strong expression, respectively. The transcriptome evidence scores (GeneX) are illustrated as red, light, medium and dark blue representing low, medium and high expression, respectively. No evidence (NE) is illustrated as grey color.
doi:10.1371/journal.pcbi.1002518.g003

mismatches between our hepatocyte-specific metabolic network and HepatoNet1 are accompanied by experimental evidence in favour of the choices made by our algorithm.

Finally we clustered the 69 plus 16 metabolic networks according to their similarity in terms of shared metabolic genes using unsupervised hierarchical clustering with average linkage and multiscale bootstrap resampling [27] (10,000 repetitions) implemented in the R statistical software (see Figure S1). The clustering shows, as it could be expected, local grouping of closely related cell types on the basis of cell anatomy (e.g. spleen in red pulp and white pulp cluster together). Interestingly, the cancers are separated into three different clusters, one containing liver, colorectal, breast and endometrial cancer, another minor cluster including cervical and head cancer and a third one containing the remaining ten cancers. It is also of interest to note that only 189 reactions (4.1% of the total number of reactions) are unique to a single cell or cancer type, while there is a larger core-set of 501 reactions (11.0% of total number of reactions) that are in common to all cells. Figure S2 shows the enrichment of some important metabolic pathways in the models.

Identification of cancer specific metabolic features

Since the Warburg effect was observed at the beginning of the 20th century, it is known that cancer cells show characteristic metabolic features that make them different from healthy cells [28]. This supposed metabolic similarity between cancer cells justified the development of a generic cancer genome-scale metabolic model which was used to identify potential drug targets against cancer proliferation [29]. Here we have inferred active metabolic networks for 16 different cancer types, which can be compared with the 24 healthy cell types that they come from (there are several healthy cell types for some of the tissues associated to the cancers) in order to identify metabolic features that are characteristic of cancer. A hypergeometric test was used to identify genes and reactions that tend to be present in most of the cancer specific active metabolic networks and absent in most of the original healthy cell types (see Tables S5 and S6). The p-values obtained from the hypergeometric test were used to identify Reporter Metabolites [30] that are significantly more involved in the metabolism of cancer cells (see Table S7). The sets of genes, reactions and metabolites showing enrichment in the cancer active metabolic networks with p-values lower than 1e-4 are listed in the supplementary material. These lists of genes, reactions and metabolites are cancer specific features that are likely to be playing a specific role in proliferation of cancer cells and could be potential drug targets. Our comparative analysis between two sets of active metabolic networks can be seen as a high throughput hypothesis generation method. These hypotheses are not based on mere correlations between cancer and the presence of a particular protein, but being based on the underlying metabolic network structure, and hereby our analysis provides a mechanistic interpretation about the possible role of each identified feature on the proliferation of cancer.

One of the most significant results from the Reporter Metabolites analysis is a much more pronounced metabolism of polyamines (PAs) such as spermidine, spermine, and putrescine in cancer cells. PAs play a variety of roles, of which several are related to oxidative stress prevention and suppression of necrosis [31]. PAs have long been known to be of particular importance for rapidly proliferating cells, and as such its transport and synthesis have been thoroughly investigated as anti-cancer drug targets [32]. Inhibition of single enzymes in the PA synthesis pathway has proved disappointing, due to extensive regulation of the system and use of exogenous PAs by the cancer cells. Second generation drugs instead work by targeting the transport system, by structural homology to the PAs themselves, or by linking other aneoplastic drugs to the PAs [33].

Another high-ranking target is the isoprenoid biosynthesis pathway, in particular the intermediate geranylgeranyl diphosphate. This metabolite has been shown to promote oncogenic events due to its role in prenylation of important cancer proteins such as Ras and Rho GTPases [34]. Several drugs have therefore been developed to target the prenylation process [35] or the biosynthesis of geranylgeranyl diphosphate [36].

A third prominent group among the Reporter Metabolites is prostaglandins and leukotrienes together with the intermediate HPETE. These autocrine compounds are synthesized from arachidonic acid and are elevated in connection with inflammation. They have been shown to aid in cancer progression by promoting metastasis and by influencing the immune system [37]. Of particular interest is prostaglandin E2, where both the synthesis and degradation have been investigated as promising targets for drug development [38].

The fact that so many of the identified targets correspond to well known and used drug targets, indicates that the method is able to generate biologically relevant hypotheses. Of particular interest are therefore the Reporter Metabolites that are currently not targeted in cancer treatment. Among the top-scoring Reporter Metabolites we identified biliverdin and bilirubin (Figure 4). Biliverdin reductase and the reactions catalyzed by this enzyme also appear among the genes and reaction most enriched in the cancer networks. Biliverdin reductase is known to be a major physiologic cytoprotectant against oxidative stress [39]. Cancer cells are known to be exposed to high oxidative stress resulting from the hydrogen peroxide generated during the oxidation of polyamines and other products of amino acid breakdown taking place in the peroxisome. Bilirubin is oxidized to biliverdin by hydrogen peroxide and subsequently reduced back to bilirubin by biliverdin reductase. This mechanism has been proven to be a major relief system for oxidative stress and could be considered a potential target against cancer proliferation. One of the hydrogen peroxide generating reactions taking place in the peroxisomes is the transformation of aminoacetone, which is an intermediate in the degradation of glycine, into methylglyoxal. Another source of methylglyoxal in cancer cells is from gluconeogenesis [40]. Methylglyoxal is known to be a toxic compound [41] that has been proven to induce apoptosis in some cancer cell lines [42]. Methylglyoxal also appeared among our top scoring reporter metabolites and both the gene coding for lactoylglutathione lyase (an enzyme that transforms methylglyoxal and glutathione into lactoylglutathione) and its associated reactions appear among the most enriched genes and reactions in the cancer active metabolic networks. Lactoylglutathione is further transformed into glutathione and lactic acid by the enzyme lactoylglutathione hydrolase (which also shows a significant enrichment in cancer metabolic networks with a p-value of 2e-3). Lactic acid is a well known metabolite produced by cancer cells. The mentioned two enzymes seem to be playing a relevant role in relieving the toxicity generated by methylglyoxal and could be potential drug targets against cancer proliferation. Targeting these enzymes would have the same effect on cancer cells as using methylglyoxal as a drug, but the advantage is that there would be no toxicity effects of methylglyoxal on healthy tissues.

Conclusions and perspectives

We here present a method that is able to integrate different sources of biological evidence to generate high quality cell type specific metabolic networks. We used this method to generate

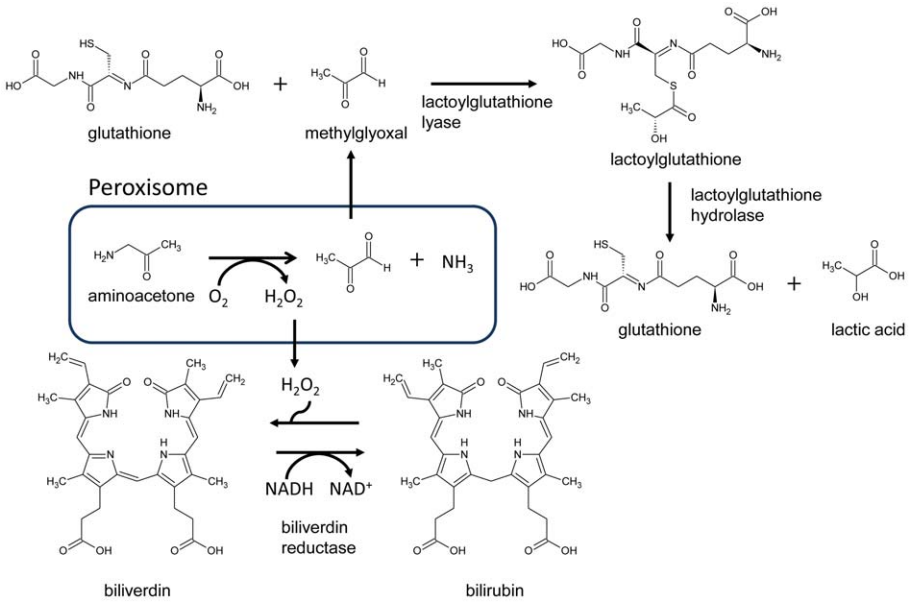


Figure 4. Example of a metabolic sub-network that was identified as being significantly more present in cancer tissues compared to their corresponding healthy tissues. Aminoacetone, which is a toxic by-product of amino acid catabolism, is converted to toxic methylglyoxal in a reaction that also result in hydrogen peroxide. The toxicity of methylglyoxal is relieved by two reaction steps involving ligation to glutathione and resulting in lactic acid. The generated hydrogen peroxide is taken care of by the enzyme biliverdin reductase. This is an example of how network-based analysis can lead to a more mechanistic interpretation of data.
doi:10.1371/journal.pcbi.1002518.g004

genome-scale metabolic networks for 69 different human cell types and 16 cancer types, and this is the first step towards the establishment of a Human Metabolic Atlas, which may become a central portal for further advancing human metabolic models with the capability of performing tissue-level and organism-level metabolic simulations, allowing for a better understanding of complex diseases. The Human Metabolic Atlas and will be made publicly accessible for the medical and scientific community and may hereby become a valuable resource in the development of personalized medicine based on system-level analysis. An example of system-level analysis is the identification of cancer specific metabolic features that we have performed by comparing the networks generated using the INIT algorithm.

Methods

Database construction

In order to have an unambiguous characterization of metabolites and reactions, KEGG and InChI identifiers were used for standardization. Metabolites lacking identifiers to external databases were left out of the HMR database together with their corresponding reactions. The metabolite identifiers were used to infer if two reactions coming from different sources were the same. Each reaction was assigned to one or several of the eight compartments included in the HMR database: nucleus, cytosol, endoplasmatic reticulum, Golgi apparatus, peroxisomes, lysosomes, mitochondria and extracellular. In cases where the subcellular localization was absent from the template models it was inferred from immunohistochemical staining in the HPA. For enzymes that were not in the HPA, Swiss-Prot and GO were used to infer localization (see Table S8 for database versions). After removing the compounds that lack identifiers, the database contained 9,922 reactions, 2,366 genes, and 9,581 metabolites for the eight different compartments (3,547 unique metabolites and 6,319 unique reactions when compartmentalization is not considered). There are 338 of these metabolites which, even if they have KEGG identifiers, are generic compounds such as "Lipid" or "2-oxoacid". Such compounds can lead to reactions that are elementally unbalanced. These problematic metabolites were removed, together with the 418 reactions in which they were involved after a detailed manual curation process. 38 reactions with wrong or unbalanced stoichiometries were also substituted by balanced versions during the curation process. In order to avoid problems associated with proton balancing, which arise from undefined protonation states of many metabolites, free exchange of protons was allowed in the models. Finally, all reactions unable to carry flux under any circumstance were removed. The reason for removing these unconnected reactions was that the algorithm requires a connected model as input. After this filtering, our template model contains 4,137 metabolites (3,397 unique) and 5,535 reactions (4,144 unique), which are associated to 1,512 metabolic genes (based on the Ensemble gene catalogue). This template model is referred to as *iHuman1512*. The numbers of genes maintained in each of the above mentioned steps are listed in Table S9. The discrepancy between the large number of reactions and the relatively small number of genes, which is also seen in previously published metabolic networks, is due to the fact that many reactions are included in the template networks based on literature studies or for connectivity reasons. In addition, some enzymes catalyze a large number of reactions and some enzymes catalyze reactions in several compartments. A comparison between *iHuman1512* and some published human metabolic networks is available in Table S10.

Algorithm for the generation of tissue-specific models

Several algorithms aiming to obtain a tissue or condition-specific active set of metabolic reactions from a generic model have been previously developed. The first of these algorithms was the Gene Inactivity Moderated by Metabolism and Expression (GIMME) algorithm [43], which uses mRNA expression data as

input. Two other algorithms were developed with the specific aim of generating human tissue specific metabolic networks. The first of those [10],_ENREF_10 developed by Shlomi and co-workers, used transcriptomic data as its sole input. The second one [11] was developed by the same authors in order to obtain a functional model for human hepatocytes and is able to integrate also metabolomic and proteomic data.

The INIT (Integrative Network Inference for Tissues) algorithm is formulated as a mixed integer-linear problem (MILP) and is specially tailored to use the evidence from the HPA as input. The problem is formulated so that all reactions in the resulting model are able to carry flux. The stoichiometric matrix S contains the stoichiometric coefficients for each internal metabolite in each reaction. By multiplying the stoichiometric matrix by the vector of reaction rates we obtain a vector of net accumulation or consumption rates for each internal metabolite. Instead of imposing the steady state condition for all the internal metabolites, as it is usually done, we allow for a small positive net accumulation rate. The net productions of metabolites will be given positive weights in the optimization. The reason for this choice is that we prefer to have a network able to synthesize molecules such as NADH or NADPH, rather than only being able to use them as cofactors. If a metabolite is present in a cell type (according to the HMDB) a positive net production of this metabolite will be imposed to the network in order to assure that all the reactions necessary for its synthesis are included in the tissue specific model.

Up to this date there is not a human biomass equation available in the literature (for example Recon1 incorporates a mouse biomass equation). On the other hand, human cells (with the exception of cancer cells), in contrast to microorganisms, do not tend to proliferate or do so slowly in comparison with the rest of their metabolic functions. This makes the biomass equation less relevant, unless the aim is to model cancer proliferation. Also human cells secrete into the blood a much broader spectrum of compounds than microbial cells secrete into their environment (which are mainly fermentation products). We therefore chose to generate networks allowing for secretion (or accumulation) of all their metabolites. If we had used the stricter steady state constraint, many reactions would have been removed from the models just because they were leading to dead end metabolites. These end metabolites could in fact be added to biomass or just be secreted into the blood stream, therefore we have aimed for a more flexible approach by allowing secretion (or net accumulation) of metabolites.

The MILP used in INIT can be specified as:

$$\begin{aligned}
&\max\left(\sum_{i\in R} w_i y_i + \sum_{j\in M} x_j\right)\\
&S\vec{v}=\vec{b}\\
&|v_i|\le 1000 y_i\\
&|v_i|+1000(1-y_i)\ge\varepsilon\\
&v_i\ge 0, i\in \textit{irreversible rxns}\\
&b_j\le 1000 x_i\\
&b_j+1000(1-x_i)\ge\varepsilon\\
&b_j\ge 0\\
&x_j=1, j\in \textit{present}\\
&y_i,x_j\in\{0,1\}
\end{aligned} \qquad (1)$$

The parameter ε is an arbitrarily small positive number. The weights of the binary variables corresponding to the reactions account for the evidence of their presence or absence. When the corresponding enzyme has been characterized in the HPA we have used values of w_i of 20, 15, 10 and −8 for high, medium, low and absent proteins respectively. These scores are arbitrary and have been chosen to quantify the evidence colour codes that appear in the HPA. We have tested the sensitivity of the algorithm to the variation in these weights by perturbing them by 20% up and down and the impact on the output of the algorithm resulted only in small changes of the resulting networks. If the evidence comes from gene expression levels which were retrieved from BioGPS [24] and the publicly dataset "Human Body Index – Transcriptional Profiling" (GSE7307), we have used weights calculated as follows:

$$w_{i,j}=5\log\left(\frac{Signal_{i,j}}{Average_i}\right) \qquad (2)$$

The signal of gene i in tissue j is divided by the average signal across all the tissues. If the signal in a particular tissue is higher than its average across all the tissues the weight will be positive, if it is lower it will have a negative weight.

For the reactions that are related to several genes or proteins the highest evidence score is used. If no gene is associated to a particular reaction, or there is no proteomic or transcriptomic evidence, a weight of −2 is used in order to avoid adding unnecessary reactions without evidence and keep the network as parsimonious as possible. If a reaction linked to several genes is added to the final tissue specific network, only the genes showing a positive evidence score are kept in the tissue specific reaction-gene association.

The MILP problem was solved using MOSEK (www.mosek.com) and its Matlab interface.

Supporting Information

Figure S1 Clustering of 69 predicted cell type specific genome-scale metabolic models for normal tissues together with 16 for cancer tissues. A dendrogram generated by unsupervised hierarchical clustering of the models based on predicted gene presence and absence is shown.
(PDF)

Figure S2 The relative pathway enrichment profiles, based on KEGG pathways, for each of the models. Blue corresponds to underrepresentation and red to overrepresentation. Note that it is the number of enzymes present for each pathway that underlie the comparison, not the abundances of the proteins.
(PDF)

Table S1 Evaluation of the models by comparison to curated tissue-specific enzymes. For each model, the set of genes is compared to the set of genes annotated as existing in the corresponding tissue in BRENDA. The p-values are derived from hypergeometric distribution.
(PDF)

Table S2 Investigation of the 80 genes that were present in HepatoNet1 but missing in HMR. The 80 missing genes were associated with 746 reactions in HepatoNet1, of which 597 metabolic and transport reactions were related to the sinusoidal space compartment. 117 metabolic reactions existed in HMR with different gene or no gene association and 32 (5 unique) reactions were altogether absent in HMR. KEGG reaction identifiers or

Transporter Classification database identifiers (TCDB) are provided for the missing reactions.
(PDF)

Table S3 Investigation of the 76 genes that were removed during the pre-processing steps. 76 genes which were present in both HepatoNet1 and the HMR database were removed in order to get a fully connected input network for the INIT algorithm. This table summarizes which genes were removed during each of the pre-processing steps (see Table S2 for details).
(PDF)

Table S4 Investigation of the 105 genes which are present in HepatoNet1 but missing in *iHepatocyte1154* due to the INIT algorithm. The 105 missing genes were associated with 182 reactions in HepatoNet1, of which 5 metabolic reactions are related to the sinusoidal space compartment. 108 metabolic reactions existed in *iHepatocyte1154* with different gene or no gene association and 69 (60 unique) reactions are absent in *iHepatocyte1154* due to INIT algorithm. KEGG reaction identifiers are provided for the missing associated reactions to the genes.
(PDF)

Table S5 List of reactions that were significantly more present in cancer tissues compared to their corresponding normal tissues (p-value<10e-4).
(PDF)

Table S6 List of genes for which their corresponding reactions were significantly more present in cancer tissues compared to their corresponding normal tissues (p-value<10e-4).
(PDF)

Table S7 List of Reporter Metabolites (p-value<10e-4).
(PDF)

Table S8 Versions of the databases used in the creation of the Human Metabolic Reaction database (HMR).
(PDF)

Table S9 Number of the genes after each pre-processing step during the generation of *iHuman1512*. Since the INIT algorithm provides a connected and functional model, reactions that are unconnected in the template model can never be included. In order to separate between reactions that were excluded due to connectivity reasons and those that were excluded due to negative evidence, a number of preprocessing steps were performed on the data in the HMR database. In the first step reactions that contain very generic metabolites such as "lipid" or "alcohol" were removed. Reactions that were elementally unbalanced were fixed or removed. Simulations were performed to ensure that the network could not gain carbon, energy or redox power in an unbalanced manner. In the second step reactions where directionality information was lacking were removed. In the third step reactions where one or more of the substrates could not be synthesized through some other reaction (unconnected reactions) were removed. Finally, in the fourth step reactions that couldn't carry flux when the model had access to all exchange metabolites (as defined in the EHMN) were removed. Consequently, *iHuman1512*, a connected human network with 5,535 reactions, associated with 1512 protein coding genes, was generated. The number of genes associated to the remaining reactions after each pre-processing steps are presented below.
(PDF)

Table S10 Comparison between iHuman1512 and some other published human metabolic networks.
(PDF)

Acknowledgments

We acknowledge valuable discussions with Prof. Bernhard Palsson and Prof. Eytan Ruppin.

Author Contributions

Wrote the paper: RA SB IN JN. Conceived the project: JN. Supervised the work: SB JN. Developed the INIT algorithm: RA. Performed the comparative analysis between cancers and healthy cell types: SB. Performed the comparison of generated models with previously published models: AM. Developed the HMR database and generated the website for the Human Metabolic Atlas: NP. Performed analysis of the generated cell type models: IN. Edited the paper: RA SB AM NP IN JN.

References

1. Caveney E, Caveney BJ, Somaratne R, Turner JR, Gourgiotis L (2011) Pharmaceutical interventions for obesity: a public health perspective. Diabetes Obes Metab 13: 490–497.
2. Rokholm B, Baker JL, Sorensen TI (2010) The levelling off of the obesity epidemic since the year 1999–a review of evidence and perspectives. Obes Rev 11: 835–846.
3. Nielsen J (2009) Systems biology of lipid metabolism: from yeast to human. FEBS Lett 583: 3905–3913.
4. Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5: 93–121.
5. Feist AM, Palsson BO (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol 26: 659–667.
6. Osterlund T, Nookaew I, Nielsen J (2011) Fifteen years of large scale metabolic modeling of yeast: Developments and impacts. Biotechnol Adv E-pub ahead of print.
7. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A 104: 1777–1782.
8. Hao T, Ma HW, Zhao XM, Goryanin I (2010) Compartmentalization of the Edinburgh Human Metabolic Network. BMC Bioinformatics 11: 393.
9. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, et al. (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 6: R2.
10. Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. Nat Biotechnol 26: 1003–1010.
11. Jerby L, Shlomi T, Ruppin E (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Mol Syst Biol 6: 401.
12. Gille C, Bolling C, Hoppe A, Bulik S, Hoffmann S, et al. (2010) HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. Mol Syst Biol 6: 411.
13. Chang RL, Xie L, Bourne PE, Palsson BO (2010) Drug off-target effects predicted using structural analysis in the context of a metabolic network model. PLoS Comput Biol 6: e1000938.
14. Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, et al. (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain. Nat Biotechnol 28: 1279–1285.
15. Bordbar A, Jamshidi N, Palsson BO (2011) iAB-RBC-283: A proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and patho-physiological states. BMC Syst Biol 5: 110.
16. Bordbar A, Lewis NE, Schellenberger J, Palsson BO, Jamshidi N (2010) Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. Mol Syst Biol 6: 422.
17. Olivares-Hernandez R, Bordel S, Nielsen J (2011) Codon usage variability determines the correlation between proteome and transcriptome fold changes. BMC Syst Biol 5: 33.
18. Berglund L, Bjorling E, Oksvold P, Fagerberg L, Asplund A, et al. (2008) A genecentric Human Protein Atlas for expression profiles based on antibodies. Mol Cell Proteomics 7: 2019–2027.
19. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, et al. (2010) Towards a knowledge-based Human Protein Atlas. Nat Biotechnol 28: 1248–1250.
20. Uhlen M, Bjorling E, Agaton C, Szigyarto CA, Amini B, et al. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. Mol Cell Proteomics 4: 1920–1932.

21. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062–6067.
22. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, et al. (2007) HMDB: the Human Metabolome Database. Nucleic Acids Res 35: D521–526.
23. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36: D480–D484.
24. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, et al. (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 10: R130.
25. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19: 524–531.
26. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, et al. (2010) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Res 39: D507–13.
27. Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22: 1540–1542.
28. Koppenol WH, Bounds PL, Dang CV (2011) Otto Warburg's contributions to current concepts of cancer metabolism. Nat Rev Cancer 11: 325–337.
29. Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, et al. (2011) Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol 7: 501.
30. Patil KR, Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. Proc Natl Acad Sci U S A 102: 2685–2689.
31. Eisenberg T, Knauer H, Schauer A, Buttner S, Ruckenstuhl C, et al. (2009) Induction of autophagy by spermidine promotes longevity. Nat Cell Biol 11: 1305–1314.
32. Seiler N (2003) Thirty years of polyamine-related approaches to cancer therapy. Retrospect and prospect. Part 1. Selective enzyme inhibitors. Curr Drug Targets 4: 537–564.
33. Seiler N (2003) Thirty years of polyamine-related approaches to cancer therapy. Retrospect and prospect. Part 2. Structural analogues and derivatives. Curr Drug Targets 4: 565–585.
34. Sebti SM, Hamilton AD (2000) Farnesyltransferase and geranylgeranyltransferase I inhibitors and cancer therapy: lessons from mechanism and bench-to-bedside translational studies. Oncogene 19: 6584–6593.
35. Philips MR, Cox AD (2007) Geranylgeranyltransferase I as a target for anti-cancer drugs. J Clin Invest 117: 1223–1225.
36. Dudakovic A, Tong H, Hohl RJ (2011) Geranylgeranyl diphosphate depletion inhibits breast cancer cell migration. Invest New Drugs 29: 912–920.
37. Schneider C, Pozzi A (2011) Cyclooxygenases and lipoxygenases in cancer. Cancer Metastasis Rev 30: 277–294.
38. Eruslanov E, Kaliberov S, Daurkin I, Kaliberova L, Buchsbaum D, et al. (2009) Altered expression of 15-hydroxyprostaglandin dehydrogenase in tumor-infiltrated CD11b myeloid cells: a mechanism for immune evasion in cancer. J Immunol 182: 7548–7557.
39. Baranano DE, Rao M, Ferris CD, Snyder SH (2002) Biliverdin reductase: a major physiologic cytoprotectant. Proc Natl Acad Sci U S A 99: 16093–16098.
40. Titov VN, Dmitriev LF, Krylin VA (2010) [Methylglyoxal–test for biological dysfunctions of homeostasis and endoecology, low cytosolic glucose level, and gluconeogenesis from fatty acids]. Ter Arkh 82: 71–77.
41. Kalapos MP (1994) Methylglyoxal toxicity in mammals. Toxicol Lett 73: 3–24.
42. Kang Y, Edwards LG, Thornalley PJ (1996) Effect of methylglyoxal on human leukaemia 60 cell growth: modification of DNA G1 growth arrest and induction of apoptosis. Leuk Res 20: 397–405.
43. Becker SA, Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol 4: e1000082.

Paper V

Integration of clinical data with a genome-scale metabolic model of the human adipocyte

Mardinoglu, A., **Agren, R**., Kampf, C., Asplund, A., Nookaew, I., Jacobson, P., Walley, A.J., Froguel, P., Carlsson, L. M., Uhlen, M., Nielsen, J.

Mol Syst Biol (2013), 9, p. 649

Molecular Systems Biology 9; Article number 649; doi:10.1038/msb.2013.5
Citation: ***Molecular Systems Biology*** **9:649**
www.molecularsystemsbiology.com

molecular systems biology

Integration of clinical data with a genome-scale metabolic model of the human adipocyte

Adil Mardinoglu[1], Rasmus Agren[1], Caroline Kampf[2], Anna Asplund[2], Intawat Nookaew[1], Peter Jacobson[3], Andrew J Walley[4], Philippe Froguel[4,5], Lena M Carlsson[3], Mathias Uhlen[6] and Jens Nielsen[1,*]

[1] Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden, [2] Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden, [3] Department of Molecular and Clinical Medicine and Center for Cardiovascular and Metabolic Research, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden, [4] Department of Genomics of Common Diseases, School of Public Health, Imperial College London, Hammersmith Hospital, London, UK, [5] Unité Mixte de Recherche 8199, Centre National de Recherche Scientifique (CNRS) and Pasteur Institute, Lille, France and [6] Department of Proteomics, School of Biotechnology, AlbaNova University Center, Royal Institute of Technology (KTH), Stockholm, Sweden
* Corresponding author. Department of Chemical and Biological Engineering, Chalmers University of Technology, Kemivägen 10, Gothenburg 41128, Sweden. Tel.: +46 3 1772 3804; Fax: +46 3 1772 3801; E-mail: nielsenj@chalmers.se

Received 24.10.12; accepted 11.2.13

We evaluated the presence/absence of proteins encoded by 14 077 genes in adipocytes obtained from different tissue samples using immunohistochemistry. By combining this with previously published adipocyte-specific proteome data, we identified proteins associated with 7340 genes in human adipocytes. This information was used to reconstruct a comprehensive and functional genome-scale metabolic model of adipocyte metabolism. The resulting metabolic model, *iAdipocytes1809*, enables mechanistic insights into adipocyte metabolism on a genome-wide level, and can serve as a scaffold for integration of omics data to understand the genotype–phenotype relationship in obese subjects. By integrating human transcriptome and fluxome data, we found an increase in the metabolic activity around androsterone, ganglioside GM2 and degradation products of heparan sulfate and keratan sulfate, and a decrease in mitochondrial metabolic activities in obese subjects compared with lean subjects. Our study hereby shows a path to identify new therapeutic targets for treating obesity through combination of high throughput patient data and metabolic modeling.
Molecular Systems Biology **9:** 649; published online 19 March 2013; doi:10.1038/msb.2013.5
Subject Categories: cellular metabolism; molecular biology of disease
Keywords: adipocyte; flux balance analysis; genome-scale metabolic model; obesity; proteome

Introduction

White adipose tissue (WAT) serves as an important buffer for handling daily lipid flux by uptake of fatty acids (FAs) in the post-prandial state and release in the post-absorptive state (Frayn, 2002). Furthermore, studies in recent years have revealed that WAT is a major endocrine organ sending out hormones and signaling molecules that regulate and coordinate energy homeostasis, insulin sensitivity, lipid metabolism, substrate selection, satiety and appetite (Cristancho and Lazar, 2011). WAT also cooperates with other tissues including liver, muscle, pancreas, heart and brain through the release of FAs (lipokines), secretory factors (adipokines) and pro-inflammatory cytokines. WAT dysfunction or overload of its lipid storage capacity can lead to wide range of diseases (e.g., immunological and inflammatory diseases), including metabolic diseases such as obesity and its adverse outcomes (Lago *et al*, 2007; Auffray *et al*, 2009). Obesity can lead to preventable cause of death and increases the likelihood of coronary heart disease, diabetes and several forms of cancer (Cao, 2010) and ~33% of the adults older than 20 years living in the United States and 20% of European adult population are obese. Obesity is therefore considered to be one of the greatest threats to global human health (Caveney *et al*, 2011). However, obesity is not directly related with particular etiological factors and it is therefore essential to enable stratification and prediction of disease risks among obese subjects to ensure early treatment either through diet intervention, exercise or surgery.

An increased understanding of the mechanisms behind obesity and related diseases will provide valuable insights into their etiology, pathogenesis and may lead to new treatment strategies. It is inherently difficult to find the exact disease onset, since such systemic diseases are caused by a combination of different genetic and environmental factors and often result in similar disease phenotypes. Therefore, it remains challenging to identify the cellular and molecular mechanisms associated with obesity, in particular as the metabolism of the single cell involves thousands of metabolites and interconnected chemical reactions that occur simultaneously. Proper understanding of such a complex system requires a holistic approach. For this purpose, the so-called genome-scale metabolic models (GEMs) are suitable as they allow for analysis of metabolism at the genome scale but at the same

time ensures identification of specific pathways, enzymes and metabolites associated with specific phenotypes (Nielsen, 2009; Thiele and Palsson, 2010).

GEMs integrate biochemical and physiological data on genes and enzymatic reactions and allow the study of relationships between networks, functions, diseases and patients at a systems level. Reconstructed large-scale GEMs can serve as scaffolds for data analysis and identify biologically meaningful correlations and mechanistic relationships between components of different sub-systems (Patil and Nielsen, 2005). Hereby, GEMs can be employed for understanding the underlying mechanisms of complex diseases, and hence be used for identification of novel therapeutic and drug targets and discovery of new biomarkers. The diseases that we are discussing are not difficult to diagnose (e.g., obesity and diabetes) but efficient biomarkers would be of interest to predict prognosis and outcome of the disease and thereby enabled patient stratification which will be the basis for developing personalized medicine (Mardinoglu and Nielsen, 2012; Nielsen, 2012; Figure 1).

Two generic literature-based GEMs of human metabolism, Recon 1 (Duarte *et al*, 2007) and the Edinburgh human metabolic network (EHMN; Ma *et al*, 2007) have been reconstructed previously. Although these first reconstructions represent a major advancement, human metabolism is specialized in different cell types, and hence there is a need for reconstruction of cell type or tissue-specific GEMs. In this context, tissue-specific GEMs have been developed for liver

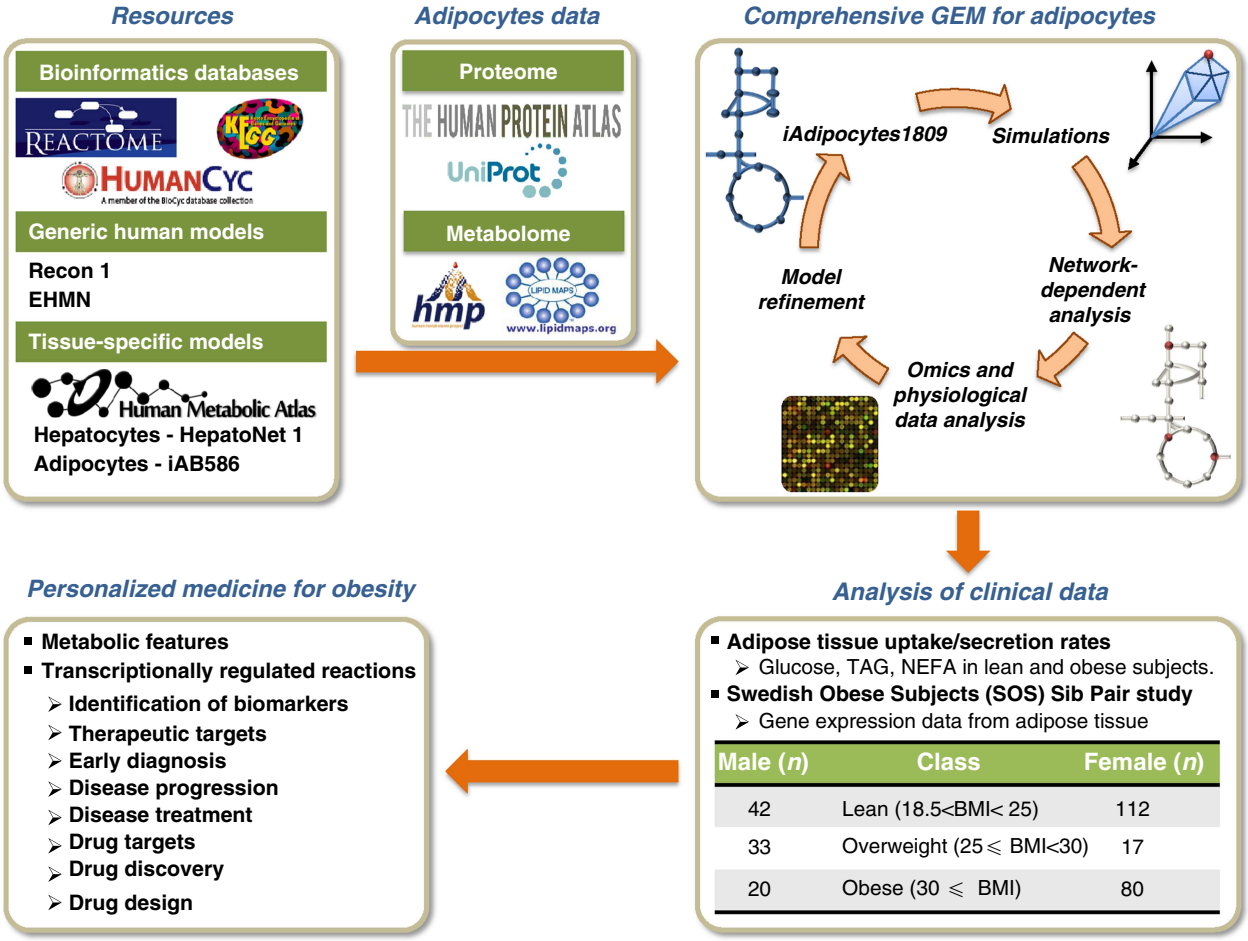


Figure 1 Genome-scale metabolic models (GEMs) provide a scaffold for integrative analysis of clinical data. Schematic illustration of how a comprehensive and functional GEM for adipocytes may provide links between specific molecular processes and subject phenotypes and hereby contribute to the development of personalized medicine for obesity. Here, the GEM *iAdipocytes1809* was reconstructed to bridge the gap between genotype and phenotype through the use of proteome, metabolome, lipidome and transcriptome data, literature-based models (Recon 1, Edinburg Human Metabolic Network (EHMN) and HepatoNet1) and public resources (Reactome, HumanCyc, KEGG and the Human Metabolic Atlas). We first performed global protein profiling of adipocytes encoded by 14 077 genes to study adipocyte biology at the genome-wide level using antibodies generated within the Human Protein Atlas (HPA). We further used information on metabolome and lipidome data from the Human Metabolome Database (HMDB) and LIPID MAPS Lipidomics Gateway, respectively. Model driven simulations, network-dependent analysis and condition-specific transcriptome data allowed for model refinement. *iAdipocytes1809* was used for the analysis of gene expression data obtained from subjects with different body mass indexes in the Swedish Obese Subjects (SOS) Sib Pair study and other adipose tissue relevant clinical data such as uptake/secretion rates in lean and obese subjects. Employing *iAdipocytes1809* for integration of transcriptome and fluxome enables understanding of adipocyte metabolism in obese subjects compared with lean subjects. Furthermore, the results of the study lead to identification of molecular mechanisms underlying obesity and its adverse outcomes. This can be useful for identification of new therapeutic and drug targets and discovery of new biomarkers for predicting prognosis and outcome of the disease, developing a system-oriented drug design strategy, obtaining novel diagnostic and therapeutic techniques and eventually determining effective personalized medicines for treatment of obesity-related diseases.

(hepatocytes) (Gille *et al*, 2010; Jerby *et al*, 2010), cardiomyocytes (Karlstaedt *et al*, 2012), kidney (Chang *et al*, 2010), brain (Lewis *et al*, 2010) and alveolar macrophage (Bordbar *et al*, 2010) and recently, three small cell type-specific GEMs of hepatocytes, myocytes and adipocytes (iAB586) were developed (Bordbar *et al*, 2011). Furthermore, using the Integrative Network Inference for Tissues (INIT) algorithm, we previously reconstructed cell type-specific draft GEMs for 69 different cell types and 16 cancer types (Agren *et al*, 2012).

With the objective of gaining new insight into adipocyte metabolism at the genome level, we first used human antibodies to evaluate the presence/absence of 14 077 proteins in adipocytes found in breast and two different soft tissue samples and checked their presence call with previously published adipocyte proteome data. Second, we manually reconstructed a high-quality, simulation ready GEM for adipocytes by using all adipocyte-specific proteome data. The model is based on previously published GEMs but also on publicly available databases on metabolism. Third, we employed the functional GEM for the analysis of microarrays that profile the gene expression from subcutaneous adipose tissue (SAT) of subjects from the Swedish Obese Subjects (SOS) Sib Pair Study, which includes nuclear families with body mass index (BMI)–discordant sibling pairs (BMI difference $\geqslant 10\,kg/m^2$). The male and female participants of the study were divided into three different groups based on their BMIs: lean, overweight and obese. Analysis of transcription data from this study recently demonstrated that there are differences in mitochondrial function between men and women (Nookaew *et al*, 2012), but there was not performed any analysis on the effect of obesity. We incorporated differentially expressed genes between obese and lean subjects into the GEM. Besides the gene expression data from the SOS Sib Pair Study, additional clinical data (e.g., plasma and WAT lipid concentrations) were also incorporated into the model. By integrating gene expression data and adipose tissue uptake/secretion rates with the reconstructed GEM, we identified metabolic differences between individuals with different BMIs by using the concept of Reporter Metabolites (Patil and Nielsen, 2005) and transcriptionally controlled reaction fluxes (Bordel *et al*, 2010; Figure 1).

Results

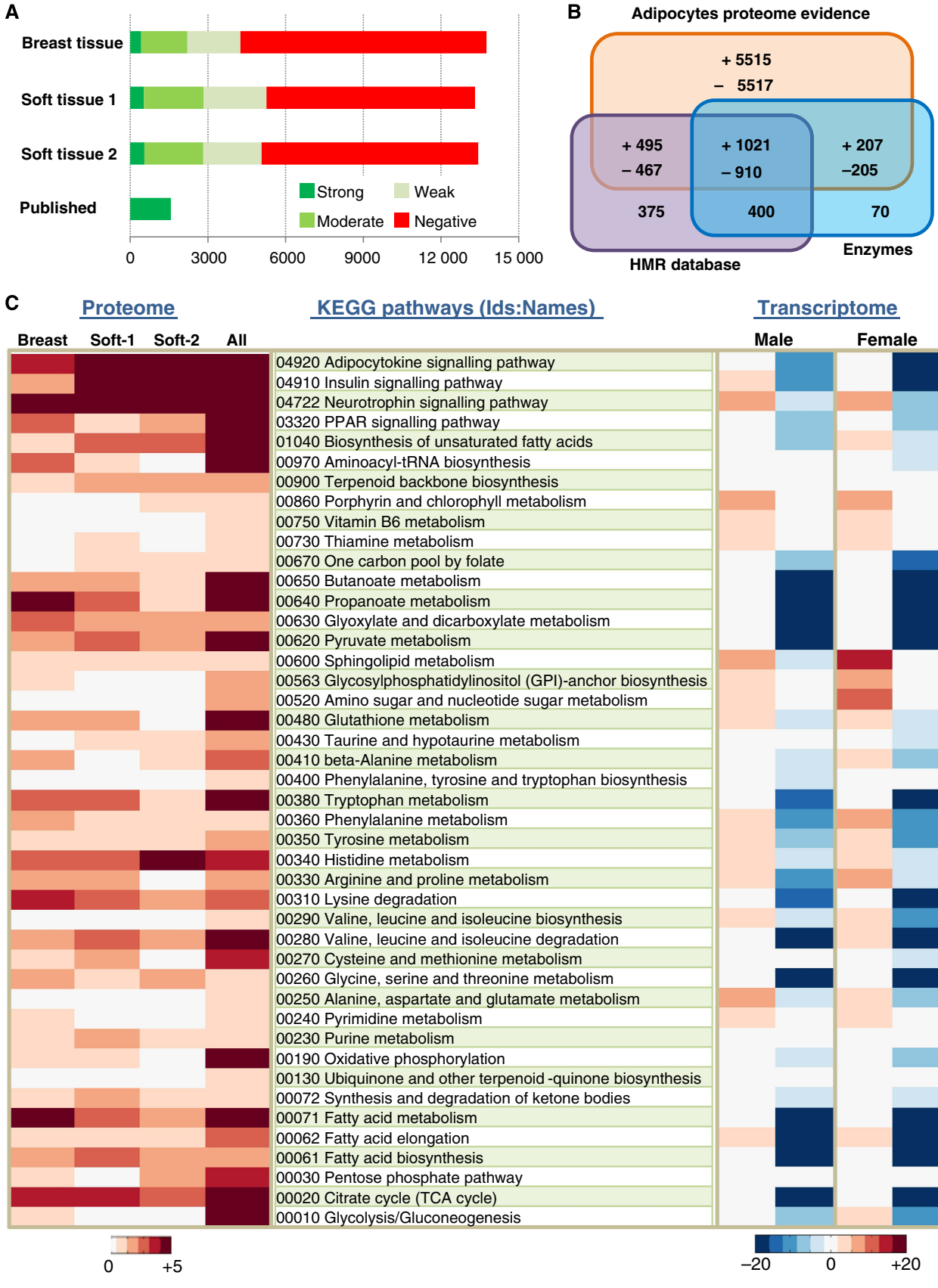
Immunohistochemistry-based proteomics of human adipocytes

Several studies have reported proteomics data of human WAT that include not only adipocytes but also the connective tissue matrix, nerve tissue, stromal vascular cells and immune cells (Peinado *et al*, 2012). Recently, Xie *et al* (2010) characterized the proteome of human adipocytes and reported existence of proteins encoded by 1574 genes in subcutaneous abdominal adipocytes taken from three healthy lean subjects. Although this first adipocyte-specific proteome study is promising to understand the unique characteristics of the adipocytes, it is still necessary to expand the protein coverage to be able to study adipocyte biology at the genome-wide level.

The Human Protein Atlas (HPA) is a knowledge-based portal that cover the annotated protein expression feature for protein targets analyzed with one or more antibodies and the main subcellular localization of protein targets in all major human cell types and cell lines (Uhlen *et al*, 2010). Here, we expanded the coverage of the HPA so as to define the protein profiles of adipocytes found in breast and two different soft tissues and examined the spatial distribution and the relative abundance of proteins encoded by 14 077 genes in adipocytes (Supplementary Dataset 1a). Soft tissue cores with 1 mm diameter were sampled from different locations of the human body and included in tissue microarrays (TMAs) together with breast tissue core as previously described (Kampf *et al*, 2012). A total of 17 296 high-throughput generated affinity-purified antibodies (Supplementary Dataset 1a) against 14 077 proteins were arranged according to abundance of the corresponding protein target. Immunohistochemically stained sections from TMAs blocks were scanned in high-resolution scanners and separated to individual spot images to represent each core. Here, proteins with strong, moderate and weak relative expression were included in the reconstruction process of the GEM for adipocytes. It is important to point out that the absolute levels of each protein have not been determined and could vary by orders of magnitude. The high-resolution images together with annotation of the presence or absence of a particular protein target in adipocytes are publically available through the HPA (http://www.proteinatlas.org).

The proteome data generated for adipocytes found in breast and two different soft tissues were merged with previously published proteome data as presented in Figure 2A (Supplementary Dataset 1b). In total, we have proteome evidence for the presence/absence of proteins associated with 14 337 genes in adipocytes and together with the genes in our Human Metabolic Reaction (HMR) database (see later), we cover 98% of the enzymes (Supplementary Dataset 1c) reported in the HPA database (Figure 2B). Our proteomics analysis resulted in provision of evidence for presence of proteins associated with 7340 genes in adipocytes that could be used for our GEM reconstruction. Each protein that has evidence for presence/absence in adipocytes was annotated with UniProt id (Apweiler *et al*, 2011) and the corresponding encoding genes were annotated with Ensembl (Flicek *et al*, 2011) id for standardization.

Since adipocytes obtained from breast and two different soft tissues were used for the protein profiling, there was some variation between the samples. To estimate the effect of these variations on the functionality of the adipocytes, we used the functional annotation tool DAVID to calculate the enrichment in KEGG pathways (Huang *et al*, 2009). The results are presented in Figure 2C for breast, the two types of soft tissues, as well as for all proteomics data used for the GEM reconstruction. The analysis demonstrated that for all the adipocyte-specific proteome data, there is enrichment in terms of metabolic pathways including FA metabolism, elongation and biosynthesis as well as major signaling pathways including adipocytokine, insulin, neurotrophin and PPAR signaling pathways. The figure shows that the samples from different tissues exhibit some differences, but that the overall pattern is similar. We therefore decided to reconstruct a general GEM for adipocytes by incorporating all proteins that were expressed in any of the three tissues.



Subcellular localization of the proteins and associated reactions

Metabolism is distributed across subcellular compartments of the cell and precise coordination is important for overall cellular function and maintenance of metabolic homeostasis. For instance, β-oxidation of very long chain FAs (VLCFAs) is a process that starts in the peroxisome and ends in the mitochondria. The subcellular localization of reactions has large implications on the functionality of GEMs, as only a portion of metabolites can be transported between compartments. Furthermore, compartments can be individually redox and/or energy balanced. It is therefore important to understand the locations of individual enzymes as well as their associated reactions when metabolism is going to be studied at the genome scale. The HPA includes subcellular profiling data using immunofluorescence-based confocal microscopy in three human cancer cell lines of different origin to be used for further in-depth functional studies.

Here, proteins were classified into eight different compartments following our HMR database standard (Agren *et al*, 2012): cytosol, nucleus, endoplasmic reticulum (ER), Golgi apparatus (GA), peroxisome, lysosome, mitochondria and extracellular space. Reactions were assigned to compartments through their association with proteins in these different compartments. We assigned a confidence score ranging from one to three for each protein due the availability of knowledge in HPA and Uniprot (Supplementary Dataset 2a). Proteins localized in the aggresome, centrosome and cytoskeletons were assigned to the cytoplasm, whereas proteins in cell junctions and focal adhesions were assigned to the extracellular compartment (Supplementary Dataset 2b). Lysosome and peroxisome were merged into vesicles in the HPA data, but proteins associated with vesicles in our model were separated based on Uniprot data. In the case a protein has different locations in the two resources, we chose to assign the protein in the location reported based on the HPA data. In cases where the subcellular localization information was absent in HPA, Uniprot and literature (indirect physiological evidence), proteins and their associated reactions were assigned to be present in the cytosol.

Transcriptome data for SAT and its correlation with proteome data

Gene expression in SAT of 304 subjects involved in the SOS Sib Pair Study was analyzed by using Affymetrix U133 Plus 2.0 microarrays (Affymetrix, Santa Clara, CA, USA). The participants of the study, 209 female and 95 male subjects, were divided into three different groups according to their BMI: lean ($18.5 < \text{BMI} < 25$), overweight ($25 \leqslant \text{BMI} < 30$) and obese ($30 \leqslant \text{BMI}$) (Supplementary Table S1). Our study considers the differences in male and female obese subjects compared with lean male and female subjects independently. Differentially expressed genes between obese, overweight and lean subjects for male and female subjects were identified and differentially expressed genes on male and female obese subjects were incorporated during the reconstruction process of GEM for adipocytes to analyze gene expression data from the SOS Sib Pair Study.

Although the generated proteome data for adipocytes cover the entire set of cellular processes (e.g., signaling, metabolism and cell cycle), GEMs are applicable only for the study of metabolism. To get a general overview of the global changes between obese, overweight and lean subjects, the enrichment of differentially expressed genes was calculated for KEGG pathways (Supplementary Figures S1 and S2) and for biological process Gene Ontology (BP:GO) terms (Supplementary Figures S3 and S4). This was done using DAVID and for male and female subjects (Huang *et al*, 2009). To check the correlation of the genome-wide transcription data of SAT with the proteome data, the enrichment of differentially expressed genes in male and female obese subjects was calculated for the most significant KEGG pathways from the analysis of the proteome data (Figure 2C). This was done as the transcriptome data for SAT represent not only adipocytes but also other cell types; including immune cells and preadipocytes linked with different BMIs.

The analysis demonstrates that some of the metabolic and signaling pathways found to be enriched in adipocytes based on the proteome data also show significant changes in gene expression between lean and obese subjects, both in males and in females. Similarly, we find that BP:GO terms that are enriched based on the proteome data (assessed with DAVID; Huang *et al*, 2009) also show enrichment based on the transcriptome data (Supplementary Dataset 3). Thus, enriched BP:GO terms such as post-translational protein modification, cellular protein metabolic process, lipid metabolic process, cellular lipid metabolic process and FA metabolic process are found both from the adipocyte-specific proteome data and from comparison of expression data for lean and obese subjects.

Reconstruction of the adipocytes GEM *iAdipocytes1809*

GEMs are reconstructed based on high-throughput data such as genome, transcriptome, proteome, metabolome and

Figure 2 Adipocytes proteome data and its correlation with transcriptome data. (**A**) Proteins encoded by 14 077 genes were annotated for their presence or absence in adipocytes found in breast and two different soft tissues sampled from different locations within the body using antibodies generated within the Human Protein Atlas (HPA) project. The proteome evidence for adipocytes in each tissue sample and previously published adipocytes proteome data were merged and there are proteome evidence for the presence/absence of proteins associated with 14 337 genes in adipocytes. (**B**) The here generated proteome data combined with published adipocytes specific proteome data encodes for a total of 7340 genes that were used to reconstruct a comprehensive genome-scale metabolic model (GEM) for adipocytes, *iAdipocytes1809*. The Venn diagram shows how these genes overlap with genes in an updated version of the Human Metabolic Reaction (HMR) database and enzymes in the HPA. (+) indicates the presence of protein encoded genes whereas (−) means the absence of protein encoded genes in adipocytes based on the here generated HPA data. (**C**) Enrichment of KEGG pathways is presented (*P*-value from hypergeometric distribution is used) to check the coverage of here generated adipocyte-specific proteome data found in breast, soft tissue 1 (Soft-1), soft tissue 2 (Soft-2) and all adipocyte-specific proteome data. The enrichment of differentially expressed genes in subcutaneous adipose tissue (SAT) of male and female obese subjects compared with lean subjects was identified in the enriched KEGG pathways in all adipocyte-specific proteome data.

fluxome and they provide an excellent scaffold for integrative analysis of this kind of data (Patil and Nielsen, 2005; Cakir *et al*, 2006). To reconstruct a large-scale comprehensive GEM for adipocytes, biochemical and genetic evidence were combined with data on protein expression and localization. Besides the proteins associated with 7340 genes identified from our proteomics analysis, we further used differentially expressed genes in SAT of obese male and female subjects for our reconstruction process. However, SAT contains not only adipocytes but also other cell types and it is therefore necessary to check the functionality of the genes before including into the GEM.

HepatoNet1, a GEM for hepatocytes which is reconstructed based on the manual evaluation of the original scientific literature (Gille *et al*, 2010), was used as a starting point for our reconstruction process and used to generate an initial candidate list of network components (Figure 3A). First, metabolism of lipids and lipoproteins in Reactome, a manually curated and peer-reviewed pathway database (Croft *et al*, 2011), was merged into HepatoNet1. Second, the resulting network was combined with the evidence-based generic human models Recon1 (Duarte *et al*, 2007) and the compartmentalized EHMN (Hao *et al*, 2010). This combined reaction list resulted in an updated version of our HMR database that is available at http://www.metabolicatlas.org. The HMR database contains 6000 metabolites in 8 different compartments (3160 unique metabolites), 8100 reactions and 3668 genes associated with those reactions. HMR includes all of the genes and gene-associated reactions in HepatoNet1, Recon1 and EHMN and other isolated gene-associated reactions, and represents the most comprehensive human reaction database for genome-scale modeling (Figure 3B). Third, the existence of each protein coding genes associated with a reaction in HMR was assessed for the presence or absence in adipocytes using previously published and the here generated adipocyte-specific proteome data. Differentially expressed genes and associated reactions in obese subjects compared with lean subjects were also included in the model by checking their functionality. This process provided us with a list of reactions that occur in adipocytes. Gaps in the resulting network were filled using our recently developed *iHuman1512* metabolic network (Agren *et al*, 2012), public databases such as KEGG (Kanehisa *et al*, 2010) and HumanCyc (Romero *et al*, 2005) and manual evaluation of the literature about adipocyte metabolism. This gap filling resulted in generation of *iAdipocytes1809* that is a functional and fully connected GEM for adipocytes (Figure 3A). Reactions were included in the GEM depending on evidence from previously published models and databases (Supplementary Table S2) or on the availability of specific experimental evidence, for example, enzyme assay or protein identification, for the occurrence of the corresponding reaction in adipocytes. Reaction directionality in the model was treated as in the original resource and the directionality was only changed if it was necessary to perform successful simulations of known biological functions of adipocytes. *iAdipocytes1809* contains 6160 reactions and 4550 metabolites in 8 different compartments (2497 unique metabolites) (Supplementary Table S3) and it is available in the Systems Biology Mark-up Language (SBML) format at http://www.metabolicatlas.org. There are 1809 genes in the model and 80% of the reactions are associated with one or more genes. In *iAdipocytes1809*, individual metabolites rather than generic pool metabolites for 59 FAs (Supplementary Table S4) have been used and this allowed us to incorporate measured concentrations of different FAs in human plasma and adipocytes into the model.

Adipocyte-specific metabolome data from the Human Metabolome Database (HMDB) (Wishart *et al*, 2009) and a comprehensive database for lipid biology, Lipidomics Gateway (Harkewicz and Dennis, 2010), were used in the reconstruction process. To ensure standardization, each metabolite in the GEM was assigned at least one of the following: HMDB ids, Lipidomics Gateway ids, KEGG ids (Kanehisa *et al*, 2010), Chemical Entities of Biological Interest (ChEBI) (Degtyarenko *et al*, 2008), or International Chemical Identifiers (InChI). In the model, new genes were assigned to the reactions in *iAdipocytes1809* using EC numbers from UniProt (Apweiler *et al*, 2011) and the Lipid Map proteome database (Cotter *et al*, 2006).

The reconstruction of *iAdipocytes1809* involved a comprehensive review of lipid metabolism and gene products related to the lipid metabolism in publicly available databases were closely examined for their existence in adipocytes. In all, 1235 protein encoding genes with proteome evidence in adipocytes and 244 genes in *iAB586* were included in *iAdipocytes1809* (Figure 3C). Another 137 genes were included in the model due to positive evidence for the presence of these proteins from the transcriptome data. These 137 genes were significantly higher expressed both in male and in female obese subjects compared with lean subjects, and are hence relevant for the analysis of gene expression data from the SOS Sib Pair Study. Another 193 genes and their associated reactions were included in the model due to connectivity constraints and known function of adipocytes and these genes are mainly associated with transport across membranes. In all, 73 of these genes do not have any negative evidence whereas 120 of these genes have negative HPA scores. These 120 genes are likely to be false negatives in the HPA, and they can be explained by either poor antibody hybridization to adipocytes or condition-dependent protein expression. We also compared *iAdipocytes1809* with *iAB586* and the previously published generic human network, *iHuman1512*, and we found that *iAdipocytes1809* contains all of the genes and associated reactions in *iAB586* and adipocyte-specific reactions and associated genes of *iHuman1512* (Figure 3D).

In *iAdipocytes1809*, 59 different common long and very long chain FAs (Supplementary Table S4) in human plasma can be taken up as NEFAs and lipoproteins (Supplementary Figures S5–S7). Cholesterol and its 59 different CEs can also be taken up from low-density lipoproteins (LDLs) and high-density lipoproteins (HDLs). In the post-prandial state, FAs and CEs can be incorporated into LD structures and in the post-absorptive state LDs can be broken down to FAs and CEs. The comprehensive *iAdipocytes1809* covers all metabolic pathways known to exist in adipocytes as well as extensive knowledge of lipid metabolism in human cells and adipocytes in particular. Lipid metabolism in adipocytes involve uptake of FAs from two potential sources: non-esterified FAs (NEFAs)

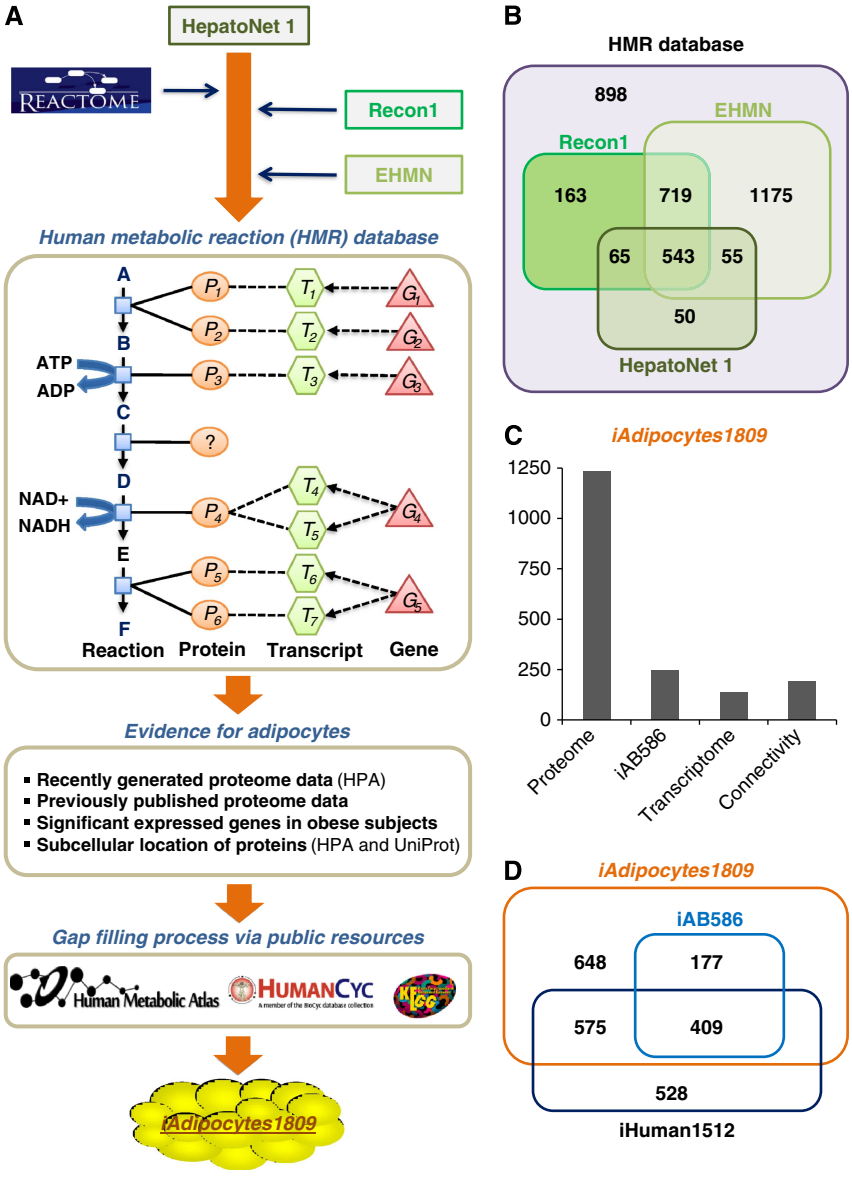


Figure 3 Illustration of the reconstruction process for *iAdipocytes1809*. (**A**) Human Metabolic Reaction (HMR) database was updated through the use of literature and previously published GEMs like HepatoNet1, Recon1 and EHMN and the Reactome database, with focus on the extensive lipid metabolism in adipocytes. HMR provides gene-transcript protein reaction (GTPR) links and GTPR represents functional relationships between genes/proteins and the corresponding reactions that they catalyze or control. Genes are first mapped to their transcripts, accounting for alternative splicing, the transcripts are mapped to the proteins and the proteins are then mapped to reactions based on the current knowledge of their effects on the reactions. In HMR, 75% of the reactions have GTPR and this allow us to incorporate proteome and transcriptome evidence in the adipocyte-specific network. The resulting gaps in the adipocyte-specific network were identified using the RAVEN Toolbox and filled manually using publicly available resources like KEGG, HumanCyc and Human Metabolic Atlas. Based on this, the comprehensive functional genome-scale metabolic model (GEM) for adipocytes, *iAdipocytes1809*, was reconstructed. (**B**) The Venn diagram illustrating the comparison of genes associated with reactions in the HMR and previously published literature-based models such as Recon 1, EHMN and HepatoNet 1. HMR includes all of the genes and associated reactions in previous published models. (**C**) Protein encoding genes and associated reactions with those genes included in *iAdipocytes1809* based on proteome (1235 genes) and transcriptome (137 genes) data and evidence from previously published small-scale adipocytes model, iAB586 (244 genes). Another 193 genes and their associated reactions were included in the model due to connectivity constraints and known function of adipocytes. (**D**) *iAdipocytes1809* is compared with iAB586 and fully connected generic human network, *iHuman1512*. *iAdipocytes1809* contains all of the genes and associated reaction in iAB586 and all of the adipocyte-specific genes and associated reactions in *iHuman1512*.

and lipoproteins (Supplementary Table S5) including chylomicrons, very low-density lipoprotein (VLDL), LDL and HDL through LPL. *iAdipocytes1809* also cover the transport of FAs, cholesterol and cholesterol esters (CEs) across the plasma membrane, small amount of *de novo* FA and cholesterol synthesis, FA transport into mitochondria and peroxisomes for β-oxidation, FA esterification into triacylglycerols (TAGs), lipolysis and lipid droplet (LD) formation and maintenance (see Supplementary information).

Model validation and formation of lipid droplets

Due to the large size of GEMs (typically thousands of metabolic reactions), proper validation is essential to ensure

that a reconstructed model has predictive ability (Mardinoglu and Nielsen, 2012). A common problem is lack of mass balancing of the individual reactions, and this typically involves metabolites with ill-defined formulas, such as polymers or metabolite pools, but different protonation states can also be problematic. This type of problem results in a situation where the model can take up or excrete metabolites in an unbalanced manner. The model was tested so that all reactions except pool reactions were mass balanced. A second issue is that reactions can violate thermodynamic constraints by being written in the wrong direction or where irreversible reactions are written as reversible. This was tested by making sure that the model could not generate high-energy compounds from low-energy compounds (such as ATP and water from ADP and phosphate or any organic compound from CO_2 and water). A third issue is that reactions may be dead-ends, meaning that they cannot carry flux, either because the model is not able to produce/take up some of the substrates or because it is not able to consume/excrete some of the products. Dead-end reactions tend to propagate as one unconnected reaction can give rise to many more. Therefore, many GEMs contain a large proportion of dead-end reactions. Considerable efforts were spent in making sure that the number of dead-end reactions was kept to a minimum. Furthermore, with the simulation ready *iAdipocytes1809*, the production of all metabolites in the model was checked with minimum input to the model (Supplementary Dataset 4) to ensure the connectivity. Lastly, even a well-connected, thermodynamically correct and balanced model may not be able to perform all relevant metabolic functions, or it may be able to perform functions that it should not do (such as synthesis of essential amino acids or FAs). The model was therefore validated for 250 known metabolic functions of adipocytes, adapted from the definitions provided in connection with setting up HepatoNet1 (Gille *et al*, 2010; Supplementary Dataset 5). All validation steps and simulations were performed using the RAVEN Toolbox (Agren *et al*, 2013).

Formation of LDs (Figure 4) is included in *iAdipocytes1809*. LDs protect the cell from the lipotoxic effects of unesterified lipids, but they are also associated with the development of metabolic diseases, such as obesity, type 2 diabetes, atherosclerosis and liver steatosis that are characterized by the extensive accumulation of LDs. LDs represent a significant energy storage that can be mobilized by catabolism and this process is highly regulated by hormones and signaling pathways. Although LDs are recognized as an organelle called adiposomes (Farese and Walther, 2009), we represented it as a

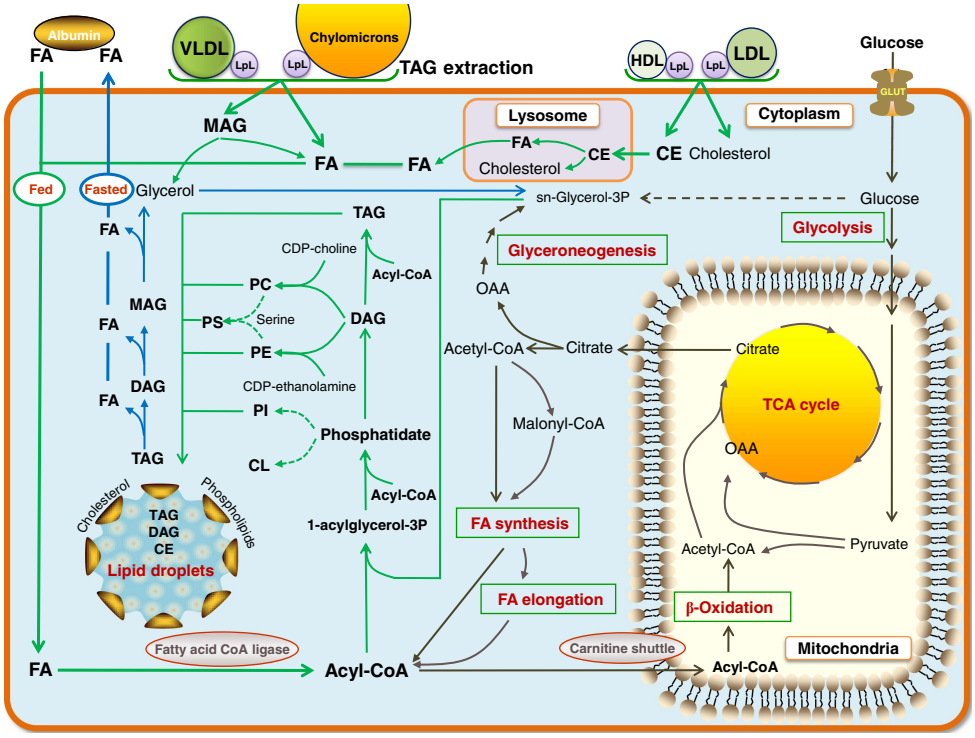


Figure 4 Uptake of fatty acids through NEFA and lipoproteins and formation of lipid droplet in adipocytes. Adipocytes in white adipose tissue (WAT) store lipid mainly in the form of triacylglycerols (TAGs) and cholesterol esters (CEs) and form lipid droplets (LDs) in the post-prandial state (green arrows) and release them by degrading LDs in the post-absorptive state (blue arrows) to provide energy for other tissues. The released fatty acids (FAs) from adipocytes are transported to other tissues by albumin. The FAs are taken up from non-esterified FAs (NEFAs) and lipoproteins, including chylomicrons, very low-density lipoprotein (VLDL) and CEs together with cholesterol are taken up with low-density lipoproteins (LDLs) and high-density lipoproteins (HDLs) through LPL. CEs taken up from lipoproteins are degraded to cholesterol and FAs in lysosomes and transported to the endoplasmic reticulum and cytosol to be stored in LDs. Adipocytes also take up glucose to be used in the *de novo* synthesis of FAs (black arrows) that occurs at low level in adipocytes. LDs are rich in TAGs, CEs and an unknown neutral lipid that migrated between CEs and TAGs, ether neutral lipid monoalk(en)yl diacylglycerol (MADAG). LDs also contain small amounts of free FAs, cholesterol and phospholipids including phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidylinositol (PI), ether-linked phosphatidylcholine (ePC), ether-linked phosphatidylethanolamine (ePE), lyso phosphatidylcholine (LPC), lysophosphatidylethanolamine (LPE), phosphatidylserine (PS) and sphingomyelin (SM). Formation of ePC, ePE, LPC, LPE and SM is included in the genome-scale metabolic model for adipocytes, *iAdipocytes1809* and are not shown in figure.

large composite metabolite in our model. LDs have polar surfaces that contain a variety of proteins, PLs and sterols and have a non-polar core that contain CEs, DAGs and TAGs (Gross *et al*, 2011; Supplementary Table S6). The diameter of LDs varies from 50 nm to 200 μm and LDs may grow to enormous sizes with different mechanisms including localized synthesis of lipids, transport of lipids and coalescence of LDs.

The function of *iAdipocytes1809* was examined by estimating the formation of LDs and by means of available physiological, biochemical and genetic evidence in lean and obese subjects and clinically observed data. Recently, McQuaid *et al* (2011) measured the delivery and transport of FAs in adipose tissue using multiple and simultaneous stable-isotope FA tracers in lean and obese subject groups over 24-h period. Even though abdominally obese subjects have greater adipose tissue mass than control lean subjects, the rates of delivery of NEFAs were downregulated in obese subjects. Clinical data on Supplementary Table S1, NEFA release, TAG extraction, glucose uptake (Supplementary Dataset 6; McQuaid *et al*, 2011) and amino-acid uptake rates (Supplementary Dataset 7; Patterson *et al*, 2002) and FA concentrations in plasma and adipocytes (Supplementary Dataset 8) were incorporated into the model to predict the formation of LDs. Based on measurements of the uptake of glucose and TAG and the release of NEFAs over a 24-h period (Figure 5A and 5B), we simulated the change in LD size (Figure 5C; Supplementary Dataset 9a). We found from our simulations that lean subjects have large dynamic changes in LD formation compared with obese subjects, which is in agreement with experimental data (Arner *et al*, 2011). Furthermore, we predicted a lower acetyl-CoA production in obese subjects, as shown in Figure 5D (Supplementary Dataset 9b).

Identification of obesity-specific metabolic features

Modeling using *iAdipocytes1809* can be applied to predict metabolic states under various perturbations, study regulation of the adipocytes, identify potential therapeutic targets and discover novel biomarkers for the development of more effective therapies. Here, we used the model to identify Reporter Metabolites (Patil and Nielsen, 2005) of male and female obese subjects compared with lean subjects using gene expression data obtained from the SOS Sib Pair Study. Reporter Metabolites are metabolite nodes in the metabolic network around which there are significant transcriptional changes. Here, 20 statistically significant Reporter Metabolites are presented for upregulated and downregulated genes in male and female obese subjects through the employment of *iAdipocytes1809* (Figure 6). The most significant results from our Reporter Metabolites analysis for upregulated and downregulated genes are correlated with the KEGG pathways enrichment results of significantly expressed genes in the obese subject groups (Supplementary Figures S1 and S2).

To illustrate the improvement of *iAdipocytes1809* over the published *iAB586,* the Reporter Metabolites were also calculated for male and female obese subjects by using *iAB586* (Supplementary Figures S8). Reporter Metabolites involved in the mitochondrial dysfunction as well as different amino acids were identified to be similar to the Reporter Metabolite analysis using *iAdipocytes1809.* However, *iAB586* could not detect several of the most significant and in our view most interesting Reporter Metabolites identified when using *iAdipocytes1809* due to the increase in number of reactions, metabolites and genes (see Discussion). Thus, the Reporter Metabolite analysis with *iAdipocytes1809 and iAB586* provides an unbiased confirmation that *iAdipocytes1809* represent significant advancement of the adipocyte metabolic network compared with *iAB586*.

Identification of obesity-specific transcriptionally regulated reactions by random sampling

We identified changes in metabolic fluxes in response to obesity and which of these changes are likely to be associated with transcriptional changes using *iAdipocytes1809*. We defined a region of feasible flux distributions using uptake rates for TAGs and glucose and NEFA release rates for lean and obese subjects, and used these to calculate a set of possible flux distributions using a random sampling algorithm (Bordel *et al*, 2010). The average values and standard deviations for each of the fluxes in *iAdipocytes1809* were calculated and the changing fluxes were compared with the significance of change in gene transcription for the corresponding enzymes. This allowed us to identify specific reactions for which flux changes are likely to be transcriptionally regulated.

The results from this analysis showed that the following pathway fluxes were transcriptionally downregulated in obese subjects: uptake of glucose, uptake of FAs, oxidative phosphorylation, mitochondrial and perixomal β-oxidation, FA metabolism and tricarboxylic acid (TCA) cycle. This analysis is consistent with the findings from the Reporter Metabolites analysis and KEGG enrichment pathway analysis (Supplementary Table S7). Furthermore, fluxes associated with beta-alanine metabolism were found to be transcriptionally downregulated in obese subjects (Figure 7). Previously, it has been reported that blood flow, glucose uptake, release of NEFA and the extraction of TAG from plasma were significantly lower in abdominally obese subjects compared with lean subjects (McQuaid *et al*, 2011). Our random sampling results clearly indicate that all metabolic pathways in mitochondria are downregulated (Kusminski and Scherer, 2012) similar to the mitochondrial decline that is a hallmark of different diseases associated with aging.

Discussion

We evaluated the presence/absence of 14 077 proteins in adipocytes using human antibodies and presented a high-quality, simulation ready and functional GEM for adipocytes, *iAdipocytes1809,* based on proteome, metabolome and transcriptome data. *iAdipocytes1809* was used to analyze clinically observed transcriptome and fluxome data to understand the mechanistic changes in adipocyte metabolism in response to obesity. Gene expression and associated metabolites in obese and lean subjects were studied using a systems biology approach, which allowed us to understand how the transcriptome data in SAT represent different metabolic functions in

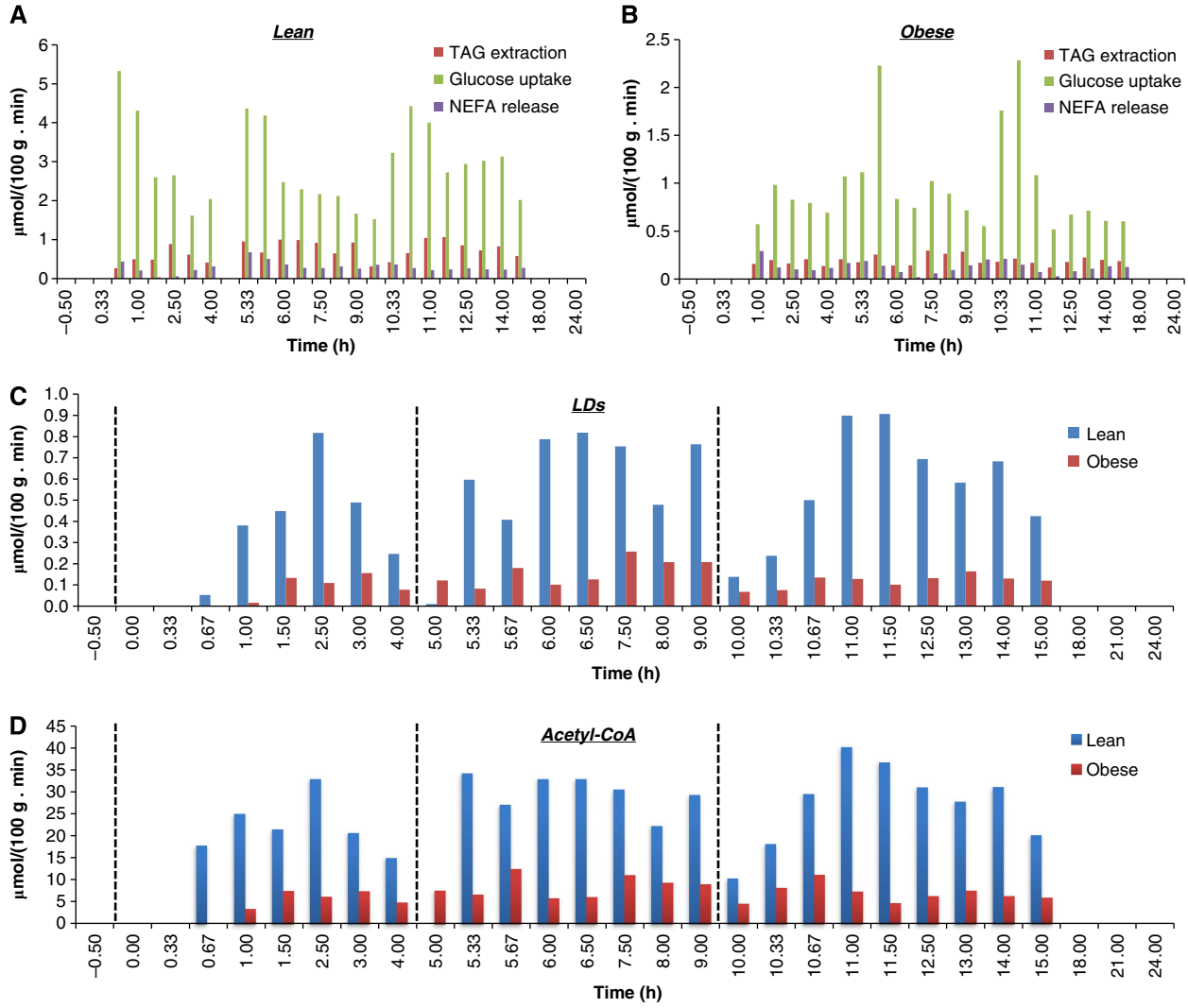


Figure 5 Simulation results for lipid droplets and acetyl-CoA production. Formation of lipid droplets (LDs) and acetyl-CoA that is the central node in the mitochondria has been simulated and qualitative amounts were predicted through the use of *iAdipocytes1809*. Acetyl-CoA is central metabolite in the mitochondria and the reduced dynamics in LD levels in obese clearly results in lower fluxes through this key metabolite in obese subjects. The uptake rates for glucose and triacylglycerols (TAGs) and release rates for non-esterified fatty acids (NEFAs) in adipocytes for lean (**A**) and obese (**B**) subjects have been used as lower and upper bounds for input reactions (McQuaid *et al*, 2011) and the amount of LDs (**C**) and acetyl-CoA (**D**) is generated for 24 h. In addition, the NEFA concentrations in adipocytes and plasma and FA content of lipid structures are incorporated into the model during the generation of pool metabolites. The dashed lines at time = 0, 5, 10 represent breakfast, lunch and dinner, respectively, for each participants of the study. *iAdipocytes1809* cannot produce LDs and acetyl-CoA at some time points since adipocytes degrade the LDs at post-absorptive state to provide FAs to other tissues as energy sources. *iAdipocytes1809* cannot produce LDs at time = 5 in lean subject groups, whereas it can produce LDs in the obese subject groups as in clinical data.

obesity. An increased understanding of metabolic pathways altered in complex diseases may contribute to the identification of novel therapeutic targets (Cheong *et al*, 2012) and here potential therapeutic targets for obesity were discovered through Reporter Metabolite analysis and random sampling of flux states. Moreover, protein coding genes related to the therapeutic targets may be used as potential drug targets.

Through our analysis, it was observed that expression of genes involved in metabolic pathways including mitochondrial and perixomal β-oxidation, FA synthesis, amino-acid metabolism, pyruvate metabolism, oxidative phosphorylation and TCA cycle are downregulated in obese subjects. Most of these pathways are linked with the mitochondrial dysfunction and different therapeutic interventions including antioxidants and chemical uncoupler treatments have proven to improve the mitochondrial dysfunction (Kusminski and Scherer, 2012). Recently, Canto *et al* (2012) have reported that increasing the NAD+ levels, identified as therapeutic targets in our study, through supplements (NAD+ precursors) enhances the oxidative metabolism and protects the cells against high-fat diet-induced obesity.

Mitochondrial acetyl-CoA has a central role in different pathways in the mitochondria and it reacts with oxaloacetate to form citrate, which can be transported from the mitochondria to the cytosol where it is participating in FA synthesis (Dean *et al*, 2009). Acetyl-CoA derived through other principal sources, including degradation of amino-acid and ketone bodies, and FA oxidation processes are insufficient for FA

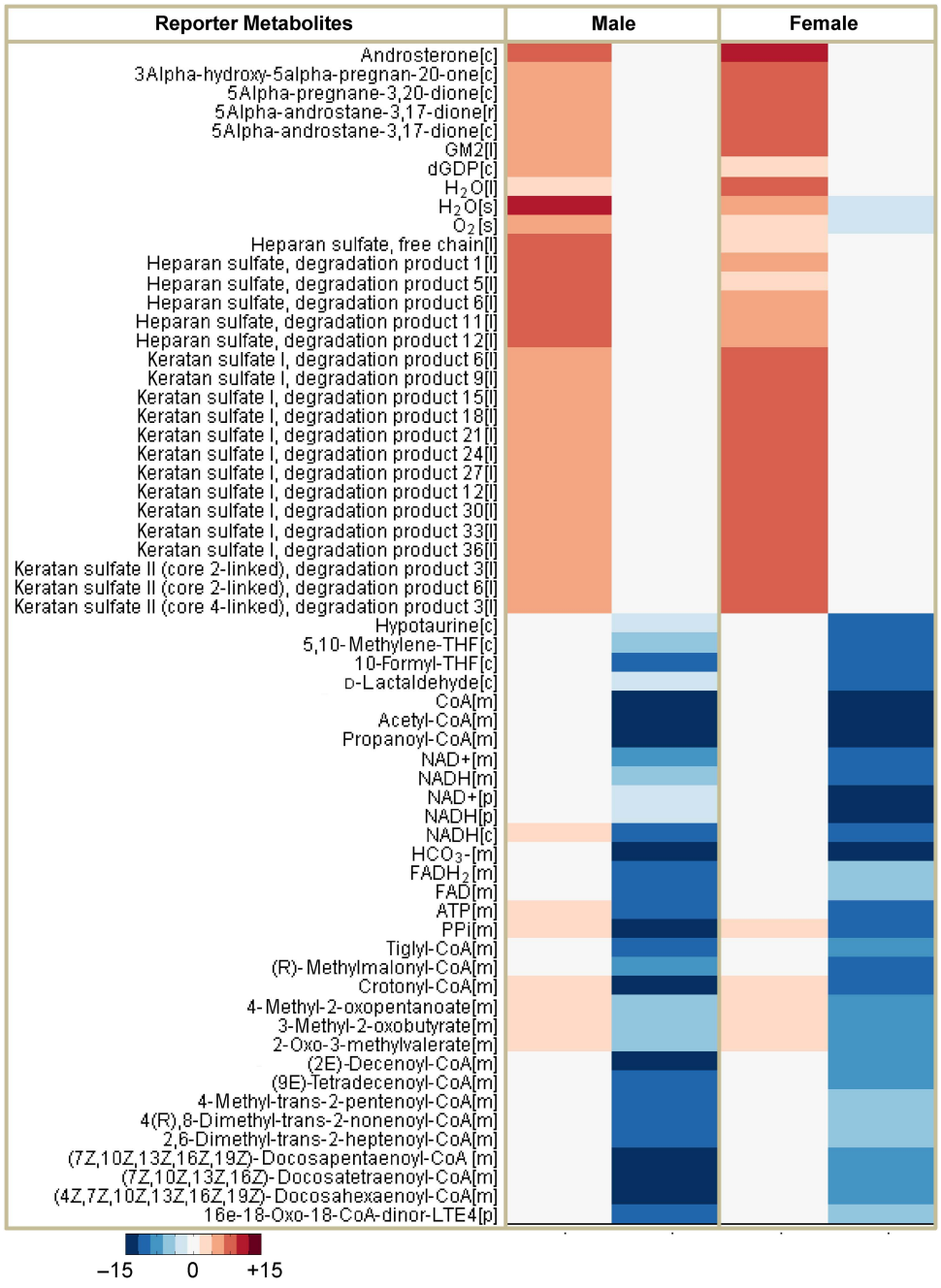


Figure 6 Representation of the Reporter Metabolites for male and female obese subjects. The Reporter Metabolites algorithm marks the regions in metabolism around which significant transcriptional changes occur. Reporter Metabolites are obtained using the *P*-values calculated from the comparison of obese subjects with lean subjects in male and female subjects separately. Top-scoring reporter metabolites androsterone, ganglioside GM2 and degradation products of heparan sulfate and keratan sulfate are associated with the upregulated differentially expressed genes and mitochondrial metabolites are associated with downregulated differentially expressed genes in obese subjects comparing with lean subjects. Top 20 metabolites associated with upregulated and downregulated genes in male and female obese subject were presented in the figure.

synthesis. Increasing the acetyl-CoA concentration and eventually FA synthesis in adipose tissue of obese subjects results in whole body regulation of metabolism, including stimulation of muscle insulin action and suppression of hepatosteatosis, as reported by Cao *et al* (2008). Here, we propose to boost the metabolic activity of mitochondria in the adipocytes of obese subjects by increasing the mitochondrial acetyl-CoA which is another therapeutic target identified in our study. Acetyl-CoA formation in lean and obese subjects was predicted through use of *iAdipocytes1809* in lean and obese subjects and its concentration can be increased through the uptake of beta-alanine. The effect of beta-alanine as a dietary supplement was previously examined in football players and it is reported that it has effect on lean tissue accruement and body fat

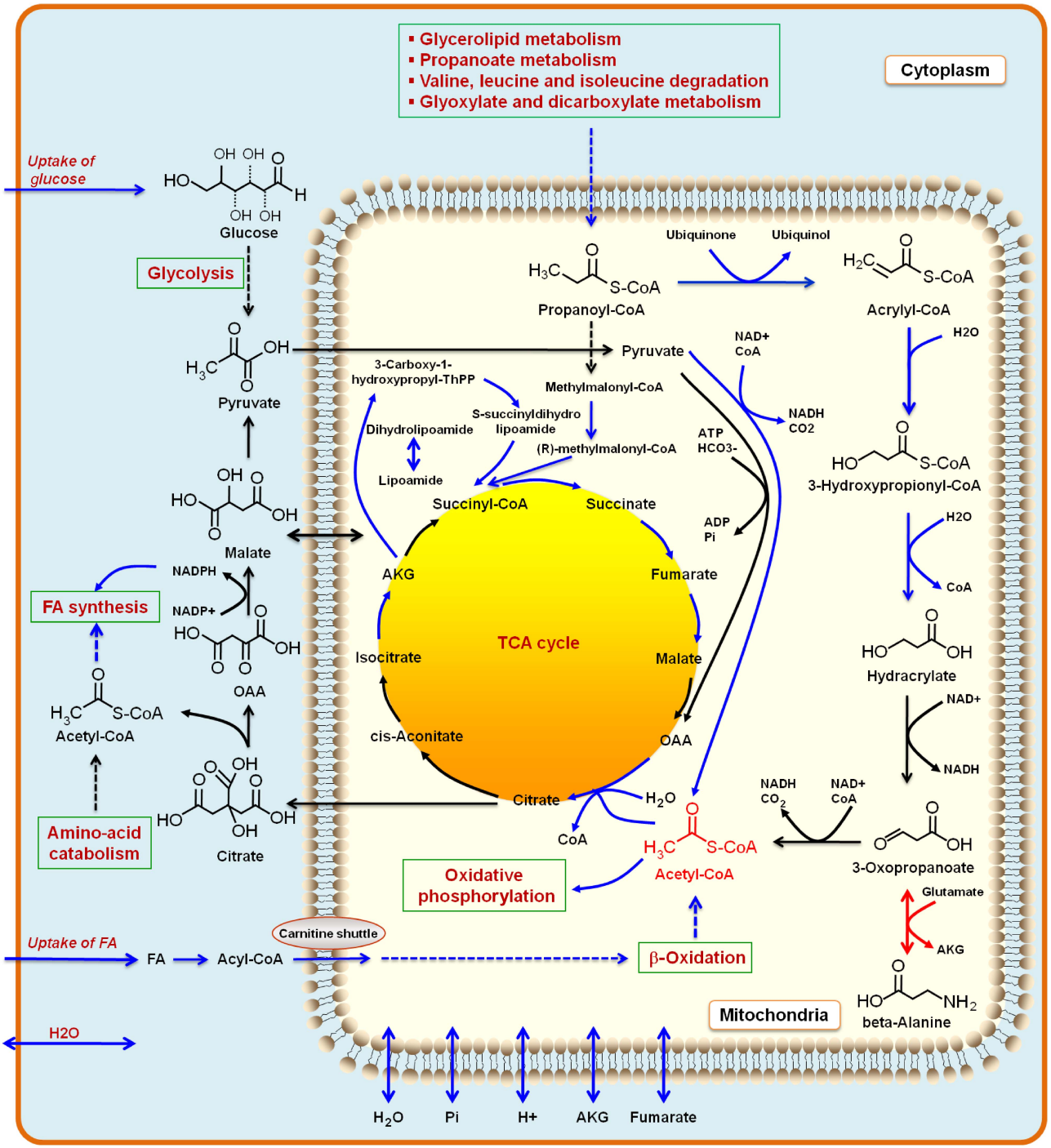


Figure 7 Illustration of the reactions and metabolites obtained from random sampling. The random sampling algorithm identifies set of reactions that are transcriptionally regulated and enzymes with transcriptional regulation showed enrichment. This suggests that the regulation of metabolism in adipocytes in obese and lean subjects has evolved to contain a few flux-regulating potential transcription factors that could be the target for genetic manipulations to redirect fluxes. The random sampling algorithms indicate that uptake of glucose, fatty acids (FAs) and water, FA synthesis and mitochondrial metabolic processes such as oxidative phosphorylation, β-oxidation, TCA cycle and β-alanine metabolism is downregulated (blue arrows in figure). Increasing the metabolic activity of mitochondria by increasing the concentration of mitochondrial acetyl-CoA may provide new strategies for treatment of obesity. One way to increase the mitochondrial acetyl-CoA concentration may be through increasing the uptake of beta-alanine as found from our random sampling results. Some of the major known biological pathways of adipocytes are shown in green boxes.

composition (Hoffman *et al*, 2006). Furthermore, rat studies reported that beta-alanine decreases the LPL enzyme activity in adipose tissue (Prabha *et al*, 1988) that may help to decrease the uptake of FAs to be stored in adipocytes. Our results suggest that increasing the level of beta-alanine in obese subjects may help to decrease the fat composition in obese subjects.

Furthermore, we have identified new candidates for potential therapeutic targets for obesity and their association with the obesity has been reported in different studies. One of the most significant results from the Reporter Metabolite analysis for upregulated genes was androsterone and its precursor 5alpha-androstane-3,17-dione. In obese subjects, increased secretion of androsterone was reported in serum concentrations using various indices calculated from urinary steroid excretion rates (Vierhapper *et al*, 2004). The measurement of androsterone-to-etiocholanolone ratio in urine is commonly used as an indicator of activities of 5α- and 5β-reductases in both male and female subjects and the ratio increased with indexes of insulin resistance, which is an adverse outcome of obesity (Tomlinson *et al*, 2008). Our results indicated that metabolism of obese subjects around androsterone increases and the detection of androsterone level in plasma may represent a marker for altered metabolism in obese subjects. 5alpha-pregnane-3,20-dione and 3alpha-hydroxy-5alpha-pregnan-20-one obtained from Reporter Metabolites analysis are other metabolites involved in steroid hormone synthesis specifically in pathway of progesterone metabolism. Several studies reported that sex steroids affect the adipose tissue deposition through the regulation of preadipocyte proliferation and/or differentiation as well as lipogenesis and/or lipolysis of differentiated adipocytes (Anderson *et al*, 2001). Zhang *et al* (2009) previously studied the comparison of progesterone metabolite formation in preadipocytes and lipid-storing adipocytes and they reported that the production of progesterone metabolites was significantly increased. Our study signified that metabolism around these two metabolites increase in obese subjects relative to metabolism in lean subjects.

Another high-ranking target for upregulated genes is ganglioside GM2 (GM2). Gangliosides, one of the major glycosphingolipids in mammals, have major roles as mediators for cell to cell or cell to matrix recognition and regulate the transmembrane signal transducers and cell proliferation. Gangliosides in adipose tissues are also associated with insulin signaling mechanisms and it is reported that series of gangliosides GM2, GM1 and GD1a are dramatically increased in adipose tissues of obese mice (Tanabe *et al*, 2009). The action of GM2 synthase also directs the metabolic flow of a-series gangliosides toward GD1a in mouse studies.

A third prominent group among the Reporter Metabolites for upregulated genes is the degradation products of heparan sulfate proteoglycans (HSPG) and keratan sulfate. These compounds are classified as glycosaminoglycans and attach to cell surface or extracellular matrix proteins. It has been reported that catalytically active adipose tissue LPL attaches to HSPG at the luminal surface of vascular endothelium (Olivecrona and Beisiegel, 1997; Lafontan, 2008) and hydrolyze the TAGs for uptake of FAs into the cell. The LPL moves between individual HSPG chains within the layer and this creates a high concentration of LPL along the surface layer of HSPG chains (Lookene *et al*, 1996). In the presence of heparin, more LPL is secreted and increased secretion was balanced by decreased degradation of LPL. There are special mechanisms that inhibit LPL and one mechanism is that LPL forms complexes with FAs (Bengtsson and Olivecrona, 1980). During the LPL hydrolysis and accumulation of FAs in the cells, the LPL is sequestered into enzyme FA complexes, lipolysis is reduced and eventually the binding of LPL to heparan sulfate is broken. If a high-affinity ligand (e.g., FAs, heparin and apoCII) is available, then the LPL detaches from the cell surface to heparan sulfate chains and without ligand in the medium, the LPL recycles into the cells where it is degraded. Furthermore, several studies have reported that more sulfated polysaccharide chains increase the affinity for binding of LPL (Olivecrona and Olivecrona, 2009). Keratan sulfate, a biomarker of proteoglycan degradation, can be expressed from stem cells in human SAT and its relevance with obesity has been reported earlier. Messier *et al* (2000) studied the effect of exercise and diet in weight loss in older obese adults with knee osteoarthritis and analyze the levels of total proteoglycan, keratan sulfate and interleukin-1 beta in their synovial fluid. It is reported that all participants of the study lost weight by changing their diet and exercise regimens and the level of keratan sulfate in synovial fluid decreased.

We envisage that *iAdipocytes1809* is a key step to better enable links between molecular processes and patient phenotypes and hereby enable patient stratification through identification of specific molecular mechanisms in adipocyte metabolism. Here, we demonstrated this by predicting differences in the formation of LDs and acetyl-CoA in lean and obese subjects. Besides enabling patient stratification, this may also lead to identification of novel therapeutic targets for obesity through combination of our model with high throughput patient data. *iAdipocytes1809* can also be used as a scaffold for a comprehensive whole adipocyte model that accounts for all of the annotated gene functions identified in adipocytes (Karr *et al*, 2012). Compared with the previously described adipose model, the here presented model is significantly larger in terms of metabolites/genes and reactions and this allows for identification of metabolic biomarkers that cannot be identified with *iAB586*. Furthermore, our model has undergone a thorough validation process where 250 metabolic functions were simulated and as our model also included a description of individual FAs and sterolesters it can much better simulate lipid metabolism, including simulation of lipid droplet formation. In conclusion, we demonstrated the high quality adipocyte GEM *iAdipocytes1809* is very well suited for integration of omics data and hereby result in a comprehensive understanding of adipocytes biology in response to obesity.

Materials and methods

Proteome data for adipocytes

The proteomic profiling of adipocytes in breast and two different soft tissues using immunohistochemistry was performed as previously described (Uhlen *et al*, 2005). Representative formalin paraffin-embedded material from donor blocks was punched (1 mm in diameter) and placed in a recipient block TMA that includes 46 normal tissues, 20 types of cancer and 47 cell lines (Kampf *et al*, 2012). Thereafter, 4-μm TMA sections were cut using a microtome and placed on super frost glass slides. Breast and two different soft tissue cores together with 708 previously prepared tissue cores were analyzed for protein expression using IHC (Kampf *et al*, 2004). The variability introduced by the individual experimental staining protocol, including the choice of antibody dilution and antigen retrieval methods, was addressed by the use of TMAs (Uhlen *et al*, 2005). Information from

databases, for example, Uniprot, ENSEMBL, and published data from PubMed were used as guides to establish if immunohistochemical results represent expected protein expression of intended target protein. In addition, other validation strategies were used to ensure antibody validity, for example, protein arrays, western blot and immunofluorescence experiments. When applicable, paired antibodies raised against separate, non-overlapping epitopes on the same target protein were used for extended validation (Uhlen *et al*, 2010). IHC-stained tissues were scanned and digitalized at $\times 20$ magnification. Annotation of high-resolution images was manually performed by certified pathologists. Relative expression was indicated with four different color codes ranging from strong (red), moderate (orange), weak (yellow) and no expression (white) (Kampf *et al*, 2004). Localization information of proteins was inferred from manually curated Uniprot data (Apweiler *et al*, 2011) and recently generated HPA data on intracellular localization of proteins (Lundberg and Uhlen, 2010).

Transcriptome data for SAT

Gene expression in SAT of subjects involved in SOS Sib Pair Study which includes nuclear families with BMI–discordant sibling pairs (BMI difference $\geqslant 10\,kg/m^2$) was measured by microarray. The study group consisted of 304 subjects (209 female and 95 male) divided into three different groups according to their BMI: lean ($18.5 < BMI < 25$), overweight ($25 \leqslant BMI < 30$) and obese ($30 \leqslant BMI$) (Supplementary Table S1). Total RNA was prepared as previously described (Carlsson *et al*, 2009) and gene expression in human SAT was measured using Affymetrix U133 Plus 2.0 microarrays (Affymetrix). Probe hybridization intensity values were summarized using to Probe Logarithmic Intensity Error (PLIER) and the robust quantile method was applied to get normalized expression values. The normalization of the microarrays was carried out using the Expression Console software from Affymetrix and the quality assessment was carried out using R Statistical and Computing language and the Bioconductor software (Gentleman *et al*, 2004).

Data availability

The annotation of the presence or absence of protein targets in adipocytes together with the high-resolution images is publically available through the HPA (http://www.proteinatlas.org).

GEM for adipocytes, *iAdipocytes1809* and HMR database is publically available in the Systems Biology Mark-up Language (SBML) format at Human Metabolic Atlas (http://www.metabolicatlas.org).

Gene expression from SAT of subjects from the SOS Sib Pair Study is publically available in Gene Expression Omnibus (GEO) database with the accession number GSE27916 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token = phadhqigowykote&acc = GSE27916.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We acknowledge Prof Fredrik Karpe and Dr Barbara A. Fielding, Oxford Centre for Diabetes, Endocrinology and Metabolism for providing adipose tissue uptake rates for lean and obese subjects. This work was supported by grants from the Knut and Alice Wallenberg Foundation, the Chalmers Foundation, the Wellcome Trust (GR079534) and the Swedish Research Council (K2010-55X-11285-13).

Author contributions: AM reconstructed *iAdipocytes1809* and validated the model together with RA. AM and RA performed the analysis of clinical data. CK and AA generated the proteome data for adipocytes. PJ, AJW, PF and LMC generated gene expression data for adipose tissue and IN normalized the gene expression data. JN and MU conceived the project. AM, RA and JN wrote the paper and all authors were involved in editing the paper.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput Biol* **8:** e1002518

Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J (2013) The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. *PLoS Comput Biol* doi:10.1371/journal.pcbi.1002980

Anderson LA, McTernan PG, Barnett AH, Kumar S (2001) The effects of androgens and estrogens on preadipocyte proliferation in human adipose tissue: influence of gender and site. *J Clin Endocr Metab* **86:** 5045–5051

Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Antunes R, Barrell D, Bely B, Bingley M, Binns D, Bower L, Browne P, Chan WM, Dimmer E, Eberhardt R, Fazzini F, Fedotov A, Foulger R, Garavelli J, Castro LG *et al* (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* **39:** D214–D219

Arner P, Bernard S, Salehpour M, Possnert G, Liebl J, Steier P, Buchholz BA, Eriksson M, Arner E, Hauner H, Skurk T, Ryden M, Frayn KN, Spalding KL (2011) Dynamics of human adipose lipid turnover in health and metabolic disease. *Nature* **478:** 110–113

Auffray C, Chen Z, Hood L (2009) Systems medicine: the future of medical genomics and healthcare. *Genome Med* **1:** 2

Bengtsson G, Olivecrona T (1980) Lipoprotein-lipase - mechanism of product inhibition. *Eur J Biochem* **106:** 557–562

Bordbar A, Feist AM, Usaite-Black R, Woodcock J, Palsson BO, Famili I (2011) A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Syst Biol* **5:** 180

Bordbar A, Lewis NE, Schellenberger J, Palsson BO, Jamshidi N (2010) Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Mol Syst Biol* **6:** 422

Bordel S, Agren R, Nielsen J (2010) Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Comput Biol* **6:** e1000859

Cakir T, Patil KR, Onsan Z, Ulgen KO, Kirdar B, Nielsen J (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. *Mol Syst Biol* **2:** 50

Canto C, Houtkooper RH, Pirinen E, Youn DY, Oosterveer MH, Cen Y, Fernandez-Marcos PJ, Yamamoto H, Andreux PA, Cettour-Rose P, Gademann K, Rinsch C, Schoonjans K, Sauve AA, Auwerx J (2012) The NAD(+) precursor nicotinamide riboside enhances oxidative metabolism and protects against high-fat diet-induced obesity. *Cell Metab* **15:** 838–847

Cao HM, Gerhold K, Mayers JR, Wiest MM, Watkins SM, Hotamisligil GS (2008) Identification of a lipokine, a lipid hormone linking adipose tissue to systemic metabolism. *Cell* **134:** 933–944

Cao YH (2010) Adipose tissue angiogenesis as a therapeutic target for obesity and metabolic diseases. *Nat Rev Drug Discov* **9:** 107–115

Carlsson LM, Jacobson P, Walley A, Froguel P, Sjostrom L, Svensson PA, Sjoholm K (2009) ALK7 expression is specific for adipose tissue, reduced in obesity and correlates to factors implicated in metabolic disease. *Biochem Biophys Res Commun* **382:** 309–314

Caveney E, Caveney BJ, Somaratne R, Turner JR, Gourgiotis L (2011) Pharmaceutical interventions for obesity: a public health perspective. *Diabetes Obes Metab* **13:** 490–497

Chang RL, Xie L, Xie L, Bourne PE, Palsson BO (2010) Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput Biol* **6:** e1000938

Cheong H, Lu C, Lindsten T, Thompson CB (2012) Therapeutic targets in cancer cell metabolism and autophagy. *Nat Biotechnol* **30:** 671–678

Cotter D, Maer A, Guda C, Saunders B, Subramaniam S (2006) Lmpd: Lipid Maps Proteome Database. *Nucleic Acids Res* **34:** D507–D510

Cristancho AG, Lazar MA (2011) Forming functional fat: a growing understanding of adipocyte differentiation. *Nat Rev Mol Cell Bio* **12:** 722–734

Croft D, O'Kelly G, Wu GM, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H *et al* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* **39:** D691–D697

Dean JT, Tran L, Beaven S, Tontonoz P, Reue K, Dipple KM, Liao JC (2009) Resistance to diet-induced obesity in mice with synthetic glyoxylate shunt. *Cell Metab* **9:** 525–536

Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* **36:** D344–D350

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA* **104:** 1777–1782

Farese RV, Walther TC (2009) Lipid droplets finally get a little R-E-S-P-E-C-T. *Cell* **139:** 855–860

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E *et al* (2011) Ensembl 2011. *Nucleic Acids Res* **39:** D800–D806

Frayn KN (2002) Adipose tissue as a buffer for daily lipid flux. *Diabetologia* **45:** 1201–1210

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge YC, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G *et al* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5:** R80

Gille C, Bolling C, Hoppe A, Bulik S, Hoffmann S, Hubner K, Karlstadt A, Ganeshan R, Konig M, Rother K, Weidlich M, Behre J, Holzhutter HG (2010) HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol Syst Biol* **6:** 411

Gross DA, Zhan C, Silver DL (2011) Direct binding of triglyceride to fat storage-inducing transmembrane proteins 1 and 2 is important for lipid droplet formation. *Proc Natl Acad Sci USA* **108:** 19581–19586

Hao T, Ma HW, Zhao XM, Goryanin I (2010) Compartmentalization of the Edinburgh Human Metabolic Network. *Bmc Bioinformatics* **11:** 393

Harkewicz R, Dennis EA (2010) Applications of mass spectrometry to lipids and membranes. *Annu Rev Biochem* **80:** 301–325

Hoffman J, Ratamess N, Kang J, Mangine G, Faigenbaum A, Stout J (2006) Effect of creatine and beta-alanine supplementation on performance and endocrine responses in strength/power athletes. *Int J Sport Nutr Exerc Metab* **16:** 430–446

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4:** 44–57

Jerby L, Shlomi T, Ruppin E (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol* **6:** 401

Kampf C, Andersson AC, Wester K, Björling E, Uhlén M, Pontén F (2004) Antibody-based tissue profiling as a tool for clinical proteomics. *Clin Proteomics* **1:** 285–299

Kampf C, Olsson I, Ryberg U, Sjöstedt E, Pontén F (2012) Production of tissue microarrays, immunohistochemistry staining and digitalization within the human protein atlas. *J Vis Exp* **63:** e3620

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38:** D355–D360

Karlstaedt A, Fliegner D, Kararigas G, Ruderisch HS, Regitz-Zagrosek V, Holzhutter HG (2012) CardioNet: a human metabolic network suited for the study of cardiomyocyte metabolism. *BMC Syst Biol* **6:** 114

Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival Jr B, Assad-Garcia N, Glass JI, Covert MW (2012) A whole-cell computational model predicts phenotype from genotype. *Cell* **150:** 389–401

Kusminski CM, Scherer PE (2012) Mitochondrial dysfunction in white adipose tissue. *Trends Endocrinol Metab* **23:** 435–443

Lafontan M (2008) Advances in adipose tissue metabolism. *Int J Obes* **32:** S39–S51

Lago F, Dieguez C, Gomez-Reino J, Gualillo O (2007) Adipokines as emerging mediators of immune response and inflammation. *Nat Clin Pract Rheumatol* **3:** 716–724

Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, Cheng JK, Patel N, Yee A, Lewis RA, Eils R, Konig R, Palsson BO (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat Biotechnol* **28:** 1279–1285

Lookene A, Chevreuil O, Ostergaard P, Olivecrona G (1996) Interaction of lipoprotein lipase with heparin fragments and with heparan sulfate: stoichiometry, stabilization, and kinetics. *Biochemistry* **35:** 12155–12163

Lundberg E, Uhlen M (2010) Creation of an antibody-based subcellular protein atlas. *Proteomics* **10:** 3984–3996

Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, Goryanin I (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* **3:** 135

Mardinoglu A, Nielsen J (2012) Systems medicine and metabolic modelling. *J Intern Med* **271:** 142–154

McQuaid SE, Hodson L, Neville MJ, Dennis AL, Cheeseman J, Humphreys SM, Ruge T, Gilbert M, Fielding BA, Frayn KN, Karpe F (2011) Downregulation of adipose tissue fatty acid trafficking in obesity a driver for ectopic fat deposition? *Diabetes* **60:** 47–55

Messier SP, Loeser RF, Mitchell MN, Valle G, Morgan TP, Rejeski WJ, Ettinger WH (2000) Exercise and weight loss in obese older adults with knee osteoarthritis: A preliminary study. *J Am Geriatr Soc* **48:** 1062–1072

Nielsen J (2009) Systems biology of lipid metabolism: from yeast to human. *FEBS Lett* **583:** 3905–3913

Nielsen J (2012) Translational and systems medicine introduction. *J Intern Med* **271:** 108–110

Nookaew I, Svensson PA, Jacobson P, Jernas M, Taube M, Larsson I, Andersson-Assarsson JC, Sjostrom L, Froguel P, Walley A, Nielsen J, Carlsson LM (2012) Adipose tissue resting energy expenditure and expression of genes involved in mitochondrial function are higher in women than in men. *J Clin Endocrinol Metab* **98:** E370

Olivecrona G, Beisiegel U (1997) Lipid binding of apolipoprotein CII is required for stimulation of lipoprotein lipase activity against apolipoprotein CII-deficient chylomicrons. *Arterioscl Throm Vas* **17:** 1545–1549

Olivecrona T, Olivecrona G (2009) The Ins and Outs of adipose tissue. *Cell Lipid Metab* 315–369

Patil KR, Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc Natl Acad Sci USA* **102:** 2685–2689

Patterson BW, Horowitz JF, Wu GY, Watford M, Coppack SW, Klein S (2002) Regional muscle and adipose tissue amino acid metabolism in lean and obese women. *Am J Physiol Endocrinol Metab* **282:** E931–E936

Peinado JR, Pardo M, de la Rosa O, Malagon MM (2012) Proteomic characterization of adipose tissue constituents, a necessary step for understanding adipose tissue complexity. *Proteomics* **12:** 607–620

Prabha ANL, Leelamma S, Kurup PA (1988) Similar effects of beta-alanine and taurine in cholesterol-metabolism. *J Biosci* **13:** 263–268

Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* **6:** R2

Tanabe A, Matsuda M, Fukuhara A, Miyata Y, Komuro R, Shimomura I, Tojo H (2009) Obesity causes a shift in metabolic flow of gangliosides in adipose tissues. *Biochem Bioph Res Co* **379:** 547–552

Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* **5:** 93–121

Tomlinson JW, Finney J, Gay C, Hughes BA, Hughes SV, Stewart PM (2008) Impaired glucose tolerance and insulin resistance are associated with increased adipose 11 beta-hydroxysterold dehydrogenase type 1 expression and elevated hepatic 5 alpha-reductase activity. *Diabetes* **57:** 2652–2660

Uhlen M, Bjorling E, Agaton C, Szigyarto CA, Amini B, Andersen E, Andersson AC, Angelidou P, Asplund A, Asplund C, Berglund L, Bergstrom K, Brumer H, Cerjan D, Ekstrom M, Elobeid A, Eriksson C, Fagerberg L, Falk R, Fall J *et al* (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics* **4:** 1920–1932

Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Bjorling L, Ponten F (2010) Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* **28:** 1248–1250

Vierhapper H, Nowotny P, Waldhausl W (2004) Production rates of cortisol in obesity. *Obes Res* **12:** 1421–1425

Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, Mandal R, Sinelnikov I, Xia JG, Jia L, Cruz JA, Lim E, Sobsey CA, Shrivastava S, Huang P, Liu P *et al* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* **37:** D603–D610

Xie XT, Yi ZP, Bowen B, Wolf C, Flynn CR, Sinha S, Mandarino LJ, Meyer C (2010) Characterization of the human adipocyte proteome and reproducibility of protein abundance by one-dimensional gel electrophoresis and HPLC-ESI-MS/MS. *J Proteome Res* **9:** 4521–4534

Zhang YH, Nadeau M, Faucher F, Lescelleur O, Biron S, Daris M, Rheaume C, Luu-The V, Tchernof A (2009) Progesterone metabolism in adipose cells. *Mol Cell Endocrinol* **298:** 76–83

Paper VI

Genome-scale metabolic modeling of hepatocytes leads to identification of serine deficiency in non-alcoholic fatty liver disease

Mardinoglu, A.[*], **Agren, R.**[*], Kampf, C., Uhlen, M. and Nielsen, J.

Submitted

[*]Authors contributed equally

Genome-scale metabolic modeling of hepatocytes leads to identification of serine deficiency in non-alcoholic fatty liver disease

Running title: Serine deficiency in NAFLD

Adil Mardinoglu[1, #], Rasmus Agren[1, #], Caroline Kampf[2], Anna Asplund[2], Mathias Uhlen[3] and Jens Nielsen[1,*]

[1] Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

[2] Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

[3] Department of Proteomics, School of Biotechnology, AlbaNova University Center, Royal Institute of Technology (KTH), Stockholm, Sweden

[*]Corresponding author

[#] These authors contributed equally to this work.

E-mail: nielsenj@chalmers.se
Tel: +46 31 772 3804
Fax: +46 31 772 3801

KEYWORDS: Metabolism; genome-scale metabolic modeling; hepatocytes; non-alcoholic fatty liver disease (NAFLD); non-alcoholic steatohepatitis (NASH)

ABSTRACT

Several liver disorders result from perturbations in the metabolism of hepatocytes and the underlying mechanisms of such disorders can be outlined through the use of genome-scale metabolic models (GEMs). Here we reconstructed a consensus GEM for hepatocytes, iHepatocytes2260, by including extensive description of lipid metabolism and expanding on previously published models. iHepatocytes2260 is reconstructed based on the updated Human Metabolic Reaction database and large scale proteomics data, which provides experimental evidence for the reactions incorporated into the model. The reconstruction process also enabled reevaluation and improved annotation of the proteomics data by using the network centric view made possible by iHepatocytes2260. The resulting GEM was employed for the analysis of transcriptomics data obtained from non-alcoholic fatty liver disease (NAFLD) patients, with the purpose of identifying biomarkers and therapeutic targets. Blood concentrations of 5-methyltetrahydrofolate, 5-formyltetrahydrofolate, chondroitin and heparan sulfates were predicted to be suitable for diagnosing non-alcoholic steatohepatitis (NASH) and staging NAFLD. Furthermore, we identified potential therapeutic targets (PSPH, SHMT1 and BCAT1) for treatment of NASH.

Introduction

Hepatocytes have a wide range of physiological functions including production of bile and hormones, removal of toxic substances, homeostatic regulation of the plasma constituents and synthesis of most plasma proteins (1). The hepatocytes, the most metabolically active cell types in human, play a major role in overall human metabolism. Deficiency or alterations in the metabolism of hepatocytes can lead to complicated disorders such as hepatitis, nonalcoholic fatty liver disease (NAFLD), cirrhosis and liver cancer, which are serious threats to public health (2). NAFLD is considered as the hepatic manifestation of obesity and metabolic syndrome, and encompasses a spectrum of pathological changes; ranging from simple fatty liver (steatosis) to non-alcoholic steatohepatitis (NASH) (3).

Even though it is well known that lipid accumulation in the liver is a hallmark of NAFLD (4), the underlying mechanisms leading to steatosis and further transition to NASH still remain elusive. It is therefore difficult to track the onset and progression or to diagnose and design effective therapeutic techniques. The adverse outcomes of this pathology may possibly be prevented once the molecular mechanisms involved in the metabolism of hepatocytes are deciphered (5). However, this requires understanding of the coordinated behavior of a very large number of interconnected metabolic reactions and metabolites. Relating this behavior with disease and patients have been a major focus in biomedicine (6). A systems biology approach, based on the employment of genome-scale metabolic models (GEMs), can be used to extend our understanding of these molecular mechanisms, which in turn may enable future therapeutic discoveries (7).

GEMs represent the current knowledge of metabolism generated through the integration of genetic and biochemical studies coupled with cellular, physiological and clinical data (8). Several generic (non-cell type-specific) GEMs for human metabolism have been previously constructed (9-12). However, neither of these generic networks contain extensive lipid metabolism, which is necessary in order to study the effect of lipids on the underlying molecular mechanism of NAFLD. Recently, a large scale GEM for adipocytes, iAdipocytes1809 with a strong focus on lipid metabolism was presented (13) and this model can provide a base for further integration of lipid metabolism into generic networks.

There is currently no efficient treatment for NASH (14) and new therapeutic approaches are in great demand. This study represents an attempt to

rationally identify biomarkers and therapeutic targets using genome-scale metabolic modeling. In order to reconstruct a high-quality model for hepatocytes, we combined clinical, biochemical and genetic studies such as expression, localization and functional characteristics of the proteins. We first significantly expanded the content of our Human Metabolic Reaction (HMR) database by including extensive lipid metabolism. This represents an important step forward since lipids have major effect on the development of NAFLD and other metabolic diseases (15). Secondly, we reconstructed a consensus GEM for hepatocytes, iHepatocytes2260, though the use of the HMR database and proteomics data. We also merged previously published hepatocytes models in order to cover the entire known metabolic functions of hepatocytes and incorporated additional clinical data (e.g. liver tissue and plasma fatty acid contents in lipid structures). During the reconstruction process, we reevaluated the hepatocyte proteomics data after identifying proteins which were included in the model to ensure network connectivity, but were assessed as absent in hepatocytes in the Human Protein Atlas (HPA) (16). Finally, we employed iHepatocytes2260 for the analysis of differential gene expression data from liver tissues of subject groups with NAFLD. This lead to new insights into the molecular mechanisms involved in NASH which were used for identification of possible metabolic biomarkers and therapeutic targets for treatment of NASH (Figure 1).

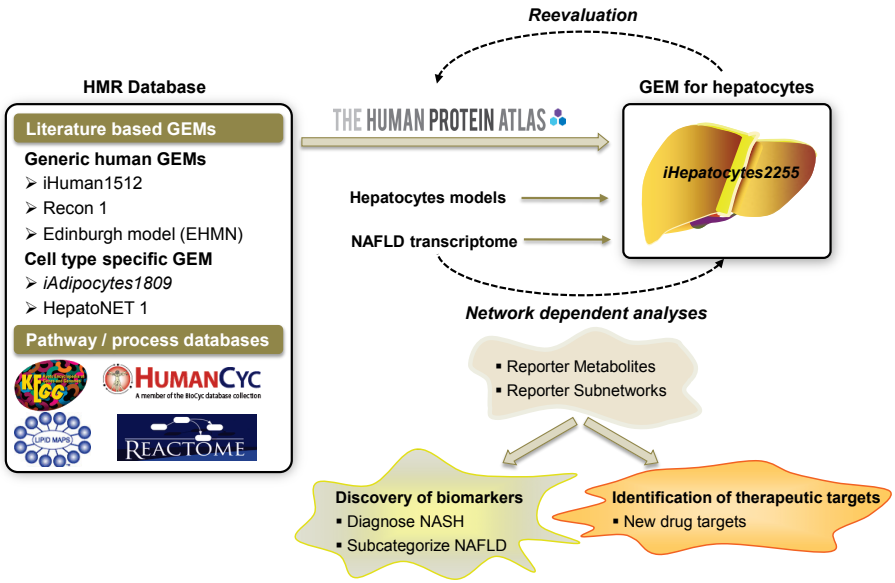


Figure 1. Effective therapeutic approach through genome-scale metabolic modelling. Schematic illustration of how a consensus genome-scale metabolic model (GEM) for hepatocytes, iHepatocytes2260 may contribute to the development of effective therapeutic approaches for nonalcoholic fatty liver disease (NAFLD) patients. The Human Metabolic Reaction (HMR) database was constructed through the use of previously published GEMs as well as pathway databases including KEGG, HumanCyc, Reactome and LIPIDMAPS Lipidomics Gateway. Elements of lipid metabolism were included in the HMR database in order to understand the effect of the lipids and their interactions during the appearance of NAFLD. The HMR database was used for reconstruction of iHepatocytes2260 based on proteomics data in the Human Protein Atlas (HPA), transcriptomics data in NAFLD patients as well as previously published hepatocytes models. During the reconstruction process, iHepatocytes2260 was employed for the improvement of proteomics data through identification and re-annotation of putative false negative proteins. The resulting GEM was used for the analysis of clinical data obtained from NAFLD patients in order to investigate the alterations in their hepatocytes metabolism and eventually for discovery of biomarkers and identifying therapeutic targets. Through our systems biology based analysis, potential biomarkers for diagnosing non-alcoholic steatohepatitis (NASH) and for subcategorizing NAFLD patients were discovered. Furthermore, a list of candidate therapeutic targets was identified in order to develop efficient treatments for NASH patients.

Results

Expansion of Human Metabolic Reaction (HMR) database

In order to provide a resource for automated and semi-automated reconstruction of cell type specific GEMs, we previously constructed the HMR database (12). This comprehensive database, together with the INIT algorithm, have been employed for automated generation of cell type specific GEMs (12). These models form the basis for the Human Metabolic Atlas (http://www.metabolicatlas.org), which is a web-based resource for human metabolism. Here, we expanded the HMR database by incorporating extensive lipid metabolism, which accounts for individual fatty acids (FAs) rather than relying on generic pool metabolites. The HMR database is formulated using 59 FAs (Table S1), which enables mapping and integration of lipidomics data. Integration of lipid metabolism (e.g. formation of lipid droplets and lipoproteins) may allow not only for understanding the contribution of lipids to the

development of diseases, but also allow for study of the relationship between lipid metabolism and cellular molecular mechanisms (13).

Reactions are included in the HMR database depending on evidence from previously published models and databases (Table S2) or on the availability of specific experimental evidence for the occurrence of the reaction. The HMR database is the largest biochemical reaction database for human metabolism in terms of number of reactions/genes/metabolites as well as in terms of covering most parts of metabolism.

Consensus genome-scale metabolic model for hepatocytes and improved annotation of proteome

Cell-type specific GEMs can be employed for the analysis of high throughput patient (-omics) data, simulation of the metabolic differences under health and disease states and eventually for predicting the cellular phenotype (17). Previously, several GEMs for hepatocytes including HepatoNET 1 (1), iLJ1046 (18), iAB676 (19) and iHepatocyte1154 (12) have been reconstructed. Here, we generated a consensus GEM for hepatocytes, iHepatocytes2260 based on proteomics data (Dataset 1) and the updated HMR database. iHepatocytes2260 contains all of the protein coding genes and associated reactions in previously published liver models (Figure 2a). In addition to the proteomics data and previous models, protein coding genes are also included in iHepatocytes2260 based on transcriptomics data and connectivity (Figure 2b). Reactions and associated proteins were assigned into eight different compartments following our HMR database standard (Agren et al., 2012a) based on the subcellular localization of the proteins in HPA and Uniprot. The protein localization information in HPA and Uniprot were assigned to relevant compartments in the HMR database (Dataset 2). A confidence score for each protein was calculated based on the availability of knowledge in HPA and Uniprot (Dataset 3). Furthermore, the connectivity in the model was checked carefully, such that all metabolites consumed in one reaction should be able to be produced by another reaction or they should be taken up from the plasma. Finally additional clinical data for plasma and hepatocyte lipid concentrations for individual FAs were incorporated into the model (Dataset 4).

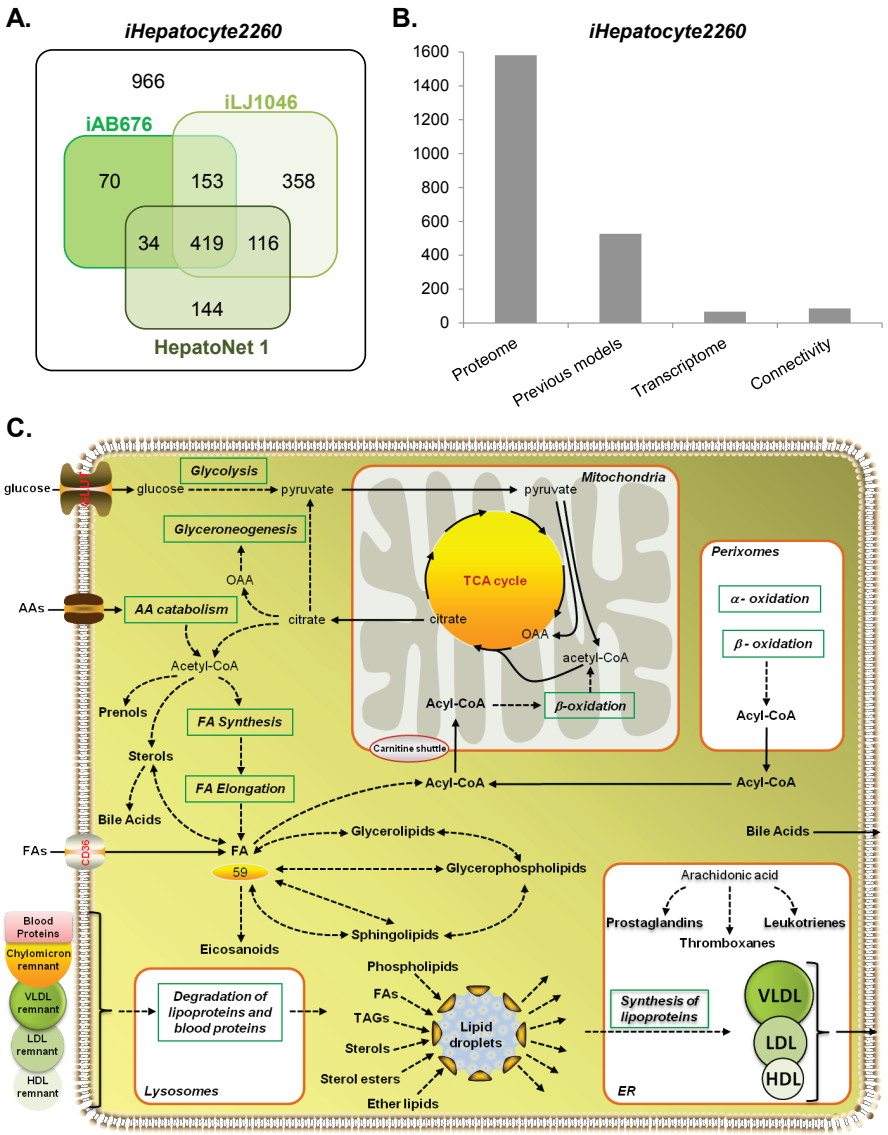


Figure 2. Consensus genome-scale metabolic model for hepatocytes, iHepatocytes2260. A) iHepatocytes2260 is reconstructed through the use of the Human Metabolic Reaction database and improved annotation of proteomics data in the HPA. This high-quality model includes the structure of major lipid metabolism in hepatocytes as well as all of the reactions and associated genes to those reactions in previously published GEM for hepatocytes. The overlapping of the genes in iHepatocytes2260 and previously published models are presented. 966 new protein coding genes were included into iHepatocytes2260 primarily based on proteomics evidence provided by HPA. B) Genes and associated reactions in iHepatocytes2260 are included into the model based on the high quality proteome, transcriptome, previously published models as well as the connectivity. The overall distribution is shown. C) iHepatocytes2260 contains extensive lipid metabolism that is known to exist in hepatocytes, in addition to other known metabolic pathways. In the model, 59 different individual fatty acids are used in order to allow the integration of high quality lipidomics data rather than generic pool names. The model can uptake the remnants of chylomicrons, very-low-density lipoprotein (VLDL), low-density lipoproteins (LDL) and high-density lipoproteins (HDL) and can form and degrade lipid droplets (LDs). Moreover the model can synthesis VLDL, LDL and HDL and secrete it to the blood. Some of the important elements of lipid metabolism are shown.

iHepatocytes2260 differs from previously published hepatocyte GEMs primarily in terms of coverage in lipid metabolism. Among the new lipid related functions are uptake of the remnants of lipoproteins (chylomicrons, very-low-density lipoprotein (VLDL), low-density lipoproteins (LDL) and high-density lipoproteins (HDL)), the formation and degradation of lipid droplets (LDs) and secretion of synthesized lipoproteins (VLDL, LDL, HDL) (Figure 2c).

We tested iHepatocytes2260 by simulating 256 different biologically defined metabolic functions (e.g. the synthesis of FAs, amino acids, cholesterol and bile acids) (Dataset 5) that is known to occur in hepatocytes using the RAVEN Toolbox (20). Furthermore, the ability of iHepatocytes2260 for performing gluconeogenesis was demonstrated using experimentally measured secretion rates for glucose and albumin and uptake rates for glycerol, lactate,

amino acids and FAs in primary rat hepatocytes (6) (Dataset 6).

The HPA covers the annotated expression of proteins and their subcellular localization in major human cell types, cancer and cell lines (16). Relative abundance of proteins encoded by 14,078 genes in hepatocytes were analyzed with 17,296 high-throughput generated affinity-purified antibodies (Dataset 1). The model reconstruction process was in excellent agreement with the protein profiling of hepatocytes in HPA. During the implementation of the metabolic tasks in iHepatocytes2260, merely 61 (~1,7%) out of 3,673 proteins in the HMR database and associated reactions had to be integrated into the model to maintain the functionality even though they have been reported to be non-expressed in hepatocytes according to the HPA. We re-analyzed the immunohistochemistry (IHC) data of these 61 proteins and found that 20 (0.5%) of these proteins actually should display presence in liver (Dataset 7). Initial disconcordant data were due to the suboptimal titration of the antibody, misinterpretation of weak IHC staining or due to interference with other cell types besides hepatocytes present in liver (e.g. kupffer cells and sinusoids). Nine (0,2%) of the investigated proteins showed more concordant results to the mathematic model when re-analyzed using another antibody targeting the same protein. 15 proteins (0,4%) with negative IHC data were kept as negative in HPA data since limited literature was available, and/or concordant results were seen in subsets of the remaining panel of tissues included in the HPA high-throughput set up. The remaining 17 proteins (0,5%) are believed to be inaccurately assessed by IHC due to technical issues, such as antigen recognition due to antigen conformational changes, fixation or sub optimal antibody. Thus, clearly our analysis showed the power of using reconstructed metabolic networks for improving the annotation of experimental expression data.

Discovery of biomarkers for non-alcoholic steatohepatitis (NASH)

NAFLD is progressively diagnosed worldwide (21), is tightly associated with obesity, type 2 diabetes, insulin-resistance, hypertension and represents a severe risk for development of cirrhosis and hepatocellular carcinoma (HCC) (22). Despite its severe drawbacks, liver biopsy is still the most common procedure for diagnosing NASH (14). Thus, there is a need for identifying metabolic biomarkers to diagnose NASH, as well as to subcategorize the NAFLD patients without taking biopsies. A metabolic biomarker can be defined as a metabolite which is secreted to the blood where its level differs between two different states.

We analyzed the liver gene expression data obtained from NASH (severe stage of NAFLD) patients (Figure S1) to understand the multi-factorial nature of its appearance by using iHepatocytes2260 as a scaffold for data analysis. We identified clear metabolic differences by performing a pair wise analysis of the gene expression of subjects with NASH with and without fatty liver (FL) versus healthy samples using the Reporter Metabolite algorithm (23).

Reporter Metabolite analysis allows for identification of areas in the metabolic network with significant enrichment of gene expression changes, and Reporter Metabolites represent key regions of the metabolic network that are significantly perturbed between the compared conditions (24). The analyses for two NASH patient groups were performed independently and a total of 50 statistically significant ($p<0.05$) Reporter Metabolites for NASH with and without FL versus healthy samples were identified. The association of the Reporter Metabolites with up and down regulated genes and their metabolic subsystems classified in the HMR database are presented in Figure 3. In addition to the known subsystems involved in the progression of the NASH, e.g. cholesterol biosynthesis, folate, vitamin B6, porphyrin, nucleotide, eicosanoid and amino acid metabolism (25, 26), new Reporter Metabolites involved in the N-glycan metabolism and biosynthesis of chondroitin sulfate (CS), a proteoglycan (PG) were identified. PGs are composed of glycosaminoglycans including CS and heparan sulfate (HS) and core proteins. The biosynthesis of PGs starts with the xylosylation of serine residues in core proteins.

In order to gain more knowledge about the metabolic differences around PGs, the detailed CS and HS biosynthesis pathway and the gene expression changes in NASH with and without FL patients versus healthy subjects are presented in Figure S2. It is observed that the expression of the genes involved in the CS biosynthesis are up regulated whereas the expression of genes involved in the biosynthesis of HS are down regulated (Table S3). CS and HS are implicated in cancer progression (27), one of the most severe outcomes of NASH. It is therefore, we predict that these changes in gene expression, in particular as it involves complete metabolic pathways, may correlate with a change in blood concentration of the pathway associated metabolites. Hence, the blood level of CS and HS can be regarded as a potential biomarker for diagnosing NASH.

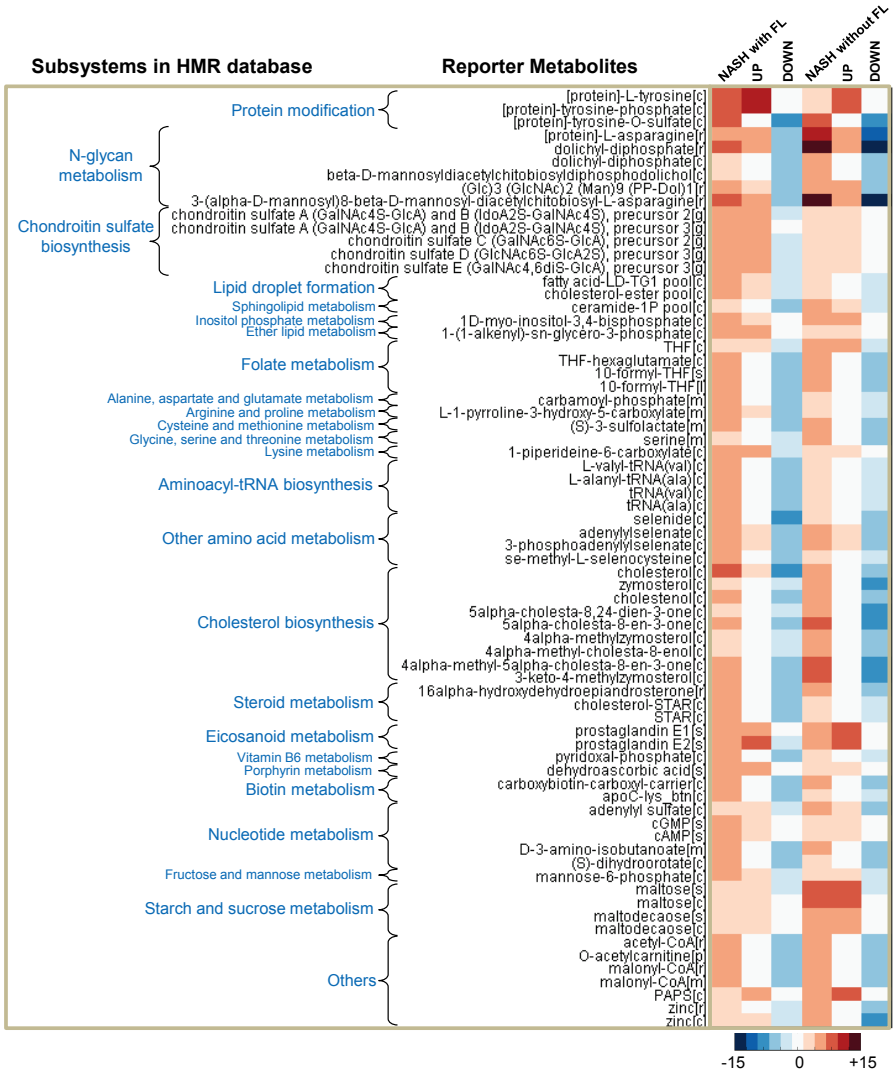


Figure 3. Reporter Metabolites through analysis of high-throughput data. Reporter metabolites that can provide global analysis of high-throughput data were identified through the analysis of transcriptomics data obtained from NAFLD patients. iHepatocytes2260 is employed in order to investigate the metabolic differences between the NASH patients and healthy subjects. In addition to the known pathways involved in the appearance of NASH, new subsystems including N-glycan metabolism and chondroitin sulfate metabolism have been discovered. The blood level of chondroitin sulfate which can be secreted and taken up by the blood is identified as a candidate biomarker for diagnosing NASH and staging NAFLD.

Identification of potential therapeutic targets for non-alcoholic steatohepatitis (NASH)

The Reporter Subnetwork algorithm identifies set of metabolic reactions which exhibit transcriptional correlation after a perturbation (in this case NASH) (24). We applied this algorithm to gain more insights into the molecular mechanisms involved in the appearance of NASH. After removing highly connected metabolites (e.g. cofactors) (Table S4) in iHepatocytes2260, the involved subnetworks in either NASH with or without FL were identified and they are presented in Figure 4A and Figure S3. The enzymes involved in the reactions are also represented in Figure 4A and related p-values and fold changes of their expression are shown in Table S5.

The Reporter Subnetwork analysis showed that the non-essential amino acids serine, glycine, glutamate, glutamine, aspartate, asparagine, alanine and the essential amino acid valine and methionine seem to be involved in the appearance of NASH. For reasons discussed later, serine, glycine and glutamate are of particular interest. Several metabolites involved in folate metabolism (e.g. tetrahydrofolate (THF), 5-methyl-THF, 5-formyl-THF and 5,10-methenyl-THF, 5,10-methylene-THF) were also identified in the Reporter Subnetwork analysis, and these metabolites are involved in the interconversion of serine, glycine and glutamate. The metabolism around THF changed in NASH patients and this

difference may be dependent on the uptake of 5-methyl-THF and 5-formyl-THF. We therefore propose that these metabolites could be used as biomarkers for diagnosis of NASH.

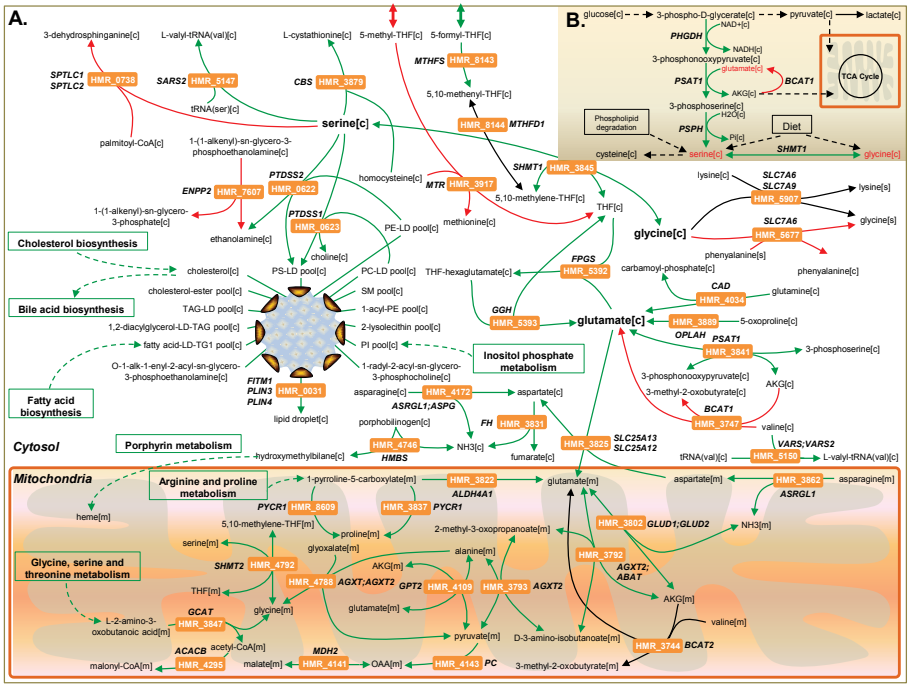


Figure 4. Reporter Subnetworks through mapping of high-throughput data. Reporter Subnetworks are transcriptionally correlated parts of the metabolism after a direct or indirect perturbation in response to NASH and significantly correlated connected subgraphs within the enzyme-interaction graph. Amino acids in particular serine, glutamate and glycine takes role in the appearance of NASH and the significant changes around these amino acids have been presented. Through our analysis, two candidate biomarkers (5-methyltetrahydrofolate and 5-formyltetrahydrofolate) as well as three different drug targets (PSPH, SHMT1 and BCAT1) have been identified.

Moreover, phosphatidylserine (PS), an essential component for formation of lipid droplets (LDs), was identified through our analysis. LDs have diverse roles in the cell, such as serving as storage for TAG and CEs or protecting the cell from excess lipids or lipophilic substances that may be toxic (28). The enzymes phosphatidylserine synthases (PTDSS1) and (PTDSS2) that catalyze the production of PS by condensation of phosphatidylcholine (PC) and phosphatidylethanolamine (PE), respectively, were significantly down regulated in NASH patients. The significant changes in the level of PS in cirrhotic (severe stage of NASH) livers was previously reported in a study on changes in lipid species in subjects with cirrhotic livers compared with healthy controls (29). Given that PS is essential for hepatocytes, we hypothesize that decreased activity of these enzymes may be associated with a decrease in the endogenous level of serine, which is the second most connected node in our identified Reporter Subnetworks (Figure 4A).

Serine is endogenously biosynthesized from a glycolytic intermediate, 3-phospho-D-glycerate. This three-step process is catalyzed by phosphoglycerate dehydrogenase (PHGDH), phosphoserine aminotransferase 1 (PSAT1) and phosphoserine phosphatase (PSPH), as shown in Figure 4B. An alternative synthesis pathway is via the reversible interconversion with glycine through hydroxymethyltransferases (SHMT1) and (SHMT2). Serine can also be derived from the diet and the degradation of protein and/or phospholipids. Serine plays a key role in the central metabolism, where it is involved in the formation of macromolecules, including lipids (sphingosine and phosphatidylserine), and other building blocks and cofactors, such as protein (glycine and cysteine), creatine, porphyrins, glutathione and nucleotides (30).

Through differential analysis of transcriptomics data from the NASH patients, it was also observed that gene expression of several enzymes that either use serine as substrate or produce it as a product, including CBS (cysteine synthesis), SARS2 (aminoacyl-tRNA biosynthesis), SHMT1 and SHMT2 (glycine synthesis) were significantly down regulated (p-values < 0.05) whereas SPTLC1 and SPTLC2 (sphingosine synthesis) were significantly up regulated (Table S5). Down regulation of CBS that catalyzes the conversion of serine and homocysteine to L-cystathionine and up regulation of MTR that condensates homocysteine to methionine through the use of 5-methyl-THF

indicate that there are metabolic changes around homocysteine in NASH patients. Notably, it has been earlier reported that the plasma homocysteine level can be used for diagnosing NASH and classifying steatosis and NASH patients (31). It is not always straight forward to relate blood concentrations to gene expression levels of the involved enzymes, but our model-based analysis suggests a mechanistic explanation for this.

Taken together, the results suggest that the changes in the level of PS in liver (29) as well as the relative increase in the homocysteine blood level (31) is caused by decreased level of endogenous serine. In order to test this hypothesis, we checked the expression level of enzymes that catalyze the biosynthesis of serine in the liver of NASH patients, and it was observed that the expression levels of PHGDH, PSAT1, PSPH in serine synthesis pathway (SSP) and SHMT1 and SHMT2 enzymes were significantly down regulated (Table S6). Decreased levels of serine in NASH patients was supported by the plasma profiling of amino acids in NASH patients, and it was reported that the serine (15 % decrease, p-value=0.0568) level in the plasma is relatively decreased (32).

Equimolar amounts of serine and 2-Oxoglutarate (AKG) are synthesized in the SSP, and down regulation of reactions in SSP decrease the anaplerosis of glutamate to the TCA cycle in the form of AKG (33). Decreased level of serine also causes an accumulation of upstream glycolytic intermediates (34), and a decreased flux of mitochondrial AKG is compensated by an increased flux of pyruvate to oxaloacetate in a healthy cell. In order to investigate the occurrence of this mechanism in NASH patients, we examined all mitochondrial reactions involving pyruvate as reactant in iHepatocytes2260. We found that the corresponding genes were down regulated for five out of seven such reactions (Figure 5 and Table S7). Furthermore, we investigated the expression of level of mitochondrial pyruvate carriers (MPC1 and 2) and mitochondrial AKG/malate carrier (SLC25A11) and it was observed that their expression levels were down regulated in NASH patients. These indicate that the mitochondrial metabolic activity (TCA cycle) of hepatocytes is decreased in NASH patients comparing to healthy subjects. This is in agreement with findings in our previous study where we investigated the metabolic changes in the case of fat accumulation in adipocytes in response to obesity (13).

In the appearance of NASH, glutamate, which is the most connected node in our Reporter Subnetwork, plays a significant role as well. All enzymes linked to glutamate except the branched chain amino-acid transaminase 1 (BCAT1) that convert AKG and valine to glutamate in the cytosol are down regulated. In NASH patients, up regulation of BCAT1 is previously reported (25) and our study identifies its mechanism in the appearance of NASH. The simultaneous up regulation of BCAT1 and down regulation of PSPH could point to an imbalance in intercellular level of AKG and glutamate. This could possible result in an accumulation of intracellular glutamate, which would then be compensated for by reduced uptake from/increased export to the blood. Notably, a previous study detected a significantly higher level of glutamate (60% increase, p=9.808E-09) in the plasma profiling of the amino acids in NASH patients (32). At the same time, this may result in a higher demand for valine in hepatocytes, and given that valine is an essential amino acid, this would arguably correlate with an increased uptake. Indeed, the same study (32) reported that the plasma valine concentration displayed significant changes (10% increase, p-value=0.012).

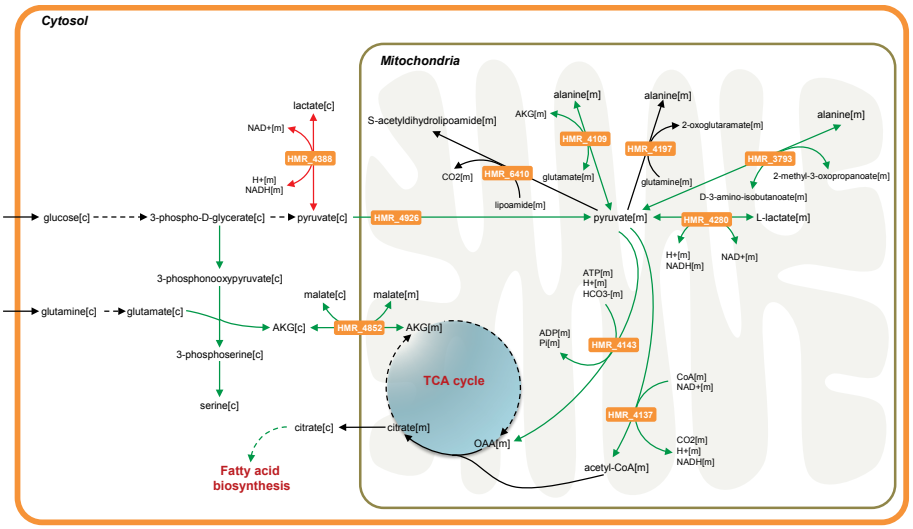


Figure 5. Mitochondrial dysfunction in NASH patients. Mitochondrial reactions involving pyruvate as a reactant in iHepatocytes2260. Arrows are colored based on the direction of change in expression of the corresponding genes in NASH with and without fatty liver samples versus healthy samples. Red arrows indicate over expression of the associated genes, whereas green arrows indicate under expression. Non-significant changes (p-value>0.05) is indicated with black arrows.

Discussions

In order to gain insights into the underlying molecular mechanisms of NASH, we reconstructed a consensus GEM for hepatocytes which allowed for study of the interactions between lipids and other cellular metabolic functions. This highly-curated GEM reconstructed though the use of the HMR database enables interpretation of systemic effects, provides deeper insight into omics data for better understanding of the genotype-phenotype relationship in NASH subjects, and allows for application of constraint-based modeling techniques to distinguish the NASH specific metabolic features. Based on liver transcriptomics data of NASH patients and systems biology based approaches, we proposed potential biomarkers and identified candidate therapeutic targets for NASH.

Based on our analysis, increasing the serine level in hepatocytes through the uptake of serine as a dietary supplement could be beneficial for NASH patients. Likewise activity loss of PHGDH in SSP in the brain, which causes low serine and glycine levels and affects neuronal function, is reversed by serine supplementation (35). The toxicity and the dosage of serine during its uptake through diet have been previously studied. Furthermore, long-term serine treatment decreased the homocysteine level in animal studies (36) and in humans in a single dose situations (37).

One other possible way to increase the serine level in order to offer the possibility for therapeutic interventions is activation of the enzymes in SSP or SHMT1 and SHMT2 that converts glycine to serine. Three different enzymes constitute the SSP and it is earlier reported that PSPH is the rate-controlling enzyme for the SSP in liver (38). Activation of the SSP through the amplification of PSPH may also decrease the flux through pyruvate and lactate formation in cytosol since increased pyruvate and lactate levels were previously reported in NASH patients (32). Recently, Frayling et al. (39) performed a genome wide association study using 1,004 non-diabetic individuals and identified eight common genetic variants relevant to insulin sensitivity and type 2 diabetes that are strongly linked to NASH phenotype. Their findings are in agreement with our results and they reported that three variants were associated with serine levels, out of which one is in the PHGDH gene and other two are independently in the PSPH gene.

Boosting the serine level through SSP will also increase the flux on unregulated anaplerotic reactions that drive glutamine derived carbon in to the TCA cycle through increased level of AKG and result in increased TCA cycle flux (33). TCA cycle intermediates, besides being involved in driving energy production, are also used for biosynthesis of lipids (citrate), porphyrin (succinyl-CoA) and amino acids (AKG and oxaloacetate). AKG concentration in the cytosol can also be increased by inhibition of BCAT1, which converts AKG and valine to glutamate. Increasing the AKG level by either over expressing PSPH or inhibiting BCAT1 may change the NASH specific patterns to healthy patterns and it may potentially be used in order to develop an effective treatment for NASH.

In conclusion, we reconstructed a consensus GEM for hepatocytes, showed how reconstructed metabolic networks can be used for improving the annotation of experimental data and employed the model to gain more insight into the metabolic transformations associated with the development of NASH. Our analysis suggests that it is possible to diagnose NASH through identified metabolic biomarkers such as 5-methyl-THF, 5-formyl-THF,

CS and HS levels in blood. Furthermore, the development of therapeutics techniques based on the enhancement of endogenous serine and AKG levels may correct the underlying etiology of NASH. This could be achieved by activation (or elevated expression) of PSPH and SHMT1 and inhibition of BCAT1. This study demonstrates that a deeper understanding of the metabolic changes obtained through genome-scale metabolic modeling may allow for elucidating the unknown etiology of NASH, discovery of novel biomarkers, identification of drug targets and eventually development efficient treatment strategies.

Experimental procedures

Human Metabolic Reaction (HMR) database

The Human Metabolic Reaction (HMR) database was constructed (12) by integrating the elements of stoichiometric networks of human metabolism, Recon 1 (9) and EHMN (10, 11), as well as the KEGG database (40). In order to generate a generic network for studying the effect of the lipids on the cellular metabolism we expanded the coverage of the HMR database (HMR2.0). We first merged metabolism of lipids and lipoproteins in Reactome, a manually curated and peer-reviewed pathway database, (41) and the literature-based GEMs including Recon 1 (9), EHMN (10, 11) and HepatoNET 1, a manually reconstructed GEM for hepatocytes, (1). Extensive lipid metabolism involving 59 different fatty acids (FAs) in the comprehensive database for lipid biology for mammalian cells, Lipidomics Gateway, (42) were included in the network and the gaps in the resulting network were filled using public databases such as KEGG (40) and HumanCyc (43).

To ensure the standardization of the HMR database, all model components were extensively annotated with database identifiers. HMDB, Lipid Map (42), KEGG and ChEBI identifiers were assigned for each metabolite and KEGG ids and EC numbers were assigned for each reaction. Alternative genes associated to reactions were assigned using Uniprot (44) and Lipid Map proteome database (45) using EC numbers. The resulting HMR database contains 3,673 genes, 6,000 metabolites (3,160 unique metabolites) and 8,110 reactions and 74% of the reactions associated to one or more genes. The generated HMR database is the most comprehensive resource for the human related biochemical reactions and includes all of the genes, metabolites and reactions in the recently published models (Table S8). In the HMR database, proteins encoded by genes are classified into eight different compartments including cytosol, nucleus, endoplasmic reticulum (ER), Golgi apparatus (GA), peroxisome, lysosome, mitochondria and extracellular space. The HMR2.0 database is available at http://www.metabolicatlas.com in SBML format.

In order to construct a simulation ready HMR database, firstly it was tested so that all individual reactions except pool reactions were mass-balanced. Secondly, it was guaranteed that high-energy compounds cannot be generated from low-energy compounds (such as ATP from ADP). This allowed us to test the thermodynamic constraints and the reversibility of the reactions. Thirdly the gap identification and gap filling capabilities of the RAVEN Toolbox (20) were used to guide targeted literature studies in order to keep the number of dead-end reactions to a minimum. The production of all metabolites in the model was tested using artificial reactions (Dataset 8). Artificial reactions were used to ensure the connectivity and were not included during the simulations and network dependent analysis.

Consensus genome-scale metabolic model for hepatocytes

GEMs provide biologically meaningful mechanistic basis for the genotype-phenotype relationships, yet it is necessary to have functional cell type GEMs in order to identify the metabolic differences between different states. We reconstructed iHepatocytes2260 by merging recently generated GEM for hepatocytes, iHepatocytes1901 and previously published liver models (1, 12, 18, 19). iHepatocytes1901 was generated from the HMR database using the tINIT (Task-driven Integrative Network Inference for Tissues) algorithm, which allows for automated reconstruction of functional GEMs based on a task-driven reconstruction approach and primarily relies on the cell type specific proteome in Human Protein Atlas (HPA, http://www.proteinatlas.org) (16).

We incorporated differentially expressed genes (p-values<0.001) in NAFLD patients in our reconstruction process, since iHepatocytes2260 is used for the analysis of NAFLD patient data. Moreover, we integrated biochemical knowledge about hepatocyte metabolism and a large number of additional clinical data into the model. The resulting iHepatocytes2260 is the largest cell/tissue type specific GEM and contains 2,260 genes, 5,693 metabolites in eight different compartments and 7,870 reactions. In the model, 74% of the reactions are associated to one or more genes. iHepatocytes2260 was validated with 256 known biological functions of hepatocytes, based on the definitions by Gille et al. (1), by using the checkTasks function in the RAVEN Toolbox (20).

Transcriptome data for NAFLD

In order to study the appearance of human NAFLD through the changes in global gene expression,

microarray data for liver samples were retrieved from ArrayExpress public repository under the accession number E-MEXP-3291. The data include samples from 45 different subjects and the samples were diagnosed as healthy (n=19), steatotic (n=10), NASH with fatty liver (FL) (n=9) and NASH without FL (n=7) (46). NASH with and without FL samples were characterized by >5% fat deposition and <5% fat deposition within hepatocytes, respectively, and both were accompanied by inflammation and fibrosis. Steatosis was diagnosed by >10% fat deposition without inflammation or fibrosis. The severity can be ordered as NASH without FL > NASH with FL > steatotic > healthy. Clinical information of these human liver samples has been described previously (47) and the age, gender and disease state of the patients have been included in Dataset 9.

The steatotic samples did not demonstrate significant gene expression changes compared to normal samples, as similarly reported in the plasma metabolic profiling of subjects with NAFLD, steatosis and NASH (32) (Figure S1). Hence, we performed the pair wise comparison analysis of the gene expression to compare NASH with and without FL samples versus healthy samples using Piano R package (48). Identified metabolic differences between the NAFLD patients through iHepatocytes2260 provided detailed information comparing the enrichment of differentially expressed genes in the KEGG pathways (Figure S4). Differentially expressed genes in NASH with and without FL samples versus healthy samples enriched in metabolism related KEGG pathways including steroid biosynthesis, oxidative phosphorylation, valine, leucine and isoleucine degradation, peroxisome, pyrimidine metabolism, pentose phosphate pathway and fatty acid biosynthesis.

SUPPLEMENTAL INFORMATION

Supplemental Information includes 4 figures, 8 tables and 9 datasets and can be found with this article online at http://www.metabolicatlas.com.

DATA AVAILABILITY

Human Metabolic Reaction (HMR) database and Genome-scale metabolic model for hepatocytes, iHepatocytes2260 is publically available in SBML format at Human Metabolic Atlas (http://www.metabolicatlas.org).

The annotation of the presence or absence of protein targets in hepatocytes together with the high-resolution images is publically available through the Human Protein Atlas (http://www.proteinatlas.org).

ACKNOWLEDGEMENTS

This work was financially supported by the Knut and Alice Wallenberg Foundation and Chalmers Foundation.

AUTHOR CONTRIBUTIONS

AM and RA updated the Human Metabolic Reaction database and reconstructed iHepatocytes2260. RA and AM performed the analysis of clinical data. CK and AA re-evaluated the proteomics data for hepatocytes. JN and MU conceived the project. AM, RA and JN wrote the paper and all authors were involved in editing the paper.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Gille C, *et al.* (2010) HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol Syst Biol* 6:411.
2. Baffy G, Brunt EM, & Caldwell SH (2012) Hepatocellular carcinoma in non-alcoholic fatty liver disease: An emerging menace. *J Hepatol* 56(6):1384-1391.
3. Neuschwander-Tetri BA & Caldwell SH (2003) Nonalcoholic steatohepatitis: summary of an AASLD Single Topic Conference. *Hepatology* 37(5):1202-1219.
4. Bedogni G, Kahn HS, Bellentani S, & Tiribelli C (2010) A simple index of lipid overaccumulation is a good marker of liver steatosis. *BMC Gastroenterol* 10:98.
5. Lanpher B, Brunetti-Pierri N, & Lee B (2006) Inborn errors of metabolism: the flux from Mendelian to complex diseases. *Nat Rev Genet* 7(6):449-460.
6. Chan SY & Loscalzo J (2012) The emerging paradigm of network medicine in the study of human disease. *Circ Res* 111(3):359-374.
7. Mardinoglu A & Nielsen J (2012) Systems medicine and metabolic modelling. *J Intern Med* 271(2):142-154.
8. Nielsen J (2009) Systems biology of lipid metabolism: from yeast to human. *FEBS Lett* 583(24):3905-3913.
9. Duarte NC, *et al.* (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A* 104(6):1777-1782.
10. Ma H, *et al.* (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* 3:135.
11. Hao T, Ma HW, Zhao XM, & Goryanin I (2010) Compartmentalization of the Edinburgh Human Metabolic Network. *BMC Bioinformatics* 11:393.
12. Agren R, *et al.* (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput Biol* 8(5):e1002518.
13. Mardinoglu A, *et al.* (2013) Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Mol Syst Biol* 9:649.
14. Machado MV & Cortez-Pinto H (2012) Non-invasive diagnosis of non-alcoholic fatty liver disease. A critical appraisal. *J Hepatol*.
15. Newgard CB (2012) Interplay between lipids and branched-chain amino acids in development of insulin resistance. *Cell Metab* 15(5):606-614.
16. Uhlen M, *et al.* (2010) Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* 28(12):1248-1250.
17. Mardinoglu A, Gatto F, & Nielsen J (2013) Genome-scale modeling of human metabolism. *Biotechnology journal*.
18. Jerby L, Shlomi T, & Ruppin E (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol* 6:401.
19. Bordbar A, *et al.* (2011) A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Syst Biol* 5:180.
20. Agren R, *et al.* (2013) The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum. *PLoS Comput Biol* 9(3):e1002980.
21. Rector RS, Thyfault JP, Wei Y, & Ibdah JA (2008) Non-alcoholic fatty liver disease and the metabolic syndrome: an

update. *World journal of gastroenterology : WJG* 14(2):185-192.
22. Ascha MS*, et al.* (2010) The incidence and risk factors of hepatocellular carcinoma in patients with nonalcoholic steatohepatitis. *Hepatology* 51(6):1972-1978.
23. Patil KR & Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc Natl Acad Sci U S A* 102(8):2685-2689.
24. Patil KR, Rocha I, Forster J, & Nielsen J (2005) Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* 6:308.
25. Greco D*, et al.* (2008) Gene expression in human NAFLD. *Am J Physiol Gastrointest Liver Physiol* 294(5):G1281-1287.
26. Anstee QM & Day CP (2012) S-adenosylmethionine (SAMe) therapy in liver disease: a review of current evidence and clinical utility. *J Hepatol* 57(5):1097-1109.
27. Afratis N*, et al.* (2012) Glycosaminoglycans: key players in cancer cell biology and treatment. *Febs Journal* 279(7):1177-1197.
28. Farese RV, Jr. & Walther TC (2009) Lipid droplets finally get a little R-E-S-P-E-C-T. *Cell* 139(5):855-860.
29. Gorden DL*, et al.* (2011) Increased Diacylglycerols Characterize Hepatic Lipid Changes in Progression of Human Nonalcoholic Fatty Liver Disease; Comparison to a Murine Model. *PLoS One* 6(8).
30. Maddocks OD*, et al.* (2013) Serine starvation induces stress and p53-dependent metabolic remodelling in cancer cells. *Nature* 493(7433):542-546.
31. Gulsen M*, et al.* (2005) Elevated plasma homocysteine concentrations as a predictor of steatohepatitis in patients with non-alcoholic fatty liver disease. *J Gastroenterol Hepatol* 20(9):1448-1455.
32. Kalhan SC*, et al.* (2011) Plasma metabolomic profile in nonalcoholic fatty liver disease. *Metabolism* 60(3):404-413.
33. Possemato R*, et al.* (2011) Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. *Nature* 476(7360):346-350.
34. Chaneton B*, et al.* (2012) Serine is a natural ligand and allosteric activator of pyruvate kinase M2. *Nature* 491(7424):458-462.
35. de Koning TJ*, et al.* (2004) Prenatal and early postnatal treatment in 3-phosphoglycerate-dehydrogenase deficiency. *Lancet* 364(9452):2221-2222.
36. Girard-Globa A, Robin P, & Forestier M (1972) Long-term adaptation of weanling rats to high dietary levels of methionine and serine. *J Nutr* 102(2):209-217.
37. Verhoef P*, et al.* (2004) Dietary serine and cystine attenuate the homocysteine-raising effect of dietary methionine: a randomized crossover trial in humans. *American Journal of Clinical Nutrition* 80(3):674-679.
38. Lund K, Merrill DK, & Guynn RW (1985) The reactions of the phosphorylated pathway of L-serine biosynthesis: thermodynamic relationships in rabbit liver in vivo. *Arch Biochem Biophys* 237(1):186-196.
39. Frayling TM (2011) Metabolite Quantitative Trait Loci (mQTL) and Their Role in Type 2 Diabetes and Insulin Sensitivity. *American Diabetes Association Meeting*:pp LB1–LB47.
40. Kanehisa M, Goto S, Furumichi M, Tanabe M, & Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 38(Database issue):D355-360.
41. Croft D*, et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 39(Database issue):D691-697.
42. Harkewicz R & Dennis EA (2011) Applications of mass spectrometry to lipids and membranes. *Annu Rev Biochem* 80:301-325.
43. Romero P*, et al.* (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* 6(1):R2.
44. O'Donovan C & Apweiler R (2011) A guide to UniProt for protein scientists. *Methods Mol Biol* 694:25-35.
45. Cotter D, Maer A, Guda C, Saunders B, & Subramaniam S (2006) LMPD: LIPID MAPS proteome database. *Nucleic Acids Res* 34(Database issue):D507-510.
46. Lake AD*, et al.* (2011) Analysis of global and absorption, distribution, metabolism, and elimination gene expression in the progressive stages of human nonalcoholic fatty liver disease. *Drug Metab Dispos* 39(10):1954-1960.
47. Fisher CD*, et al.* (2009) Hepatic cytochrome P450 enzyme alterations in humans with progressive stages of nonalcoholic fatty liver disease. *Drug Metab Dispos* 37(10):2087-2094.
48. Varemo L, Nielsen J, & Nookaew I (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res*.