Decomposing complex reaction networks using random sampling, principal component analysis and basis rotation

Barrett, Christian L; Herrgard, Markus J; Palsson, Bernhard

Published in:
B M C Systems Biology

Link to article, DOI:
10.1186/1752-0509-3-30

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

SYSTEMS BIOLOGY

Reconstruction of biochemical networks in microorganisms

Adam M. Feist, Markus J. Herrgård*‡, Ines Thiele*, Jennie L. Reed§ and Bernhard Ø. Palsson*||*

Abstract | Systems analysis of metabolic and growth functions in microbial organisms is rapidly developing and maturing. Such studies are enabled by reconstruction, at the genomic scale, of the biochemical reaction networks that underlie cellular processes. The network reconstruction process is organism specific and is based on an annotated genome sequence, high-throughput network-wide data sets and bibliomic data on the detailed properties of individual network components. Here we describe the process that is currently used to achieve comprehensive network reconstructions and discuss how these reconstructions are curated and validated. This Review should aid the growing number of researchers who are carrying out reconstructions for particular target organisms.

**Department of Bioengineering, University of California, San Diego, La Jolla, California 92093, USA. ‡Synthetic Genomics, 11149 N. Torrey Pines Road, La Jolla, California 92037, USA. §Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. ||Centre for Systems Biology, University of Iceland, Vatnsmyrarvegi 16, IS-101 Reykjavik, Iceland. Correspondence to B.O.P. e-mail: palsson@ucsd.edu*
doi:10.1038/nrmicro1949
Published online
31 December 2008

Reconstructed networks of biochemical reactions are at the core of systems analyses of cellular processes. Such networks form a common denominator for both experimental data analysis and computational studies in systems biology. The conceptual basis for the reconstruction process has been outlined[1], and computational methods and tools used to characterize them have been reviewed[2,3]. Furthermore, the number of available, well-curated organism-specific network reconstructions is increasing (Supplementary information S1 (table)) and the spectrum of their uses is broadening[4].

This Review describes the detailed work flows that form the basis of the reconstruction process and provide key procedural information needed for the increasing number of researchers who are performing organism-specific reconstructions. We describe the procedures in which various experimental data types are integrated to reconstruct biochemical networks, the current status of network reconstructions and how network reconstructions can be used in a prospective manner to discover new interactions and pathways. We will focus on the networks that underlie three key cellular processes: metabolism, transcription and translation, and transcriptional regulation. The reconstruction process for genome-scale metabolic networks is well developed, whereas the process for the reconstruction of transcriptional regulation and for transcriptional and translational processes at the genome-scale is only now developing. In addition, we will briefly discuss the impact of network content on modelling and integration of these types of networks, as well as the prospects of reconstructing other types of networks, such as signalling and small RNA (sRNA) pathways.

Metabolic networks

Before annotated genomic sequences were available, primary literature and biochemical characterization of enzymes provided the main sources of information for reconstructing metabolic networks in a select number of organisms. Accordingly, some of the earliest metabolic reconstructions that were subsequently used in modelling applications were for *Clostridium acetobutylicum*[5], *Bacillus subtilis*[6] and *Escherichia coli*[7–10].

Today, with the ability to sequence and annotate whole genomes, we can generate metabolic network reconstructions at a genome scale, even for organisms for which little direct biochemical information is available in the published literature. To implement the metabolic reconstruction process, we need to answer the following questions for each of the enzymes in a metabolic network: what substrates and products does an enzyme act on; what are the stoichiometric coefficients for each metabolite that participates in the reaction (or reactions) catalysed by an enzyme; are the outlined reactions reversible; and where does the reaction occur in the cell (for example, the cytoplasm or periplasm)? These data come from a range of sources. Establishing a set of the chemical reactions that constitute a reaction network culminates in a database of chemical equations. Each reaction also

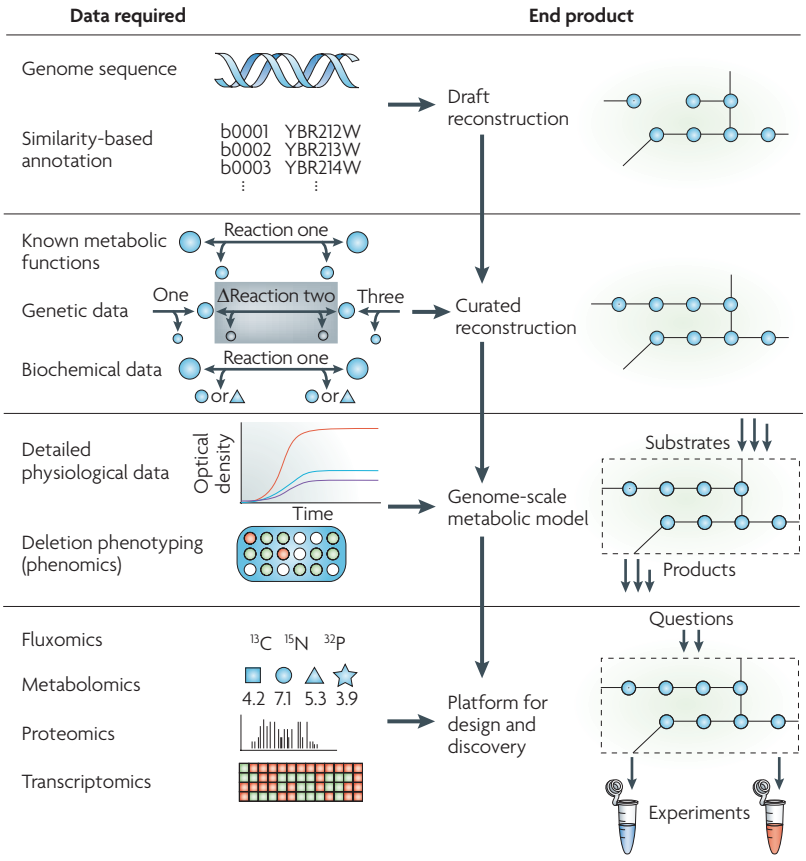


Figure 1 | **Phases and data used to generate a metabolic reconstruction.** Genome-scale metabolic reconstruction can be divided into four major phases, each of which builds from the previous one. An additional characteristic of the reconstruction process is the iterative refinement of reconstruction content that is driven by experimental data from the three later phases. For each phase, specific data types are necessary that range from high-throughput data types (for example, phenomics and metabolomics) to detailed studies that characterize individual components (for example, biochemical data for a particular reaction). For example, the genome annotation can provide a parts list of a cell, whereas genetic data can provide information about the contribution of each gene product towards a phenotype (for example, when removed or mutated). The product generated from each reconstruction phase can be used and applied to examine a growing number of questions, with the final product having the broadest applications.

has additional information associated with it, such as its cellular localization, thermodynamics, and genetic or genomic information. The genome-scale metabolic network reconstruction process comprises four fundamental steps (FIG. 1).

Step one: automated genome-based reconstruction. The starting point for reconstructions is the annotated genome for a particular target organism and strain (BOX 1). Genome annotations can be found in organism-specific databases, such as EcoCyc[11] for *E. coli* and SGD (*Saccharomyces* Genome Database)[12] or CYGD (Comprehensive Yeast Genome Database)[13] for *Saccharomyces cerevisiae*, or in databases with collections of genome annotations, such as EntrezGene[14], CMR (Comprehensive Microbial Resource)[15], Genome Reviews (through EBI; European Bioinformatics Institute)[16] or IMG (Integrated Microbial Genomes)[17] (see Further information). The genome annotation provides unique identifiers for the reconstruction, lists the metabolic enzymes that are thought to be present in the target organism and indicates how the gene products interact (as subunits, protein complexes or isozymes) to form active enzymes that catalyse metabolic reactions. The next step in the reconstruction process is to determine which biochemical reactions these enzymes carry out, which can be determined manually or by using automated tools.

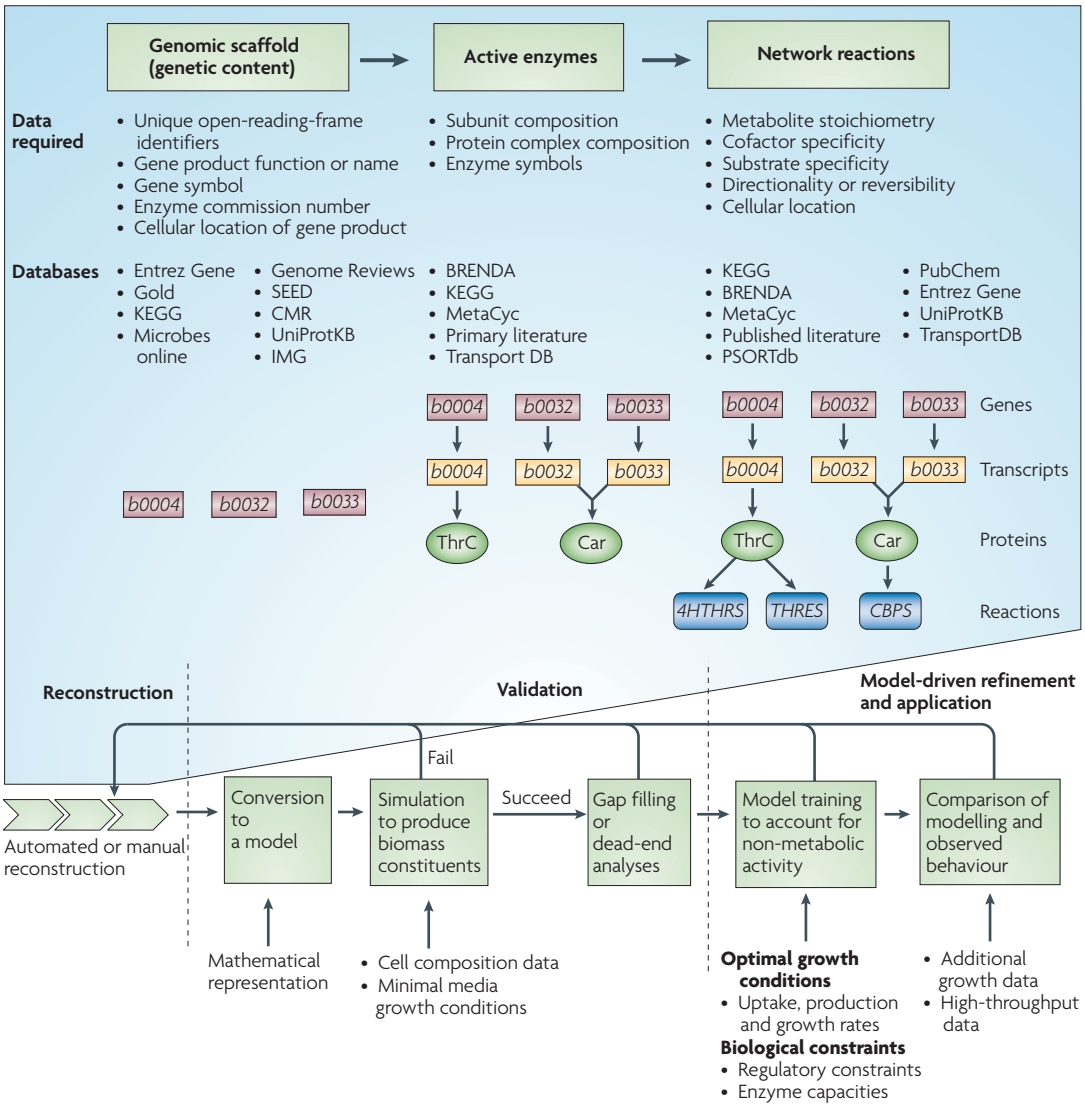
Metabolic databases, such as KEGG (Kyoto Encyclopedia of Genes and Genomes)[18], BRENDA[19], MetaCyc[20], SEED[21] and Transport DB[22] (see Further information), contain collections of metabolic and transport reactions that have been shown to occur in a range of different organisms. Many of these databases link enzyme commission (EC) numbers or transport commission (TC) numbers to individual or sets of reactions that have been observed biochemically in other organisms. However, substrate specificities and enzyme activities can vary between enzymes with the same EC or TC number, and therefore the actual reactions that are catalysed by the enzyme in the target organism may differ from that of the analogous enzyme in a reference organism. In addition, some information that is needed for the metabolic reconstruction, such as subcellular localization and reaction directionality, might be missing (Supplementary information S2 (table)).

Information from metabolic databases can be extracted manually, either by examining each active enzyme and reaction for a given organism or by using automated tools to piece together reactions from the metabolic databases. A number of such automated tools that facilitate the reconstruction process have been developed (BOX 2).

Step two: curating the draft reconstruction. Although the automated extraction of metabolic reactions from databases provides an initial set of candidate biochemical reactions encoded on a genome, it cannot establish certain organism-specific features, such as substrate or cofactor specificity and subcellular localization. Such information requires domain-specific knowledge of the organism. Therefore, the draft network reconstruction needs to be manually curated, ideally with input from organism-specific experts. An automatically reconstructed metabolic network will be incomplete, will have gaps and may also contain mistakenly included reactions that do not actually occur in the target organism. Manual curation is thus necessary to add and correct information that the automatic procedures miss or misplace in the initial network reconstruction. Although the automated reconstruction step is rapid, the manual curation process is labour intensive and sometimes tedious.

Organism-specific databases, textbooks[23–26], primary publications, review articles and experts familiar with the legacy data for an organism are the main sources of information for the manual curation step. These detailed sources contain information about

Box 1 | **Reconstruction, validation and utilization of a metabolic reconstruction**



The process of metabolic reconstruction can be performed in a sequential manner (see the figure). The process is initiated by obtaining the genetic content (that is, a parts list of the cell) from the genome annotation. Active enzymes on this scaffold are associated with the genetic content by using information from databases and published literature. The metabolic reactions that these enzymes catalyse are then delineated and a gene to protein to reaction association is ultimately generated. Automated reconstruction tools are available (BOX 2) to aid in this process and several databases possess the necessary information for each data type (see Further information and below).

Following the initial reconstruction process, a reconstruction is converted to a model in a mathematical format that can be used for computation. Further in the validation phase, the ability of the organism to produce biomass constituents and grow is examined using a biomass objective function (BOX 3). This analysis functionally tests the reconstruction for an experimentally observed phenomenon. A dead-end analysis should follow, for which computational algorithms are available (see the main text), to examine reactions on a pathway basis for their physiological role.

For predictions of physiological behaviour, a training data set is needed to examine non-metabolic energy needs and organism-specific components (for example, the electron transport system). In this phase, additional known key network properties can be applied in addition to the metabolic functions outlined in the reconstruction (for example, key regulatory interactions under a given condition) to improve predictive capabilities. For prospective use, high- and low-throughput data can also be compared with modelling simulations to validate the content and make predictions or find specific areas of disagreement between the functionality of the currently characterized content and experimental observations.

Available metabolic and transport databases include: BRENDA[19], CMR[15], Entrez Gene[14], Genome Reviews[16], GOLD[102], IMG[17], KEGG[18], MetaCyc[20], Microbes Online[103], PSORTdb[104], PubChem[105], SEED[106], Transport DB[22] and UniProtKB[107].

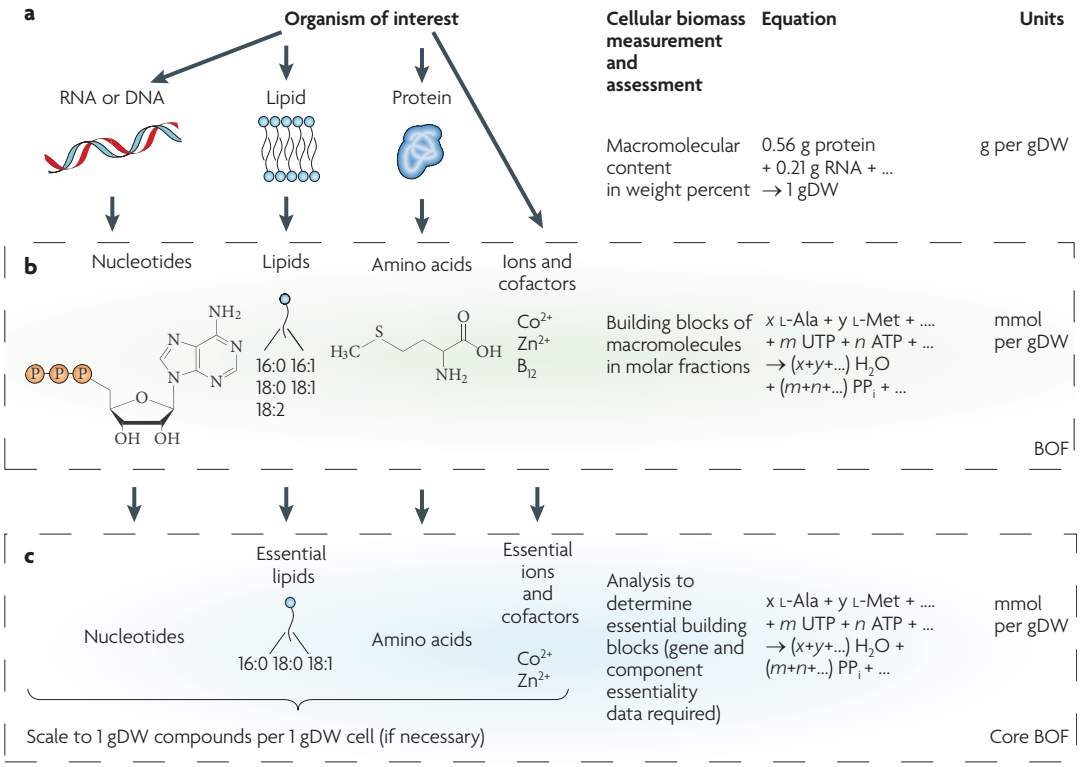
Box 2 | **Automated reconstruction of metabolic networks**

Problem	Description	Methods
Genome annotations		
Annotations are not continuously updated with new information	As new genes are found, older genome annotations are not updated, resulting in incorrectly annotated genes. For example, in most databases, *slr0788* in *Synechocystis* spp. is annotated as a pre-B-cell enhancing factor (a mammalian function assigned to a bacterial gene), but in SEED[21], is correctly annotated as nicotinamide phosphoribosyltransferase.	Automated annotation pipelines can be used to reanalyse older genome annotations[117].
Incorrect annotations	Incorrect annotations can be due to either missing genes (from sequencing or gene-finding algorithm errors) or incorrect gene annotations. This can occur for a number of reasons. For example, when new sequences are not used to update older genome annotations or when weak homology is used as sole evidence for functional assignment.	Analysis of reconstructed networks can help identify some of these errors[44,93,113,114].
Missing functionalities	Approximately 30% of enzyme activities with enzyme commission numbers lack sequence data[118].Therefore, not all reactions will be associated with gene or protein sequences. For example, in 2005, the 6-phosphogluconlactonase gene (*pgl*) in *Escherichia coli* was discovered[119]. Prior to this, there was no *pgl* gene in the genome annotation even though the enzymatic activity was observed in cell extracts.	Automated tools have been developed to find missing reactions (for example, SMILEY algorithm[44], GapFind (or GapFill)[113] PathoLogic[114] and topology-based methods[120]).
Transporter specificity	Annotations for transporters often lack sufficient detail to determine what substrate (or substrates) they transport, even though the mechanism (for example, proton symport or ATP hydrolysis) is known.	Methods for improving transporter functional annotations are needed.
Databases		
Gene-protein-reaction (GPR) associations	Relationships between genes, enzymes and reactions are not always clearly defined (for example, subunits compared with isozymes).	Can be automated based on comparisons of sequences and known GPRs[31].
Reaction specificity	Reactions are often characterized through their actions on a general class of compounds, which can result in ambiguous connections in a network. Common general classes include electron carriers (for example, quinones, NAD compared with NADP) and alcohols (for example, ethanol and methanol compared with butanol).	Changes in databases are needed or automated tools need to be developed.
Reaction imbalances	Reactions are not elementally balanced for H, C, P, N, O or S. This means that substrates and products are missing from imbalanced reactions. For example, analysis of the KEGG database[96] in 2004 found that only 51% of the reactions were balanced for C, P, N, O, H and S.	Automated procedures are available to check elemental reaction balancing[96].
Reaction directionality	Reactions are generally defined as reversible. This can be a problem; for example, if cycles between reactions allow the free conversion of ADP into ATP (free-energy equivalents).	Automated procedures have been developed[115,116].
Compound protonation states	Reactions are generally written for the neutral form of molecules and do not account for the protonation state of compounds (for example, carboxylic acid groups are deprotonated at pH 7). This affects the stoichiometric coefficients for protons across the network.	pKa prediction software is available, and therefore automation is possible.
Coenzyme availability	Enzymes often need coenzymes (for example, pyridoxal 5-phosphate, vitamin B12 and biotin). For enzymes to be functional, the cell must be able to produce them or get them from the environment. BRENDA[19] contains this type of information, and is available for download.	Automation is possible now that data are becoming available.
Organism-specific pathways	The cell membrane (or membranes) is composed of macromolecules (for example, phospholipids and peptidoglycans) that can vary across organisms and species. As a result, the biosynthesis pathways for these compounds are often unique.	Would require experimental data and is therefore unlikely to become subject to automation.

To date, most metabolic reconstructions have been generated based on a combination of genome annotation (see the table), and data from databases and literature, with heavy reliance on genome annotations for less studied organisms. Methods have been developed to help automate this process, but the resulting reconstructions still require manual curation if the goal is to convert them to mathematical models[31,108].

A number of automated methods have been produced that facilitate the reconstruction process. Some of these are used to map genes in the genome to reactions, thereby forming a draft metabolic network (for example, PathwayTools[109], GEM System[110], metaShark[111], SEED[21] and others[31,32,112]), whereas others are used to refine the networks by filling in missing reactions (for example, SMILEY algorithm[44], GapFind (also known as GapFill)[113] and PathoLogic[114]) or by evaluating reaction directionality[115,116]. Methods that refine the networks improve draft reconstructions built from gene-to-reaction mapping from databases, as they can correct incorrect or missing information from metabolic databases and/or genome annotations. Because automated methods rely heavily on metabolic and transport databases, together with genome annotations, errors will propagate into reconstructed networks. A table of common problems encountered during automated network reconstruction is provided as a guide for the use of such methods and should enable further advancement of such tools.

Box 3 | **Procedure to generate a biomass objective function**



An organism-specific biomass objective function (BOF) can be used to test the functionality of a network by examining the fundamental property of cellular growth and regeneration (see the figure). The BOF, a known growth-supporting media condition and a reconstruction in a mathematical format are necessary for this test. Starting with the organism of interest, the macromolecular weight percent contribution of each component is determined (see the figure, part **a**). These data can be generated using readily available assay kits. Each macromolecule is then broken down into the cellular building blocks that constitute the macromolecule or those that are necessary to synthesize the macromolecule in terms of molar fractions (see the figure, part **b**). The building block will often be physiologically present in the network (for example, lipid molecules), but in some cases, the most appropriate metabolite in the network is used to generate the BOF (for example, protein is broken down into individual amino acids and the net product of protein synthesis, water). With the availability of gene and/or component essentiality data, a core BOF can be generated that possesses different metabolites compared with the wild-type BOF. In formulating the core BOF, gene essentiality data are used together with the pathway context to determine the most basic macromolecule that is necessary for cell viability (see the figure, part **c**). Alternatively, published data that determine minimally essential biomass components can be incorporated to generate the core BOF. A core BOF can be used in simulations to more accurately examine essential components or aspects of the network. This process ultimately results in a BOF (or BOFs) in mmol per gram of dry weight (gDW) that can be used to evaluate an organism-specific network.

BiGG knowledge base
The collection of established biochemical, genetic and genomic data (BiGG) represented by a network reconstruction.

Genome-scale network reconstruction
(GENRE). A two-dimensional genome annotation (for example, a metabolic reconstruction) that contains a list of all the chemical transformations known to take place in a particular network (usually the entire metabolic network of a particular organism; for example, a GENRE of *E. coli*). These transformations can be represented by a stoichiometric matrix. A genre is updated as the BiGG knowledge base expands.

properties such as reaction directionality and location that are not always found in more general databases. For example, protein localization studies[27] can be used to assign metabolic reactions to subcellular compartments. Similarly, biochemical studies of enzymes from the target organism (or a closely related organism) can provide information on reversibility and substrate specificity that is specific to that organism. These sources of information provide more direct evidence for the inclusion of specific reactions in the metabolic reconstruction. The availability of such sources for a particular organism is highly variable[28]. The goal of manual reconstruction is to fill in gaps or holes in the network by inference or through direct evidence from the available literature on the organism or its close relatives. Gap-filling is further discussed below and examples in metabolic networks are presented in BOX 2 and Supplementary information S2 (table).

A high-quality network reconstruction is therefore based on a combination of automated genome-based procedures coupled with detailed and laborious literature-based manual curation. This process creates a biochemically, genomically and genetically (BiGG) structured knowledge base that is both organism specific and available to all researchers working with the target organism. All the reactions placed in a BiGG knowledge base form a genome-scale network reconstruction (GENRE). GENREs are formed in an iterative manner (for example, *E. coli*[29,30]) as the corresponding BiGG knowledge base grows for the target organism, based on new experimental data or new genome annotation.

Table 1 | **Systematic data-driven discovery of new pathways or enzymes**

Data type	Discovery type	Refs
Growth in diverse media conditions	New substrate utilization pathways	44
Deletion-strain growth phenotyping and synthetic lethal interactions	Alternative pathway discovery	97,126
Systematic *in vitro* enzymatic assays	New metabolic reactions and pathways	127
Metabolomics	New metabolite utilization or production pathways	128
Proteomics, transcriptomics and genomic neighbourhood	Candidate genes for filling network gaps	120,129, 130

Step three: converting a genome-scale reconstruction to a computational model. Before a reconstruction can be used for computations of network and/or physiological capabilities, a subtle, but crucial step must be made in which a reconstruction is converted to a mathematical representation[31,32] (BOX 1). This conversion translates a GENRE into a mathematical format that becomes the basis for a genome-scale model (GEM). Subsequent computations serve as a way to interrogate data consistency and to compute which functions a reconstructed network can and cannot carry out.

Representation of a network in a mathematical format enables the deployment of a large range of computational tools to analyse network properties. These computational tools focus on the evaluation of network systemic properties and the functions that a network can perform under the physicochemical constraints placed on the cell. This approach has led to the so-called constraint-based reconstruction and analysis (COBRA) framework[1] for the target organism. Various computational platforms have been developed that apply constraint-based methods to metabolic GEMs[3,33,34]. In addition to stoichiometric representation, metabolic networks are commonly analysed as graphs[35] or using a pathway or sub-system-based approach[36], although these essentially non-parametric approaches are not discussed further here.

With a mathematical representation and computational platform, the generation of a biomass objective function is necessary to compute the ability of a network to support growth (BOX 3). Here, the macromolecular composition of the cell (and the building blocks that are used to generate them) is used to define a necessary functionality that the network must be able to execute. A useful consistency check performed on reconstructed networks is to use them to compute growth rates under a given condition. The set of experimental data that is necessary to perform such analysis includes: the composition of cellular biomass; the composition of the minimal growth media that is necessary to support growth *in vivo*; and a training data set that includes growth rate and substrate-uptake rates. Phenotypic data (growth, uptake and secretion rates) can be obtained through growth experiments in minimal or complex media by monitoring media components. These data are typically available in published cell-characterization studies, but may need to be generated for a specific organism of interest. Cellular biomass composition data can be obtained through assays that determine overall cellular composition and further catalogue the breakdown of each macromolecule of the cell (this information has been catalogued extensively for *E. coli*[37]). With essentiality data (gene and/or cellular content), this equation can be refined[30]. Genome-scale gene essentiality data sets are being produced for model organisms (listed in REF. 38), and these data sets are often available through specific projects or organism-specific databases, such as SGD[39]. Overall, the analysis and testing of a network's ability to produce biomass components are often used to curate metabolic networks (Supplementary information S1 (table)).

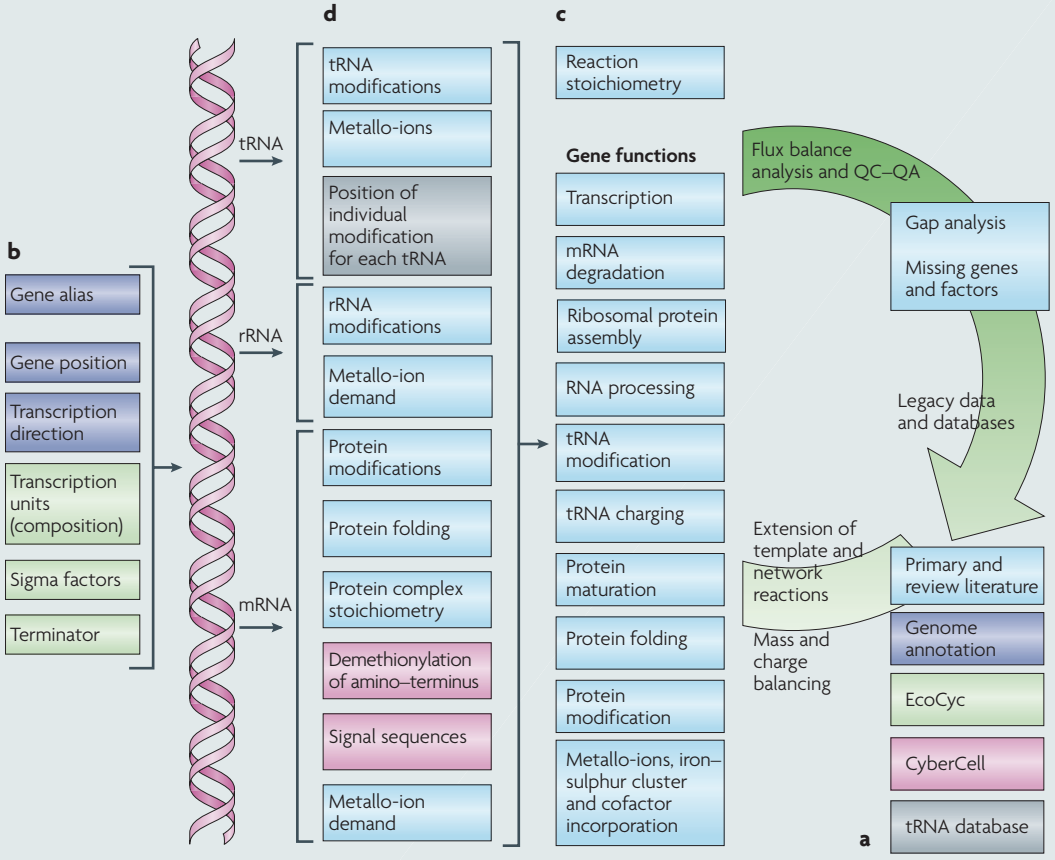
Aside from simulations to produce biomass constituents, additional gap-filling analyses can be performed to add missed pathways or remove those that have been incorrectly included from the automated reconstruction process, and additional cellular objective functions can be evaluated computationally to understand cellular behaviour[40,41]. Our current ability to gap-fill in metabolic networks has been recently reviewed[42].

Once gap-filling analyses are complete, additional steps are necessary to account for strain-specific parameters and non-metabolic activities in modelling simulations. In this phase, growth data are necessary to understand and quantify these key physiological parameters. Two major factors to consider during this phase are the stoichiometry for translocation (or energy-coupling) reactions and maintenance parameters[30,43]. Translocation reactions differ from other reactions in the network because the mass and energy balances around these ion-pumping components are difficult to measure experimentally. Characterizing reactions of this type is therefore challenging, but can be accomplished if the proper experimental data are available (Supplementary information S2 (table)). After this phase is complete, a model can be applied to study the specific growth condition from which the training data were based and can be used to explore additional environmental conditions.

Step four: reconstruction uses and integration of high-throughput data. High-throughput data sets that evaluate a large number of interactions across different growth or genetic conditions can be used to refine and expand the metabolic content of a network. These types of comparisons and analyses have the potential to truly evaluate genome-scale 'omics data sets in an integrated manner by placing them in a functional and structured context. Several successful studies have been conducted for microbial species to uncover new metabolic knowledge using systematic data-driven discovery (TABLE 1). The necessary data types to support studies of discovery and expansion, as well as pilot studies for discovery, have been recently reviewed[42]. Briefly, these studies fall into three categories: studies that use a reconstruction to examine topological network properties, studies that use a reconstruction in constraint-based modelling for quantitative or qualitative analyses and studies that are purely data driven.

Genome-scale model
A network reconstruction in a mathematical format that can be computationally interrogated and can be subsequently used for experimental design.

Box 4 | **Reconstruction of transcriptional and translational networks**



The reconstruction of a transcriptional and translational (TR–TR) network can be performed in an algorithmic manner, as depicted in the figure (illustrated for *Escherichia coli*). First, the network components responsible for every transcriptional or translational step need to be identified from different resources (for example, from primary and review literature, genome annotations and databases) (see the figure, part **a**). For each component, functions are then translated into a stoichiometric, and mass- and charge-balanced reaction based on primary and review literature. The resulting set of reactions can be separated into two groups: component-specific reactions (for example, the dimerization reaction of a protein) and template reactions (for example, a transcription initiation reaction). Template reactions can be formulated because polymerization reactions are similar for most genes. However, they need to be specified for each gene by considering the information listed (see the figure, part **b**) to produce active gene products for the different subsystems or pathways (see the figure, part **c**). The active form of some gene products may require post-translational modifications, protein folding, covalent binding of metallo-ions or coenzymes (see the figure, part **d**). The resulting reaction list is subsequently converted into a mathematical format (for example, a stoichiometric matrix) and tested for functionality, completeness, correctness and predictive potential compared with known cellular phenotypes. Discrepancies are elucidated by repeating the entire procedure again. The overall structure of the reconstructed TR–TR network resembles that of a metabolic reconstruction (see the main text). The quality control and quality assurance (QC–QA) procedures help to guaranty consistency and correctness of the network through mass and charge balancing of all possible network reactions, analysis and filling of network gaps and functionally testing for the production of every network component and its intermediate form. In the figure, the different resources in part **a** are colour coded according to their use in parts **b**–**d**.

One example of systematic data-driven discovery integrated a number of data types and GEM modelling to annotate unknown gene functions in *E. coli*[44]. In this analysis, an iterative process was used first to identify discrepancies between modelling predictions and high-throughput growth phenotyping data (using data from Biolog; see Further information), second to determine potential reactions that remedy disagreements (and the open reading frames (ORFs) that might encode proteins to catalyse them) through a computational analysis and third to characterize targeted ORFs experimentally to confirm their function. To drive discovery, this approach analysed a range of data types (phenotyping, gene expression and enzyme activity) to propose and validate computational predictions. This example shows the promise of integrating modelling results and experimental data. Such integration will probably become a key approach that will allow us to expand current metabolic knowledge and aid our discovery of new components and interactions in cellular processes.

Box 5 | **Challenges in network reconstruction**

A highly systematic process is now used to build a metabolic network reconstruction and model for any given microbial organism that starts with an annotated genome and ends with a predictive model of microbial physiology. For well characterized model organisms, this process has already enabled models to be produced that have helped discover new metabolic functionalities. However, there are many organisms of practical interest for which only initial steps towards the building of comprehensive metabolic network reconstructions and models have been taken. These include pathogens, such as *Plasmodium falciparum* and *Staphylococcus aureus*, as well as many microorganisms that are relevant to bioprocessing or bioenergy applications. Below, we outline some of the unique challenges that must be addressed when building metabolic network reconstructions for these organisms, using the malaria parasite *P. falciparum* as an example to discuss these challenges[121].

The most fundamental type of challenge for reconstruction is that in which a genome encodes for proteins that have a low degree of sequence homology to any other organism (for example, owing to severe sequence biases, such as high A+T content, as for *P. falciparum*). For these organisms, automated, homology-based, function prediction tools will result in a highly incomplete initial reconstruction of metabolic networks with numerous gaps. Although more sophisticated sequence-analysis methods allow more complete initial reconstructions for organisms such as *P. falciparum*, these initial networks still require manual curation to define the comprehensive set of metabolic capabilities that the organism possesses[111]. The experimental identification of metabolic functions in *P. falciparum* is further complicated by inefficient methods for genetic manipulation. However, these methods have improved in recent years, enabling systematic validation of putative metabolic functions and the development of strain collections that can be used for general functional genomics studies[122].

Our understanding of the metabolic physiology of *P. falciparum* and many other organisms of practical importance is also limited by our inability to culture these organisms in defined conditions. In most challenging cases, the organism cannot be readily cultured outside the host organism. *P. falciparum* can be cultured *in vitro* in red blood cells, but the presence of two different cell types in the culture poses problems to our understanding of the physiology of the parasite. For example, the transport of nutrients from the media to the parasite is only partially understood[123]. Further complications arise from the fact that typical *in vitro* culturing conditions require the use of non-specific media components, such as serum albumin. This makes it challenging to perform the types of auxotrophy experiments that are commonly used to establish metabolic functions in microorganisms, such as *Escherichia coli* or yeast. Even if well defined *in vitro* cultivation conditions can be established, it is likely that the metabolic behaviour in these conditions would fail to capture relevant features of *in vivo* physiology. This was shown to be the case for *P. falciparum* when *in vivo* expression profiles derived from patient blood samples were compared with expression profiles obtained from *in vivo* cultures of the parasite[124].

Despite these challenges, much progress has been made in our understanding of the metabolic physiology of pathogens such as *P. falciparum*. Development of metabolic network reconstructions and models for these challenging organisms has enabled systematic evaluation of current knowledge gaps and the use of model-based gap-filling strategies discussed in the main text. Progress in reconstructing other types of networks, including transcriptional regulatory networks, for pathogens is more severely affected by the lack of facile genetic systems. For example, despite extensive profiling with gene and protein-expression technologies, the mechanisms that regulate gene and protein expression in *P. falciparum* have remained elusive[125].

Transcription and translation processes

Reconstructions of transcriptional and translational (TR–TR) networks at a genome scale follow a similar procedure as that established for metabolism. TR–TR network reconstructions can be generated using a genome annotation and the genome sequence as a scaffold. A TR–TR network reconstruction contains sequence-specific synthesis reactions for every included gene and gene product that participate in transcriptional and translational functions (BOX 4). Such TR–TR reconstructions do not contain transcriptional regulators and their functions (discussed in the next section). Furthermore, the presented stoichiometric TR–TR reconstructions are different from kinetic, small-scale or sequence-independent formulations of transcriptional and/or translational networks[45–49], which are not discussed here. The scope of the TR–TR reconstruction is the synthesis of all proteins, tRNAs and ribosomal RNAs that are involved in the functions listed in BOX 4. This scope ranges from the metabolites that are consumed by the network to the functional proteins (for example, ribosomes), mRNAs and tRNAs. This type of TR–TR network has recently been developed for *E. coli*[131].

Step one: automated genome-based reconstruction. Information about the components of the TR–TR network can be directly extracted from the genome annotation. This step should provide details for gene function, gene type (for example, protein coding and tRNAs), start and stop codons, direction of transcription and transcription unit association (for prokaryotes). Some genome annotations and databases (for example, RegulonDB[47] and BioCyc[50]; see Further information) provide information about the type of transcription terminator (for example, Rho dependency and attenuation) and sigma factors for transcriptional initiation (for example, σ^{70} and σ^{H}). TR–TR reactions can be formulated in an automated manner using this information, the genome sequence and template reactions (BOX 4). These manually formulated template reactions can be used because the TR–TR reactions are similar for most genes or gene products. For example, for transcriptional initiation in *E. coli*, the holoenzyme RNA polymerase ($\alpha_2\beta\beta'$) must bind to a sigma factor (for example, σ^{70}) and this complex must then bind to a promoter site of a gene with a recognition site for this sigma factor. By contrast, gene-specific information, such as nucleotide triphosphate (NTP)

composition of an mRNA, needs to be specified in the template reaction. This subsequently allows the accurate formulation of the synthesis reactions in a gene-specific manner by using information about sigma factors, amino acids and NTPs together with template reactions.

Step two: curation and formulation based on bibliomic data. By using data from primary literature articles, template reactions must be manually formulated and curated. Manual curation is also required for information about protein-complex stoichiometry and the presence and stoichiometry of metallo-ions or coenzymes (for example, flavins), as many databases do not contain this information. Challenges that are unique to reconstruction of TR–TR networks include reaction mechanisms of certain modifications (for example, tRNA modifications[51]) or pathways (for example, iron–sulphur cluster biogenesis) that are not well established[131] (BOX 5). These reactions and pathways need to be tracked in the reconstruction (for example, by using notes or a confidence score) to allow their update as new information becomes available.

Step three: converting a genome-scale reconstruction to a computational model. The reactions list generated in steps one and two can be readily converted into a mathematical format using bioinformatically driven programming that extracts the stoichiometric coefficients from each network reaction and transfers them into the matrix. The network boundaries in TR–TR networks typically border metabolism: metabolic components are imported or exported across these boundaries. The uptake constraints for these metabolites can be derived from experimental data (for example, overall protein content) as a function of growth rate. These parameters have been directly measured for *E. coli* cells with 40-minute doubling times[52].

Step four: reconstruction uses and integration of high-throughput data. The reconstruction of TR–TR networks is a first step towards a new generation of cellular network models that will account quantitatively for mRNA and protein abundance. These models could increase the scope of modelling and therefore our understanding of cellular processes. For example, such models would allow us to calculate ribosome production at different growth rates and determine functional interactions of the network proteins by detecting functional modules. Furthermore, such TR–TR networks will increase our understanding of the relationship between mRNA and protein abundance and will allow us to consider the cost of the cellular machinery synthesis through *in silico* modelling. The reconstruction of TR–TR networks will also enable quantitative integration of high-throughput data to both expand and refine our knowledge of TR–TR networks and its components. However, there is a need to develop approaches to map relative or absolute molecule concentration data onto network reactions. Although it might be easier to integrate transcriptomic and proteomic data, an integration procedure that uses chromatin immunoprecipitation followed by microarray hybridization (ChIP–chip) to quantify the binding affinities of the RNA polymerase or other transcription factors needs to be established. Lastly, integration of the TR–TR network with other cellular processes should enable a mechanistically detailed and comprehensive description of the capabilities of different organisms.

Transcriptional regulatory networks

The basic structure of transcriptional regulatory networks (TRNs) involves the interactions between transcription factors and their target promoters that lead to activation or repression of transcription. This definition of a network boundary does not include upstream environmental and intracellular signals that regulate transcription-factor activity or any additional regulatory mechanisms that might influence gene-expression levels (for example, DNA is compacted by various proteins that influence DNA structure such that it cannot be efficiently transcribed). Most of the experimental and computational activities to elucidate TRNs have so far focused on mapping the basic structure of the network, and therefore this Review will concentrate on the network of transcription-factor–promoter interactions. ChIP–chip has also been used to map genome-wide locations of proteins that are involved in the packaging of DNA[53,54] (for example, histones and histone-like proteins), and it is expected that future reconstructions of TRNs will include global regulation of DNA accessibility and thus transcription in addition to local regulation at specific promoters by specific transcription factors.

Step one: automated reconstruction. In contrast to metabolic networks, for which experimental methods to measure system-wide levels of metabolites and fluxes are not yet fully developed, methods for large-scale measurement of TRN interactions and components are already well established. This has enabled the development of top-down approaches for TRN reconstruction that integrate multiple high-throughput data sets to reconstruct TRNs. The types of experimental approaches that are used for high-throughput studies of TRNs are typically multiplexed versions of classical, low-throughput assays for gene expression, *in vitro* DNA binding and *in vivo* DNA binding.

The most direct way to experimentally map TRNs is to determine genome-wide *in vivo* binding sites of a transcription factor using high-throughput versions of the ChIP assay. The most commonly used method is ChIP–chip, which uses a microarray-based approach to detect genomic loci if a given transcription factor binds under a given condition[55]. ChIP–chip data have now been generated in diverse microorganisms and for numerous transcription factors, allowing comprehensive mapping of TRNs, especially in yeast[56]. However, challenges remain in applying ChIP–chip (for example, transcription-factor antibody availability). To fully map TRNs, ChIP–chip experiments need to be performed for the same transcription factor under multiple conditions, as the set of target genes can vary from one condition to another[57]. Analogous to the development of multiplexed ChIP assays, high-throughput *in vitro*

DNA-binding assays that use both microarray[58] and microfluidic platforms[59] have been developed. *In vitro* methods reveal potential transcription-factor binding sites in a condition-independent manner. These *in vitro* methods require purified proteins and therefore can be challenging to apply in practice. However, they have been shown to provide valuable complementary data to *in vivo* experiments[60].

Array-based, genome-wide gene-expression profiling-based approaches are perhaps the most widely used methods to characterize TRN function. Expression profiling studies of strains in which specific transcription factors have been deleted[61,62] or overexpressed are particularly useful[63]. In addition, large compendia of gene-expression data measured in response to different genetic and environmental perturbations can be used to identify candidate regulatory interactions[64] and transcriptional modules, as well as potential regulators for these modules[65]. However, gene-expression profiling alone is not sufficient to differentiate between direct transcription-factor binding on a given promoter and indirect effects.

A major challenge remains in integrating all the available experimental data types, as well as *cis*-regulatory motif information derived from sequence conservation, to systematically reconstruct TRNs[56,66]. ChIP–chip data alone are sufficient to reconstruct the connectivity of the TRN, but expression profiling data on transcription-factor deletion or overexpression or time-course expression profiling studies are required to establish the mode of regulation (activation or repression). Furthermore, combinatorial interactions between transcription factors on promoters can only be mapped by performing expression profiling experiments in multiple transcription-factor deletion strains[61] or by performing ChIP–chip experiments for one transcription factor in strains in which another transcription factor has been deleted[67].

Fully automated TRN reconstruction would require ChIP–chip experiments that target all major transcription factors and allow gene-expression profiling of transcription-factor deletion strains under a set of representative experimental conditions. If these types of data are available and are of sufficiently high quality, TRN reconstruction can be done in a largely automated manner. Recent developments in massively parallel sequencing technologies promise to further improve our ability to automatically reconstruct TRNs by providing higher resolution, sensitivity and quality data on both gene expression[68] and DNA binding[69] compared with array-based methods. As an alternative to full mapping of transcription factor–target interactions, a number of approaches have been developed to identify condition-dependent co-regulated gene clusters or modules based on large gene-expression data sets and assign regulators to these clusters based on a combination of ChIP–chip data, expression response to transcription-factor deletions, *cis*-regulatory motifs and time-dependent gene-expression profiling data[65,70]. These types of approaches do not always allow all individual regulatory interactions to be mapped, but they substantially reduce the complexity of the TRN reconstruction problem.

Step two: reconstruction based on bibliomic data. Analogous to the reconstruction of metabolic networks, TRNs can be reconstructed in a bottom-up way based on both genomic and bibliomic data. Genomic data can be used to identify potential transcription factors as well as potential transcription-factor target sites through comparative genomics of closely related species[71]. However, genomic information alone is insufficient to obtain predictions of transcription-factor functions or targets[72] and thus substantial amounts of additional experimental information is required. The reconstruction of TRNs based on bibliomic data relies on individual studies on transcriptional regulation of single promoters that typically aim to dissect the role of different binding sites on the promoter using gene-expression assays (for example, northern blots, reverse transcription PCR or reporter gene approaches) in response to transcription-factor deletions or partial deletions of promoter regions, *in vivo* DNA-binding assays (for example, ChIP) and *in vitro* DNA-binding assays. The challenge in using literature data is that only a subset of all the promoters have been extensively characterized, and even in well-characterized organisms, such as *E. coli*, the conditions, methods and strains used can be variable. For these reasons, bottom-up reconstructions are only expected to provide a partial picture of the full TRN, and their main role in most species would be to provide validation data for more comprehensive top-down reconstruction approaches. A limited number of databases currently store literature-derived information on transcriptional regulation: the most comprehensive is RegulonDB for *E. coli*[73].

Step three: converting a genome-scale reconstruction to a computational model. TRNs reconstructed using either automated or bibliomic methods are typically represented in two alternative ways: as graphs on which each transcription-factor node is connected to its target gene nodes by a directed edge or as co-regulated gene modules with candidate transcription factors and environmental (for example, carbon source) regulators associated with each module. However, for the network to predict expression responses to environmental or genetic perturbations, these network reconstructions must be converted to computational models using one of the possible modelling frameworks. Although stochastic and kinetic models provide a good starting point for small-scale regulatory network modelling, these approaches do not scale up to larger and genome-scale networks. Most large-scale regulatory network models built so far have used Boolean network approaches and a range of probabilistic modelling frameworks, including simplified additive kinetic modelling approaches using, for example, log-linear kinetics[66,74,75].

The choice of modelling framework is largely determined by the type of network reconstruction that is used as a starting point to build the model, the type of data that are available to parameterize the model and what types of predictions one wants to make. Boolean representations provide a good starting point for building qualitative models based on TRNs reconstructed primarily using bibliomic data[61,76]. Boolean models have been built

Bibliomic data
Legacy data that are contained in peer-reviewed scientific publications. The 'omic designation represents a comprehensive assessment of legacy data for a target organism.

for *E. coli* and yeast, and these representations can be further converted to a matrix formalism that allows more straightforward integration with metabolic network models[77]. Many different probabilistic modelling frameworks, including probabilistic Boolean networks[74] physical network models[66] and more complex types of models[75], have been applied to reconstruct large-scale TRNs. However, most of these approaches have been used as tools for systematic data-based TRN reconstruction and have not yet been used to build large-scale predictive models.

By contrast, recent studies have used additive kinetic modelling approaches to model genome-scale TRNs either in settings in which the network structure is known; for example, based on ChIP–chip or bibliomic data[78] or in conjunction with methods that identify co-regulated gene clusters[70]. Unlike Boolean models, these simplified kinetic models can be used to predict quantitative dynamic expression changes, but substantial amounts of time-course gene-expression data are usually needed to parameterize the models. Recently, a predictive, additive kinetic model of the *Halobacterium salinarum* TRN GENRE was built using a combination of computational methods[70]. First, condition-dependent regulatory modules were built using bi-clustering of a well-designed gene-expression data compendium together with *cis*-regulatory motif information, and then quantitative effects of transcription factors and environmental factors on expression of these modules were identified based on dynamic gene-expression data. The *H. salinarum* study also showed that predictive TRN models can be built even for species with poorly characterized TRNs, as sufficient quantities of relevant high-throughput data can be generated in a systematic manner.

Step four: applications of TRN models. Analyses performed using TRN models have identified novel regulatory interactions and predicted general patterns of cellular behaviour. For example, a previous effort combined a comprehensive literature-based reconstruction of the TRN that controls metabolism in *E. coli* with expression profiling of single and double transcription-factor deletion strains to improve the ability of an integrated regulatory and metabolic network model to predict phenotypes and expression changes[61]. Similarly, when predictions from the *H. salinarum* model[70] were compared with experimental data, a number of novel regulators for key cellular processes in this archaeon were identified.

Technologies for mapping TRNs are maturing rapidly and promise to allow largely automated reconstruction of these types of networks in the near future. Major challenges still remain in modelling TRNs in a physicochemically realistic fashion and in integrating TRNs with other cellular processes. The signalling pathways that lead to the activation of transcription factors are also less understood than the TRN itself and the experimental techniques for mapping these pathways are not as well developed as TRN mapping methods (discussed below).

Expansion of reconstruction efforts

Together, the metabolic, transcriptional regulation, translation and transcription processes represent a sizable fraction of the genes in a microbial genome. However, other networks are also being intensively studied and will probably be the subject of future network reconstruction efforts. Such efforts are likely to develop according to four-step reconstruction processes that parallel those described above.

Two-component signalling systems are an example of this type of network. Current models of TRNs in *E. coli* already include some of the known two-component signalling pathways that respond to metabolic stimuli[61]. The components of two-component signalling pathways (histidine kinases and response regulators) can be identified easily by sequence homology, but the connectivity of these pathways is not completely known, even in *E. coli*. Progress has recently been made to systematically map the connectivity of two-component pathways in *E. coli*[79] and other bacteria[80] using a range of experimental methods. It is expected that in the future comprehensive reconstructions of two-component systems can be achieved by combining literature-based information with these types of high-throughput data[81].

The second type of network that has attracted increasing attention in recent years is the translational regulatory network controlled by sRNAs. The most common mechanism for sRNA action is the repression of specific-mRNA translation through the binding of translation initiation regions, although other mechanisms, including the regulation of protein expression or activity, also exist[82]. It has been estimated that typical bacterial genomes carry up to 300 sRNA genes and that these sRNAs play a crucial part in the control of cellular functions, including metabolism and virulence[83]. The process of finding sRNAs in bacterial genomes is reasonably well-established[84], but finding mRNA targets for these sRNAs is still challenging. A number of experimental and computational techniques have been devised to determine the targets of sRNAs at the genome-wide scale (reviewed in REFS 82,85), which will accelerate the process of mapping comprehensive sRNA regulatory networks. Initial systems studies of the known sRNA regulatory network in *E. coli* have indicated that sRNA regulation acts in concert with transcriptional regulation to provide mechanisms that allow tight, condition-dependent regulation of target protein levels[86].

Integration of network reconstructions

Once two or more of the five different types of networks described above have been reconstructed for a target organism, they can be integrated to form computational GENREs and computational GEMs that span a high number of cellular activities.

The integration of TRN and metabolic networks has received the most attention to date because these two network types have been most comprehensively reconstructed[61,76,87] (see Supplementary information S1 (table) for metabolic networks). TRNs regulate metabolism by modulating active enzyme concentrations, and subsequently by controlling the maximum

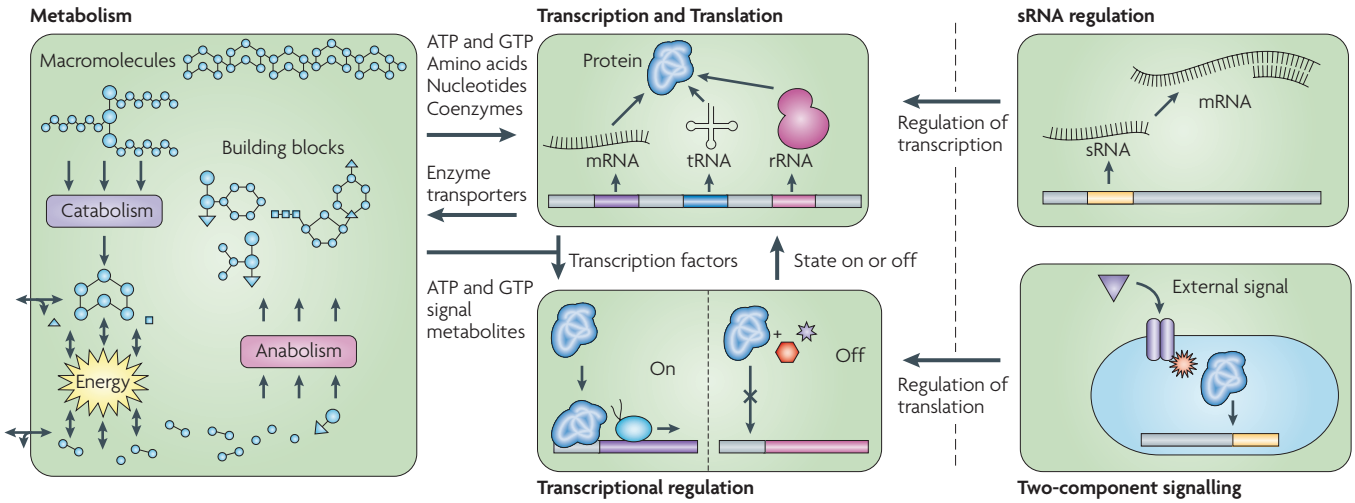


Figure 2 | **Network integration: the interface between different types of reconstruction.** The ultimate goal of network reconstruction is to fully represent every component of the cell and define the interactions between them. Reconstruction of metabolism, transcriptional regulation, and transcriptional and translational networks is currently possible (see the main text), but to date most emphasis has been put on metabolic reconstruction. Incorporation of small RNAs (sRNA) and two-component signalling interactions are future areas of reconstruction in which reconstruction technologies and development are needed. For integration of networks, the interplay between each of the processes needs to be defined to fully connect each of the major cellular functions.

flux levels through reactions. The levels of metabolites, in turn, regulate gene expression and thus the two networks are an integrated process.

Integration of metabolism and transcription and translation processes is straightforward in principle. On the one hand, transcription and translation requires energy and building blocks, such as nucleotides and amino acids, as inputs, and therefore these processes are constrained by the ability of the metabolic network to produce these precursors[48]. On the other hand, the transcription and translation processes exert demands on the metabolic network function and thus limit other metabolic functions[48]. Furthermore, the TR–TR network feeds back to the metabolic network by controlling the levels of the enzymes in the metabolic network.

Although GENREs that include three or more different networks have not been produced, they should be achievable. In principle, each network can be described by a stoichiometric matrix once the underlying reactions have been determined. Stoichiometric matrices for metabolism have been produced ('M' matrix), and a stoichiometric matrix format for TR–TR reactions ('E' matrix), which form the expression state of networks, is also achievable[131]. TRN reconstruction ('O' matrix) in bacteria can be based on the operon structure of a genome and could also be described by a corresponding stoichiometric matrix, once the underlying chemical reactions have been defined. Given that stoichiometric matrices can be integrated in a one-step process, an 'OME' matrix that describes the integrated network can be formulated. Currently, however, TRNs are described by a set of logistical statements and, although a matrix format has been developed for Boolean statements[77], which has enabled network integration, we ultimately need to seek chemical representation of TRNs. Working towards this aim, a small-scale integration of the three networks has been produced[87], foreshadowing what is to come at the genomic scale.

Stoichiometric matrix
A matrix that contains the stoichiometric coefficients for the reactions that constitute a network. The rows represent the compounds, the columns represent the chemical transformations and the entries represent the stoichiometric coefficients.

Conversion to a computational model

Integrated network reconstructions, which are essentially two-dimensional annotations[88], can be used to build GEMs that represent the functions of integrated networks to make phenotypic predictions (FIG. 2). This conversion mathematically describes the reactions that have been shown to take place in a network of interest[1] and therefore represents the conversion of a BiGG knowledge base into a GEM. The use of computational approaches to interrogate the properties of GEMs has been described[2]. GEMs have been used for experimentation in three ways[4]: to discover missing content in a reconstruction, to understand integrated physiological process and to prospectively design experiments and physiological processes. The first topic is germane to this Review, as it is aimed at systematically discovering the missing content of a reconstruction.

The effects of missing network content

An important issue in the conversion of a network reconstruction into a predictive computational model is the coverage and accuracy of available data from which the network was reconstructed. Therefore, it is important to understand the impact and influence network components can have on computational results. Intended-use examples of *in silico* models are used to help understand this issue.

Qualitative predictions obtained using GEMs (for example, will an organism grow during a particular environmental or genetic perturbation, or does gene expression increase or decrease) are likely to be less sensitive than quantitative predictions (for example, what is the cellular

growth rate or what level of gene expression is expected) to errors in network content. This is because qualitative predictions are compared with binary outcomes (digital outcomes), rather than a range of numerical values (analogue outcomes). If qualitative predictions regarding growth phenotypes are being generated, omitting an individual reaction from a network will not affect the results. For example, individually removing approximately 87% of the 2,077 reactions from an *E. coli* metabolic model (*i*AF1260 (REF. 30)) did not affect the qualitative growth predictions for a particular environmental condition.

In-depth studies have been performed to assess the influence of individual network components, input parameters and the querying methods that are used to probe GEMs on computational predictions. The results from these studies can be used to gauge the influence of the content of reconstructions. For example, these analyses examine input and output values[30,43,89,90], biomass objective function composition[30,40,43,91,92], querying methods[93,94] and network components[90,95–100]. TRNs and TR–TR networks are less developed and are expected to be more difficult to assess than metabolic networks owing to the highly nonlinear structure of some components of the network, the higher number of interactions that are typically observed per component and the greater amount of missing knowledge in these networks. Missing regulatory interactions between transcription factors and metabolic target genes in the network would thus be expected to have a moderate effect on predictive abilities, as the regulation is likely to be highly redundant.

These initial studies demonstrate the necessity to identify the scope and intention of GENRE applications *a priori* and further show how computational analysis can help identify missing components and errors when computational results are compared with biological functions. This model-driven gap-filling approach is expected to continue to develop and lead to GEMs with improved predictive capabilities.

Conclusions

The reconstruction process relies on work flows that organize and integrate various data types and other relevant information about the network of interest. Over the past 10 years, the development of work flows for genome-scale metabolic networks has increased to the point at which they represent BiGG knowledge bases and are in wide use. More recently, similar methods have been developed for other cellular processes, such as transcriptional regulation, transcription and translation. The implementation of these work flows for a growing number of organisms should accelerate systems analysis in a single organism, in communities of organisms and phyla. The work flows reviewed here have been implemented and have enabled a wide range of analyses[4]. To facilitate wider use and the development of additional analysis procedures, improvements in the distribution of GENREs are needed. Two areas that will aid distribution and usage are the standardization of a reconstruction format (for example, SBML[101] (Systems Biology Markup Language; see Further information) and available reconstruction databases, if accessible.

It is expected that the reconstruction process will continue to grow in scope, depth and accuracy, and enable a broadening spectrum of basic and applied studies. The availability of high-quality comprehensive reconstructions will accelerate the implementation of the systems biology paradigm (biological components to networks to computational models to phenotypic studies), which will help us realize the broad transformative potential of this paradigm in the life sciences. Network reconstructions are necessary for us to build a mechanistic genotype–phenotype relationship. To date, quantitative genotype–phenotype relationships have been best established for bacterial metabolism[4], and this Review should help new practitioners build such relationships for their target organisms.

1. Reed, J. L., Famili, I., Thiele, I. & Palsson, B. O. Towards multidimensional genome annotation. *Nature Rev. Genet.* **7**, 130–141 (2006).
A review of the conceptual basis for network reconstruction.
2. Price, N. D., Reed, J. L. & Palsson, B. O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Rev. Microbiol.* **2**, 886–897 (2004).
A comprehensive and succinct review of COBRA methods.
3. Becker, S. A. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. *Nature Protoc.* **2**, 727–738 (2007).
4. Feist, A. M. & Palsson, B. O. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nature Biotechnol.* **26**, 659–667 (2008).
A review of the history of applications of the genome-scale *E. coli* metabolic reconstruction.
5. Papoutsakis, E. T. Equations and calculations for fermentations of butyric acid bacteria. *Biotechnol. Bioeng.* **26**, 174–187 (1984).
6. Papoutsakis, E. & Meyer, C. Fermentation equations for propionic-acid bacteria and production of assorted oxychemicals from various sugars. *Biotechnol. Bioeng.* **27**, 67–80 (1985).
7. Papoutsakis, E. & Meyer, C. Equations and calculations of product yields and preferred pathways for butanediol and mixed-acid fermentations. *Biotechnol. Bioeng.* **27**, 50–66 (1985).
8. Majewski, R. A. & Domach, M. M. Simple constrained optimization view of acetate overflow in *E. coli*. *Biotechnol. Bioeng.* **35**, 732–738 (1990).
9. Varma, A., Boesch, B. W. & Palsson, B. O. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Appl. Environ. Microbiol.* **59**, 2465–2473 (1993).
10. Varma, A., Boesch, B. W. & Palsson, B. O. Biochemical production capabilities of *Escherichia coli*. *Biotechnol. Bioeng.* **42**, 59–73 (1993).
11. Karp, P. D. *et al.* Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Res.* **35**, 7577–7590 (2007).
12. Christie, K. R. *et al. Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.* **32**, D311–D314 (2004).
13. Guldener, U. *et al.* CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.* **33**, D364–D368 (2005).
14. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **35**, D26–D31 (2007).
15. Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K. & White, O. The Comprehensive Microbial Resource. *Nucleic Acids Res.* **29**, 123–125 (2001).
16. Stoesser, G., Tuli, M. A., Lopez, R. & Sterk, P. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **27**, 18–24 (1999).
17. Markowitz, V. M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34**, D344–D348 (2006).
18. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
19. Schomburg, I. *et al.* BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* **32**, D431–D433 (2004).
20. Krieger, C. J. *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **32**, D438–D442 (2004).
21. DeJongh, M. *et al.* Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinformatics* **8**, 139 (2007).
An innovative approach for semi-automatic reconstruction of genome-scale metabolic networks that combines automated genome annotation tools with model-based gap filling.
22. Ren, Q., Chen, K. & Paulsen, I. T. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res.* **35**, D274–D279 (2007).
23. Neidhardt, F. C. (ed.) Escherichia coli *and* Salmonella*: Cellular and Molecular Biology* (ASM Press, Washington DC, 1996).

24. Dickinson, J. R. & Schweizer, M. (eds) *The Metabolism and Molecular Physiology of* Saccharomyces cerevisiae 2nd edn (Taylor & Francis, London; Philadelphia, 2004).
25. Marre, R. *et al.* (eds) Legionella (ASM Press, Washington DC, 2001).
26. Mobley, H. L. T., Mendz, G. L. & Hazell, S. L. Helicobacter pylori (ASM Press, Washington DC, 2001).
27. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
28. Janssen, P., Goldovsky, L., Kunin, V., Darzentas, N. & Ouzounis, C. A. Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications. *EMBO Rep.* **6**, 397–399 (2005).
29. Reed, J. L. & Palsson, B. O. Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. *J. Bacteriol.* **185**, 2692–2699 (2003).
30. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**, 121 (2007).
31. Notebaart, R. A., van Enckevort, F. H., Francke, C., Siezen, R. J. & Teusink, B. Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics* **7**, 296 (2006).
32. Borodina, I. & Nielsen, J. From genomes to *in silico* cells via metabolic networks. *Curr. Opin. Biotechnol.* **16**, 350–355 (2005).
33. Lee, S. Y. *et al.* Systems-level analysis of genome-scale *in silico* metabolic models using MetaFluxNet. *Biotechnol. Bioprocess Eng.* **10**, 425–431 (2005).
34. Klamt, S., Saez-Rodriguez, J. & Gilles, E. D. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst. Biol.* **1**, 2 (2007).
35. Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Rev. Genet.* **5**, 101–113 (2004).
36. Kwast, K. E. *et al.* Genomic analyses of anaerobically induced genes in *Saccharomyces cerevisiae*: functional roles of Rox1 and other factors in mediating the anoxic response. *J. Bacteriol.* **184**, 250–265 (2002).
37. Neidhardt, F. C. & Umbarger, H. E. in Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (ed. Neidhardt, F. C.) 13–16 (ASM Press, Washington DC, 1996).
38. Joyce, A. R. *et al.* Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. *J. Bacteriol.* **188**, 8259–8271 (2006).
39. Cherry, J. M. *et al.* SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* **26**, 73–79 (1998).
40. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* **3**, 119 (2007).
41. Knorr, A. L., Jain, R. & Srivastava, R. Bayesian-based selection of metabolic objective functions. *Bioinformatics* **23**, 351–357 (2007).
42. Breitling, R., Vitkup, D. & Barrett, M. P. New surveyor tools for charting microbial metabolic maps. *Nature Rev. Microbiol.* **6**, 156–161 (2008).
A review of available computational tools that can improve and expand biological network reconstructions.
43. Varma, A. & Palsson, B. O. Parametric sensitivity of stoichiometric flux balance models applied to wild-type *Escherichia coli* metabolism. *Biotechnol. Bioeng.* **45**, 69–79 (1995).
44. Reed, J. L. *et al.* Systems approach to refining genome annotation. *Proc. Natl Acad. Sci. USA* **103**, 17480–17484 (2006).
The first demonstration of the gap-filling process: a network-guided discovery process.
45. Roussel, M. R. & Zhu, R. Stochastic kinetics description of a simple transcription model. *Bull. Math. Biol.* **68**, 1681–1713 (2006).
46. Mehra, A. & Hatzimanikatis, V. An algorithmic framework for genome-wide modeling and analysis of translation networks. *Biophys. J.* **90**, 1136–1146 (2006).
47. Weitzke, E. L. & Ortoleva, P. J. Simulating cellular dynamics through a coupled transcription, translation, metabolic model. *Comput. Biol. Chem.* **27**, 469–480 (2003).
48. Allen, T. E. & Palsson, B. O. Sequenced-based analysis of metabolic demands for protein synthesis in prokaryotes. *J. Theor. Biol.* **220**, 1–18 (2003).
49. Drew, D. A. A mathematical model for prokaryotic protein synthesis. *Bull. Math. Biol.* **63**, 329–351 (2001).
50. Karp, P. D. *et al.* Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**, 6083–6089 (2005).
51. Sprinzl, M. & Vassilenko, K. S. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **33**, D139–D140 (2005).
52. Neidhardt, F. C., Ingraham, J. L. & Schaechter, M. *Physiology of The Bacterial Cell: a Molecular Approach* (Sinauer Associates, Sunderland, Massachusetts, 1990).
53. Cho, B. K., Knight, E. M., Barrett, C. L. & Palsson, B. Ø. Genome-wide analysis of Fis binding in *Escherichia coli* indicates a causative role for A-/AT-tracts. *Genome Res.* **18**, 900–910 (2008).
54. Dion, M. F. *et al.* Dynamics of replication-independent histone turnover in budding yeast. *Science* **315**, 1405–1408 (2007).
55. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
56. Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
57. Buck, M. J. & Lieb, J. D. A chromatin-mediated mechanism for specification of conditional transcription factor targets. *Nature Genet.* **38**, 1446–1451 (2006).
58. Mukherjee, S. *et al.* Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genet.* **36**, 1331–1339 (2004).
59. Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
60. Liu, X., Lee, C. K., Granek, J. A., Clarke, N. D. & Lieb, J. D. Whole-genome comparison of Leu3 binding *in vitro* and *in vivo* reveals the importance of nucleosome occupancy in target site selection. *Genome Res.* **16**, 1517–1528 (2006).
61. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. & Palsson, B. O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96 (2004).
This study shows the value of literature-based reconstruction of TRNs for well-studied organisms, as well as the integration of metabolic-network and TRN models.
62. Hu, Z., Killion, P. J. & Iyer, V. R. Genetic reconstruction of a functional transcriptional regulatory network. *Nature Genet.* **39**, 683–687 (2007).
63. Chua, G. *et al.* Identifying transcription factor functions and targets by phenotypic activation. *Proc. Natl Acad. Sci. USA* **103**, 12045–12050 (2006).
64. Faith, J. J. *et al.* Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5**, e8 (2007).
65. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genet.* **34**, 166–176 (2003).
66. Workman, C. T. *et al.* A systems approach to mapping DNA damage response pathways. *Science* **312**, 1054–1059 (2006).
References 65 and 67 present alternative statistical approaches for mapping TRNs from large-scale experimental data sets (gene expression or ChIP–chip) obtained for well- characterized model organisms.
67. Zeitlinger, J. *et al.* Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113**, 395–404 (2003).
68. Kim, J. B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481–1484 (2007).
69. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651–657 (2007).
This study presents a comprehensive approach to the building of predictive models of large-scale TRNs for even poorly understood organisms using a low number of genome-scale experiments.
70. Bonneau, R. *et al.* A predictive model for transcriptional control of physiology in a free living cell. *Cell* **131**, 1354–1365 (2007).
71. Perez-Rueda, E. & Collado-Vides, J. The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.* **28**, 1838–1847 (2000).
72. Price, M. N., Dehal, P. S. & Arkin, A. P. Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLoS Comput. Biol.* **3**, 1739–1750 (2007).
73. Salgado, H. *et al.* RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.* **34**, D394–D397 (2006).
74. Shmulevich, I., Dougherty, E. R., Kim, S. & Zhang, W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* **18**, 261–274 (2002).
75. Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451**, 535–540 (2008).
76. Herrgard, M. J., Lee, B. S., Portnoy, V. & Palsson, B. O. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res.* **16**, 627–635 (2006).
77. Gianchandani, E. P., Papin, J. A., Price, N. D., Joyce, A. R. & Palsson, B. O. Matrix formalism to describe functional states of transcriptional regulatory systems. *PLoS Comput. Biol.* **2**, e101 (2006).
78. Kao, K. C. *et al.* Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis. *Proc. Natl Acad. Sci. USA* **101**, 641–646 (2004).
79. Yamamoto, K. *et al.* Functional characterization *in vitro* of all two-component signal transduction systems from *Escherichia coli*. *J. Biol. Chem.* **280**, 1448–1456 (2005).
80. Skerker, J. M., Prasol, M. S., Perchuk, B. S., Biondi, E. G. & Laub, M. T. Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. *PLoS Biol.* **3**, e334 (2005).
Combinations of systematic *in vivo* and *in vitro* profiling approaches, such as those used in this work, will be needed to decipher the connectivity and function of bacterial two-component systems.
81. Seshasayee, A. S., Bertone, P., Fraser, G. M. & Luscombe, N. M. Transcriptional regulatory networks in bacteria: from input signals to output responses. *Curr. Opin. Microbiol.* **9**, 511–519 (2006).
82. Vogel, J. & Wagner, E. G. Target identification of small noncoding RNAs in bacteria. *Curr. Opin. Microbiol.* **10**, 262–270 (2007).
83. Romby, P., Vandenesch, F. & Wagner, E. G. The role of RNAs in the regulation of virulence-gene expression. *Curr. Opin. Microbiol.* **9**, 229–236 (2006).
84. Vogel, J. & Sharma, C. M. How to find small non-coding RNAs in bacteria. *Biol. Chem.* **386**, 1219–1238 (2005).
85. Altuvia, S. Identification of bacterial small non-coding RNAs: experimental approaches. *Curr. Opin. Microbiol.* **10**, 257–261 (2007).
86. Shimoni, Y. *et al.* Regulation of gene expression by small non-coding RNAs: a quantitative view. *Mol. Syst. Biol.* **3**, 138 (2007).
87. Lee, J. M., Gianchandani, E. P., Eddy, J. A. & Papin, J. A. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput. Biol.* **4**, e1000086 (2008).
88. Palsson, B. O. Two-dimensional annotation of genomes. *Nature Biotechnol.* **22**, 1218–1219 (2004).
89. Cakir, T. *et al.* Flux balance analysis of a genome-scale yeast model constrained by exometabolomic data allows metabolic system identification of genetically different strains. *Biotechnol. Prog.* **23**, 320–326 (2007).
90. Vemuri, G. N., Eiteman, M. A., McEwen, J. E., Olsson, L. & Nielsen, J. Increasing NADH oxidation reduces overflow metabolism in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **104**, 2402–2407 (2007).
91. Pramanik, J. & Keasling, J. D. Stoichiometric model of *Escherichia coli* metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnol. Bioeng.* **56**, 398–421 (1997).
92. Pramanik, J. & Keasling, J. D. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnol. Bioeng.* **60**, 230–238 (1998).
93. Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl Acad. Sci. USA* **99**, 15112–15117 (2002).
94. Shlomi, T., Berkman, O. & Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl Acad. Sci. USA* **102**, 7695–7700 (2005).

95. Feist, A. M., Scholten, J. C. M., Palsson, B. O., Brockman, F. J. & Ideker, T. Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Mol. Syst. Biol.* **2**, 1–14 (2006).
96. Pharkya, P., Burgard, A. P. & Maranas, C. D. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.* **14**, 2367–2376 (2004).
97. Harrison, R., Papp, B., Pal, C., Oliver, S. G. & Delneri, D. Plasticity of genetic interactions in metabolic networks of yeast. *Proc. Natl Acad. Sci. USA* **104**, 2307–2312 (2007).
A combination of genome-scale modelling and experimentation was used to study condition-dependent genetic interactions and identify novel alternative pathways in yeast.
98. Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z. N. & Barabasi, A. L. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **427**, 839–843 (2004).
99. Burgard, A. P., Vaidyaraman, S. & Maranas, C. D. Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments. *Biotechnol. Prog.* **17**, 791–797 (2001).
100. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264–276 (2003).
101. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
102. Bernal, A., Ear, U. & Kyrpides, N. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.* **29**, 126–127 (2001).
103. Alm, E. J. *et al.* The MicrobesOnline Web site for comparative genomics. *Genome Res.* **15**, 1015–1022 (2005).
104. Rey, S. *et al.* PSORTdb: a protein subcellular localization database for bacteria. *Nucleic Acids Res.* **33**, D164–D168 (2005).
105. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**, D13–D21 (2008).
106. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).
107. Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, D115–D119 (2004).
108. Borodina, I., Krabben, P. & Nielsen, J. Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res.* **15**, 820–829 (2005).
109. Karp, P. D., Paley, S. & Romero, P. The Pathway Tools software. *Bioinformatics* **18**, S225–S232 (2002).
110. Arakawa, K., Yamada, Y., Shinoda, K., Nakayama, Y. & Tomita, M. GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. *BMC Bioinformatics* **7**, 168 (2006).
111. Pinney, J. W., Shirley, M. W., McConkey, G. A. & Westhead, D. R. 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 (2005).
112. Goesmann, A., Haubrock, M., Meyer, F., Kalinowski, J. & Giegerich, R. PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics* **18**, 124–129 (2002).
113. Satish Kumar, V., Dasika, M. S. & Maranas, C. D. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **8**, 212 (2007).
114. Green, M. L. & Karp, P. D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5**, 76 (2004).
115. Henry, C. S., Jankowski, M. D., Broadbelt, L. J. & Hatzimanikatis, V. Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. *Biophys. J.* **90**, 1453–1461 (2006).
116. Kümmel, A., Panke, S. & Heinemann, M. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics* **7**, 512 (2006).
117. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
118. Pouliot, Y. & Karp, P. D. A survey of orphan enzyme activities. *BMC Bioinformatics* **8**, 244 (2007).
119. Thomason, L. C., Court, D. L., Datta, A. R., Khanna, R. & Rosner, J. L. Identification of the *Escherichia coli* K-12 *ybhE* gene as *pgl*, encoding 6-phosphogluconolactonase. *J. Bacteriol.* **186**, 8248–8253 (2004).
120. Fuhrer, T., Chen, L., Sauer, U. & Vitkup, D. Computational prediction and experimental verification of the gene encoding the $NAD^+/NADP^+$-dependent succinate semialdehyde dehydrogenase in *Escherichia coli*. *J. Bacteriol.* **189**, 8073–8078 (2007).
121. Pinney, J. W. *et al.* Metabolic reconstruction and analysis for parasite genomes. *Trends Parasitol.* **23**, 548–554 (2007).
A detailed review of the challenges that are encountered in the reconstruction of metabolic networks for parasites such as *P. falciparum*.
122. Balu, B. & Adams, J. H. Advancements in transfection technologies for *Plasmodium*. *Int. J. Parasitol.* **37**, 1–10 (2007).
123. Kirk, K. & Saliba, K. J. Targeting nutrient uptake mechanisms in *Plasmodium*. *Curr. Drug Targets* **8**, 75–88 (2007).
124. Daily, J. P. *et al.* Distinct physiological states of *Plasmodium falciparum* in malaria-infected patients. *Nature* **450**, 1091–1095 (2007).
125. Deitsch, K. *et al.* Mechanisms of gene regulation in *Plasmodium*. *Am. J. Trop. Med. Hyg.* **77**, 201–208 (2007).
126. Shlomi, T. *et al.* Systematic condition-dependent annotation of metabolic genes. *Genome Res.* **17**, 1626–1633 (2007).
127. Saito, N. *et al.* Metabolomics approach for enzyme discovery. *J. Proteome Res.* **5**, 1979–1987 (2006).
128. Chiang, K. P., Niessen, S., Saghatelian, A. & Cravatt, B. F. An enzyme that regulates ether lipid signaling pathways in cancer annotated by multidimensional profiling. *Chem. Biol.* **13**, 1041–1050 (2006).
129. Popescu, L. & Yona, G. Automation of gene assignments to metabolic pathways using high-throughput expression data. *BMC Bioinformatics* **6**, 217 (2005).
130. Rodionov, D. A. *et al.* Genomic identification and *in vitro* reconstitution of a complete biosynthetic pathway for the osmolyte di-myo-inositol-phosphate. *Proc. Natl Acad. Sci. USA* **104**, 4279–4284 (2007).
131. Thiele, I., Jamshidi, N., Fleming, R. M. T. & Palsson, B. O. Genome-scale reconstruction of *E. coli*'s transcriptional and translational machinery: a knowledge-base and its mathematical formulation. *PLoS Comput. Biol.* (in the press).

Acknowledgements
The authors thank A. Osterman and N. Jamshidi for their insights. A.M.F. and I.T. were supported by National Institutes of Health (NIH) grant R01 GM057089 and M.J.H. was supported by NIH grant R01 GM071808. B.O.P. serves on the scientific advisory board of Genomatica.

DATABASES
Entrez Genome Project: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
Bacillus subtilis | *Clostridium acetobutylicum* | *Escherichia coli* | *Halobacterium salinarum* | *Saccharomyces cerevisiae* | *Staphylococcus aureus* | *Plasmodium falciparum*

FURTHER INFORMATION
Bernhard Ø. Palsson's homepage: http://systemsbiology.ucsd.edu/
BioCyc: http://biocyc.org
Biolog: http://www.biolog.com/
BRENDA: http://www.brenda-enzymes.info/
CMR: http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi
CYGD: http://mips.gsf.de/genre/proj/yeast/
EBI: http://www.ebi.ac.uk/
EcoCyc: http://ecocyc.org/
EntrezGene: http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
Genome Reviews: http://www.ebi.ac.uk/GenomeReviews/
IMG: http://img.jgi.doe.gov/cgi-bin/pub/main.cgi
KEGG: http://www.genome.jp/kegg/
MetaCyc: http://metacyc.org/
RegulonDB: http://regulondb.ccg.unam.mx/http:/regulondb.ccg.unam.mx/
SBML: http://sbml.org/Main_Page
SEED: http://www.theseed.org/wiki/Main_Page
SGD: http://www.yeastgenome.org/
Transport DB: http://www.membranetransport.org/

SUPPLEMENTARY INFORMATION
See online article: S1 (table) | S2 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF