Towards multidimensional genome annotation

Jennifer L. Reed, Iman Famili‡, Ines Thiele* and Bernhard O. Palsson**

Abstract | Our information about the gene content of organisms continues to grow as more genomes are sequenced and gene products are characterized. Sequence-based annotation efforts have led to a list of cellular components, which can be thought of as a one-dimensional annotation. With growing information about component interactions, facilitated by the advancement of various high-throughput technologies, systemic, or two-dimensional, annotations can be generated. Knowledge about the physical arrangement of chromosomes will lead to a three-dimensional spatial annotation of the genome and a fourth dimension of annotation will arise from the study of changes in genome sequences that occur during adaptive evolution. Here we discuss all four levels of genome annotation, with specific emphasis on two-dimensional annotation methods.

One-dimensional annotation
Details the position of genes within the genome and describes the cellullar function of gene products.

Two-dimensional annotation
Accounts for the cellular components that are identified in a one-dimensional annotation as well as their chemical and physical interactions.

Network reconstruction
A description of the network components and their interactions.

**Department of Bioengineering, University of California, San Diego, La Jolla, California 92093, USA.*
‡Genomatica Inc., 5405 Morehouse Drive, Suite 210, San Diego, California 92121, USA.
Correspondence to B.O.P.
e-mail: palsson@ucsd.edu
doi:10.1038/nrg1769

The growing number of fully sequenced genomes and high-throughput data sets has led to the identification and condition-dependent use of cellular components on a genome scale. Information about the function of cellular components, their interactions, spatial location and alterations over evolutionary time can be represented in the different levels of genome annotation (FIG. 1). One-dimensional genome annotation involves the identification of genes in the genome and the assignment of either predicted or known functionality to the identified gene products. This one-dimensional annotation is commonly referred to as genome annotation; however, other levels of detail can be annotated as well.

Two-dimensional genome annotation specifies the cellular components and their interactions (for example, protein–protein interactions, regulatory interactions and metabolite transformations). The delineation of chemical and physical interactions between cellular components leads to a network reconstruction that effectively represents two-dimensional information. If such networks are used to provide a structured basis for studying the genotype–phenotype relationship, they need to be biochemically, genomically and genetically accurate.

As we learn more about the spatial orientation of network components and evolution of genomes, new levels of annotation will be used to describe the genomes (FIG. 1). Knowledge about the intracellular arrangement of chromosomes and other cellular components will lead to a three-dimensional annotation of the genome, as cellular packing and localization of the genome can have an important role in its function. Four-dimensional genome annotation might arise from the study of changes in genome sequences that occur during adaptive evolution.

Here we discuss all four levels of genome annotation, with an emphasis on how to generate two-dimensional annotations for biochemical-reaction networks. Our focus is motivated by an increased interest in generating two-dimensional reconstructions, which are useful for evaluating one-dimensional annotations, for analysing and interpreting experimental results (such as gene-expression data), and for their promise to systematically drive biological discovery. The analysis and evaluation of these interaction networks and their accompanying models will provide insights into human diseases[1,2], such as cancer and diabetes, enable formal comparative genomic analyses, identify drug targets for human pathogens[3,4], and can be used to design industrially or environmentally useful organisms[5–11].

1D annotations: network components

Advances in high-throughput and computational technologies have resulted in the genome sequencing of hundreds of organisms across all three domains of life[12]. One-dimensional annotation of sequenced genomes involves the identification of genes, followed by functional assignment using various computational tools. The bioinformatics methods that are used to derive the one-dimensional annotations have been reviewed elsewhere[13]. These methods include gene-finding algorithms such as GLIMMER[14], GlimmerM[15] and GENSCAN[16] and sequence-homology search tools such as BLAST[17,18], FASTA[19] and HMMER[20].

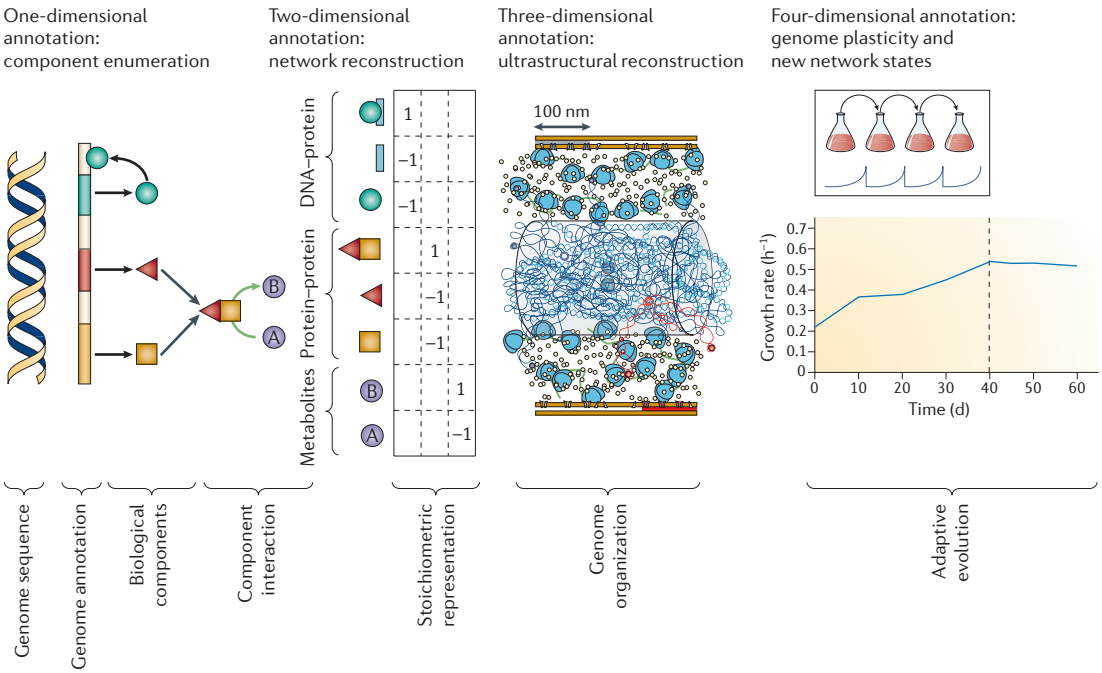


Figure 1 | **Four levels of annotation.** One-dimensional genome annotation provides a list of network components. The interaction between network components can be represented using a two-dimensional annotation (where a matrix of stoichiometric coefficients is used to represent component interactions). The structural organization of the genome can also be represented spatially in a three-dimensional annotation. Changes in genome sequence can be characterized in a four-dimensional annotation. One- and two-dimensional annotations are reproduced, with permission, from *Nature Biotechnology* REF. 102 © (2004) Macmillan Publishers Ltd. The three-dimensional annotation is reproduced, with permission, from REF. 103 © (2002) Blackwell Publishing Ltd. Data for the four-dimensional annotation are from REF. 104.

Additionally, non-homology-based algorithms such as gene neighbour[21], gene cluster[22,23], Rosetta stone[24–26] and phylogenetic profiles[27,28] are used to assign functionality on the basis of patterns across multiple genomes. More recently experimental data, including gene-expression[29–33] and protein-interaction maps[34], have identified functionally related proteins. Even with all of these methods a large fraction of the genes in the genome have an unknown function.

2D annotations: component interactions

Two-dimensional genome annotation builds on one-dimensional annotation by accounting for cellular components and their interactions. Components can physically and/or chemically interact with one or more other components. These interactions often lead to an altered state of the component, such as a phosphorylated or bound state of a protein or a biochemical transformation. The components can be arranged as rows in a table and each identified interaction among the components can be represented by numerical entries in a column (FIG. 1). Therefore, component-interaction maps can be represented by a table that contains two-dimensional information. Two-dimensional annotation allows the information in a one-dimensional annotation to be placed into a biological context, and in some cases can lead to a one-dimensional genome re-annotation[35–39].

An example of a two-dimensional annotation is a metabolic-network reconstruction, which is basically a genetically, genomically and biochemically structured database that can be queried using various computational methods (reviewed in REF. 40). Below we describe how to generate, represent and validate genome-scale reconstructions of biochemical-reaction networks, thereby creating a two-dimensional annotation for a genome. The fundamental goal of a reconstruction is to accurately define the chemical transformations that take place among chemical components in a network.

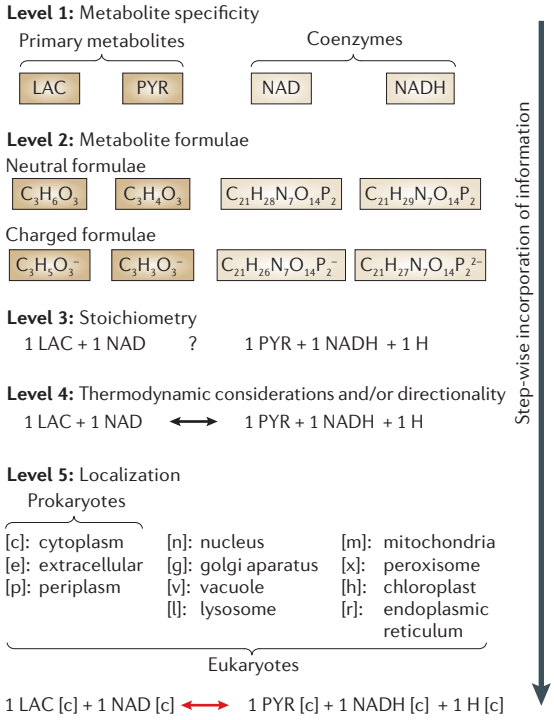
Although the emphasis here is on metabolic networks as the literature on them is well developed, other biological networks can also be represented in much the same way as they fundamentally obey many of the same chemical rules[41]. Protein-interaction, signalling and regulatory networks are often represented qualitatively. Although the available reconstructions of these networks describe the components and their interactions, they currently lack the biochemical details of metabolic reconstructions. Therefore, many of the reconstruction details that are presented in this review are transferable to these networks if the details (BOX 1), such as stoichiometry, are known. A stoichiometric reconstruction of the JAK–STAT signalling network was recently published[42]. It includes data on the stoichiometry of network components (ligands, proteins and ATP) that participate in individual signalling events (ligand binding, protein dimerization and phosphorylation).

Three-dimensional annotation
Details the spatial location of genes (rather than the gene products) within the cell as a result of genome packaging.

Four-dimensional annotation
Details changes in genome sequence that result from adaptive evolution.

Box 1 | **Defining metabolic reactions**

Different levels of information are needed to obtain a detailed description of a biochemical transformation. Biochemical accuracy is especially important if the mathematical representation of the reconstruction is to be used for subsequent computations, otherwise the calculated network properties are likely to be incorrect. The first level defines the metabolite specificity of a gene product. Although primary metabolites are often the same for homologous enzymes across organisms, the use of coenzymes might vary. In the case of lactate dehydrogenase in *Escherichia coli* (see figure), NAD serves as an electron acceptor for lactate (LAC) resulting in the formation of pyruvate (PYR) and NADH. The second level of detail accounts for the charged molecular formula of each metabolite at a physiological pH. The knowledge of the chemical formula leads to the third level of detail, the stoichiometric coefficients of the reaction. By balancing out the elements and charge in the reaction, the overall stoichiometry of the reaction can be defined. It is here that protons and water molecules are often added to balance the chemical equation. The directionality of the reaction represents the fourth level, at which biochemical studies and thermodynamic properties define the *in vivo* reaction directionality. At the fifth level, the cellular compartment in which the reaction takes place has to be determined. See supplementary information S1 (box) for more details.



How to reconstruct metabolic networks. Although high gene- or protein-sequence homology implies a similar function for gene products, a one-dimensional annotation that is based purely on sequence homology is an hypothesis[43] that needs biochemical verification. Several details need to be considered for translating a one-dimensional annotation of a gene into a set of defined biochemical reactions (BOX 1). Scientists who want to reconstruct biochemical-reaction networks should pay attention to the issues that are outlined below and summarized in BOX 1.

As a first step in generating enzyme-specific biochemical reactions, the substrate specificity of an enzyme has to be determined. In general, enzymes can be classified into two groups on the basis of substrate specificity: those that function only on one or a few highly similar substrates and those with a broader substrate specificity that can function on a class of compounds with similar functional groups (for example, alcohol dehydrogenase). The substrates that are recognized by either type of these enzymes might differ across organisms. The substrate specificity can differ for primary metabolites, as well as coenzymes (such as NADH versus NADPH and ATP versus GTP). BRENDA[44], an online database, contains detailed information about enzyme substrate specificities for a number of organisms and links to relevant publications.

Once the molecular formulae have been determined for the participating metabolites, the stoichiometry of the reaction can be specified. Here the overall charge and every element (including C, H, N, O, S and P) of the substrates and products have to be balanced. The stoichiometry for the metabolites is generally available in biochemical databases (TABLE 1), although protons and water molecules are often left out of the reactions in these databases. The directionality or reversibility of a reaction, which is a function of the thermodynamics of the reaction, also needs to be defined. Biochemical characterization studies will sometimes test the reversibility of enzyme reactions, but the directionality can differ between *in vitro* and *in vivo* environments owing to differences in temperature, pH and metabolite concentrations.

Reactions and proteins need to be assigned to specific cellular compartments. This task is relatively easy for prokaryotes, which have only a small number of cellular compartments, but becomes challenging for eukaryotes, which have significantly more subcellular compartments (BOX 1). Incorrect assignment of the location of a reaction can lead to further gaps in the metabolic network and misrepresentation of the network properties. In the absence of experimental data, proteins can be assumed to reside in the cytosol[45].

Algorithms, such as PSORT[46] and SubLoc[47], predict the cellular localization of proteins on the basis of nucleotide or amino-acid sequences (see REF. 48 for a review of the algorithms). Additionally, high-throughput experimental approaches have been developed for determining the cellular localization of proteins, such as immunofluorescence[49] and GFP tagging[50] of individual proteins. In multicellular organisms, the expression of individual genes can vary across cell types[33]; in these cases tissue-specific reconstructions might be more functionally relevant.

Table 1 | **Information that is contained in non-organism-specific databases**

	KEGG	BRENDA	UniProtKB	Entrez Gene	PubChem	MetaCyc	Transport DB	TIGR	PSORTdb
Information about the definition of metabolic reactions									
Substrate specificity	✓	✓	✓			✓	✓		
Metabolite formulae	✓	✓			✓	✓	✓		
Stoichiometry	✓	✓	✓			✓			
Reaction directionality	✓	✓				✓	✓		
Subcellular localization				✓		✓			✓
Other information about metabolic-reaction properties									
Genome sequence and annotation	✓		✓	✓				✓	
GPR associations	✓	✓				✓	✓		
Literature	✓	✓	✓	✓		✓	✓		

GPR associations, gene–protein–reaction associations.

Databases do not generally contain all the different types of information that have been discussed above. Consequently, various sources must be used to comprehensively capture the relationship between different components (BOX 2). The amount and types of available data can vary widely for different organisms, which affects the quality of reconstruction efforts. The one-dimensional annotation is the primary data source for a genome-scale reconstruction as it provides a list of enzymes and transporters. With large numbers of ORFs still having unassigned functions, it is important to remember that the sequence-derived list of metabolic enzymes is not complete.

More detailed metabolic information can be found in biochemistry textbooks, scientific literature and online databases. A measure of the amount of scientific literature that is available for an organism is its species-knowledge-index (SKI) value[12], which is calculated as the number of abstracts per species in Medline divided by the number of genes in the genome. Reconstructions for organisms with the top SKI values have already appeared (TABLE 2), including *Escherichia coli*, *Homo sapiens* and *Staphylococcus aureus*. A number of online databases contain information that is available in the biochemical literature, including ExPASy Proteomics Server[51], KEGG — Kyoto Encyclopedia Genes and Genomes[52], and BRENDA[44], all of which detail enzymatic activities (TABLE 1). For many organisms, organism-specific databases are being developed that detail the metabolic capabilities and collate the available data for a specific organism (EcoCyc[53], WIT[54], SGD[55], MetaCyc[56] and others).

Some types of information are more reliable than others, leading to multiple levels of confidence in the biochemical reactions[57]. Reactions that are based on biochemical characterizations of enzymes are more reliable than those that are based solely on sequence similarity. The least reliable reactions are those that were added without any genetic or biochemical evidence to fill in the gaps in the metabolic pathways (this is discussed in more detail below). Assigning confidence levels to the reactions in a reconstruction will aid in the network evaluation, which is described in a later section.

Assembly and representation of metabolic-network reconstruction. The previous sections described how annotations for individual metabolic genes can be translated into a list of metabolic reactions detailing stoichiometry, directionality and localization. The next step in generating a two-dimensional network reconstruction involves assembling the metabolic reactions into a network and representing the reconstruction mathematically (BOX 3), allowing for the quantitative analysis of network properties.

Assembling a reconstructed metabolic network involves: analysing traditional biochemical pathways (such as glycolysis and amino-acid biosynthesis); filling in missing metabolic activities that are not represented in the one-dimensional annotation; and adding reactions that do not fit into defined biochemical pathways but are supported by the one-dimensional annotation. By starting with central metabolism, the cellular fueling reactions[58] that are present in all organisms, and then moving on to the biosynthesis of individual macromolecular building blocks (for example, amino acids, nucleotides and lipids), the reconstructed network can be assessed in a step-wise fashion. For example, if proline is a non-essential amino acid for an organism then the metabolic network should contain a complete proline-biosynthesis pathway, even if some of the enzymes are not in the current one-dimensional annotation (this can sometimes lead to one-dimensional re-annotation[37,38]). Once all the main metabolic pathways or subsystems are assembled, several enzymatic activities, which do not participate in traditional biochemical pathways, that are included in the one-dimensional annotation need to be added to the reconstruction. These enzymes might be involved in the use of other carbon sources or connect different pathways.

Metabolite connectivity
The number of reactions a given metabolite participates in.

Systemic reactions
Mathematically derived reactions which represent overall or dominant types of chemical transformation in a given network.

Isozymes
Proteins encoded by different genes that catalyse the same reaction.

A reconstructed metabolic network can be represented in several ways: textually, graphically, and mathematically as a matrix. A textual representation, such as a list or database entry, can be easily shared and queried. A graphical representation, such as a map of nodes and edges, can be useful for analysing topological features of a network. An advantage of using a matrix representation is that it can be readily used to study quantitative properties of a network by a growing number of computational methods (reviewed in REF. 40). Each row in a matrix representation corresponds to a particular chemical state of a network component and each column to a chemical transformation among the network components (BOX 3). The elements of the matrix correspond to the stoichiometric coefficients of the metabolites in the individual chemical reactions. The resulting matrix is called the stoichiometric matrix, and is denoted by S. This matrix is a compact mathematical representation of a reconstructed network. Several systemic properties can be readily identified from the stoichiometric matrix, such as metabolite connectivity[59] and systemic reactions[60].

The assembled reaction network depends on the genome of the organism. Gene–protein–reaction (GPR) associations formally connect reactions in the metabolic network to proteins and genes in the organism. These GPR associations indicate which genes encode which proteins and which enzymatic reactions these proteins catalyse (BOX 3). Once constructed, GPR associations can be used to relate various data types, including genomic, transcriptomic, proteomic and flux data. GPR associations need to distinguish between isozymes, enzyme complexes, enzyme subunits, and single and multifunctional enzymes[39,45] so that they capture the complexity and diversity of the biological relationships. GPR associations are available for several reconstructed organisms[4,39,45,53,61,105,106].

Automating network reconstruction. The manual reconstruction process is laborious and can take up to a year for a typical bacterial genome, depending on the amount of literature available. Recently, efforts have been made to automate the reconstruction process. Databases, such as KEGG[52], can provide an automatic metabolic reconstruction, where reactions that are associated with enzyme-commission numbers present in a genome are included in the metabolic network. Pathway Tools[62] is a program that can automate a network reconstruction using metabolic reactions that are associated with enzyme-commission numbers and/or enzyme names from one-dimensional genome annotation; it makes use of known metabolic pathways to evaluate reactions and pathways in a reconstruction. Defined pathways are scored by the program and included in the reconstruction on the basis of the number of enzymes in a pathway that are found in the one-dimensional annotation. Pathway Tools will include missing reactions in a pathway if a significant fraction of the other enzymes in the pathway are present in the one-dimensional annotation. A number of these automated reconstructions have been generated using Pathway Tools[3,63–68] and are available through BioCyc[69].

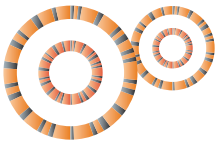
The results of these informatics approaches are limited by the quality of the one-dimensional annotation that they operate on. Therefore, automated reconstructions need detailed evaluation to assure their accuracy. Potential problems with these automated reconstructions involve incorrect substrate specificity, reaction reversibility, cofactor usage, treatment of enzyme subunits as separate enzymes and missing reactions that have no assigned ORFs[36].

Although an initial list of genes and reactions can be easily obtained by using the automated methods that are mentioned above, a good reconstruction of a metabolic or regulatory network requires the understanding of properties and characteristics of the organism or the cell. Because the number of experimentally verified gene products and reactions is limited for most

Box 2 | **Sources of information**

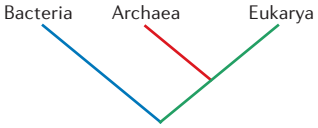
Several data sources provide the information that is required to define metabolic reactions (BOX 1). The amount of data that is available will vary from organism to organism. The genome sequence and one-dimensional annotation is one of the most important sources of information, as it contains the most comprehensive list of the cellular components. Organism-specific literature is often available, providing information on biochemical characterization of enzymes, gene essentiality, minimal medium requirements and favourable growth environments. Physiological data are needed for evaluating the reconstruction and can be found in the literature or generated experimentally. Phylogenetic data are useful when a particular organism is not well studied but a close relative is; in these cases information can be inferred from a close relative. Organism-specific and non-organism-specific databases (TABLE 1) contain a vast amount of data about gene function and associated metabolic activities. Cellular localization of enzymes can be predicted by several algorithms or can be taken from experimental data.
OD 600, optical density at a wavelength of 600 nm.

Genome sequence and annotation

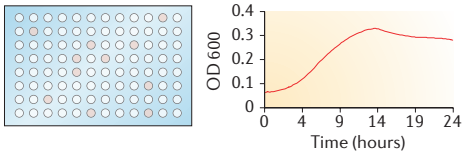


Available literature

Phylogenetic data

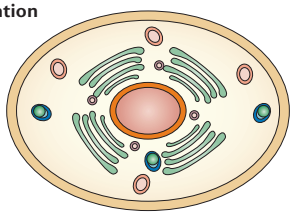


Physiological data



Databases

Localization



Signal sequences: *PLLLLPISGSALP*

Table 2 | **Organisms and network properties for which genome-scale metabolic reconstructions have been generated***

Organism	Genes	SKI	N_g	N_m	N_r	Status	Refs
Bacteria							
Bacillus subtilis	4,225	4.8	614	637	754	C, E	95
Escherichia coli	4,405	55.1	904 720 961	625 438 NA	931 627 1,107	C, E C, E C	39 92 53
Francisella tularensis	1,804	ND	350‡	NA	429	C	68
Geobacter sulfurreducens	3,530	ND	588	541	523	C, E	105
Haemophilus influenzae	1,775	8.9	296 400	343 451	488 461	C, E C, E	96 97
Helicobacter pylori	1,632	13	341 291 301‡‖	485 340 442‖	476 388 533‖	C, E C, E C	61 98 63
Lactococcus lactis	2,310	ND	358	422	621	C,E	99
Mannheimia succiniciproducens	2,463	ND	335	352	373	C, E	100
Pseudomonas aeruginosa	5,640	5.7	546 718	467 623	542 800	C, E C	¶ 67
Staphylococcus aureus	2,702	16	619	571	641	C, E	4
Streptomyces coelicolor	8,042	0.13	700	500	700	C, E	36
Archaea							
Methanococcus jannaschii	1,821	0.3	436‡	510	609	C	64
Methanosarcina barkeri	5,072	ND	692	558	619	C, E	106
Eukarya							
Arabidopsis thaliana	28,848	ND	1,418	NA	894	C	66
Homo sapiens	28,783	48.5	2,709‡	661	1,093	C	65
Mus musculus	28,287	15.6	1,156§	872	1,220	C, E	94
Plasmodium falciparum	5,342	ND	737‡	525	697	C	3
Saccharomyces cerevisiae	6,183	10.6	750 708	646 584	1,149 1,175	C, E C, E	45 93

*Several non-curated automated reconstructions are also available from KEGG[52] and BioCyc[69]. ‡Only enzyme numbers were reported. §Genes as reported[101]. ‖Latest numbers from HpCyc[63]. ¶J. Edwards, personal communication. C, a curated network; E, a network that is evaluated using computational modelling methods that are based on a stoichiometric matrix; NA, not available; ND, not determined; N_g, number of genes; N_m, number of metabolites; N_r, number of reactions; SKI, reported species knowledge index[12].

organisms, knowledge about the metabolic capabilities of the organism (such as non-essential amino acids for multicellular organisms or minimal media requirements for bacteria) is crucial. A combination of both automated and manual reconstruction efforts are needed to quickly generate accurate reconstructions.

Evaluation of a network reconstruction. Once the biological components are put together in a network reconstruction and their interactions are formally de-scribed, basic network properties can be evaluated using a computational model. These are organism-specific genome-scale models that are built through successive iterations to increase their scope and coverage. Results of *in silico* evaluations are then compared with available biological, biochemical and physiological information. An *in silico* model can be used to relate component interactions to network functionalities that often represent observable phenotypic states. Such a model can 'bring genomes to life' by formally representing the genotype–phenotype relationship.

Network evaluation can be done sequentially by first examining if the model can generate the precursor metabolites, biomass components and metabolites that the organism is known to produce or degrade, and then identifying network gaps and completing metabolic pathways on the basis of physiological information, and finally comparing the network behaviour with various experimental observations (BOX 4). These experimental observations can include gene-expression data, P/O ratio, energy-maintenance requirements and cellular phenotypes (see supplementary information S1 (box) for a more detailed discussion of network-evaluation methods). Network evaluation, although labour intensive, often leads to network adjustments, refinements and/or expansions.

Even genomes of well-studied organisms harbour genes of unknown function (for example, 20% for *E. coli*[70]). As a result, metabolic networks that are

Computational model
A set of equations that mathematically represents network reconstruction and is used to predict the behaviour of a system.

Precursor metabolites
Metabolites that are generated by catabolic pathways and used by anabolic pathways to generate biomass components.

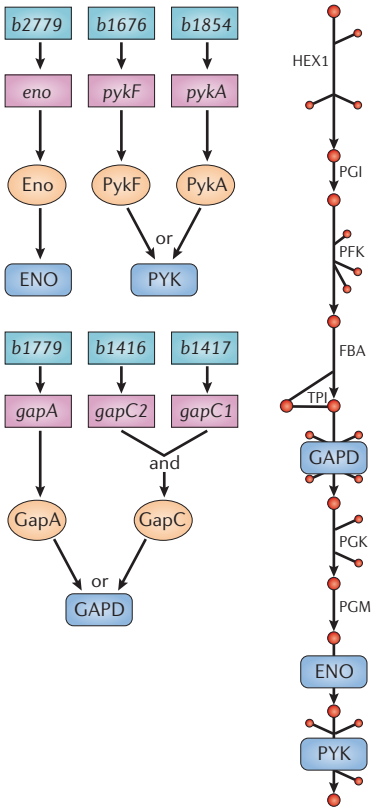
Biomass components
The macromolecules (proteins, carbohydrates, lipids and nucleotides), vitamins, cofactors, metals and minerals that make up a cell.

Box 3 | **Assembly and representation**

A list of charge and elementally balanced metabolic reactions can be represented in a stoichiometric matrix (S), where rows and columns correspond to metabolites and reactions and the elements are the stoichiometric coefficients. In genome-scale metabolic networks these stoichiometric matrices contain few non-zero elements, as relatively few metabolites participate in a given reaction. Connections between genes and reactions can be represented as gene–protein–reaction (GPR) associations by using Boolean rules or visualized using graphic images. In the GPR scheme, the first level (teal) corresponds to genetic loci, the second level (pink) to transcripts, the third level (orange) to functional proteins, and the fourth level (blue) to reactions. [c], cytoplasmic reactions.

Abbreviation	Glycolytic reactions	Genes
HEX1	[c]GLC + ATP → G6P + ADP + H	*glk*
PGI	[c]G6P ↔ F6P	*pgi*
PFK	[c]ATP + F6P → ADP + FDP + H	*pfkA*, *pfkB*
FBA	[c]FDP ↔ DHAP + G3P	*fbaA*, *fbaB*
TPI	[c]DHAP ↔ G3P	*tpiA*
GAPD	[c]G3P + NAD + PI ↔ 13DPG + H + NADH	*gapA*, *gapC1*, *gapC2*
PGK	[c]13DPG + ADP ↔ 3PG + ATP	*pgk*
PGM	[c]3PG ↔ 2PG	*gpmA*, *gpmB*
ENO	[c]2PG ↔ H_2O + PEP	*eno*
PYK	[c]ADP + H + PEP → ATP + PYR	*pykA*, *pykF*

	HEX1	PGI	PFK	FBA	TPI	GAPD	PGK	PGM	ENO	PYK
ATP	−1	0	−1	0	0	0	1	0	0	1
GLC	−1	0	0	0	0	0	0	0	0	0
ADP	1	0	1	0	0	0	−1	0	0	−1
G6P	1	−1	0	0	0	0	0	0	0	0
H	1	0	1	0	0	1	0	0	0	−1
F6P	0	1	−1	0	0	0	0	0	0	0
FDP	0	0	1	−1	0	0	0	0	0	0
DHAP	0	0	0	1	−1	0	0	0	0	0
G3P	0	0	0	1	1	−1	0	0	0	0
NAD	0	0	0	0	0	−1	0	0	0	0
PI	0	0	0	0	0	−1	0	0	0	0
13DPG	0	0	0	0	0	1	−1	0	0	0
NADH	0	0	0	0	0	1	0	0	0	0
3PG	0	0	0	0	0	0	1	−1	0	0
2PG	0	0	0	0	0	0	0	1	−1	0
PEP	0	0	0	0	0	0	0	0	1	−1
H_2O	0	0	0	0	0	0	0	0	1	0
PYR	0	0	0	0	0	0	0	0	0	1



Boolean rules
Logic statements that use Boolean operators (and, or, not) to evaluate the 'on/off' state of a variable.

P/O ratio
The number of ATP molecules (P) that are formed per oxygen atom (O) consumed during respiration.

Network gap
One or more reaction that is missing from the network reconstruction owing to the lack of direct genetic or biochemical evidence.

Blocked reactions
Reactions that, at steady state, can have no net flux (reactions that involve dead-end metabolites are blocked reactions).

Pathway holes
Missing reactions from defined metabolic pathways such as glycolysis and amino-acid biosynthesis.

constructed solely on the basis of genomic and biochemical evidence often contain many network gaps. Network gaps can be identified by analysing the ability of the network to generate individual biomass components that are needed for growth. For example, if a metabolic network is unable to generate a non-essential amino acid owing to missing steps in the biosynthetic pathways, the network gaps can be closed by completing the pathway with the missing reactions.

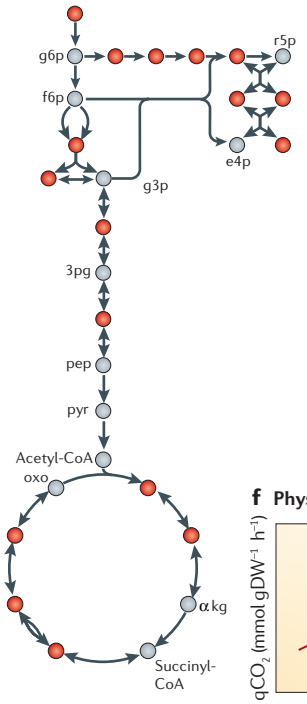
Physiological data, such as the growth capabilities of an organism, can be used to identify missing reactions or refine existing pathways. For example, metabolic pathways that are involved in the use of a carbon source can be added to a network reconstruction even in the absence of genomic or biochemical information if the organism can grow on the compound. The growth requirements of an organism therefore provide important evidence for improving, refining and expanding the quality and the content of the reconstructed networks. Reactions that are added to the network at this stage should be assigned low confidence scores because there are no genetic or biochemical data to confirm them.

Analytical tools can also be used to identify network gaps that involve reactions (blocked reactions or pathway holes) or metabolites (dead-end metabolites) that are isolated from the rest of the network. Isolated reactions can be identified computationally using flux-coupling analysis[71] (or Pathway Tools) and isolated metabolites can be identified through metabolite connectivity[39]. Addition of any reaction to the reconstructed network to fill network gaps should be supported, if possible, by previous observations and/or presence in phylogenetically related organisms. Subsequently, for each added reaction, putative genes can be identified using homology-based and context-based computational techniques (such as those that are described in the section on one-dimensional annotation)[36,37,68]. Such added reactions and putative assignments form a set of testable hypotheses that are subject to further experimental investigation. Reactions that cause network gaps can be removed from the network; for example, pathways that have many gaps might not occur in an organism and the functional assignment of associated genes should be re-examined[38]. On the other hand, gaps that were included on the basis of biochemical data indicate missing metabolic knowledge and should remain.

Discrepancies between predicted and experimental phenotypic data for genetic perturbations (either knockouts or knockdowns through small interfering RNA) on defined growth conditions can also be used to evaluate the content of the metabolic network. As described above, false negatives (for example, experimental growth but no predicted *in silico* growth) can indicate that reactions are missing from the metabolic network or the existence of isozymes[35,45]. False positives (for example, growth that is predicted *in silico* without

Box 4 | **Network evaluation**

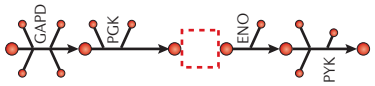
a **Precursor metabolite formation**



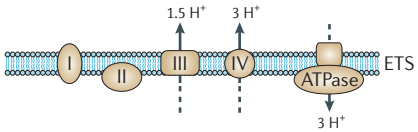
b **Incorporating biomass composition**

D = 0.1	% (w/w)
Proteins	
Amino acids	45.0
Free amino acids	1.1
Carbohydrates	
Monosaccharides	–
Disaccharides	
Trehalose	0.8
Oligosaccharides	–
Polysaccharides	
Glycogen	8.4
Mannan	13.1
Other carbohydrates	18.4
Nucleotides	
RNA	6.3
DNA	0.4
Lipids	2.9
Ash	5.0
Total	**101.4**

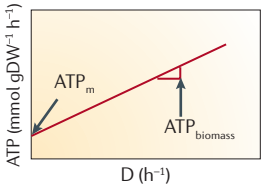
c **Filling network gaps**



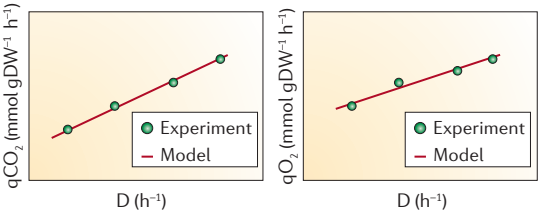
d **P/O ratio calculation**



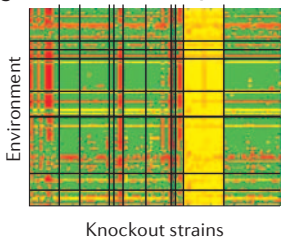
e **ATP-maintenance calculation**



f **Physiological-data comparison**



g **Knockout-data comparison**



Once a network is reconstructed and mathematically represented, the basic network capabilities must be evaluated in two steps: by evaluating the ability of the network to meet growth and physiological demands, and by incorporating and examining organism-specific network properties. A metabolic network must be able to generate all the precursor metabolites that are required for the synthesis of biomass components (**a**). In addition, biosynthetic pathways that are required for the formation of biomass components must also be present (**b**). The ability of the network to make the biosynthetic components implies that certain network gaps must be closed, even in the absence of direct genetic, biochemical or physiological data (**c**). If experimental data are available for the P/O ratio and the stoichiometry of the electron-transport system (ETS) in the cell, the efficiency of the network for making ATP through oxidative phosphorylation can be calculated (**d**). The energy maintenance that is required for growth ($ATP_{biomass}$)-associated and non-growth (ATP_m)-associated activities must also be incorporated into the network reconstruction (**e**). ATP-maintenance values can be extrapolated from growth data. Once the energy maintenance is determined, the ability of the network to establish uptake and secretion rates for molecules such as CO_2 or O_2 (qCO_2 or qO_2) can be calculated and compared with experimental measurements (**f**). Evaluating the inconsistency between model and experimental-knockout results can lead to experimentation and biological discovery and increase network accuracy (**g**). Although network evaluation is more or less a sequential procedure in the order described, many of the steps might need to be repeated iteratively following changes to the network that arise from its evaluation. See supplementary information S1 (box) for more details. D, dilution rate; gDW, grams dry weight. Panel **g** is reproduced, with permission, from *Nature* REF. 35 © (2004) Macmillan Publishers Ltd.

experimental growth) can implicate reactions that are incorrectly included in the metabolic reconstruction[35,45]. Reactions that have a low confidence score (which indicates that there is no genetic, biochemical or physiological evidence) that cause false positives should probably be removed from the network. However, false positives can also be attributed to low expression or activity of an enzyme, and can point to potential kinetic or transcriptional regulation. This type of analysis is particularly important for organisms for which limited genomic or biochemical information is available.

Network evaluation is highly dependent on the availability of data, especially physiological data, which can often be the most limiting factor. The amount of data is highly variable and depends on the organism being studied. As a result, substantial experimental and computational efforts will need to be combined for poorly characterized organisms to generate accurate one-dimensional and two-dimensional genome annotations.

Iterative network reconstruction and model building. The mathematical representation of metabolic networks allows for network evaluation not only for a specific metabolic pathway, but also for the large-scale network as a whole. Network reconstructions iteratively evolve as a result of network evaluation[72], genome re-annotation[43]

Dead-end metabolites
A metabolite that is either only produced or only consumed by the metabolic network (pathway holes, network gaps and blocked reactions involve dead-end metabolites).

Flux-coupling analysis
A computational method that determines how fluxes through a pair of reactions are related.

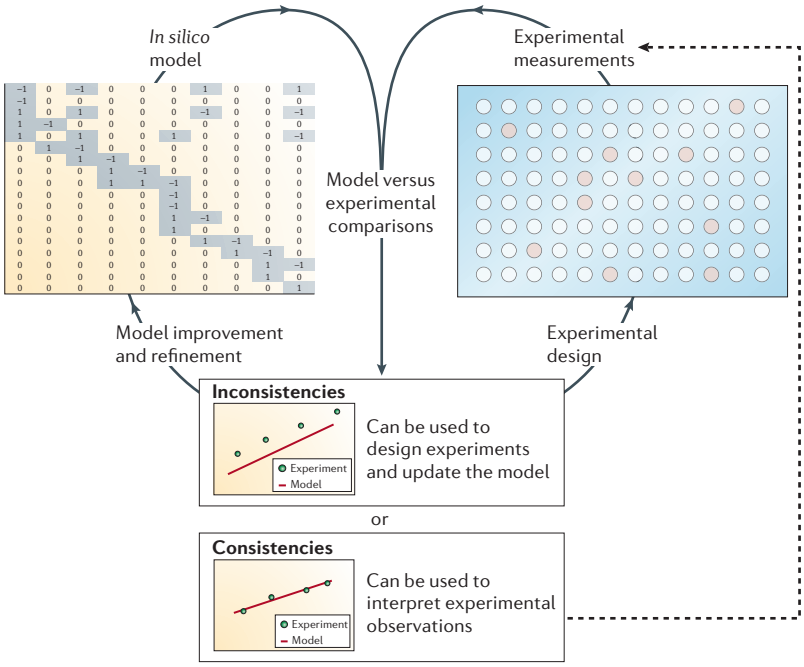


Figure 2 | **Model-guided network expansion.** By comparing model predictions with experimental data, consistencies and inconsistencies can be identified. Consistencies can aid in the interpretation of data, whereas inconsistencies generate hypotheses about the organism, by the identification or elimination of components and/or interactions. An iterative process of model development, through comparison with experimental data, will not only lead to improved models but also expand our knowledge of the networks and organisms that are being studied.

and the availability of new experimental data (FIG. 2). Analysing high-throughput data such as transcriptomic, proteomic and metabolomic data in the context of network reconstruction (that is, 'putting content into context') provides the means to evaluate the accuracy of the network reconstruction, to generate hypotheses about any inconsistencies and to evaluate experimental data within the context of functional roles. Combined analysis of transcriptomic, proteomic and protein-interaction data has provided detailed description and a better understanding of galactose metabolism in *Saccharomyces cerevisiae*[73]. A genome-scale metabolic and regulatory model of *E. coli* was used to generate hypotheses about new metabolic reactions and regulatory interactions in the organism, by analysing gene-expression and physiological growth data[35]. A network reconstruction can therefore be used not only to improve and refine the accuracy of the network, but also as a tool to evaluate the consistency of various heterogeneous data sets within the context of a biological network, and to generate testable hypotheses that drive experimental discovery.

3D annotations: genome spatial orientation

The one-dimensional annotation conveys a list of genes and their functions, which can be translated into a table of gene products and their known interactions (two-dimensional annotation). The genome itself must operate within the three-dimensional confines of a cell. A growing number of studies indicate that both the genomic location (that is, the linear allelic address) and the spatial localization (that is, the position within the cell) of a gene is important for genome function (for review see REFS 74,75). The origin and terminus of replication in the *Caulobacter crescentus* genome are orientated towards opposite ends of the replicating cell and individual loci around the chromosome are localized linearly along the cell axis[76]. These observations support the hypothesis that genome location is not random[74,75]. Analysis of the genomic parameters shows location-dependent patterning of properties such as gene-expression levels[77,78] and location of essential genes[79–81]. In addition, the metabolic energy state of a cell might influence the organization of a genome and the expression level of genes[82,83]. All of these studies indicate that the three-dimensional organization of the genome is important for proper cellular function. Annotating three-dimensional information to a genome at this level is just emerging and we can expect progress to be made in this direction, especially when we can use technologies that allow us to resolve DNA organization on the 100-nm-length scale in a cell.

4D annotations: evolutionary changes

Genomes can undergo short-term adaptive changes. Therefore, one can think of a fourth dimension to genome function — time. Such adaptive changes can have an epigenetic or a genetic basis. These mechanisms and how they function during adaptation have been studied for individual loci (such as *arcB*[84], *glpR*[85], *mglD* and *mglO* (REF. 86) in *E. coli*, and *PDR1* (REF. 87) in *S. cerevisiae*) but have not yet been elucidated on a genome scale, with the exception of genome rearrangements. There has been a growing realization that the genome sequences we have are 'snap-shots' of a genome that is continually changing. A fuller understanding of the plasticity and adaptation of genomes on a genome scale is needed. Full genome re-sequencing could provide information on the genetic basis of genome adaptation, allowing us to fully determine all the sequence changes that occur in genomes. High-accuracy mass spectroscopy and new low-cost sequencing methods have been used to re-sequence small and large portions of *E. coli* genomes that have undergone adaptive evolution[88,89]. Analysis of these sequence changes should provide insights into the mechanisms and functions of these adaptive evolutionary changes.

Future directions

The four dimensions of genome annotation are important for describing and capturing the functional capabilities of a cell. A detailed quality-controlled and quality-assessed process for genome-scale network reconstruction (an example of a two-dimensional annotation) has developed over the past 5–10 years. It is a laborious and detailed process that involves manual curation of a wide range of data types. Similar to sequence assembly and one-dimensional genome annotation, this process of two-dimensional annotation is iterative, involving the successive addition

of more and more detailed data for a particular organism. These high-quality reconstructions can be used as the basis for computation of phenotypic traits, and they represent a key step in the development of the burgeoning field of systems biology[90].

This entire process, from reaction definition to network evaluation, yields highly curated genome-scale network reconstructions[40], which have been reconciled with heterogeneous data sets. Genome-scale network reconstructions are biochemically, genomically and genetically structured databases that have defined confidence scores for the components and their interactions. New data can be used to expand these reconstructions, resulting in a history of iteratively built reconstructions such as those for *E. coli*[91].

Several reconstructions have been made for model organisms, including *E. coli*[39,53,92], *S. cerevisiae*[45,93], *Mus musculus*[94] and *H. sapiens*[65] (TABLE 2). More organisms are likely to follow suit including other multicellular model organisms (*Caenorhabditis elegans*, *Drosophila melanogaster* and *Rattus norvegicus*), human pathogens and organisms that are of industrial interest. Existing network reconstructions will continue to grow iteratively in content, scope and detail. A manual human reconstruction effort is currently underway (progress is tracked on the Human Metabolic Network Reconstruction web site), which will include cellular compartmentalization and track tissue distribution of enzymes. It will also account for transcript variations that arise from alternate splicing. When completed, it will be used to generate tissue-specific reconstructions to further study human disease states.

As genomic sciences continue to evolve we can anticipate that multiple dimensions in genome annotation will appear as we characterize genome-scale functions. The expansion in dimensionality of genome annotation allows for the formalization of our knowledge about genomes, their attributes and functions. Such efforts will show the breadth of backgrounds that are needed to master genome science and how they will serve to focus and integrate disciplines that until recently have been separate. One-dimensional annotation is primarily based on bioinformatic analysis, two-dimensional annotation is the domain of network biology, three-dimensional annotation involves ultrastructural studies on yet-to-be explored length scales and four-dimensional annotation will entail experimental study of genome-scale sequence changes during adaptive evolution. Even higher-order annotation of genomes might eventually emerge. Well-curated multidimensional annotation of genomes is fundamental to systems biology and genomic science. We currently have the methods and information needed to generate one-dimensional and two-dimensional annotations; as we learn more about the structural arrangement of genomes within the cell and how these genomes adaptively evolve we can begin to generate higher levels of annotation.

1. Thiele, I., Price, N. D., Vo, T. D. & Palsson, B. O. Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. *J. Biol. Chem.* **280**, 11683–11695 (2005).
2. Jamshidi, N., Wiback, S. J. & Palsson, B. O. *In silico* model-driven assessment of the effects of single nucleotide polymorphisms (SNPs) on human red blood cell metabolism. *Genome Res.* **12**, 1687–1692 (2002).
3. Yeh, I., Hanekamp, T., Tsoka, S., Karp, P. D. & Altman, R. B. Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. *Genome Res.* **14**, 917–924 (2004).
4. Becker, S. A. & Palsson, B. O. Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiol.* **5**, 8 (2005).
5. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647–657 (2003).
6. Alper, H., Jin, Y. S., Moxley, J. F. & Stephanopoulos, G. Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metab. Eng.* **7**, 155–164 (2005).
7. Alper, H., Miyaoku, K. & Stephanopoulos, G. Construction of lycopene-overproducing *E. coli* strains by combining systematic and combinatorial gene knockout targets. *Nature Biotechnol.* **23**, 612–616 (2005).
8. Fong, S. S. *et al. In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol. Bioeng.* **91**, 743–748 (2005).
9. Carlson, R., Fell, D. & Srienc, F. Metabolic pathway analysis of a recombinant yeast for rational strain development. *Biotechnol. Bioeng.* **79**, 121–134 (2002).
10. Pharkya, P., Burgard, A. P. & Maranas, C. D. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.* **14**, 2367–2376 (2004).
11. Liao, J. C., Hou, S. Y. & Chao, Y. P. Pathway analysis, engineering and physiological considerations for redirecting central metabolism. *Biotechnol. Bioeng.* **52**, 129–140 (1996).
12. 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).
13. Stein, L. Genome annotation: from sequence to biology. *Nature Rev. Genet.* **2**, 493–503 (2001).
This article provides a thorough review of one-dimensional annotation methods that involve gene finding and gene-functional assignment, as well as placing genes in the context of biological processes.
14. Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26**, 544–548 (1998).
15. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24–31 (1999).
16. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
17. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
19. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85**, 2444–2448 (1988).
20. Eddy, S. HMMER: profile HMMs for protein sequence analysis. *HMMER: sequence analysis using pofile hidden Markov Models web site* [online], < http://hmmer.wustl.edu > (2003).
21. Bowers, P. M. *et al.* Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* **5**, R35 (2004).
This article describes several context-based methods for identifying genes that are functionally related. The article also announces the creation of the Prolinks database that includes results for several genomes.
22. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol.* **1**, 93–108 (1999).
23. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA* **96**, 2896–2901 (1999).
24. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
25. Marcotte, E. M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
26. Marcotte, C. J. & Marcotte, E. M. Predicting functional linkages from gene fusions with confidence. *Appl. Bioinformatics* **1**, 93–100 (2002).
27. Wu, J., Kasif, S. & DeLisi, C. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* **19**, 1524–1530 (2003).
28. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288 (1999).
29. Kharchenko, P., Vitkup, D. & Church, G. M. Filling gaps in a metabolic network using expression information. *Bioinformatics* **20** (Suppl. 1), I178–I185 (2004).
30. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
31. Walker, M. G., Volkmuth, W., Sprinzak, E., Hodgson, D. & Klingler, T. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res.* **9**, 1198–1203 (1999).

32. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
33. Zhang, W. *et al.* The functional landscape of mouse gene expression. *J. Biol.* **3**, 21 (2004).
34. Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnol.* **23**, 561–566 (2005).
35. 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 article describes an iterative model-building approach for identifying new regulatory interactions that is based on gene-expression data. The work also resulted in the identification of knowledge gaps in metabolism and regulation from analysis of mutant phenotyping data.
36. Borodina, I., Krabben, P. & Nielsen, J. Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res.* **15**, 820–829 (2005).
This article describes a metabolic reconstruction that is generated by automated methods followed by manual curation for *Streptomyces coelicolor*. It discusses problems that are associated with automated reconstructions and provides examples where two-dimensional annotation enhanced one-dimensional annotation by finding genes for missing metabolic enzymes.
37. Green, M. L. & Karp, P. D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5**, 76 (2004).
This article presents a method for identifying the genes responsible for encoding enzymes that are missing from pathways in current metabolic reconstructions. This method was applied to reconstructions from three different organisms and led to new putative assignments for about half the missing enzymes.
38. Karp, P. D., Krummenacker, M., Paley, S. & Wagg, J. Integrated pathway-genome databases and their role in drug discovery. *Trends Biotechnol.* **17**, 275–281 (1999).
39. Reed, J. L., Vo, T. D., Schilling, C. H. & Palsson, B. O. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biol.* **4**, R54 (2003).
40. 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).
This review provides a comprehensive overview of developed methods for interrogating reconstructions using a constraint-based modelling approach.
41. Papin, J. A., Hunter, T., Palsson, B. O. & Subramaniam, S. Reconstruction of cellular signalling networks and analysis of their properties. *Nature Rev. Mol. Cell Biol.* **6**, 99–111 (2005).
42. Papin, J. A. & Palsson, B. O. The JAK–STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys. J.* **87**, 37–46 (2004).
43. Ouzounis, C. A. & Karp, P. D. The past, present and future of genome-wide re-annotation. *Genome Biol* **3**, COMMENT2001 (2002).
44. Schomburg, I. *et al.* BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* **32**, D431–D433 (2004).
45. Duarte, N. C., Herrgard, M. J. & Palsson, B. O. Reconstruction and validation of *Saccharomyces cerevisiae i*ND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14**, 1298–1309 (2004).
46. Gardy, J. L. *et al.* PSORTb v. 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **21**, 617–623 (2005).
47. Hua, S. & Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17**, 721–728 (2001).
48. Schneider, G. & Fechner, U. Advances in the prediction of protein targeting signals. *Proteomics* **4**, 1571–1580 (2004).
49. Ross-Macdonald, P. *et al.* Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418 (1999).
50. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
51. Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788 (2003).
52. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
53. Keseler, I. M. *et al.* EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.* **33**, D334–D337 (2005).
54. Overbeek, R. *et al.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28**, 123–125 (2000).
55. 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).
56. Krieger, C. J. *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **32**, D438–D442 (2004).
57. Vo, T. D., Greenberg, H. J. & Palsson, B. O. Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J. Biol. Chem.* **279**, 39532–39540 (2004).
58. Neidhardt, F. C., Ingraham, J. L. & Schaechter, M. *Physiology of the bacterial cell* (Sinauer Associates, Sunderland, Massachusetts, 1990).
59. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
60. Famili, I. & Palsson, B. O. Systemic metabolic reactions are obtained by singular value decomposition of genome-scale stoichiometric matrices. *J. Theor. Biol.* **224**, 87–96 (2003).
61. Thiele, I., Vo, T. D., Price, N. D. & Palsson, B. O. An expanded metabolic reconstruction of *Helicobacter pylori* (*i*IT341 GSM/GPR): An *in silico* genome-scale characterization of single and double deletion mutants. *J. Bacteriol.* **187**, 5818–5830 (2005).
62. Karp, P. D., Paley, S. & Romero, P. The Pathway Tools software. *Bioinformatics* **18** (Suppl. 1), S225–S232 (2002).
63. Paley, S. M. & Karp, P. D. Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. *Bioinformatics* **18**, 715–724 (2002).
64. Tsoka, S., Simon, D. & Ouzounis, C. A. Automated metabolic reconstruction for *Methanococcus jannaschii*. *Archaea* **1**, 223–229 (2004).
65. Romero, P. *et al.* Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* **6**, R2 (2005).
66. Zhang, P. *et al.* MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol.* **138**, 27–37 (2005).
67. Romero, P. & Karp, P. PseudoCyc, a pathway-genome database for *Pseudomonas aeruginosa*. *J. Mol. Microbiol. Biotechnol.* **5**, 230–239 (2003).
68. Larsson, P. *et al.* The complete genome sequence of *Francisella tularensis*, the causative agent of tularemia. *Nature Genet.* **37**, 153–159 (2005).
69. Karp, P. D. *et al.* Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**, 6083–6089 (2005).
70. Serres, M. H. *et al.* A functional update of the *Escherichia coli* K-12 genome. *Genome Biol.* **2**, RESEARCH0035 (2001).
71. Burgard, A. P., Nikolaev, E. V., Schilling, C. H. & Maranas, C. D. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.* **14**, 301–312 (2004).
72. Palsson, B. The challenges of *in silico* biology. *Nature Biotechnol.* **18**, 1147–1150 (2000).
73. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
This article illustrates how the combination of experimental measurements and model predictions can be used to identify new network interactions. The experiments were carried out to better understand and generate new hypotheses concerning galactose utilization in yeast.
74. Thanbichler, M., Viollier, P. H. & Shapiro, L. The structure and function of the bacterial chromosome. *Curr. Opin. Genet. Dev.* **15**, 153–162 (2005).
This review discusses studies that relate to the topological (three-dimensional) structure of bacterial chromosomes. It describes recent evidence that the organization of bacterial chromosomes is non-random and that during replication the position of the genome within the cell is spatially arranged.
75. Chakalova, L. *et al.* Replication and transcription: shaping the landscape of the genome. *Nature Rev. Genet.* **6**, 669–677 (2005).
76. Viollier, P. H. *et al.* Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. *Proc. Natl Acad. Sci. USA* **101**, 9257–9262 (2004).
77. Allen, T. E. *et al.* Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. *J. Bacteriol.* **185**, 6392–6399 (2003).
78. Jeong, K. S., Ahn, J. & Khodursky, A. B. Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biol.* **5**, R86 (2004).
79. Gerdes, S. Y. *et al.* Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684 (2003).
80. Rocha, E. P. & Danchin, A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.* **31**, 6570–6577 (2003).
81. Rocha, E. P. & Danchin, A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nature Genet.* **34**, 377–378 (2003).
82. Hatfield, G. W. & Benham, C. J. DNA topology-mediated control of global gene expression in *Escherichia coli*. *Annu. Rev. Genet.* **36**, 175–203 (2002).
83. Travers, A. & Muskhelishvili, G. DNA supercoiling — a global transcriptional regulator for enterobacterial growth? *Nature Rev. Microbiol.* **3**, 157–169 (2005).
84. Flores, N. *et al.* Adaptation for fast growth on glucose by differential expression of central carbon metabolism and *gal* regulon genes in an *Escherichia coli* strain lacking the phosphoenolpyruvate: carbohydrate phosphotransferase system. *Metab. Eng.* **7**, 70–87 (2005).
85. Raghunathan, A. & Palsson, B. O. Scalable method to determine mutations that occur during adaptive evolution of *Escherichia coli*. *Biotechnol. Lett.* **25**, 435–441 (2003).
86. Notley-McRobb, L. & Ferenci, T. Adaptive *mgl*-regulatory mutations and genetic diversity evolving in glucose-limited *Escherichia coli* populations. *Environ. Microbiol.* **1**, 33–43 (1999).
87. Anderson, J. B. *et al.* Mode of selection and experimental evolution of antifungal drug resistance in *Saccharomyces cerevisiae*. *Genetics* **163**, 1287–1298 (2003).
88. Honisch, C., Raghunathan, A., Cantor, C. R., Palsson, B. O. & van den Boom, D. High-throughput mutation detection underlying adaptive evolution of *Escherichia coli*-K12. *Genome Res.* **14**, 2495–2502 (2004).
89. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
This article describes a new non-electrophoretic DNA-sequencing method for rapid whole-genome sequencing and provides results for the DNA sequence of an adaptively evolved strain of *E. coli*.
90. Palsson, B. O. *Systems Biology: Properties of Reconstructed Networks* (Cambridge Univ. Press, 2006).
91. Reed, J. L. & Palsson, B. O. Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. *J. Bacteriol.* **185**, 2692–2699 (2003).
92. Edwards, J. S. & Palsson, B. O. The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. USA* **97**, 5528–5533 (2000).
93. Forster, J., Famili, I., Fu, P., Palsson, B. O. & Nielsen, J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* **13**, 244–253 (2003).
94. Sheikh, K., Forster, J. & Nielsen, L. K. Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*. *Biotechnol. Prog.* **21**, 112–121 (2005).
95. Park, S. M., Schilling, C. H. & Palsson, B. O. *Compositions and methods for modeling* Bacillus subtilis *metabolism* (US Patent and Trademark Office, 2003).
96. Schilling, C. H. & Palsson, B. O. Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J. Theor. Biol.* **203**, 249–283 (2000).
97. Edwards, J. S. & Palsson, B. O. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* **274**, 17410–17416 (1999).

98. Schilling, C. H. *et al.* Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* **184**, 4582–4593 (2002).
99. Oliveira, A. P., Nielsen, J. & Forster, J. Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiol.* **5**, 39 (2005).
100. Hong, S. H. *et al.* The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nature Biotechnol.* **22**, 1275–1281 (2004).
101. Eppig, J. T. *et al.* The Mouse Genome Database (MGD): from genes to mice — a community resource for mouse biology. *Nucleic Acids Res.* **33**, D471–D475 (2005).
102. Palsson, B. O. Two-dimensional annotation of genomes. *Nature Biotechnol.* **22**, 1218–1219 2004).
103. Woldringh, C. L. The role of co-transcriptional translation and protein translocation (transertion) in bacterial chromosome segregation. *Mol. Microbiol.* **45**, 17–29 (2002).
104. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* **420**, 186–189 (2002).
105. Mahadevan, R. *et al.* Characterization of metabolism in the Fe(III)-reducing organism *Geobacter sulfurreducens* by constraint-based modeling. *Appl. Environ. Microbiol.* (in the press).
106. Feist, A. M. *et al.* Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Mol. Systems Biol.* (in the press).

Acknowledgements
The authors would like to thank T. Allen and S. Fong for useful comments on the manuscript. This work was funded in part by the US National Institutes of Health. B.O.P. serves on the scientific advisory board of Genomatica, Inc.

Competing interests statement
The authors declare competing financial interests: see web version for details.

FURTHER INFORMATION

BRENDA — the comprehensive enzyme information system: http://www.brenda.uni-koeln.de
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
ExPASy Proteomics Server: www.expasy.org
GENSCAN: http://bioweb.pasteur.fr/seqanal/interfaces/genscan.html
GlimmerM: http://www.tigr.org/software/glimmerm
Human Metabolic Network Reconstruction web site: http://gcrg.ucsd.edu/organisms/human_flyer4.pdf
KEGG — Kyoto Encyclopedia of Genes and Genomes: http://www.genome.ad.jp/kegg
MetaCyc: http://metacyc.org
Pathway Tools: http://bioinformatics.ai.sri.com/ptools
PSORTdb: http://db.psort.org
PubChem: http://pubchem.ncbi.nlm.nih.gov
The GLIMMER homepage: http://www.tigr.org/~salzberg/glimmer.html
TIGR — The Institute for Genomic Research: http://www.tigr.org
TransportDB: http://www.membranetransport.org
UniProtKB: http://us.expasy.org/uniprot

SUPPLEMENTARY INFORMATION

See online article: S1 (box)
Access to this links box is available online.